Econometric Theory, 2016, Page 1 of 33.
doi:10.1017/S0266466616000220

ADAPTIVE BAYESIAN ESTIMATION OF CONDITIONAL DENSITIES

ANDRIY NORETS
Brown University

DEBDEEP PATI
Florida State University

We consider a nonparametric Bayesian model for conditional densities. The model is a finite mixture of normal distributions with covariate dependent multinomial logit mixing probabilities. A prior for the number of mixture components is specified on positive integers. The marginal distribution of covariates is not modeled. We study asymptotic frequentist behavior of the posterior in this model. Specifically, we show that when the true conditional density has a certain smoothness level, then the posterior contraction rate around the truth is equal up to a log factor to the frequentist minimax rate of estimation. An extension to the case when the covariate space is unbounded is also established. As our result holds without a priori knowledge of the smoothness level of the true density, the established posterior contraction rates are adaptive. Moreover, we show that the rate is not affected by inclusion of irrelevant covariates in the model. In Monte Carlo simulations, a version of the model compares favorably to a cross-validated kernel conditional density estimator.

1. INTRODUCTION

Conditional distributions provide a general way to describe a relationship between a response variable and covariates. An introduction to classical nonparametric estimation of conditional distributions and applications in economics can be found in Chapters 5–6 of Li and Racine (2007). Applications of flexible Bayesian models for conditional densities include analysis of financial data and distribution of earnings in Geweke and Keane (2007), estimation of health expenditures in Keane and Stavrunova (2011), and analysis of firms leverage data in Villani, Kohn, and Nott (2012); see also MacEachern (1999), De Iorio, Muller, Rosner, and MacEachern (2004), Griffin and Steel (2006), Dunson, Pillai, and Park (2007), Dunson and Park (2008), Villani, Kohn, and Giordani (2009), Chung and Dunson (2009), Li, Villani, and Kohn (2010), Norets and Pelenis (2012), and Norets and Pelenis (2014). This literature suggests that the Bayesian approach to nonparametric conditional distribution estimation has several attractive properties. First, it does not require fixing a bandwidth or similar tuning parameters. Instead, it

We thank the editor, the co-editor, and referees for helpful comments. Dr. Pati acknowledges support for this project from the Office of Naval Research (ONR BAA 14-0001) and NSF DMS-1613156. Address correspondence to Andriy Norets, Associate Professor, Department of Economics, Brown University, Providence, RI 02912; e-mail: andriy_norets @brown.edu.

provides estimates of the objects of interest where these tuning parameters are averaged out with respect to their posterior distribution. Second, the Bayesian approach naturally provides a measure of uncertainty through the posterior distribution. Third, the Bayesian approach performs well in out-of-sample prediction and Monte Carlo exercises. The present paper contributes to the literature on theoretical properties of these models and provides an explanation for their excellent performance in applications.

We focus on mixtures of Gaussian densities with covariate dependent mixing weights and a variable number of mixture components for which a prior on positive integers is specified. Conditional on the number of mixture components, we model the mixing weights by a multinomial logit with a common scale parameter. The marginal distribution of covariates is not modeled. This model is closely related to mixture-of-experts (Jacobs, Jordan, Nowlan, and Hinton (1991), Jordan and Xu (1995), Peng, Jacobs, and Tanner (1996), Wood, Jiang, and Tanner (2002)), also known as smooth mixtures in econometrics (Geweke and Keane (2007), Villani et al. (2009), Norets (2010)). We study asymptotic frequentist properties of the posterior distribution in this model.

Understanding frequentist properties of Bayesian nonparametric procedures is important because frequentist properties, such as posterior consistency and optimal contraction rates, guarantee that the prior distribution is not dogmatic in a precise sense. It is not clear how to formalize this using other approaches, especially, in high or infinite dimensional settings. There is a considerable literature on frequentist properties of nonparametric Bayesian density estimation (Barron, Schervish, and Wasserman (1999), Ghosal, Ghosh, and Ramamoorthi (1999), Ghosal and van der Vaart (2001), Ghosal, Ghosh, and van der Vaart (2000), Ghosal and van der Vaart (2007), Huang (2004), Scricciolo (2006), van der Vaart and van Zanten (2009), Rousseau (2010), Kruijer, Rousseau, and van der Vaart (2010), Shen, Tokdar, and Ghosal (2013)). There are fewer results for conditional distribution models in which the distribution of covariates is left unspecified. Norets (2010) studies approximation bounds in Kullback–Leibler distance for several classes of conditional density models. Norets and Pelenis (2014) consider posterior consistency for a slightly more general version of the model we consider here and kernel stick breaking mixtures for conditional densities. Pati, Dunson, and Tokdar (2013) study posterior consistency when mixing probabilities are modeled by transformed Gaussian processes. Tokdar, Zhu, and Ghosh (2010) show posterior consistency for models based on logistic Gaussian process priors. Shen and Ghosal (2016) obtain posterior contraction rates for a compactly supported conditional density model based on splines.

In this article, we show that under reasonable conditions on the prior, the posterior in our model contracts at an optimal rate up to a logarithmic factor. The assumed prior distribution does not depend on the smoothness level of the true conditional density. Thus, the obtained posterior contraction rate is adaptive across all smoothness levels. An interpretation of this is that the prior puts

sufficient amount of weight around conditional densities of all smoothness levels and, thus, the posterior can concentrate around the true density of any smoothness nearly as quickly as possible. In this particular sense, the prior is not dogmatic with regard to smoothness.

Adaptive posterior convergence rates in the context of density estimation are obtained by Huang (2004), Scricciolo (2006), van der Vaart and van Zanten (2009), Rousseau (2010), Kruijer et al. (2010), and Shen et al. (2013). If the joint and conditional densities have the same smoothness, adaptive posterior contraction rates for multivariate joint densities in van der Vaart and van Zanten (2009) and Shen et al. (2013) imply adaptive rates for the conditional densities. However, it is important to note here that when the conditional density is smoother than the joint density in the sense of Hölder, it is not clear if the optimal adaptive rates for the conditional density can be achieved with a model for the joint distribution. A closely related concern, which is occasionally raised by researchers using mixtures for modeling a joint multivariate distribution and then extracting conditional distributions of interest, is that many mixture components might be used primarily to provide a good fit to the marginal density of covariates and, as a result, the fit for conditional densities deteriorates (see, for example, Wade, Dunson, Petrone, and Trippa (2014)). In our settings, this problem does not arise as we put a prior on the conditional density directly and do not model the marginal density of the covariates. The resulting convergence rate depends only on the smoothness level of the conditional density.

An important advantage of estimating the conditional density directly is that the problem of covariate selection can be easily addressed. We show that in a version of our model the posterior contraction rate is not affected by the presence of a fixed number of irrelevant covariates. Also, an application of Bayesian model averaging to the covariate selection problem delivers posterior contraction rates that are not affected by irrelevant covariates. Thus, we can say that the posterior contraction rates we obtain are also adaptive with respect to the dimension of the relevant covariates.

Our results hold for expected total variation and Hellinger distances for conditional densities, where the expectation is taken with respect to the distribution of covariates. The use of these distances allows us to easily adapt a general posterior contraction theorem from Ghosal et al. (2000) to the case of a model for conditional distributions only. An important part of our proof strategy is to recognize that our model for the conditional density is consistent with a joint density that is a mixture of multivariate normal distributions so that we can exploit approximation results for mixtures of multivariate normal distributions obtained in De Jonge and van Zanten (2010) and Shen et al. (2013). Our entropy calculations improve considerably the bounds obtained in Norets and Pelenis (2014).

We also evaluate the finite sample performance of our conditional density model in Monte Carlo simulations. The model performs consistently with the established asymptotic properties and compares favorably to a cross-validated kernel conditional density estimator from Hall, Racine, and Li (2004).

The paper is organized as follows. Section 2 presents the assumptions on the true conditional density, the proposed prior distributions, and the main theorem on posterior convergence rates. The prior thickness results are given in Section 3. Section 4 describes the sieve construction and entropy calculations. An extension of the results to an unbounded covariate space is considered in Section 5. The presence of irrelevant covariates is analyzed in Section 6. Section 7 presents results of Monte Carlo simulations. We conclude with a discussion of the results in Section 8.

2. MAIN RESULTS

2.1. Notation

Let $\mathcal{Y} \subset \mathbb{R}^{d_y}$ be the response space, $\mathcal{X} \subset \mathbb{R}^{d_x}$ be the covariate space, and $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}$. Let $\mathcal{F}$ denote a space of conditional densities with respect to the Lebesgue measure,

$$\mathcal{F} = \left\{ f : \mathcal{Y} \times \mathcal{X} \to [0, \infty) \text{ - Borel measurable}, \int f(y|x)dy = 1, \forall x \in \mathcal{X} \right\}.$$

Suppose $(Y^n, X^n) = (Y_1, X_1, \ldots, Y_n, X_n)$ is a random sample from the joint density $f_0 g_0$, where $f_0 \in \mathcal{F}$ and g_0 is a density on $\mathcal{X}$ with respect to the Lebesgue measure. Let P_0 and E_0 denote the probability measure and expectation corresponding to $f_0 g_0$. For $f_1, f_2 \in \mathcal{F}$,

$$d_h(f_1, f_2) = \left(\int \left(\sqrt{f_1(y|x)} - \sqrt{f_2(y|x)} \right)^2 g_0(x) dy dx \right)^{1/2} \text{ and}$$

$$d_1(f_1, f_2) = \int |f_1(y|x) - f_2(y|x)| g_0(x) dy dx$$

denote analogs of the Hellinger and total variation distances correspondingly. Also, let us denote the Hellinger distance for the joint densities by d_H.

Let us denote the largest integer that is strictly smaller than β by $\lfloor \beta \rfloor$. For $L : \mathcal{Z} \to [0, \infty)$, $\tau_0 \geq 0$, and $\beta > 0$, a class of locally Hölder functions, $\mathcal{C}^{\beta, L, \tau_0}$, consists of $f : \mathbb{R}^d \to \mathbb{R}$ such that for $k = (k_1, \ldots, k_d)$, $k_1 + \cdots + k_d \leq \lfloor \beta \rfloor$, mixed partial derivative of order k, $D^k f$, is finite and for $k_1 + \cdots + k_d = \lfloor \beta \rfloor$ and $\Delta z \in \mathcal{Z}$,

$$\left| D^k f(z + \Delta z) - D^k f(z) \right| \leq L(z) ||\Delta z||^{\beta - \lfloor \beta \rfloor} e^{\tau_0 ||\Delta z||^2}.$$

Operator "$\lesssim$" denotes less or equal up to a multiplicative positive constant relation. $J(\epsilon, A, \rho)$ denotes the ϵ-covering number of the set A with respect to the metric ρ. For a finite set A, let $|A|$ denote the cardinality of A. The set of natural numbers is denoted by $\mathbb{N}$. The m-dimensional simplex is denoted by Δ^{m-1}. I_k stands for the $k \times k$ identity matrix. Let $\phi_{\mu, \sigma}$ denote a multivariate normal density with mean $\mu \in \mathbb{R}^k$ and covariance matrix $\sigma^2 I_k$ (or a diagonal matrix with squared elements of σ on the diagonal, when σ is a k-vector).

2.2. Assumptions about data generating process

First, we assume that $f_0 \in \mathcal{C}^{\beta,L,\tau_0}$. Second, we assume that $\mathcal{X} = [0,1]^{d_x}$, except for Section 5 where we consider possibly unbounded $\mathcal{X}$. Third, g_0 is assumed to be bounded above. Fourth, for all $k \leq \lfloor \beta \rfloor$ and some $\varepsilon > 0$,

$$\int_{\mathcal{Z}} \left| \frac{D^k f_0(y|x)}{f_0(y|x)} \right|^{(2\beta+\varepsilon)/k} f_0(y|x)dydx < \infty,$$
$$\int_{\mathcal{Z}} \left| \frac{L(y,x)}{f_0(y|x)} \right|^{(2\beta+\varepsilon)/\beta} f_0(y|x)dydx < \infty. \quad \textbf{(2.1)}$$

Finally, for all $x \in \mathcal{X}$, all sufficiently large $y \in \mathcal{Y}$ and some positive (c,b,τ),

$$f_0(y|x) \leq c\exp(-b||y||^{\tau}). \quad \textbf{(2.2)}$$

2.3. Prior

The prior, Π, on $\mathcal{F}$ is defined by a location mixture of normal densities

$$p(y|x,\theta,m) = \sum_{j=1}^{m} \frac{\alpha_j \exp\left\{-0.5||x-\mu_j^x||^2/\sigma^2\right\}}{\sum_{i=1}^{m} \alpha_i \exp\left\{-0.5||x-\mu_i^x||^2/\sigma^2\right\}} \phi_{\mu_j^y,\sigma}(y), \quad \textbf{(2.3)}$$

and a prior on $m \in \mathbb{N}$ and $\theta = (\mu_j^y, \mu_j^x, \alpha_j, j = 1,2,\ldots;\sigma)$, where $\mu_j^y \in \mathbb{R}^{d_y}$, $\mu_j^x \in \mathbb{R}^{d_x}$, $\alpha_j \in [0,1]$, $\sigma \in (0,\infty)$. The covariate dependent mixing weights are modeled by multinomial logit with restrictions on the coefficients and a common scale parameter σ. To facilitate simpler notations and shorter proofs, we assume σ to be the same for all components of (y,x), except for Section 6. Extensions to component-specific σ's, which would result in near optimal posterior contraction rates for anisotropic f_0, can be done along the lines of Section 5 in Shen et al. (2013).

We assume the following conditions on the prior. For positive constants $a_1, a_2, \ldots, a_9$, the prior for σ satisfies

$$\Pi\left(\sigma^{-2} \geq s\right) \leq a_1 \exp\{-a_2 s^{a_3}\} \quad \text{for all sufficiently large } s > 0 \quad \textbf{(2.4)}$$

$$\Pi\left(\sigma^{-2} < s\right) \leq a_4 s^{a_5} \quad \text{for all sufficiently small } s > 0 \quad \textbf{(2.5)}$$

$$\Pi\left\{s < \sigma^{-2} < s(1+t)\right\} \geq a_6 s^{a_7} t^{a_8} \exp\left\{-a_9 s^{1/2}\right\}, \quad s > 0, \quad t \in (0,1). \quad \textbf{(2.6)}$$

An example of a prior that satisfies (2.4)–(2.5) is the inverse Gamma prior for σ. The usual conditionally conjugate inverse Gamma prior for σ^2 satisfies (2.4) and (2.5), but not (2.6). (2.6) requires the probability to values of σ near 0 to be higher than the corresponding probability for inverse Gamma prior for σ^2. This assumption is in line with the previous work on adaptive posterior contraction rates for mixture models, see Kruijer et al. (2010); Shen and Ghosal (2016). Prior for $(\alpha_1,\ldots,\alpha_m)$ given m is Dirichlet$(a/m,\ldots,a/m)$, $a > 0$.

$$\Pi(m=i) \propto \exp(-a_{10} i (\log i)^{\tau_1}), i = 2, 3, \ldots, \quad a_{10} > 0, \tau_1 \geq 0. \tag{2.7}$$

A priori, $\mu_j = \left(\mu_j^y, \mu_j^x\right)$'s are independent from other parameters and across j, and μ_j^y is independent of μ_j^x. Prior density for μ_j^x is bounded away from 0 on $\mathcal{X}$ and equal to 0 elsewhere. Prior density for μ_j^y is bounded below for some $a_{12}, \tau_2 > 0$ by

$$a_{11} \exp\left(-a_{12} ||\mu_j^y||^{\tau_2}\right), \tag{2.8}$$

and for some $a_{13}, \tau_3 > 0$ and all sufficiently large $r > 0$,

$$1 - \Pi\left(\mu_j^y \in [-r, r]^{d_y}\right) \leq \exp\left(-a_{13} r^{\tau_3}\right). \tag{2.9}$$

2.4. Results

To prove the main result, we adapt a general posterior contraction theorem to the case of conditional densities. We define the Hellinger, total variation, and Kullback–Leibler distances for conditional distributions as special cases of the corresponding distances for the joint densities. Therefore, the proof of the following result is essentially the same as the proof of Theorem 2.1 in Ghosal and van der Vaart (2001) and is omitted here.

