

Bernoulli **22**(1), 2016, 396–420
DOI: [10.3150/14-BEJ663](https://doi.org/10.3150/14-BEJ663)

Adaptive Bayesian density regression for high-dimensional data

WEINING SHEN¹ and SUBHASHIS GHOSAL²

¹*Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, USA. E-mail: wshen@mdanderson.org*

²*Department of Statistics, North Carolina State University, 4276 SAS Hall, 2311 Stinson Drive, Raleigh, NC 27695, USA. E-mail: sghosal@ncsu.edu*

Density regression provides a flexible strategy for modeling the distribution of a response variable Y given predictors $\mathbf{X} = (X_1, \dots, X_p)$ by letting that the conditional density of Y given \mathbf{X} as a completely unknown function and allowing its shape to change with the value of \mathbf{X} . The number of predictors p may be very large, possibly much larger than the number of observations n , but the conditional density is assumed to depend only on a much smaller number of predictors, which are unknown. In addition to estimation, the goal is also to select the important predictors which actually affect the true conditional density. We consider a nonparametric Bayesian approach to density regression by constructing a random series prior based on tensor products of spline functions. The proposed prior also incorporates the issue of variable selection. We show that the posterior distribution of the conditional density contracts adaptively at the truth nearly at the optimal oracle rate, determined by the unknown sparsity and smoothness levels, even in the ultra high-dimensional settings where p increases exponentially with n . The result is also extended to the anisotropic case where the degree of smoothness can vary in different directions, and both random and deterministic predictors are considered. We also propose a technique to calculate posterior moments of the conditional density function without requiring Markov chain Monte Carlo methods.

Keywords: adaptive estimation; density regression; high-dimensional models; MCMC-free computation; nonparametric Bayesian inference; posterior contraction rate; variable selection

1. Introduction

We consider Bayesian estimation of the conditional density of a response Y given a large number of predictors $\mathbf{X} = (X_1, \dots, X_p)$, where p is possibly much larger than the sample size n . This problem is sometimes referred to as density regression and has received attention in many scientific application areas such as genome association studies. Non-Bayesian approaches to density regression usually focus on the kernel approach [14, 25],

This is an electronic reprint of the original article published by the ISI/BS in *Bernoulli*, 2016, Vol. 22, No. 1, 396–420. This reprint differs from the original in pagination and typographic detail.

which requires estimating the bandwidth using cross-validation [15], bootstrap [25] or other methods.

In the Bayesian literature, there are two common approaches to density regression. One approach models the joint density and obtains the conditional density as a by-product. The other approach directly models the conditional density while leaving the marginal distribution of \mathbf{X} unspecified. In this paper, we focus on the