THEOREM 2.1. *Let ϵ_n and $\tilde{\epsilon}_n$ be positive sequences with $\tilde{\epsilon}_n \leq \epsilon_n$, $\epsilon_n \to 0$, and $n\tilde{\epsilon}_n^2 \to \infty$, and c_1, c_2, c_3, and c_4 be some positive constants. Let ρ be d_h or d_1. Suppose $\mathcal{F}_n \subset \mathcal{F}$ is a sieve with the following bound on the metric entropy $J(\epsilon_n, \mathcal{F}_n, \rho)$*

$$\log J(\epsilon_n, \mathcal{F}_n, \rho) \leq c_1 n \epsilon_n^2, \tag{2.10}$$

$$\Pi(\mathcal{F}_n^c) \leq c_3 \exp\left\{-(c_2+4) n \tilde{\epsilon}_n^2\right\}, \tag{2.11}$$

and for a generalized Kullback–Leibler neighborhood

$$\mathcal{K}(f_0, \epsilon) = \left\{ f : \int f_0 g_0 \log(f_0/f) < \epsilon^2, \int f_0 g_0 [\log(f_0/f)]^2 < \epsilon^2 \right\},$$

$$\Pi(\mathcal{K}(f_0, \tilde{\epsilon}_n)) \geq c_4 \exp\left\{-c_2 n \tilde{\epsilon}_n^2\right\}. \tag{2.12}$$

Then, there exists $M > 0$ such that

$$\Pi\left(f : \rho(f, f_0) > M\epsilon_n | Y^n, X^n\right) \overset{P_0^n}{\to} 0.$$

Let us briefly discuss the assumptions of the theorem and their role in the proof. Condition (2.10) controls the size of the sieve $\mathcal{F}_n$ measured by the metric entropy. The left (right) hand side of the condition increases (decreases) as ϵ_n decreases and, thus, the condition provides a lower bound for the smallest posterior contraction rate the theorem can deliver. The condition implies the existence of a test ϕ^n of $f = f_0$ against $\{f \in \mathcal{F}_n : \rho(f, f_0) > M\epsilon_n\}$ with appropriately decreasing

errors of both types. This test is used in the proof to bound the expectation of the integrand in the numerator of

$$\Pi\left(f : \rho(f, f_0) > M\epsilon_n | Y^n, X^n\right) = \frac{\int_{f:\rho(f,f_0)>M\epsilon_n} \prod_i f(Y_i|X_i)/f_0(Y_i|X_i) d\Pi(f)}{\int \prod_i f(Y_i|X_i)/f_0(Y_i|X_i) d\Pi(f)} \tag{2.13}$$

multiplied by $(1-\phi^n)$ for $f \in \mathcal{F}_n$. A bound for the the remaining part of the numerator, where $f \notin \mathcal{F}_n$, is obtained from condition (2.11). Condition (2.12) requires the prior to put sufficient probability on the Kullback–Leibler neighborhoods of the true density. The left (right) hand side of the condition decreases (increases) as $\tilde{\epsilon}_n$ decreases and, thus, the condition provides a lower bound for the smallest contraction rate. In the proof of the theorem, (2.12) is used to show that the denominator in (2.13) has an appropriate lower bound with probability converging to 1.

Ghosal et al. (2000), who originally introduced a slightly more restrictive version of the theorem with $\epsilon_n = \tilde{\epsilon}_n$, argue that the theorem requires the prior to spread the mass on the model space almost "uniformly" in the following sense: Suppose all the distances are equivalent and $\mathcal{F}_n = \mathcal{F}$ so that (2.11) holds; then, (2.10) implies that the model space can be covered by $\exp(c_1 n\epsilon_n^2)$ balls of radius ϵ_n and (2.12) requires the probability of each ball to be comparable to $\exp(-c_1 n\epsilon_n^2)$. We refer the reader to Ghosal et al. (2000) for a further discussion.

THEOREM 2.2. *Under the assumptions in Sections 2.2 and 2.3, the sufficient conditions of Theorem 2.1 hold with*

$$\epsilon_n = n^{-\beta/(2\beta+d)}(\log n)^t,$$

where $t > t_0 + \max\{0, (1-\tau_1)/2\}$, $t_0 = (ds + \max\{\tau_1, 1, \tau_2/\tau\})/(2 + d/\beta)$, $d = d_y + d_x$, *and* $s = 1 + 1/\beta + 1/\tau$.

The proof of the theorem is divided into two main parts. First, we establish the prior thickness condition (2.12) in Theorem 3.1. Then, the conditions on the sieve are established in Theorems 4.1 and 4.2.

3. PRIOR THICKNESS

The prior thickness condition is formally proved in Theorem 3.1. Let us briefly describe the main steps of the proof placing it in the context of the previous literature. First, we recognize that the covariate dependent mixture defined in (2.3) is consistent with the following mixture of normals for the joint distribution of (y, x),

$$p(y, x|\theta, m) = \sum_{j=1}^{m} \alpha_j \phi_{\mu_j, \sigma}(y, x), \tag{3.1}$$

where $\mu_j = \left(\mu_j^y, \mu_j^x\right)$.

Second, we bound the Hellinger distance between conditional densities $f_0(y|x)$ and $p(y|x,\theta,m)$ by a distance between the joint densities $f_0(y|x)u(x)$ and $p(y,x|\theta,m)$, where $u(x)$ is a uniform density on $\mathcal{X}$. It is important to note that $f_0(y|x)u(x)$ has the same smoothness level as $f_0(y|x)$.

Third, we obtain a suitable approximation for the joint distribution $f_0(y|x)u(x)$ by mixtures $p(y,x|\theta,m)$ using modified results from Shen et al. (2013). The idea of the approximation argument is introduced in Rousseau (2010) in the context of approximation of a univariate density by mixtures of beta densities. Kruijer et al. (2010) use this idea for obtaining approximation results for mixtures of univariate normal densities. De Jonge and van Zanten (2010) extend the idea to approximation of multivariate functions, but the functions they approximate are not necessarily densities and their weights α_j's could be negative. Shen et al. (2013) use the same techniques with an additional step to approximate multivariate densities by mixtures with α_j's belonging to a simplex. It is not clear whether the mixing weights they obtain are actually non-negative. In Lemma A.3 in the appendix, we state a modified version of their Theorem 3 that ensures non-negativity of the weights. With a suitable approximation at hand, verification of condition (2.12) proceeds along the lines of similar results in Ghosal and van der Vaart (2001), Ghosal and van der Vaart (2007), Kruijer et al. (2010), and, especially, Shen et al. (2013), with modifications necessary to handle the case of conditional distributions.

THEOREM 3.1. *Suppose the assumptions from Sections 2.2 and 2.3 hold. Then, for any $C > 0$ and all sufficiently large n,*

$$\Pi(\mathcal{K}(f_0,\tilde{\epsilon}_n)) \geq \exp\left\{-Cn\tilde{\epsilon}_n^2\right\}, \tag{3.2}$$

where $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d)}(\log n)^t$, $t > (ds + \max\{\tau_1, 1, \tau_2/\tau\})/(2+d/\beta)$, $s = 1 + 1/\beta + 1/\tau$, and (τ, τ_1, τ_2) are defined in Sections 2.2 and 2.3.

Proof. By Lemma A.1, for $p(\cdot|\cdot,\theta,m)$ defined in (3.1),

$$\begin{aligned} d_h^2(f_0, p(\cdot|\cdot,\theta,m)) &= \int \left(\sqrt{f_0(y|x)} - \sqrt{p(y|x,\theta,m)}\right)^2 g_0(x)dydx \\ &\leq C_1 \int \left(\sqrt{f_0(y|x)u(x)} - \sqrt{p(y,x|\theta,m)}\right)^2 d(y,x) \\ &= C_1 d_H^2(f_0 u, p(\cdot|\theta,m)), \end{aligned} \tag{3.3}$$

where $u(x)$ is a uniform density on $\mathcal{X}$.

For $\sigma_n = [\tilde{\epsilon}_n/\log(1/\tilde{\epsilon}_n)]^{1/\beta}$, ε defined in (2.1), a sufficiently small $\delta > 0$, b and τ defined in (2.2), $a_0 = \{(8\beta + 4\varepsilon + 16)/(b\delta)\}^{1/\tau}$, $a_{\sigma_n} = a_0\{\log(1/\sigma_n)\}^{1/\tau}$, and $b_1 > \max\{1, 1/2\beta\}$ satisfying $\tilde{\epsilon}_n^{b_1}\{\log(1/\tilde{\epsilon}_n)\}^{5/4} \leq \tilde{\epsilon}_n$, the proof of Theorem 4 in Shen et al. (2013) implies the following three claims. First, there exists a partition of $\{z \in \mathcal{Z} : ||z|| \leq a_{\sigma_n}\}$, $\{U_j, j = 1, \ldots, K\}$ such that for $j = 1, \ldots, N$, U_j is a ball with diameter $\sigma_n\tilde{\epsilon}_n^{2b_1}$ and center $z_j = (x_j, y_j)$; for $j = N+1, \ldots, K$, U_j is a set

with a diameter bounded above by σ_n; $1 \le N < K \le C_2\sigma_n^{-d}\{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau}$, where $C_2 > 0$ does not depend on n. Second, there exist $\theta^\star = \{\mu_j^\star, \alpha_j^\star, j = 1, 2, \ldots; \sigma_n\}$ with $\alpha_j^\star = 0$ for $j > N$, $\mu_j^\star = z_j$ for $j = 1, \ldots, N$, and $\mu_j^\star \in U_j$ for $j = N+1, \ldots, K$ such that for $m = K$ and a positive constant C_3,

$$d_H(f_0 u, p(\cdot|\theta^\star, m)) \le C_3\sigma_n^\beta. \tag{3.4}$$

Third, there exists constant $B_0 > 0$ such that

$$P_0(\|z\| > a_{\sigma_n}) \le B_0\sigma_n^{4\beta+2\varepsilon+8}. \tag{3.5}$$

For θ in set

$$S_{\theta^\star} = \Big\{(\mu_j, \alpha_j,\ j = 1, 2, \ldots; \sigma) :\ \mu_j \in U_j, j = 1, \ldots, K,$$
$$\sum_{j=1}^{K} \left|\alpha_j - \alpha_j^\star\right| \le 2\tilde{\epsilon}_n^{2db_1},\ \min_{j=1,\ldots,K} \alpha_j \ge \tilde{\epsilon}_n^{4db_1}/2, \sigma^2 \in \left[\sigma_n^2/\left(1+\sigma_n^{2\beta}\right), \sigma_n^2\right]\Big\},$$

we have

$$\begin{aligned} d_H^2(p(\cdot|\theta^\star, m), p(\cdot|\theta, m)) &\le \left\| \sum_{j=1}^{K} \alpha_j^\star \phi_{\mu_j^\star, \sigma_n} - \sum_{j=1}^{K} \alpha_j \phi_{\mu_j, \sigma} \right\|_1 \\ &\le \sum_{j=1}^{K} \left|\alpha_j^\star - \alpha_j\right| \\ &\quad + \sum_{j=1}^{N} \alpha_j^\star \left[\left\| \phi_{\mu_j^\star, \sigma_n} - \phi_{\mu_j, \sigma_n} \right\|_1 + \left\| \phi_{\mu_j, \sigma_n} - \phi_{\mu_j, \sigma} \right\|_1 \right]. \end{aligned}$$

For $j = 1, \ldots, N$, $\left\| \phi_{\mu_j^\star, \sigma_n} - \phi_{\mu_j, \sigma_n} \right\|_1 \le \left\| \mu_j^\star - \mu_j \right\| / \sigma_n \le \tilde{\epsilon}_n^{2b_1}$. Also,

$$\begin{aligned} \left\| \phi_{\mu_j, \sigma_n} - \phi_{\mu_j, \sigma} \right\|_1 &\le \sqrt{d/2} \left| \frac{\sigma_n^2}{\sigma^2} - 1 - \log \frac{\sigma_n^2}{\sigma^2} \right|^{1/2} \\ &\le C_4 \sqrt{d/2} \left| \frac{\sigma_n^2}{\sigma^2} - 1 \right| \lesssim \sigma_n^{2\beta}, \end{aligned} \tag{3.6}$$

where the penultimate inequality follows from the fact that $|\log x - x + 1| \le C_4 |x-1|^2$ for x in a neighborhood of 1 and some $C_4 > 0$. Hence, $d_H(p(\cdot|\theta, m), p(\cdot|\theta^\star, m)) \lesssim \sigma_n^\beta$ and, by (3.3), (3.4) and the triangle inequality, $d_h(f_0, p(\cdot|\cdot, \theta, m)) \le C_5\sigma_n^\beta$ for some $C_5 > 0$, all $\theta \in S_{\theta^\star}$, and $m = K$.

Next, for $\theta \in S_{\theta^\star}$, let us consider a lower bound on the ratio $p(y|x, \theta, m)/f_0(y|x)$. Note that $\sup_{y,x} f_0(y|x) < \infty$ and $p(y|x, \theta, m) \ge \sigma^{d_x} p(y, x|\theta, m)$. For $z \in \mathcal{Z}$ with $\|z\| \le a_{\sigma_n}$, there exists $J \le K$ for which $||z - \mu_J|| \le \sigma_n$. Thus, for all sufficiently large n such that $\sigma_n^2/\sigma^2 \le 2$, $p(z|\theta, m) \ge \min_j \alpha_j \cdot \phi_{\mu_J, \sigma}(z) \ge [\tilde{\epsilon}_n^{4db_1}/2] \cdot \sigma_n^{-d} e^{-1}/(2\pi)^{d/2}$ and

$$\frac{p(y|x,\theta,m)}{f_0(y|x)} \geq C_6 \tilde{\epsilon}_n^{4db_1} \sigma_n^{-d_y}, \text{ for some } C_6 > 0. \tag{3.7}$$

For $z \in \mathcal{Z}$ with $\|z\| > a_{\sigma_n}$, $\left\|z - \mu_j\right\|^2 \leq 2(\|z\|^2 + \|\mu\|^2) \leq 4\|z\|^2$ for all $j = 1, \ldots, K$. Thus, for all sufficiently large n, $p(z|\theta,m) \geq \sigma_n^{-d} \exp(-4\|z\|^2/\sigma_n^2)/(2\pi)^{d/2}$ and

$$\frac{p(y|x,\theta,m)}{f_0(y|x)} \geq C_7 \sigma_n^{-d_y} \exp\left(-4\|z\|^2/\sigma_n^2\right), \text{ for some } C_7 > 0.$$

Denote the lower bound in (3.7) by λ_n and consider all sufficiently large n such that $\lambda_n < e^{-1}$. For any $\theta \in S_{\theta^\star}$,

$$\begin{aligned}
&\int \left(\log \frac{f_0(y|x)}{p(y|x,\theta,m)}\right)^2 1\left\{\frac{p(y|x,\theta,m)}{f_0(y|x)} < \lambda_n\right\} f_0(y|x) g_0(x) dy dx \\
&= \int \left(\log \frac{f_0(y|x)}{p(y|x,\theta,m)}\right)^2 1\left\{\frac{p(y|x,\theta,m)}{f_0(y|x)} < \lambda_n, ||(y,x)|| > a_{\sigma_n}\right\} f_0(y|x) g_0(x) dy dx \\
&\leq \frac{4}{\sigma_n^4} \int_{\|z\| > a_{\sigma_n}} \|z\|^4 f_0 g_0 dz \leq \frac{4}{\sigma_n^4} E_0(\|Z\|^8)^{1/2} (P_0(\|Z\| > a_{\sigma_n}))^{1/2} \leq C_8 \sigma_n^{2\beta+\varepsilon}
\end{aligned}$$

for some constant C_8. The last inequality follows from (3.5) and tail condition in (2.2). Also note that

$$\begin{aligned}
&\log \frac{f_0(y|x)}{p(y|x,\theta,m)} 1\left\{\frac{p(y|x,\theta,m)}{f_0(y|x)} < \lambda_n\right\} \\
&\leq \left\{\log \frac{f_0(y|x)}{p(y|x,\theta,m)}\right\}^2 1\left\{\frac{p(y|x,\theta,m)}{f_0(y|x)} < \lambda_n\right\}
\end{aligned}$$

and, thus,

$$\int \log \frac{f_0(y|x)}{p(y|x,\theta,m)} 1\left\{\frac{p(y|x,\theta,m)}{f_0(y|x)} < \lambda_n\right\} f_0 g_0 dz \leq C_8 \sigma_n^{2\beta+\varepsilon}.$$

By Lemma A.4, both $E_0(\log(f_0(Y|X)/p(Y|X,\theta,m)))$ and $E_0\left([\log(f_0(Y|X)/p(Y|X,\theta,m))]^2\right)$ are bounded by $C_9 \log(1/\lambda_n)^2 \sigma_n^{2\beta} \leq A\tilde{\epsilon}_n^2$ for some constant A.

Finally, we calculate a lower bound on the prior probability of $m = K$ and $\{\theta \in S_{\theta^\star}\}$. By (2.7), for some $C_{10} > 0$,

$$\begin{aligned}
\Pi(m = K) &\propto \exp\left[-a_{10} K (\log K)^{\tau_1}\right] \\
&\geq \exp\left[-C_{10} \tilde{\epsilon}_n^{-d/\beta} \{\log(1/\tilde{\epsilon}_n)\}^{d+d/\beta+d/\tau+\tau_1}\right].
\end{aligned} \tag{3.8}$$

From Lemma 10 of Ghosal and van der Vaart (2007), for some constants $C_{11}, C_{12} > 0$ and all sufficiently large n,

$$\Pi\left(\sum_{j=1}^{K}\left|\alpha_j-\alpha_j^\star\right|\geq 2\tilde{\epsilon}_n^{2db_1},\min_{j=1,\ldots,K}\alpha_j\geq\tilde{\epsilon}_n^{4db_1}/2\middle|m=K\right)$$
$$\geq\exp\left[-C_{11}K\log(1/\tilde{\epsilon}_n)\right]$$
$$\geq\exp\left[-C_{12}\tilde{\epsilon}_n^{-d/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{d/\beta+d/\tau+d+1}\right]. \tag{3.9}$$

For π_μ denoting the prior density of μ_j^y and some $C_{13},C_{14}>0$, (2.8) implies

$$\Pi(\mu_j\in U_j,j=1,\ldots,N)$$
$$\geq\left\{C_{13}\pi_\mu(a_\sigma)\mathrm{diam}(U_1)^d\right\}^N$$
$$\geq\exp\left[-C_{14}\tilde{\epsilon}_n^{-d/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{d+d/\beta+d/\tau+\max\{1,\tau_2/\tau\}}\right]. \tag{3.10}$$

Assumption (2.6) on the prior for σ, implies

$$\Pi\left(\sigma^{-2}\in\left\{\sigma_n^{-2},\sigma_n^{-2}\left(1+\sigma_n^{2\beta}\right)\right\}\right)$$
$$\geq a_8\sigma_n^{-2a_7}\sigma_n^{2\beta a_8}\exp\left\{-a_9\sigma_n^{-1}\right\}\geq\exp\left\{-C_{15}\sigma_n^{-1}\right\}. \tag{3.11}$$

It follows from (3.8)–(3.11), that for all sufficiently large n, $s=1+1/\beta+1/\tau$, and some $C_{16}>0$

$$\Pi(\mathcal{K}(f_0,A\tilde{\epsilon}_n))\geq\Pi(m=N,\theta_p\in S_{\theta_p})$$
$$\geq\exp\left[-C_{16}\tilde{\epsilon}_n^{-d/\beta}\left\{\log(1/\tilde{\epsilon}_n)\right\}^{ds+\max\{\tau_1,1,\tau_2/\tau\}}\right].$$

The last expression of the above display is bounded below by $\exp\{-Cn\tilde{\epsilon}_n^2\}$ for any $C>0$, $\tilde{\epsilon}_n=n^{-\beta/(2\beta+d)}(\log n)^t$, any $t>(ds+\max\{\tau_1,1,\tau_2/\tau\})/(2+d/\beta)$, and all sufficiently large n. Since the inequality in the definition of t is strict, the claim of the theorem follows immediately. ■

4. SIEVE CONSTRUCTION AND ENTROPY BOUNDS

For $H\in\mathbb{N}$, $0<\underline{\sigma}<\overline{\sigma}$, and $\overline{\mu},\underline{\alpha}>0$, let us define a sieve

$$\mathcal{F}=\{p(y|x,\theta,m):m\leq H,\ \alpha_j\geq\underline{\alpha},$$
$$\sigma\in[\underline{\sigma},\overline{\sigma}],\mu_j^y\in[-\overline{\mu},\overline{\mu}]^{d_y},j=1,\ldots,m\}. \tag{4.1}$$

In the following theorem, we bound the covering number of $\mathcal{F}$ in norm

$$d_{SS}(f_1,f_2)=\sup_{x\in\mathcal{X}}\|f_1(y|x)-f_2(y|x)\|_1.$$

THEOREM 4.1. *For* $0 < \epsilon < 1$ *and* $\underline{\sigma} \le 1$,

$$J(\epsilon, \mathcal{F}, d_{SS}) \le H \cdot \left\lceil \frac{16\overline{\mu} d_y}{\underline{\sigma}\epsilon} \right\rceil^{Hd_y} \cdot \left\lceil \frac{48 d_x}{\underline{\sigma}^2 \epsilon} \right\rceil^{Hd_x} \cdot H \left\lceil \frac{\log(\underline{\alpha}^{-1})}{\log(1+\epsilon/[12H])} \right\rceil^{H-1} \cdot \left\lceil \frac{\log(\overline{\sigma}/\underline{\sigma})}{\log(1+\underline{\sigma}^2\epsilon/[48\max\{d_x,d_y\}])} \right\rceil.$$

For $\underline{\alpha} \le 1/2$, *all sufficiently large* H, *large* $\overline{\sigma}$ *and small* $\underline{\sigma}$,

$$\Pi(\mathcal{F}^c) \le H^2 \exp\{-a_{13}\overline{\mu}^{\tau_3}\} + H^2 \underline{\alpha}^{a/H} + \exp\{-a_{10} H (\log H)^{\tau_1}\} + a_1 \exp\{-a_2 \underline{\sigma}^{-2a_3}\} + a_4 \exp\{-2a_5 \log \overline{\sigma}\}.$$

Proof. We will start with the first assertion. Fix a value of m. Define set $S^m_{\mu^y}$ to contain centers of $|S^m_{\mu^y}| = \lceil 16\overline{\mu} d_y/(\underline{\sigma}\epsilon)\rceil$ equal length intervals partitioning $[-\overline{\mu}, \overline{\mu}]$. Similarly, define set $S^m_{\mu^x}$ to contain centers of $|S^m_{\mu^x}| = \lceil 48 d_x/(\underline{\sigma}^2\epsilon)\rceil$ equal length intervals partitioning $[0,1]$.

For $N_\alpha = \lceil \log(\underline{\alpha}^{-1})/\log(1+\epsilon/(12m))\rceil$, define

$$Q_\alpha = \{\gamma_j,\ j = 1,\dots,N_\alpha : \gamma_1 = \underline{\alpha},\ (\gamma_{j+1}-\gamma_j)/\gamma_j = \epsilon/(12m),\ j = 1,\dots,N_\alpha - 1\}$$

and note that for any $\gamma \in [\underline{\alpha}, 1]$ there exists $j \le N_\alpha$ such that $0 \le (\gamma - \gamma_j)/\gamma_j \le \epsilon/(12m)$. Let $S^m_\alpha = \{(\tilde\alpha_1,\dots,\tilde\alpha_m) \in \Delta^{m-1} : \tilde\alpha_{j_k} \in Q_\alpha,\ 1 \le j_1 < j_2 < \cdots < j_{m-1} \le m\}$. Note that $|S^m_\alpha| \le m(N_\alpha)^{m-1}$. Let us consider an arbitrary $\alpha \in \Delta^{m-1}$. Since S^m_α is permutation invariant, we can assume without loss of generality that $\alpha_m \ge 1/m$. By definition of S^m_α, there exists $\tilde\alpha \in S^m_\alpha$ such that $0 \le (\alpha_j - \tilde\alpha_j)/\tilde\alpha_j \le \epsilon/(12m)$ for $j = 1,\dots,m-1$. Also,

$$\frac{|\alpha_m - \tilde\alpha_m|}{\min(\alpha_m, \tilde\alpha_m)} = \frac{|\alpha_m - \tilde\alpha_m|}{\alpha_m} = \frac{\sum_{j=1}^{m-1} \tilde\alpha_j (\alpha_j - \tilde\alpha_j)/\tilde\alpha_j}{\alpha_m} \le \frac{\epsilon}{12}.$$

Define $S_\sigma = \{\sigma^l, l = 1,\dots,N_\sigma = \lceil \log(\overline{\sigma}/\underline{\sigma})/(\log(1+\underline{\sigma}^2\epsilon/(48\max\{d_x,d_y\}))\rceil$, $\sigma^1 = \underline{\sigma}, (\sigma^{l+1} - \sigma^l)/\sigma^l = \underline{\sigma}^2\epsilon/(48\max\{d_x,d_y\})\}$. Then $|S_\sigma| = N_\sigma$.

Below we show that

$$S_{\mathcal{F}} = \{p(y|x,\theta,m) : m \le H,\ \alpha \in S^m_\alpha,\ \sigma \in S_\sigma,\ \mu^x_{jl} \in S^m_{\mu^x},\ \mu^y_{jk} \in S^m_{\mu^y},\ j \le m,\ l \le d_x,\ k \le d_y\}$$

provides an ϵ-net for $\mathcal{F}$ in d_{SS}. Fix $p(y|x,\theta,m) \in \mathcal{F}$ for some $m \le H, \alpha \in \Delta^{m-1}$ with $\alpha_j \ge \underline{\alpha}$, $\mu^x \in [0,1]^{d_x}$, $\mu^y \in [-\overline{\mu}, \overline{\mu}]^{d_y}$ and $\sigma \in [\underline{\sigma}, \overline{\sigma}]$ with $\sigma^l \le \sigma \le \sigma^{l+1}$. Find $\tilde\alpha \in S^m_\alpha$, $\tilde\mu^x_{jl} \in S^m_{\mu^x}$, $\tilde\mu^y_{jk} \in S^m_{\mu^y}$, and $\tilde\sigma = \sigma_l \in S_\sigma$ such that for all $j = 1,\dots,m$, $k = 1,\dots,d_y$, and $l = 1,\dots,d_x$

$$|\mu^y_{jk} - \tilde\mu^y_{jk}| \le \frac{\underline{\sigma}\epsilon}{16 d_y},\ |\mu^x_{jl} - \tilde\mu^x_{jl}| \le \frac{\underline{\sigma}^2\epsilon}{96 d_x},\ \frac{\alpha_j - \tilde\alpha_j}{\alpha_j} \le \frac{\epsilon}{12},\ \frac{|\sigma - \tilde\sigma|}{\sigma} \le \frac{\underline{\sigma}^2\epsilon}{48\max\{d_x,d_y\}}.$$

Let $K_j = \exp\{-0.5||x - \mu^x_j||^2/\sigma^2\}$. The proof of Proposition 3.1 in Norets and Pelenis (2014) implies the following inequality for any $x \in \mathcal{X}$ [1]

$$\int |p(y|x,\theta,m) - p(y|x,\tilde{\theta},m)|dy \le 2 \max_{j=1,\ldots,m} ||\phi_{\mu_j^y,\sigma} - \phi_{\tilde{\mu}_j^y,\tilde{\sigma}}||_1 + 2\left(\frac{\sum_{j=1}^m \alpha_j |K_j - \tilde{K}_j|}{\sum_{j=1}^m \alpha_j K_j} + \frac{\sum_{j=1}^m \tilde{K}_j |\alpha_j - \tilde{\alpha}_j|}{\sum_{j=1}^m \alpha_j K_j}\right).$$

It is easy to see that

$$||\phi_{\mu_j^y,\sigma} - \phi_{\tilde{\mu}_j^y,\tilde{\sigma}}||_1 \le 2\sum_{k=1}^{d_y}\left\{\frac{|\mu_{jk}^y - \tilde{\mu}_{jk}^y|}{\sigma \wedge \tilde{\sigma}} + \frac{|\sigma - \tilde{\sigma}|}{\sigma \wedge \tilde{\sigma}}\right\} \le \frac{\epsilon}{4}.$$

Also,

$$\begin{aligned}&\frac{\sum_{j=1}^m \alpha_j |K_j - \tilde{K}_j|}{\sum_{j=1}^m \alpha_j K_j} + \frac{\sum_{j=1}^m \tilde{K}_j |\alpha_j - \tilde{\alpha}_j|}{\sum_{j=1}^m \alpha_j K_j} \\ &\quad \le \max_j \frac{|K_j - \tilde{K}_j|}{K_j} + \max_j \frac{|\alpha_j - \tilde{\alpha}_j|}{\alpha_j} + \max_j \frac{|K_j - \tilde{K}_j||\alpha_j - \tilde{\alpha}_j|}{\alpha_j K_j}.\end{aligned}$$

Since $|\alpha_j - \tilde{\alpha}_j|/\alpha_j \le \epsilon/12$ and $\epsilon < 1$, the above display is bounded by $\epsilon/4$ if we can show $|K_j - \tilde{K}_j|/K_j \le \epsilon/12$. Observe that

$$\begin{aligned}&\left|\frac{||x-\mu_j^x||^2}{2\sigma^2} - \frac{||x-\tilde{\mu}_j^x||^2}{2\tilde{\sigma}^2}\right| \\ &\quad \le \frac{1}{2}\left|\frac{1}{\sigma^2} - \frac{1}{\tilde{\sigma}^2}\right| \left\|x-\mu_j^x\right\|^2 + \frac{1}{2\tilde{\sigma}^2}\left|\left\|x-\mu_j^x\right\|^2 - \left\|x-\tilde{\mu}_j^x\right\|^2\right| \\ &\quad \le \frac{||x-\mu_j^x||^2\,|(\sigma-\tilde{\sigma})/\sigma|}{\underline{\sigma}^2} + \frac{||\mu_j^x - \tilde{\mu}_j^x||\left(2||x|| + ||\mu_j^x|| + ||\tilde{\mu}_j^x||\right)}{2\underline{\sigma}^2} \\ &\quad \le \frac{\epsilon}{48} + \frac{\epsilon}{48} = \frac{\epsilon}{24},\end{aligned} \tag{4.2}$$

where the penultimate inequality follows from $||x-\mu_j^x||^2 \le d_x$, $2||x|| + ||\mu_j^x|| + ||\tilde{\mu}_j^x|| \le 4d_x^{1/2}$, and $||\mu_j^x - \tilde{\mu}_j^x|| \le d_x^{1/2}\max_l |\mu_{jl}^x - \tilde{\mu}_{jl}^x|$. Now since $|1-e^x| < 2|x|$ for $|x| < 1$,

$$\begin{aligned}\frac{\left|K_j - \tilde{K}_j\right|}{K_j} &= \left|1 - \exp\left\{\frac{||x-\mu_j^x||^2}{2\sigma^2} - \frac{||x-\tilde{\mu}_j^x||^2}{2\tilde{\sigma}^2}\right\}\right| \\ &\le 2\left|\frac{||x-\mu_j^x||^2}{2\sigma^2} - \frac{||x-\tilde{\mu}_j^x||^2}{2\tilde{\sigma}^2}\right| \le \frac{\epsilon}{12}.\end{aligned} \tag{4.3}$$

This concludes the proof for the covering number.

Next, let us obtain an upper bound for $\Pi(\mathcal{F}^c)$. From the assumptions in Section 2.3,

$$\Pi(\exists\, j \in \{1,\dots,m\}, \text{ s.t. } \mu_j^y \in [-\overline{\mu},\overline{\mu}]^{d_y}) \le m\exp(-a_{13}\overline{\mu}^{\tau_3}).$$

For all sufficiently large H,

$$\Pi(m > H) = C_1 \sum_{i=H+1}^{\infty} e^{-a_{10} i (\log i)^{\tau_1}} \le C_1 \int_H^{\infty} e^{-a_{10} r (\log H)^{\tau_1}} dr \le e^{-a_{10} H (\log H)^{\tau_1}}.$$

Observe that $\alpha_j | m \sim \text{Beta}(a/m, a(m-1)/m)$. Considering separately $a(m-1)/m - 1 < 0$ and $a(m-1)/m - 1 \ge 0$, it is easy to see that $(1-q)^{a(m-1)/m-1} \le 2$ for any $q \in [0, \underline{\alpha}]$ and $\underline{\alpha} \le 1/2$. Thus,

$$\begin{aligned}
\Pi(\alpha_j < \underline{\alpha} | m) &= \frac{\Gamma(a)}{\Gamma(a/m)\Gamma(a(m-1)/m)} \int_0^{\underline{\alpha}} q^{a/m-1}(1-q)^{a(m-1)/m-1} dq \\
&\le \frac{\Gamma(a)}{\Gamma(a/m)\Gamma(a(m-1)/m)} 2 \int_0^{\underline{\alpha}} q^{a/m-1} dq \\
&= \frac{\Gamma(a) 2 \underline{\alpha}^{a/m}}{\Gamma(a/m+1)\Gamma(a(m-1)/m)} \le e^2 2\Gamma(a+1)\underline{\alpha}^{a/m} = C(a)\underline{\alpha}^{a/m}, \qquad \textbf{(4.4)}
\end{aligned}$$

where the final inequality is implied by the following facts: $\Gamma(a/m+1) \ge \int_1^{\infty} q^{a/m} e^{-q} dq \ge e^{-1}$ and $\Gamma(a(m-1)/m) \ge \int_0^1 q^{a(m-1)/m-1} e^{-q} dq \ge me^{-1}/a(m-1)$.

Consider $\Pi(\sigma \notin [\underline{\sigma}, \overline{\sigma}]) = \Pi(\sigma^{-1} \ge \underline{\sigma}^{-1}) + \Pi(\sigma^{-1} \le \overline{\sigma}^{-1})$. Since the prior for σ satisfies (2.4) and (2.5), for sufficiently large $\overline{\sigma}$ and small $\underline{\sigma}$

$$\begin{aligned}
&\Pi(\sigma^{-1} \ge \underline{\sigma}^{-1}) \le a_1 \exp\left\{-a_2 \underline{\sigma}^{-2a_3}\right\}, \\
&\Pi(\sigma^{-1} \le \overline{\sigma}^{-1}) \le a_4 \overline{\sigma}^{-2a_5} = a_4 \exp\{-2a_5 \log \overline{\sigma}\}. \qquad \textbf{(4.5)}
\end{aligned}$$

Now observe that

$$\begin{aligned}
\Pi(\mathcal{F}^c) &\le \Pi\left(\exists m \le H, \exists j \le m, \text{ s.t. } \mu_j^y \notin \{[-\overline{\mu},\overline{\mu}]^{d_y}\}^c\right) + \Pi(m > H) \\
&\quad + \Pi(\sigma \notin [\underline{\sigma}, \overline{\sigma}]) + \Pi\left(\exists m \le H, \exists j \le m, \text{ s.t. } \alpha_j < \underline{\alpha} \middle| m\right) \\
&\le \sum_{m=1}^{H} m \Pi\left(\mu_j^y \notin \{[-\overline{\mu},\overline{\mu}]^{d_y}\}^c\right) \\
&\quad + \sum_{m=1}^{H} m \Pi(\alpha_j < \underline{\alpha} | m) + \Pi(m > H) + \Pi(\sigma \notin [\underline{\sigma}, \overline{\sigma}]) \\
&\le \frac{H(H+1)}{2} \exp\{-a_{13}\overline{\mu}^{\tau_3}\} + \frac{H(H+1)}{2} C(a)\underline{\alpha}^{a/H} \\
&\quad + \exp\{-0.5 a_{10} H (\log H)^{\tau_1}\} + \Pi(\sigma \notin [\underline{\sigma}, \overline{\sigma}]) \\
&\le H^2 \exp\{-a_{13}\overline{\mu}^{\tau_3}\} + H^2 \underline{\alpha}^{a/H} \\
&\quad + \exp\{-a_{10} H (\log H)^{\tau_1}\} + \Pi(\sigma \notin [\underline{\sigma}, \overline{\sigma}]). \qquad \blacksquare
\end{aligned}$$

THEOREM 4.2. *For $n \ge 1$, let $\epsilon_n = n^{-\beta/(2\beta+d)}(\log n)^t$, $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d)}(\log n)^{t_0}$ for $t_0 > (ds + \max\{\tau_1, 1, \tau_2/\tau\})/(2 + d/\beta)$ and define*

$\mathcal{F}_n$ as in (4.1) with $\epsilon = \epsilon_n, H = n\epsilon_n^2/(\log n)$, $\underline{\alpha} = e^{-nH}$, $\underline{\sigma} = n^{-1/(2a_3)}$, $\overline{\sigma} = e^n$, and $\overline{\mu} = n^{1/\tau_3}$. Then for all $t > t_0 + \max\{0, (1-\tau_1)/2\}$, and some constants $c_1, c_3 > 0$ and every $c_2 > 0$, $\mathcal{F}_n$ satisfies (2.10) and (2.11) for all large n.

Proof. Since $d_1 \le d_{SS}$ and $d_h \le d_1^2$, Theorem 4.1 implies

$$\log J(\epsilon_n, \mathcal{F}_n, \rho) \le c_1 H \log n = c_1 n\epsilon_n^2.$$

Also,

$$\Pi(\mathcal{F}_n^c) \le H^2 \exp\{-a_{13}n\} + H^2 \exp\{-an\} + \exp\{-a_{10}H(\log H)^{\tau_1}\} + a_1 \exp\{-a_2 n\} + a_4 \exp\{-2a_5 n\}.$$

Hence, $\Pi(\mathcal{F}_n^c) \le e^{-(c_2+4)n\tilde{\epsilon}_n^2}$ for any c_2 if $\epsilon_n^2 (\log n)^{\tau_1 - 1}/\tilde{\epsilon}_n^2 \to \infty$, which holds for $t > t_0 + \max\{0, (1-\tau_1)/2\}$. ■

5. UNBOUNDED COVARIATE SPACE

The assumption of bounded covariate space $\mathcal{X}$ in Section 2 could be restrictive in some applications. In this section, we consider a generalization of our result to the case when the covariate space is possibly unbounded. We re-formulate the assumptions on the data generating process and the prior distributions below.

5.1. Assumptions about data generating process

Let $\mathcal{X} \subset \mathbb{R}^{d_x}$. First, let us assume that there exist a constant $\eta > 0$ and a probability density function $\bar{g}_0(x)$ with respect to the Lebesgue measure such that $\eta \bar{g}_0(x) \ge g_0(x)$ for all $x \in \mathcal{X}$ and $\tilde{f}_0(y,x) = f_0(y|x)\bar{g}_0(x) \in \mathcal{C}^{\beta, L, \tau_0}$. Second, we assume g_0 satisfies

$$\int e^{\kappa \|x\|^2} g_0(x) dx \le B < \infty \tag{5.1}$$

for some constant $\kappa > 0$. Third, $\tilde{f}_0(y,x)$ is assumed to satisfy

$$\int_{\mathcal{Z}} \left| \frac{D^k \tilde{f}_0(y,x)}{\tilde{f}_0(y,x)} \right|^{(2\beta+\varepsilon)/k} \tilde{f}_0(y,x) dy dx < \infty,$$

$$\int_{\mathcal{Z}} \left| \frac{L(y,x)}{\tilde{f}_0(y,x)} \right|^{(2\beta+\varepsilon)/\beta} \tilde{f}_0(y,x) dy dx < \infty \tag{5.2}$$

for all $k \le \lfloor \beta \rfloor$ and some $\varepsilon > 0$. Finally, for all sufficiently large $(y,x) \in \mathcal{Y} \times \mathcal{X}$ and some positive (c, b, τ),

$$\tilde{f}_0(y,x) \le c \exp(-b||(y,x)||^{\tau}). \tag{5.3}$$

Let us elaborate on how the above assumptions allow for f_0 of smoothness level β. First of all, the original assumptions on the data generating process for the bounded $\mathcal{X}$ are a special case of the assumptions here with $\bar{g}_0$ being a uniform

density on $\mathcal{X}$. Second, when covariate density g_0 has a higher smoothness level than β and $\bar{g}_0 = g_0$, the assumption $\tilde{f}_0 \in \mathcal{C}^{\beta,L,\tau_0}$ essentially restricts the smoothness of f_0 only. Finally, when g_0 has a lower smoothness level than β, then our assumptions require existence of a sufficiently smooth and well behaved upper bound on g_0 in addition to f_0 having smoothness level β.

5.2. Prior

The assumption on the prior for μ_j^x is the only part of the prior from Section 2.3 that we need to modify here. Similarly to the prior on μ_j^y, we assume that the prior density for μ_j^x is bounded below for some $a_{14}, a_{15}, \tau_4 > 0$ by

$$a_{14} \exp\left(-a_{15}||\mu_j^x||^{\tau_4}\right),$$

and for some $a_{16}, \tau_5 > 0$ and all sufficiently large $r > 0$,

$$1 - \Pi\left(\mu_j^x \in [-r,r]^{d_x}\right) \geq \exp\left(-a_{16} r^{\tau_5}\right).$$

COROLLARY 5.1. *Under the assumptions in Sections 5.1 and 5.2, the posterior contracts at the rate specified in Theorem 2.2.*

Proof. We will first show that an analog of the *prior thickness* result from Theorem 3.1 holds with the same choice of $\tilde{\epsilon}_n$. By Corollary A.2 in the appendix, for $p(\cdot|\cdot,\theta,m)$ defined in (3.1),

$$d_h^2(f_0, p(\cdot|\cdot,\theta,m)) \leq C_1 d_H^2(f_0 \bar{g}_0, p(\cdot|\theta,m)). \tag{5.4}$$

Since the joint density $f_0 \bar{g}_0$ satisfies the assumptions of Theorem 4 in Shen et al. (2013), the rest of the proof of the prior thickness result is exactly the same as the proof of Theorem 3.1 except for π_μ in (3.10) would now denote the joint prior density of $\left(\mu_j^y, \mu_j^x\right)$.

Next, we will construct an appropriate *sieve*. For sequences $\underline{\alpha}, \overline{\mu}, \overline{\mu}^x, H, \underline{\sigma}, \overline{\sigma}$ to be chosen later, define:

$$\mathcal{F} = \{p(y|x,\theta,m) : m \leq H,\ \alpha_j \geq \underline{\alpha},\ \sigma \in [\underline{\sigma}, \overline{\sigma}],$$
$$\mu_j^y \in [-\overline{\mu}, \overline{\mu}]^{d_y}, \mu_j^x \in [-\overline{\mu}^x, \overline{\mu}^x]^{d_x},\ j = 1, \ldots, m\}.$$

The choice of $S_{\mathcal{F}}$, an ϵ-net of $\mathcal{F}$ is the same as in the proof of Theorem 4.1 with the following modifications. $S_{\mu^x}^m$ now contains centers of $|S_{\mu^x}^m| = \lceil 192 d_x (\overline{\mu}^x)^2/(\underline{\sigma}^2 \epsilon)\rceil$ equal length intervals partitioning $[-\overline{\mu}^x, \overline{\mu}^x]$. We also need an adjustment to S_σ here:

$$S_\sigma = \left\{\sigma^l, l = 1, \ldots, N_\sigma = \left\lceil \log(\overline{\sigma}/\underline{\sigma})/\left(\log\left(1 + \underline{\sigma}^2 \epsilon/\left(384(\overline{\mu}^x)^2 \max\{d_x, d_y\}\right)\right)\right\rceil, \sigma^1 = \underline{\sigma},\right.\right.$$
$$\left.\left.(\sigma^{l+1} - \sigma^l)/\sigma^l = \underline{\sigma}^2 \epsilon/\left(384(\overline{\mu}^x)^2 \max\{d_x, d_y\}\right)\right)\right\}.$$

Since we are dealing with possibly unbounded $\mathcal{X}$ here, we will find the covering number of $\mathcal{F}$ in d_1 instead of d_{SS}. The only part different from

the proof of Theorem 4.1 is the treatment of $\left|K_j - \tilde{K}_j\right|/K_j$. To show that $\int \left|K_j - \tilde{K}_j\right|/K_j g_0(x)dx \leq \epsilon/12$, we divide the range of integration into two parts: $\mathcal{X}_1 = \{x \in \mathcal{X} : |x_l| \leq \overline{\mu}^x, l = 1, \ldots, d_x\}$ and $\mathcal{X} \setminus \mathcal{X}_1$.

For $x \in \mathcal{X}_1$, the same argument as in the bounded covariate space (inequalities in (4.2) and (4.3)) combined with $\|x\| \leq d_x^{1/2}\overline{\mu}^x$, $||x - \mu_j^x||^2 \leq 4(\overline{\mu}^x)^2 d_x$, $2||x|| + ||\mu_j^x|| + ||\tilde{\mu}_j^x|| \leq 4\overline{\mu}^x d_x^{1/2}$, $|\sigma - \tilde{\sigma}|/\sigma \leq \underline{\sigma}^2\epsilon/(384(\overline{\mu}^x)^2 d_x)$ and $|\mu_{jl} - \tilde{\mu}_{jl}| \leq \epsilon\underline{\sigma}^2/(192 d_x \overline{\mu}^x)$ imply $\left|K_j - \tilde{K}_j\right|/K_j \leq \epsilon/24$.

For $x \in \mathcal{X} \setminus \mathcal{X}_1$, the left hand side of (4.2) is bounded above by

$$\left(2||x||^2 + 2||\mu_j^x||^2\right)\epsilon/\left(384(\overline{\mu}^x)^2 d_x\right) + \left(||x|| + d_x^{1/2}\overline{\mu}^x\right) d_x^{1/2}\epsilon/(192\overline{\mu}^x d_x) \quad \textbf{(5.5)}$$

$$\leq \epsilon/96 + ||x||^2\epsilon/\left(192(\overline{\mu}^x)^2 d_x\right) + ||x||\epsilon/\left(192\overline{\mu}^x d_x^{1/2}\right) \quad \textbf{(5.6)}$$

$$\leq \epsilon/96 + ||x||^2\epsilon/\left(96\overline{\mu}^x d_x^{1/2}\right), \quad \textbf{(5.7)}$$

where the last inequality holds for $\overline{\mu}^x \geq 1$ as $||x|| \geq \overline{\mu}^x$ for $x \in \mathcal{X} \setminus \mathcal{X}_1$. Since $|1 - e^r| \leq e^{|r|}$ for all $r \in \mathbb{R}$,

$$\left|\frac{K_j - \tilde{K}_j}{K_j}\right| \leq \exp\left(\frac{\epsilon}{96} + \frac{\epsilon\,\|x\|^2}{96 d_x^{1/2}\overline{\mu}^x}\right), \ \forall x \in \mathcal{X} \setminus \mathcal{X}_1.$$

Now,

$$\begin{aligned}\int \left|\frac{K_j - \tilde{K}_j}{K_j}\right| g_0(x)dx &\leq \int_{\mathcal{X}_1} \left|\frac{K_j - \tilde{K}_j}{K_j}\right| g_0(x)dx + \int_{\mathcal{X}\setminus\mathcal{X}_1} \left|\frac{K_j - \tilde{K}_j}{K_j}\right| g_0(x)dx \\ &\leq \frac{\epsilon}{24} + \exp\left(\frac{\epsilon}{96}\right)\int_{\mathcal{X}\setminus\mathcal{X}_1} \exp\left(\frac{\epsilon\,\|x\|^2}{96 d_x^{1/2}\overline{\mu}^x}\right) g_0(x)dx \\ &\leq \frac{\epsilon}{24} + \exp\left(\frac{\epsilon}{96}\right)\int_{\mathcal{X}\setminus\mathcal{X}_1} \exp\left(-\kappa_\epsilon\,\|x\|^2\right)\exp\left(\kappa\,\|x\|^2\right) g_0(x)dx,\end{aligned}$$

where $\kappa_\epsilon = \kappa - \epsilon/(96 d_x^{1/2}\overline{\mu}^x) \geq \kappa/2$ for small ϵ and large $\overline{\mu}^x$. Since $\|x\| \geq \overline{\mu}^x$ in $\mathcal{X} \setminus \mathcal{X}_1$, we have

$$\begin{aligned}\int \left|\frac{K_j - \tilde{K}_j}{K_j}\right| g_0(x)dx &\leq \frac{\epsilon}{24} + \exp\left(\frac{\epsilon}{96}\right)\exp\left(-\kappa(\overline{\mu}^x)^2/2\right)\int_{\mathcal{X}\setminus\mathcal{X}_1} \exp\left(\kappa\,\|x\|^2\right) g_0(x)dx \\ &\leq \frac{\epsilon}{24} + B\exp\left(\frac{\epsilon}{96}\right)\exp\left(-\kappa(\overline{\mu}^x)^2/2\right),\end{aligned}$$

where B is defined in (5.1). For $(\overline{\mu}^x)^2 \geq -(2/\kappa)\log\{\epsilon e^{-\epsilon/96}/24B\}$,

$$\int \left|\frac{K_j - \tilde{K}_j}{K_j}\right| g_0(x)dx \leq \frac{\epsilon}{12}.$$

Hence for $(\overline{\mu}^x)^2 \geq -(2/\kappa)\log\{\epsilon e^{-\epsilon/96}/24B\}$, following the proof of Theorem 4.1 we obtain,

$$J(\epsilon,\mathcal{F},d_1) \le H \cdot \left\lceil \frac{16\overline{\mu}d_y}{\underline{\sigma}\epsilon} \right\rceil^{Hd_y} \cdot \left\lceil \frac{192 d_x (\overline{\mu}^x)^2}{\underline{\sigma}^2\epsilon} \right\rceil^{Hd_x} \cdot H \left\lceil \frac{\log(\underline{\alpha}^{-1})}{\log(1+\epsilon/[12H])} \right\rceil^{H-1} \cdot \left\lceil \frac{\log(\overline{\sigma}/\underline{\sigma})}{\log(1+\underline{\sigma}^2\epsilon/[384(\overline{\mu}^x)^2\max\{d_x,d_y\}])} \right\rceil.$$

Observe that $\Pi(\mathcal{F}^c)$ is bounded above by

$$H^2\exp\{-a_{13}\overline{\mu}^{\tau_3}\} + H^2\exp\{-a_{16}(\overline{\mu}^x)^{\tau_5}\} + H^2\underline{\alpha}^{a/H} + \exp\{-a_{10}H(\log H)^{\tau_1}\} + \Pi(\sigma \notin [\underline{\sigma},\overline{\sigma}]).$$

The rest of the proof follows the argument in the proof of Theorem 4.2 with the same sequences and $\overline{\mu}^x = n^{1/\tau_5}$. ■

6. IRRELEVANT COVARIATES

In applications, researchers often tackle the problem of selecting a set of relevant covariates for regression or conditional distribution estimation. In the Bayesian framework, this is usually achieved by introducing latent indicator variables for inclusion of covariates in the model, see, for example, Bhattacharya, Pati, and Dunson (2014), Shen and Ghosal (2016), Yang and Tokdar (2015). This is equivalent to a Bayesian model averaging procedure, where every possible subset of covariates represents a model. It is straightforward to extend the results of the previous sections to a model with latent indicator variables for covariate inclusion and show that the posterior contraction rate will not be affected by the irrelevant covariates. In this section, we show that even without introduction of the indicator variables, irrelevant covariates do not affect the posterior contraction rate in a version of our model with component specific scale parameters.

Let $\theta = \{\mu_j^y, \mu_j^x, \alpha_j, j = 1,2,\ldots;\ \sigma^y = (\sigma_1^y,\ldots,\sigma_{d_y}^y), \sigma^x = (\sigma_1^x,\ldots,\sigma_{d_x}^x)\}$ and

$$p(y|x,\theta,m) = \sum_{j=1}^{m} \frac{\alpha_j \exp\left\{-0.5\sum_k \left(x_k-\mu_{jk}^x\right)^2/\left(\sigma_k^x\right)^2\right\}}{\sum_{i=1}^m \alpha_i \exp\left\{-0.5\sum_k \left(x_k-\mu_{ik}^x\right)^2/\left(\sigma_k^x\right)^2\right\}} \phi_{\mu_j^y,\sigma^y}(y).$$

Suppose f_0 depends only on the first $d_x^0 < d_x$ covariates $x_{1d_x^0} = (x_1,\ldots,x_{d_x^0})$ with the marginal density $g_{1d_x^0}$. Let us assume conditions (5.1)–(5.3) from Section 5.1 with the following change in the definition of $\tilde{f}_0$: for $\eta > 0$ and a probability density function $\bar{g}_{1d_x^0}(x_{1d_x^0})$ with respect to the Lebesgue measure such that $\eta\bar{g}_{1d_x^0}(x_{1d_x^0}) \ge g_{1d_x^0}(x_{1d_x^0})$ for all $x \in \mathcal{X}$ and $\tilde{f}_0(y,x_{1d_x^0}) = f_0(y|x_{1d_x^0})\bar{g}_{1d_x^0}(x_{1d_x^0}) \in \mathcal{C}^{\beta,L,\tau_0}$. In addition, let us assume that the tail condition (5.3) holds for $f_0 g_0$.

For $l = 1,\ldots,d_y$ and $k = 1,\ldots,d_x$, σ_l^y and σ_k^x are assumed to be independent a priori with densities satisfying (2.4)–(2.6). Other parts of the prior are assumed to be the same as in Section 5.2.

Let us briefly explain why we introduce component specific scale parameters. Our proof of the following corollary exploits the fact that when $\mu_{jk}^x = 0$ for $k > d_x^0$

and all j, covariates x_k for $k > d_x^0$ do not enter the model, and for μ_{jk}^x near zero and large σ_k^x for $k > d_x^0$ this holds approximately. At the same time, approximation arguments in the prior thickness results require σ_k^x very close to zero for $k \le d_x^0$. Thus, we need to allow scale parameters for relevant and irrelevant covariates to take different values (our assumption of different scale parameters for components of y is not essential for the result).

COROLLARY 6.1. *Under the assumptions of this section, the posterior contracts at the rate specified in Theorem 2.2 with $d = d_y + d_x$ replaced by $d^0 = d_y + d_x^0$.*

Proof. First, consider the prior thickness result in Theorem 3.1. For any θ let

$$\theta_{d_x^0} = \left\{\mu_j^y, \mu_{j1}^x, \ldots, \mu_{jd_x^0}^x, \alpha_j, j = 1, 2, \ldots;\ \sigma^y, \sigma_1^x, \ldots, \sigma_{d_x^0}^x\right\}$$

and define $p(\cdot|\cdot, \theta_{d_x^0}, m)$ and $p(\cdot|\theta_{d_x^0}, m)$ as before with $(y, x_{1d_x^0})$ as the arguments. By the triangle inequality,

$$d_h(f_0, p(\cdot|\cdot, \theta, m)) \le d_h(f_0, p(\cdot|\cdot, \theta_{d_x^0}, m)) + d_h(p(\cdot|\cdot, \theta_{d_x^0}, m), p(\cdot|\cdot, \theta, m)).$$

By Corollary A.2 in the Appendix, $d_h(f_0, p(\cdot|\cdot, \theta_{d_x^0})) \le C_1 d_H(f_0 \bar{g}_{1d_x^0}, p(\cdot|\theta_{d_x^0}))$. By the argument leading to (3.4) and (3.5), there exist $\theta_{d_x^0}^\star$ such that $d_H(f_0 \bar{g}_{1d_x^0}, p(\cdot|\theta_{d_x^0}^\star, m)) \le C_2 \sigma_n^\beta$, $P_0(\|y, x\| > a_{\sigma_n}) \le B_0 \sigma_n^{4\beta+2\epsilon+8}$, z_j and U_j are defined on the space for $(y, x_{1d_x^0})$, and $1 \le N < K \le C_2 \sigma_n^{-d^0} \{\log(1/\tilde{\epsilon}_n)\}^{d^0 + d^0/\tau}$. Let

$$\begin{aligned} S_{\theta^\star} = \Big\{ & (\mu_j, \alpha_j,\ j = 1, 2, \ldots; \sigma^y, \sigma^x) :\ \left(\mu_j^y, \mu_{j1}^x, \ldots, \mu_{jd_x^0}^x\right) \in U_j, \\ & \left\| \left(\mu_{jd_x^0+1}^x, \ldots, \mu_{jd_x}^x\right) \right\| \le \sigma_n \tilde{\epsilon}_n^{2b_1},\ j \le K; \\ & \sum_{j=1}^{K} \left|\alpha_j - \alpha_j^\star\right| \le 2\tilde{\epsilon}_n^{2d^0 b_1},\ \min_{j=1,\ldots,K} \alpha_j \ge \tilde{\epsilon}_n^{4d^0 b_1}/2; \\ & \left(\sigma_k^x\right)^2, \left(\sigma_l^y\right)^2 \in \left[\sigma_n^2/(1+\sigma_n^{2\beta}), \sigma_n^2\right],\ l \le d_y,\ k \le d_x^0; \\ & \left(\sigma_k^x\right)^2 \in \left[a_{\sigma_n}^2, 2a_{\sigma_n}^2\right],\ k = d_x^0 + 1, \ldots, d_x \Big\}. \end{aligned}$$

For $\theta \in S_{\theta^\star}$ and $m = K$, as in the proof of Theorem 3.1,

$$\begin{aligned} d_H\left(f_0 \bar{g}_{1d_x^0}, p\left(\cdot|\theta_{d_x^0}, m\right)\right) & \le d_H\left(f_0 \bar{g}_{1d_x^0}, p\left(\cdot|\theta_{d_x^0}^*, m\right)\right) \\ & \quad + d_H\left(p\left(\cdot|\theta_{d_x^0}, m\right), p\left(\cdot|\theta_{d_x^0}^*, m\right)\right) \le C_3 \sigma_n^\beta. \quad \textbf{(6.1)} \end{aligned}$$

Next, we tackle $d_h(p(\cdot|\cdot, \theta_{d_x^0}, m), p(\cdot|\cdot, \theta, m))$. Following the entropy calculations in Theorem 4.1, we have for $m = K$,

$$d_h^2(p(\cdot|\cdot,\theta_{d_x^0},m),p(\cdot|\cdot,\theta,m)) \le \int \max_{1\le j\le K} |K_j-\tilde{K}_j|/\left|K_j\right| g_0(x)dx,$$

where

$$K_j = \exp\left\{-\sum_{k=1}^{d_x}\frac{\left(x_k-\mu_{jk}^x\right)^2}{2\left(\sigma_k^x\right)^2}\right\},$$

$$\tilde{K}_j = \exp\left\{-\sum_{k=1}^{d_x^0}\frac{\left(x_k-\mu_{jk}^x\right)^2}{2\left(\sigma_k^x\right)^2}-\sum_{k=d_x^0+1}^{d_x}\frac{x_k^2}{2\left(\sigma_k^x\right)^2}\right\},$$

and $\tilde{K}_j$ is normalized in a convenient way. To show that $\int |K_j-\tilde{K}_j|/K_j g_0(x)dx \le 2\sigma_n^{2\beta}$, we divide the range of integration into two parts: $\mathcal{X}_1 = \{x\in\mathcal{X}: |x_l|\le A_n, l=d_x^0+1,\dots,d_x\}$ and $\mathcal{X}\setminus\mathcal{X}_1$, where $A_n = a_{\sigma_n}^2\log(B/\sigma_n^{2\beta})$ and B is defined in assumption (5.1). Observe that for $\theta\in S_{\theta^\star}$, $x\in\mathcal{X}_1$, and all sufficiently large n,

$$\left|\sum_{k=d_x^0+1}^{d_x}\frac{(2x_k-\mu_{jk}^x)(-\mu_{jk}^x)}{2(\sigma_k^x)^2}\right| \le \frac{2A_n(d_x-d_x^0)\sigma_n\tilde{\epsilon}_n^{2b_1}}{a_{\sigma_n}^2} \le 1$$

and, hence, using $|1-e^r|\le|r|$ for $|r|\le 1$, we obtain for $\theta\in S_{\theta^\star}$ and $x\in\mathcal{X}_1$,

$$\left|\frac{K_j-\tilde{K}_j}{K_j}\right| = \left|1-\exp\left\{\sum_{k=d_x^0+1}^{d_x}\frac{(2x_k-\mu_{jk}^x)(-\mu_{jk}^x)}{2(\sigma_k^x)^2}\right\}\right|$$

$$\le \frac{2A_n(d_x-d_x^0)\sigma_n\tilde{\epsilon}_n^{2b_1}}{a_{\sigma_n}^2} \le \sigma_n^{2\beta}. \tag{6.2}$$

For $x\in\mathcal{X}\setminus\mathcal{X}_1$ using $|1-e^r|\le e^{|r|}$,

$$\int_{\mathcal{X}\setminus\mathcal{X}_1}\max_{1\le j\le K}\left|\frac{K_j-\tilde{K}_j}{K_j}\right| g_0(x)dx \le \int_{\mathcal{X}\setminus\mathcal{X}_1} e^{\sum_{k=d_x^0+1}^{d_x}\frac{|x_k|}{a_{\sigma_n}^2}-\kappa\|x\|^2} e^{\kappa\|x\|^2} g_0(x)dx.$$

For $x\in\mathcal{X}\setminus\mathcal{X}_1$, $\kappa\|x\|^2 \ge \kappa(d_x-d_x^0)^{-1}(\sum_{k=d_x^0+1}^{d_x}|x_k|)^2 \ge 2\sum_{k=d_x^0+1}^{d_x}|x_k|/a_{\sigma_n}^2$ and hence

$$\int_{\mathcal{X}\setminus\mathcal{X}_1}\max_{1\le j\le K}\left|\frac{K_j-\tilde{K}_j}{K_j}\right| g_0(x)dx \le Be^{-A_n/a_{\sigma_n}^2} \le \sigma_n^{2\beta}. \tag{6.3}$$

From (6.2) and (6.3), it follows that $d_h(p(\cdot|\cdot,\theta_{d_x^0}),p(\cdot|\cdot,\theta)) \le 2^{1/2}\sigma_n^\beta$.

Next let us establish an analog of (3.7) when $\|(y,x)\|\le a_{\sigma_n}$. Using the argument leading to (3.7) with $\left\|(y,x_{1d_x^0})\right\| \le a_{\sigma_n}$ and $((x_k-\mu_{jk}^x)/\sigma_k^x)^2\le 4$ for $k=d_x^0+1,\dots,d_x$, $j=1,\dots,m$, we get for $\theta\in S_{\theta^\star}$ and $\|(y,x)\|\le a_{\sigma_n}$,

$$\frac{p(y|x,\theta,m)}{f_0(y|x)} \geq \frac{1}{f_0(y|x)} \min_{j=1,\ldots,K} \alpha_j \cdot \sigma_n^{-d_y} \exp\left\{ - \sum_{k=d_x^0+1}^{d_x} \frac{(x_k - \mu_{jk}^x)^2}{2(\sigma_k^x)^2} \right\}$$

$$\geq C_5 \tilde{\epsilon}_n^{4d^0 b_1} \sigma_n^{-d_y} = \lambda_n.$$

For $\|(y,x)\| \geq a_{\sigma_n}$,

$$p(y \mid x,\theta,m) \geq \min_{1\leq j\leq m} C_6 \sigma_n^{-d_y} \exp\left\{ - \frac{||y - \mu_j^y||^2}{2\sigma_n^2} \right\}$$

$$\geq C_7 \sigma_n^{-d_y} \exp\left(-C_8 \frac{a_{\sigma_n}^2}{\sigma_n^2} - C_9 \frac{\|y\|^2}{\sigma_n^2} \right)$$

implies

$$\left\{ \log \frac{f_0(y|x)}{p(y|x,\theta,m)} \right\}^2 \leq C_{10} \left(\frac{a_{\sigma_n}^4}{\sigma_n^4} + \frac{\|y\|^4}{\sigma_n^4} \right).$$

Then, following the proof of Theorem 3.1,

$$\int \left\{ \log \frac{f_0(y|x)}{p(y|x,\theta,m)} \right\}^2 1\left\{ \frac{p(y|x,\theta,m)}{f_0(y|x)} < \lambda_n \right\} f(y|x) g_0(x) dy dx$$

$$\leq C_{11} \left\{ \frac{a_{\sigma_n}^4 P_0(\|Z\| > a_{\sigma_n})}{\sigma_n^4} + \frac{E_0\left(\|Y\|^8\right)^{1/2} (P_0(\|Z\| > a_{\sigma_n}))^{1/2}}{\sigma_n^4} \right\}$$

$$\leq C_{12} \sigma_n^{2\beta+\varepsilon/2} \sigma_n^{\varepsilon/2} a_{\sigma_n}^4 \leq C_{12} \sigma_n^{2\beta+\varepsilon/2}.$$

The rest of the proof of $E_0(\log(f_0(Y|X)/p(Y|X,\theta,m))) \leq A\tilde{\epsilon}_n^2$ and $E_0([\log(f_0(Y|X)/p(Y|X,\theta,m))]^2) \leq A\tilde{\epsilon}_n^2$ goes through without any changes.

The lower bound for the prior probability of $S_{\theta^\star}$ and $m = K$ is the same as the one in Theorem 3.1, except d is replaced with d^0. The only additional calculation for σ_k^x, $k = d_x^0 + 1, \ldots, d_x$ follows from Assumption (2.6),

$$\Pi((\sigma_k^x)^{-2} \in [a_{\sigma_n}^{-2}/2, a_{\sigma_n}^{-2}]) \gtrsim a_{\sigma_n}^{-2a_7}.$$

In the definition of sieve (4.1), let us replace condition $\sigma \in [\underline{\sigma}, \overline{\sigma}]$ by

$$\sigma_l^y, \sigma_k^x \in [\underline{\sigma}, \overline{\sigma}],\ l = 1, \ldots, d_y,\ k = 1, \ldots, d_x.$$

The presence of the component specific scale parameters and the dimension of x affect only constants in the sieve entropy bound and the bound on the prior probability of the sieve complement. Thus, Theorem 4.2 holds with d replaced by d^0. ■

7. FINITE SAMPLE PERFORMANCE

In this section, we evaluate the finite sample performance of our conditional density model in Monte Carlo simulations. Specifically, we explore how the sample

size and irrelevant covariates affect estimation results. We also compare our estimator with a conditional density kernel estimator from Hall et al. (2004) that is based on a cross-validation method for obtaining the bandwidth. Hall et al. (2004) showed that irrelevant covariates do not affect the convergence rate of their estimator, and, thus, this estimator appears to be a suitable benchmark. The kernel estimation results are obtained by the publicly available R package `np` (Hayfield and Racine (2008)).

It has been established in the literature (see, for example, Villani et al. (2009)), that slightly more general specifications of covariate dependent mixture models perform better in practice. Thus, we use the following specification

$$p(y|x,\theta,m)=\sum_{j=1}^{m}\frac{\alpha_j \exp\left\{-0.5\sum_{k=1}^{d_x}\left(x_k-\mu_{jk}^x\right)^2/\left(\sigma_k^x s_{jk}^x\right)^2\right\}}{\sum_{i=1}^{m}\alpha_i \exp\left\{-0.5\sum_{k=1}^{d_x}\left(x_k-\mu_{ik}^x\right)^2/\left(\sigma_k^x s_{ik}^x\right)^2\right\}}\phi_{x'\beta_j,\sigma^y s_j^y}(y), \quad \textbf{(7.1)}$$

where $x'\beta_j$ are used instead of locations μ_j^y, x includes zeroth coordinate equal 1, and local scale parameters $\left(s_j^y, s_{jk}^x\right)$ introduced in addition to the global $\left(\sigma^y, \sigma_k^x\right)$. The prior is specified as follows,

$$\beta_j \overset{iid}{\sim} N\left(\underline{\beta}, \underline{H}_\beta^{-1}\right),\ \mu_j \overset{iid}{\sim} N\left(\underline{\mu}, \underline{H}_\mu^{-1}\right),$$
$$\left(s_j^y\right)^{-2} \overset{iid}{\sim} G\left(\underline{A}_{sy}, \underline{B}_{sy}\right),\ \left(s_{jk}^x\right)^{-2} \overset{iid}{\sim} G\left(\underline{A}_{sxk}, \underline{B}_{sxk}\right),\ k=1,\ldots,d_x,$$
$$\left(\sigma^y\right)^{-1} \overset{iid}{\sim} G\left(\underline{A}_{\sigma y}, \underline{B}_{\sigma y}\right),\ \left(\sigma_k^x\right)^{-1} \overset{iid}{\sim} G\left(\underline{A}_{\sigma xk}, \underline{B}_{\sigma xk}\right),\ k=1,\ldots,d_x,$$
$$(\alpha_1,\ldots,\alpha_m)|m \overset{iid}{\sim} D\left(\underline{a}/m,\ldots,\underline{a}/m\right),$$
$$\Pi\left(m=k\right)=(e^{\underline{A}_m}-1)e^{-\underline{A}_m\cdot k},$$

where $G(A,B)$ stands for a Gamma distribution with shape A and rate B. In Theorem A.5 in the Appendix, we show that an analog of Corollary 6.1 holds for this slightly more general setup. To obtain estimation results for this model we use an MCMC algorithm developed in Norets (2015).

We use the following (data-dependent) values for prior hyper-parameters:

$$\underline{\beta}=\left(\sum_i x_i x_i'\right)^{-1}\sum_i x_i y_i,\ \underline{H}_\beta^{-1}=\underline{c}_\beta\left(\sum_i x_i x_i'\right)^{-1}\sum_i\left(y_i-x_i'\underline{\beta}\right)^2/n,$$
$$\underline{\mu}=\sum_i x_i/n,\ \underline{H}_\mu^{-1}=\sum_i\left(x_i-\underline{\mu}\right)\left(x_i-\underline{\mu}\right)'/n,$$
$$\underline{A}_{\sigma y}=\underline{c}_\sigma/\left(\sum_i\left(y_i-x_i'\underline{\beta}\right)^2/n\right),\ \underline{B}_{\sigma y}=\underline{c}_\sigma/\left(\sum_i\left(y_i-x_i'\underline{\beta}\right)^2/n\right)^{1/2},$$
$$\underline{A}_{\sigma xl}=\underline{c}_\sigma/\left(\sum_i\left(x_{il}-\sum_i x_{il}/n\right)^2/n\right),\ \underline{B}_{\sigma xl}=\underline{c}_\sigma/\left(\sum_i\left(x_{il}-\sum_i x_{il}/n\right)^2/n\right)^{1/2},$$

$\underline{A}_{sxk} = \underline{c}_s$, $\underline{B}_{sxk} = \underline{c}_s$, $\underline{A}_{sy} = \underline{c}_s$, $\underline{B}_{sy} = \underline{c}_s$,
$\underline{a} = 15$, $\underline{A}_m = 1$,

where $\underline{c}_\beta = 100$, $\underline{c}_\sigma = 0.1$, $\underline{c}_s = 10$. Thus, a modal prior draw would have one mixture component and it would be near a normal linear regression estimated by the least squares. As Figure 1 illustrates, the prior variances are chosen sufficiently large so that a wide range of densities can be easily accommodated by the prior.

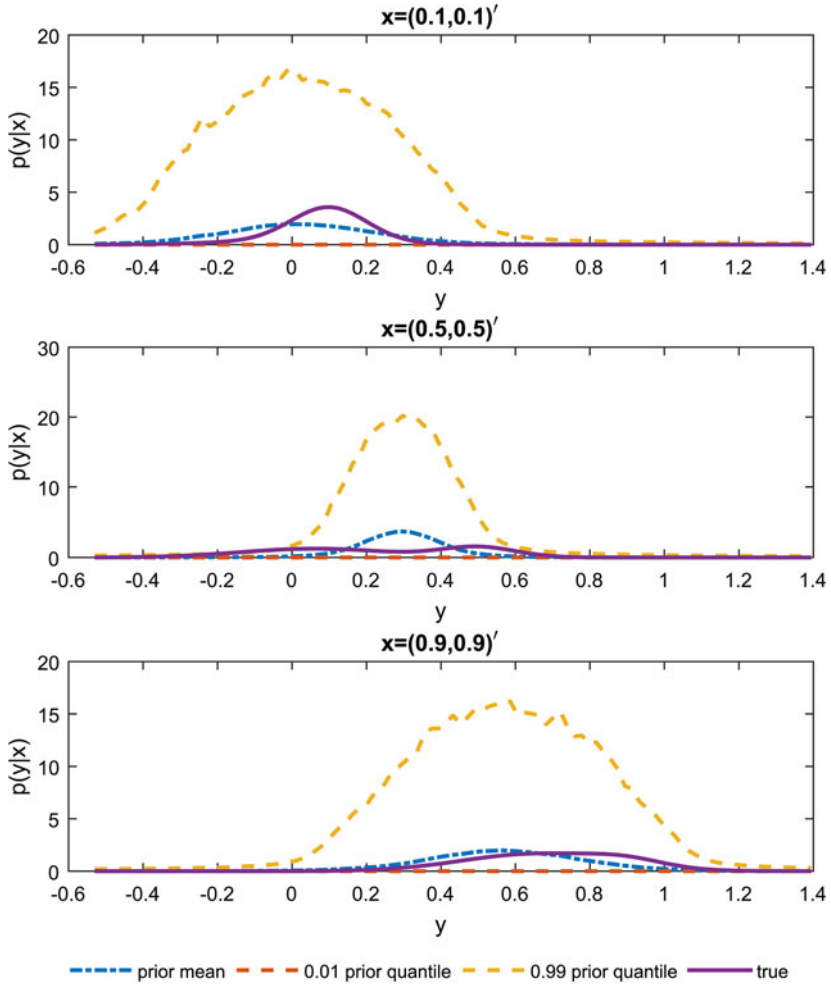

FIGURE 1. Simulated prior conditional densities for $d_x = 2$ and $x \in \{(0.1, 0.1)', (0.5, 0.5)', (0.9, 0.9)'\}$. The solid lines are the true DGP values, the dash-dotted lines are the prior means, and the dashed lines are pointwise 1% and 99% prior quantiles.

The DGP for simulation experiments is as follows: $x_i = (x_{i1}, \ldots, x_{id_x})$, $x_{ik} \sim U[0,1]$ (or $x_{ik} \sim N\left(0.5, 12^{-1/2}\right)$ for an unbounded support case) and the true conditional density is

$$f_0(y_i|x_{i1}) = e^{-2x_{i1}} N\left(y_i; x_{i1}, 0.1^2\right) + \left(1 - e^{-2x_{i1}}\right) N\left(y_i; x_{i1}^4, 0.2^2\right). \tag{7.2}$$

Note that the DGP conditional density depends only on the first coordinate of x_i, the rest of the coordinates are irrelevant. This DGP was previously used without irrelevant covariates by Dunson et al. (2007), Dunson and Park (2008), and Norets and Pelenis (2014).

The kernel estimates reported below are obtained by functions `npcdensbw` (bandwidth selection) and `npcdens` (conditional density estimation) from R package `np`. We use `np`'s default parameter values for these functions: Gaussian kernels and likelihood cross-validation for selecting a vector of bandwidths.

For each estimation exercise, we perform 5,000 MCMC iterations, of which the first 500 are discarded for burn-in. The MCMC draws of m mostly belong to $\{3, \ldots, 13\}$. Figure 2 presents Bayesian and kernel estimation results for one dataset of size $n = 1000$ and $d_x = 2$. Each panel in the figure shows the DGP densities, the kernel estimates, the posterior means, and the posterior 0.01%-quantiles conditional on a particular value of covariate x. As can be seen from the figure, the estimation results from both approaches can be pretty close.

In every Monte Carlo experiment we perform, 50 simulated datasets are used. For each dataset, the performance of an estimator is evaluated by the mean absolute error

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

where $x_j \in \{0.1, 0.5, 0.9\}^{d_x}$ and y_i belongs to a 100 points equal spaced grid on the range of simulated values for y. The results for the root mean squared error, the Hellinger distance, and the MAE are qualitatively the same, and, thus, only results for MAE are reported here.

Table 1 presents estimation results for 9 Monte Carlo experiments based on different values of n, d_x, the covariate support, and the prior. The table gives MAE for the kernel and posterior mean estimators averaged over 50 simulated datasets. It also shows the average difference between MAEs of the two estimators and the corresponding t-statistics. In all the experiments, the posterior mean estimator performs better than the kernel estimator and the differences are highly statistically significant. The first three rows of the table present the results for the covariates with bounded support, $d_x = 1$, and $n \in \{10^2, 10^3, 10^4\}$. As expected, the MAE decreases as the sample size increases for both estimators. The next five rows of the table show the results for $d_x \in \{1, 3, 5\}$ and $n = 10^3$ for covariates with bounded and unbounded support. Even though the posterior mean outperforms the kernel estimator in absolute terms, the MAE for the kernel estimator

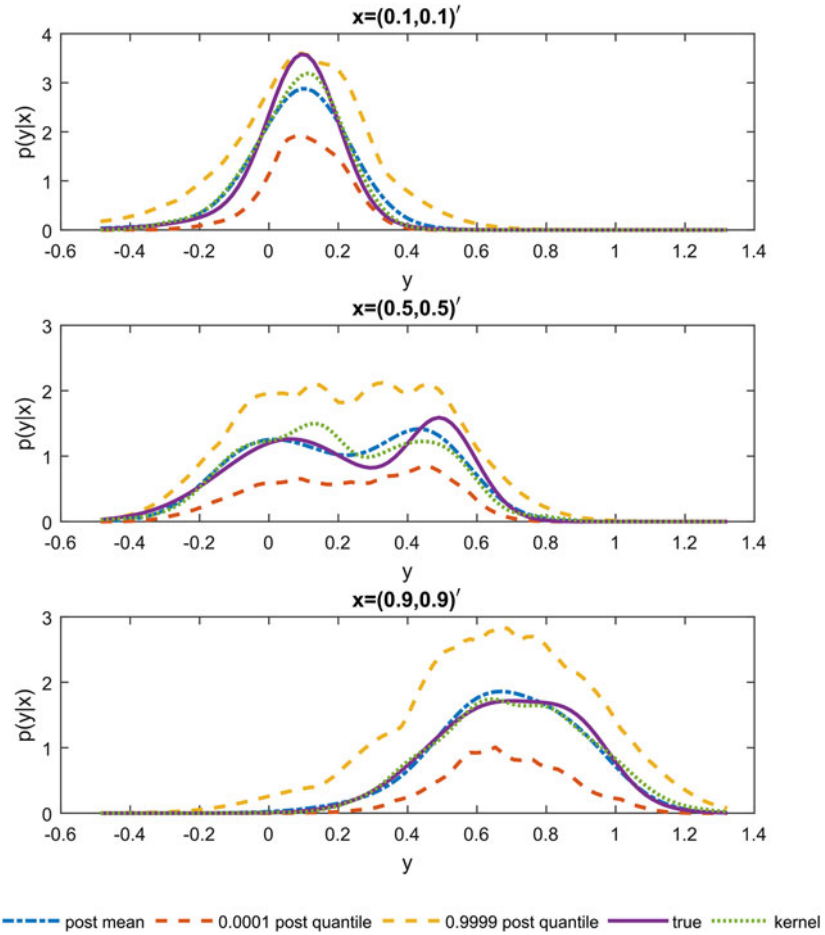

FIGURE 2. Estimated conditional densities for $d_x = 2$ and $x \in \{(0.1, 0.1)', (0.5, 0.5)', (0.9, 0.9)'\}$. The solid lines are the true values, the dash-dotted lines are the posterior means, the dotted lines are kernel estimates, and the dashed lines are pointwise 0.01% and 99.99% posterior quantiles.

increases slower in the number of irrelevant covariates: for $x_{ik} \sim N(0.5, 12^{-1/2})$ ($x_{ik} \sim U[0, 1]$), the MAE increases only by 0.0007 (0.002) as d_x changes from 1 to 5; the corresponding increase for the posterior mean estimator is 0.01 (0.022). The last row of the table shows the results for $d_x = 1$, $n = 10^3$, and the following alternative prior hyperparameters: $\underline{c}_\beta = 200$, $\underline{c}_\sigma = 0.2$, $\underline{c}_s = 15$, $\underline{a} = 12$, and $\underline{A}_m = 2$. Thus, the results are not very sensitive to reasonable variations in $(\underline{c}_\beta, \underline{c}_\sigma, \underline{c}_s, \underline{a}, \underline{A}_m)$.

The dimension of the covariates does not noticeably affect the computing time for the posterior mean estimator. The computations for the kernel estimator are

TABLE 1. MAE for kernel and posterior mean estimators

$x_{ik} \sim g_0$	d_x	n	Bayes	Kernel	B-K	%(B < K)	t-stat
$U[0,1]$	1	10^2	0.107	0.164	−0.058	1	−15.47
$U[0,1]$	1	10^4	0.032	0.040	−0.008	0.88	−8.28
$U[0,1]$	1	10^3	0.062	0.096	−0.033	1	−16.16
$U[0,1]$	3	10^3	0.074	0.097	−0.022	0.96	−13.40
$U[0,1]$	5	10^3	0.084	0.098	−0.015	0.86	−7.88
$N(0.5, 12^{-1/2})$	1	10^3	0.028	0.054	−0.026	1	−15.33
$N(0.5, 12^{-1/2})$	3	10^3	0.033	0.054	−0.021	1	−11.78
$N(0.5, 12^{-1/2})$	5	10^3	0.038	0.054	−0.017	0.92	−8.88
$U[0,1]$	1	10^3	0.060	0.096	−0.036	1	−17.72

very fast for low-dimensional covariates. They slow down considerably when d_x increases. For $d_x = 5$, the posterior mean is slightly faster to compute than the kernel estimator.

Overall, the Monte Carlo experiments suggest that the model proposed in this paper is a practical and promising alternative to classical nonparametric methods.

8. CONCLUSION

We show above that under a reasonable prior distribution, the posterior contraction rate in our model is bounded above by $\epsilon_n = n^{-\beta/(2\beta+d)}(\log n)^t$ for any

$$t > [d(1+1/\beta+1/\tau)+\max\{\tau_1, 1, \tau_2/\tau\}]/(2+d/\beta)+\max\{0,(1-\tau_1)/2\}.$$

Rate $n^{-\beta/(2\beta+d)}$ is minimax for estimation of multivariate densities when their smoothness level is β and dimension of (y,x) is d. Since the total variation distance between joint densities for (y,x) is bounded by the sum of the integrated total variation distance between the conditional densities and the total variation distance between the densities of x, the minimax rate for estimation of conditional densities of smoothness β in integrated total variation distance cannot be faster than $n^{-\beta/(2\beta+d)}$. Thus, we can claim that our Bayesian nonparametric model achieves optimal contraction rate up to a log factor. We are not aware of analogous results for estimators based on kernels or mixtures. In the classical settings, Efromovich (2007) develops an estimator based on orthogonal series that achieves minimax rates for one-dimensional y and x. In a recent paper, Shen and Ghosal (2016) consider a compactly supported Bayesian model for conditional densities based on tensor products of spline functions. They show that under suitable sparsity assumptions, the posterior contracts at an optimal rate even when the dimension of covariates increases exponentially with the sample size. An advantage of our results is that we do not need to assume a known upper bound on

the smoothness level and the boundedness away from zero for the true density. The analysis of the posterior contraction rates in our model under sparsity and increasing dimension of covariates is an important direction for future work.

NOTE

1. Norets and Pelenis (2014) use this inequality in conjunction with a lower bound on K_j, which leads to entropy bounds that are not sufficiently tight for adaptive contraction rates.

REFERENCES

Barron, A., M.J. Schervish, & L. Wasserman (1999) The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics* 27(2), 536–561.

Bhattacharya, A., D. Pati, & D. Dunson (2014) Anisotropic function estimation using multi-bandwidth Gaussian processes. *The Annals of Statistics* 42(1), 352–381.

Chung, Y. & D.B. Dunson (2009) Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* 104(488), 1646–1660.

De Iorio, M., P. Muller, G.L. Rosner, & S.N. MacEachern (2004) An ANOVA model for dependent random measures. *Journal of the American Statistical Association* 99(465), 205–215.

De Jonge, R. & J.H. van Zanten (2010) Adaptive nonparametric Bayesian inference using location-scale mixture priors. *The Annals of Statistics* 38(6), 3300–3320.

Dunson, D.B. & J.H. Park (2008) Kernel stick-breaking processes. *Biometrika* 95(2), 307–323.

Dunson, D.B., N. Pillai, & J.H. Park (2007) Bayesian density regression. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 69(2), 163–183.

Efromovich, S. (2007) Conditional density estimation in a regression setting. *The Annals of Statistics* 35(6), 2504–2535.

Geweke, J. & M. Keane (2007) Smoothly mixing regressions. *Journal of Econometrics* 138, 252–290.

Ghosal, S., J.K. Ghosh, & R.V. Ramamoorthi (1999) Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics* 27(1), 143–158.

Ghosal, S., J.K. Ghosh, & A.W. van der Vaart (2000) Convergence rates of posterior distributions. *The Annals of Statistics* 28(2), 500–531.

Ghosal, S. & A.W. van der Vaart (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* 29(5), 1233–1263.

Ghosal, S. & A.W. van der Vaart (2007) Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics* 35(2), 697–723.

Griffin, J.E. & M.F.J. Steel (2006) Order-based dependent Dirichlet processes. *Journal of the American Statistical Association* 101(473), 179–194.

Hall, P., J. Racine, & Q. Li (2004) Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* 99(468), 1015–1026.

Hayfield, T. & J.S. Racine (2008) Nonparametric econometrics: The np package. *Journal of Statistical Software* 27(5), 1–32.

Huang, T.M. (2004) Convergence rates for posterior distributions and adaptive estimation. *The Annals of Statistics* 32(4), 1556–1593.

Jacobs, R.A., M.I. Jordan, S.J. Nowlan, & G.E. Hinton (1991) Adaptive mixtures of local experts. *Neural Computation* 3(1), 79–87.

Jordan, M. & L. Xu (1995) Convergence results for the em approach to mixtures of experts architectures. *Neural Networks* 8(9), 1409–1431.

Keane, M. & O. Stavrunova (2011) A smooth mixture of tobits model for healthcare expenditure. *Health Economics* 20(9), 1126–1153.

Kruijer, W., J. Rousseau, & A. van der Vaart (2010) Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics* 4, 1225–1257.

Li, F., M. Villani, & R. Kohn (2010) Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities. *Journal of Statistical Planning and Inference* 140(12), 3638–3654.

Li, Q. & J.S. Racine (2007) *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

MacEachern, S.N. (1999) Dependent nonparametric processes. *Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55. American Statistical Association.

Norets, A. (2010) Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics* 38(3), 1733–1766.

Norets, A. (2015) Optimal retrospective sampling for a class of variable dimension models. Unpublished manuscript, Brown University.

Norets, A. & J. Pelenis (2012) Bayesian modeling of joint and conditional distributions. *Journal of Econometrics* 168, 332–346.

Norets, A. & J. Pelenis (2014) Posterior consistency in conditional density estimation by covariate dependent mixtures. *Econometric Theory* 30, 606–646.

Pati, D., D.B. Dunson, & S.T. Tokdar (2013) Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis* 116, 456–472.

Peng, F., R.A. Jacobs, & M.A. Tanner (1996) Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association* 91(435), 953–960.

Rousseau, J. (2010) Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *The Annals of Statistics* 38(1), 146–180.

Scricciolo, C. (2006) Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Annals of Statatistics* 34(6), 2897–2920.

Shen, W. & S. Ghosal (2016) Adaptive Bayesian density regression for high-dimensional data. *Bernoulli* 22(1), 396–420.

Shen, W., S.T. Tokdar, & S. Ghosal (2013) Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* 100(3), 623–640.

Tokdar, S., Y. Zhu, & J. Ghosh (2010) Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Analysis* 5(2), 319–344.

van der Vaart, A.W. & J.H. van Zanten (2009) Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics* 37(5B), 2655–2675.

Villani, M., R. Kohn, & P. Giordani (2009) Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics* 153(2), 155–173.

Villani, M., R. Kohn, & D.J. Nott (2012) Generalized smooth finite mixtures. *Journal of Econometrics* 171(2), 121–133.

Wade, S., D.B. Dunson, S. Petrone, & L. Trippa (2014) Improving prediction from Dirichlet process mixtures via enrichment. *The Journal of Machine Learning Research* 15(1), 1041–1071.

Wood, S., W. Jiang, & M. Tanner (2002) Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* 89(3), 513–528.

Yang, Y. & S.T. Tokdar (2015) Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics* 43(2), 652–674.

APPENDIX

LEMMA A.1. *Suppose* $f, f_0 \in \mathcal{F}$, $g_0(x) \leq \bar{g} < \infty$, $g(x)$ *and* $u(x)$ *are densities on* $\mathcal{X}$, $u(x) \geq \underline{u} > 0$. *Then,*

$$d_h^2(f_0, f) \leq \frac{4\bar{g}}{\underline{u}} \int \left(\sqrt{f_0(y|x)u(x)} - \sqrt{f(y|x)g(x)} \right)^2 dydx.$$

Proof. Observe that

$$d_h^2(f_0,f) = \int \left(\sqrt{f_0(y|x)} - \sqrt{f(y|x)}\right)^2 g_0(x)dydx$$
$$\leq \frac{\bar{g}}{\underline{u}} \int \left(\sqrt{f_0(y|x)u(x)} - \sqrt{f(y|x)u(x)}\right)^2 dydx$$
$$\leq \frac{2\bar{g}}{\underline{u}}(\mathrm{I}+\mathrm{II}), \tag{A.1}$$

where $\mathrm{I} = \int \left(\sqrt{f_0(y|x)u(x)} - \sqrt{f(y|x)g(x)}\right)^2 dydx$, $\mathrm{II} = \int \left(\sqrt{f(y|x)g(x)} - \sqrt{f(y|x)u(x)}\right)^2 dydx$.

Observe that

$$\mathrm{II} \leq \int \left(\sqrt{g(x)} - \sqrt{u(x)}\right)^2 dx = 2\left(1 - \int \sqrt{g(x)u(x)}dx\right) \leq \mathrm{I}. \tag{A.2}$$

The final inequality in (A.2) follows since $\int \sqrt{f_0(y|x)f(y|x)}dy \leq \frac{1}{2}\left(\int f_0(y|x)dy + \int f(y|x)dy\right) = 1$. Combining (A.1) and (A.2), we obtain

$$d_h^2(f_0,f) \leq 4\mathrm{I} = \frac{4\bar{g}}{\underline{u}} \int \left(\sqrt{f_0(y|x)u(x)} - \sqrt{f(y|x)g(x)}\right)^2 dydx. \quad \blacksquare$$

COROLLARY A.2. *Suppose $f, f_0 \in \mathcal{F}$, $g(x)$ and $\bar{g}_0(x)$ are densities on $\mathcal{X}$, with $\bar{g}_0$ satisfying $\eta\bar{g}_0(x) \geq g_0(x)$ for some constant $\eta > 0$ and all $x \in \mathcal{X}$. Then,*

$$d_h^2(f_0,f) \leq 4\eta \int \left(\sqrt{f_0(y|x)\bar{g}_0(x)} - \sqrt{f(y|x)g(x)}\right)^2 dydx.$$

To prove the corollary note that the inequality (A.1) in the proof of Lemma A.1 holds under $\eta\bar{g}_0(x) \geq g_0(x)$ with u replaced by $\bar{g}_0$ and $\bar{g}/\underline{u}$ replaced by η. The rest of the lemma's proof applies with $\bar{g}_0$ replacing u.

LEMMA A.3. *In Theorem 3 of Shen et al. (2013), replace their $g_\sigma = f_\sigma + (1/2)f_0 1\{f_\sigma < (1/2)f_0\}$ with $g_\sigma = f_\sigma + 2|f_\sigma|1\{f_\sigma < 0\}$, where notation from Shen et al. (2013) is used. Then, the claim of the theorem holds.*

Proof. With the alternative definition of g_σ, the proof of Shen et al. (2013) goes through with the following changes. First, $1 \leq \int g_\sigma(x)dx = \int f_\sigma(x)dx + 2\int |f_\sigma|1\{(f_\sigma < 0\} \leq 1 + 3\int_{A_\sigma^c} f_0(x)dx \leq 1 + K_2\sigma^{2\beta}$. Second, replace inequality $r_\sigma \leq g_\sigma$ with $(1/2)r_\sigma \leq g_\sigma$. $\blacksquare$

LEMMA A.4. *There is a $\lambda_0 \in (0,1)$ such that for any $\lambda \in (0,\lambda_0)$ and any two conditional densities $p, q \in \mathcal{F}$, a probability measure P on $\mathcal{Z}$ that has a conditional density equal to p, and d_h defined with the distribution on $\mathcal{X}$ implied by P,*

$$P\log\frac{p}{q} \leq d_h^2(p,q)\left(1 + 2\log\frac{1}{\lambda}\right) + 2P\left\{\left(\log\frac{p}{q}\right)1\left(\frac{q}{p} \leq \lambda\right)\right\},$$

$$P\left(\log\frac{p}{q}\right)^2 \leq d_h^2(p,q)\left(12 + 2\left(\log\frac{1}{\lambda}\right)^2\right) + 8P\left\{\left(\log\frac{p}{q}\right)^2 1\left(\frac{q}{p} \leq \lambda\right)\right\}.$$

Proof. The proof is exactly the same as the proof of Lemma 4 of Shen et al. (2013), which in turn, follows the proof of Lemma 7 in Ghosal and van der Vaart (2007). ■

THEOREM A.5. *Assume f_0 satisfies the assumptions in Section 6 with $d_y = 1$. Then the model (7.1) in Section 7 and the prior specifications following it leads to the same posterior contraction rate as specified in Corollary 6.1.*

Proof. In the following we will verify prior thickness condition with the same $\tilde{\epsilon}_n$ (with $d_y = 1$) as in Corollary 6.1 and modify the sieve construction accordingly. The proof proceeds along the lines of the proof of Corollary 6.1. The main difference is that the following joint density is used in bounds for the distance between conditional densities

$$\tilde{p}(y, x \mid \theta, m) = \sum_{j=1}^{m} \frac{\alpha_j \exp\{-0.5 \sum_{k=1}^{d_x} (x_k - \mu_{jk}^x)^2/(\sigma_k^x s_{jk}^x)^2\}}{\sum_{i=1}^{m} \alpha_i \exp\{-0.5 \sum_{k=1}^{d_x} (x_k - \mu_{ik}^x)^2/(\sigma_k^x s_{ik}^x)^2\}} \phi_{\mu_j^y + x'\beta_j, \sigma^y s_j^y}(y)$$
$$\cdot \sum_{j=1}^{m} \alpha_j \phi_{\mu_j^x, \sigma^x \circ s_j^x}(x),$$

where $\circ$ denotes the Hadamard product. The intercept absorbed in the notation "$x'\beta_j$" in (7.1) is denoted by μ_j^y here. Let

$$\theta_{d_x^0} = \{\mu_j^y, \mu_{j1d_x^0}^x = (\mu_{j1}^x, \ldots, \mu_{jd_x^0}^x), \alpha_j, s_j^y, s_{j1d_x^0}^x = (s_{j1}^x, \ldots, s_{jd_x^0}^x),$$
$$\beta_{j1d_x^0} = (\beta_{j1}, \ldots, \beta_{jd_x^0}),\ j = 1, 2, \ldots;\ \sigma^y, \sigma_{1d_x^0}^x = (\sigma_1^x, \ldots, \sigma_{d_x^0}^x)\},$$

$$S_{\theta^\star} = \big\{(\mu_j, \alpha_j,\ j = 1, 2, \ldots; \sigma^y, \sigma^x) :\ (\mu_j^y, \mu_{j1}^x, \ldots, \mu_{jd_x^0}^x) \in U_j,$$
$$||(\mu_{jd_x^0+1}^x, \ldots, \mu_{jd_x}^x)|| \le \sigma_n \tilde{\epsilon}_n^{2b_1}, ||\beta_j|| \le \sigma_n \tilde{\epsilon}_n^{2b_1}\ j \le K;$$
$$\sum_{j=1}^{K} \left|\alpha_j - \alpha_j^\star\right| \le 2\tilde{\epsilon}_n^{2d^0 b_1},\ \min_{j=1,\ldots,K} \alpha_j \ge \tilde{\epsilon}_n^{4d^0 b_1}/2;$$
$$(\sigma_k^x)^2, (\sigma^y)^2 \in [\sigma_n^2/(1+\sigma_n^{2\beta}), \sigma_n^2],\ k \le d_x^0;$$
$$(\sigma_k^x)^2 \in [a_{\sigma_n}^2, 2a_{\sigma_n}^2],\ k = d_x^0 + 1, \ldots, d_x;$$
$$s_{jk}^x, s_j^y \in [1, 1+\sigma_n^{2\beta}], j = 1, 2, \ldots, K; k = 1, \ldots, d_x\big\},$$

and $s_{jk}^x = s_j^y = 1$ and $\beta_{jk} = 0$ in $\theta_{d_x^0}^*$ for $k = 1, 2, \ldots, d_x^0$ and $j = 1, \ldots, K$.

Similarly to the proof of Corollary 6.1,

$$d_h(f_0, p(\cdot|\cdot, \theta, m)) \lesssim \sigma_n^\beta + d_H(\tilde{p}(\cdot|\theta_{d_x^0}^*, m), \tilde{p}(\cdot|\theta_{d_x^0}, m)) + d_h(p(\cdot|\cdot, \theta_{d_x^0}, m), p(\cdot|\cdot, \theta, m)).$$

Consider $\theta \in S_{\theta^\star}$ and let $s_{\cdot j} = \prod_{k=1}^{d_x^0} s_{jk}$. Then, $d_H(\tilde{p}(\cdot|\theta_{d_x^0}^*, m), \tilde{p}(\cdot|\theta_{d_x^0}, m))^2$ can be bounded by

$$\left\| \sum_{j=1}^{K} \alpha_j^* \phi_{\mu_j^*, \sigma_n}(\cdot) - \sum_{j=1}^{K} \alpha_j s_{\cdot j} \phi_{\mu_j^y + x'_{1d_x^0}\beta_{j1d_x^0}, s_j^y \sigma_j^y}(\cdot) \phi_{\mu_{j1d_x^0}^x, s_{j1d_x^0}^x \circ \sigma_{1d_x^0}^x}(\cdot) \frac{\sum_{j=1}^{K} \alpha_j \phi_{\mu_{j1d_x^0}^x, s_{j1d_x^0}^x \circ \sigma_{1d_x^0}^x}(\cdot)}{\sum_{j=1}^{K} \alpha_j s_{\cdot j} \phi_{\mu_{j1d_x^0}^x, s_{j1d_x^0}^x \circ \sigma_{1d_x^0}^x}(\cdot)} \right\|_1$$

$$\leq \left\| \sum_{j=1}^{K} \alpha_j^* \phi_{\mu_j^*,\sigma_n}(\cdot) - \sum_{j=1}^{K} \alpha_j s_{\cdot j} \phi_{\mu_j^y + x'_{1d_x^0}\beta_{j1d_x^0}, s_j^y \sigma^y}(\cdot) \phi_{\mu^x_{j1d_x^0}, s^x_{j1d_x^0} \circ \sigma^x_{1d_x^0}}(\cdot) \right\|_1$$
$$+ \left\| \sum_{j=1}^{K} \alpha_j s_{\cdot j} \phi_{\mu_j^y + x'_{1d_x^0}\beta_{j1d_x^0}, s_j^y \sigma^y}(\cdot) \phi_{\mu^x_{j1d_x^0}, s^x_{j1d_x^0} \circ \sigma^x_{1d_x^0}}(\cdot) \left\{ \frac{\sum_{j=1}^{K} \alpha_j \phi_{\mu^x_{j1d_x^0}, s^x_{j1d_x^0} \circ \sigma^x_{1d_x^0}}(\cdot)}{\sum_{j=1}^{K} s_{\cdot j} \alpha_j \phi_{\mu^x_{j1d_x^0}, s^x_{j1d_x^0} \circ \sigma^x_{1d_x^0}}(\cdot)} - 1 \right\} \right\|_1$$
$$\lesssim \left[\sum_{j=1}^{K} \left| \alpha_j - \alpha_j^* \right| + \sigma_n^{2\beta} + \sum_{j=1}^{K} \alpha_j^* \left(\frac{||\mu_j - \mu_j^*||}{\sigma_n} + \frac{\left\| \beta_{j1d_x^0} \right\| a_{\sigma_n}}{\sigma_n} + \sigma_n^{2\beta} \right) + \sigma_n^{2\beta} \right] \lesssim \sigma_n^{2\beta},$$

where the penultimate inequality is implied by $\left| \sigma_n^2/(s_j^y \sigma^y)^2 - 1 \right| \leq 3\sigma_n^{2\beta}$, $\left| \sigma_n^2/(s_{jk}^x \sigma_k^x)^2 - 1 \right| \leq 3\sigma_n^{2\beta}$, $|s_{\cdot j} - 1| \leq d_x^0 \sigma_n^{2\beta}$, $\int ||x_{1d_x^0}|| \phi_{\mu^x_{j1d_x^0}, s^x_{j1d_x^0} \circ \sigma^x_{1d_x^0}}(x_{1d_x^0}) d(x_{1d_x^0}) \lesssim ||\mu^x_{j1d_x^0}|| \leq a_{\sigma_n}$, and an argument similar to the one preceding (3.6).

Next note that for $\theta \in S_{\theta^\star}$,

$$d_h^2 \left(p\left(\cdot|\cdot, \theta_{d_x^0}, m\right), p(\cdot|\cdot, \theta, m) \right) \lesssim \int \max_{1 \leq j \leq m} |K_j - \tilde{K}_j| / \left| K_j \right| g_0(x) dx$$
$$+ \int \max_{1 \leq j \leq m} \frac{\left| x' \beta_j \right|}{\sigma_n} g_0(x) dx$$
$$\lesssim \sigma_n^{2\beta} + \tilde{\epsilon}_n^{2b_1} \int \|x\| g_0(x) dx \lesssim \sigma_n^{2\beta},$$

where the first part of the penultimate inequality follows similarly to the proof of Corollary 6.1.

Next, let us bound the ratio $p(y|x, \theta, m)/f_0(y|x)$. For $||(y, x)|| \leq a_{\sigma_n}$, observe that $\exp\left\{ -|y - \mu_j^y - x'\beta_j|^2/(2\sigma_n^2) \right\} \geq \exp\left\{ -|y - \mu_j^y|^2/\sigma_n^2 - |x'\beta_j|^2/\sigma_n^2 \right\}$ and $|x'\beta_j|/\sigma_n \leq a_{\sigma_n} \epsilon_n^{2b_1} \leq 1$. Thus, λ_n can be defined by (3.7).

For $\|(y, x)\| \geq a_{\sigma_n}$,

$$\left\{ \log \frac{f_0(y|x)}{p(y|x, \theta, m)} \right\}^2 \lesssim \left(\frac{a_{\sigma_n}^4}{\sigma_n^4} + \frac{|y|^4}{\sigma_n^4} + \|x\|^4 \tilde{\epsilon}_n^{4b_1} \right),$$

which implies that

$$\int \left\{ \log \frac{f_0(y|x)}{p(y|x, \theta, m} \right\}^2 1\left\{ \frac{p(y|x, \theta, m}{f_0(y|x)} < \lambda_n \right\} f(y|x) g_0(x) dy dx$$
$$\lesssim \left[\frac{a_{\sigma_n}^4 P_0(\|Z\| > a_{\sigma_n})}{\sigma_n^4} + \left\{ \frac{E_0(|Y|^8)^{1/2}}{\sigma_n^4} + \tilde{\epsilon}_n^{4b_1} E_0(\|X\|^8)^{1/2} \right\} (P_0(\|Z\| > a_{\sigma_n}))^{1/2} \right]$$
$$\lesssim \sigma_n^{2\beta + \varepsilon/2},$$

as in the proof of Corollary 6.1.

The lower bound for the prior probability of $S_{\theta^\star}$ and $m = K$ is the same as the one in Theorem 3.1, except d is replaced with d^0. The only additional calculation is as follows,

$$\Pi\left(s_j^y, s_{jk}^x \in \left[1, 1 + \sigma_n^{2\beta} \right], j = 1, 2, \ldots, K; k = 1, \ldots, d_x \right) \gtrsim \exp\{ -2\beta K d \log(1/\sigma_n) \},$$

which can be bounded from below as required by the arguments in the proof of Theorem 3.1. Thus, the prior thickness condition follows.

Finally, let us consider bounds on the sieve entropy and the prior probability of the sieve's complement. The argument here involves only minor changes in the proofs of Theorem 4.1 and Corollary 5.1. In the definition of sieve (4.1), let us add the following conditions for the β_js and the local scale parameters

$$\beta_j \in \left[-\overline{\beta}^x, \overline{\beta}^x\right]^{d_x}, \overline{\beta}^x = \overline{\mu}^x, j = 1, \ldots, m$$
$$s_j^y, s_{jk}^x \in \left[\underline{\sigma}, \overline{\sigma}\right], k = 1, \ldots, d_x,\ j = 1, 2, \ldots, m.$$

As in Corollary 5.1 and Corollary 6.1, we aim to find the covering number of $\mathcal{F}$ in d_1 instead of d_{SS}. First, let us replace the definition of S_σ in the proof of Corollary 5.1 with

$$S_\sigma = \left\{\sigma^l, l = 1, \ldots, N_\sigma = \left\lceil \log\left(\overline{\sigma}^2/\underline{\sigma}^2\right)/\left(\log\left(1+\underline{\sigma}^4\epsilon\right)/\left(2\cdot 384(\overline{\mu}^x)^2\max\{d_x, d_y\}\right)\right)\right\rceil,\right.$$
$$\left.\sigma^1 = \underline{\sigma}, \left(\sigma^{l+1} - \sigma^l\right)/\sigma^l = \underline{\sigma}^4\epsilon/\left(2\cdot 384(\overline{\mu}^x)^2\max\{d_x, d_y\}\right)\right\}$$

and use this S_σ as the grid for s_j^y, s_{jk}^x, σ^y, and σ_k^x, $k = 1, \ldots, d_x$, $j = 1, 2, \ldots, m$. Note that for $\tilde{\sigma} > \sigma$ and $\tilde{s} > s$, $|\sigma s - \tilde{\sigma}\tilde{s}|/(\sigma s) \le |\sigma - \tilde{\sigma}|/\sigma + |s - \tilde{s}|/s$ and that is why 384 is replaced by $2\cdot 384$ in the new definition of S_σ. Since $s_j^y\sigma^y, s_{jk}^x\sigma_k^x \in [\underline{\sigma}^2, \overline{\sigma}^2]$, all the bounds obtained in Corollary 5.1 now involve $(\underline{\sigma}^2, \overline{\sigma}^2)$ in place of $(\underline{\sigma}, \overline{\sigma})$.

Another difference is in the treatment of the new term $x'\beta_j$. Observe that for $\beta_j^{(1)}, \beta_j^{(2)} \in \mathcal{F}$ for $j = 1, \ldots, m$,

$$\int_{\mathcal{X}} \max_{1\le j\le m}\left|x'\beta_j^{(1)} - x'\beta_j^{(2)}\right| g_0(x)dx \le \max_{1\le j\le m}\left\|\beta_j^{(1)} - \beta_j^{(2)}\right\| \int_{\mathcal{X}} \|x\|\, g_0(x)dx.$$

Let us define S_β^m to contain centers of $|S_\beta^m| = \lceil 2\cdot 192 d_x \int_{\mathcal{X}} \|x\|\, g_0(x)dx(\overline{\beta}^x)^2/(\underline{\sigma}^4\epsilon)\rceil$ equal length intervals partitioning $[-\overline{\beta}, \overline{\beta}]$. $S_{\mu^x}^m$ now contains centers of $|S_{\mu^x}^m| = \lceil 2\cdot 192 d_x(\overline{\mu}^x)^2/(\underline{\sigma}^2\epsilon)\rceil$ equal length intervals partitioning $[-\overline{\mu}^x, \overline{\mu}^x]$.

As in the proof of Corollary 5.1, we thus obtain

$$J(\epsilon, \mathcal{F}, d_1) \le H\cdot\left\lceil\frac{16\overline{\mu}d_y}{\underline{\sigma}^2\epsilon}\right\rceil^{Hd_y}\cdot\left\lceil\frac{2\cdot 192 d_x(\overline{\mu}^x)^2}{\underline{\sigma}^4\epsilon}\right\rceil^{Hd_x}\cdot H\left\lceil\frac{\log(\underline{\alpha}^{-1})}{\log(1+\epsilon/[12H])}\right\rceil^{H-1}$$
$$\cdot\left\lceil\frac{2\cdot 192 d_x\int_{\mathcal{X}}\|x\|\, g_0(x)dx\cdot\overline{\beta}^2}{\underline{\sigma}^4\epsilon}\right\rceil^{Hd_x}$$
$$\cdot\left\lceil\frac{\log\left(\overline{\sigma}^2/\underline{\sigma}^2\right)}{\log\left(1+\underline{\sigma}^4\epsilon/\left[2\cdot 384(\overline{\mu}^x)^2\max\{d_x, d_y\}\right]\right)}\right\rceil^{d(H+1)}.$$

Observe that $\Pi(\mathcal{F}^c)$ is bounded above by

$$H^2\exp\left\{-a_{13}\overline{\mu}^{\tau_3}\right\} + H^2\exp\left\{-a_{16}(\overline{\mu}^x)^{\tau_5}\right\} + H^2\underline{\alpha}^{a/H} + \exp\left\{-a_{10}H(\log H)^{\tau_1}\right\}$$
$$+ d\,\Pi\left(\sigma^y \notin [\underline{\sigma}, \overline{\sigma}]\right) + dH\,\Pi\left(s_j^y \notin [\underline{\sigma}, \overline{\sigma}]\right).$$

The rest of the proof follows the argument in the proof of Theorem 4.2 with the same sequences, except $\underline{\sigma} = n^{-1/a_3}$ (as the prior for $\left(s_j^y\right)^2$ satisfies the same conditions ((2.4)–(2.6)) as the prior for σ^y) and $\overline{\mu}^x = n^{1/\tau_5}$. Thus, the claim of Corollary 6.1 holds. ■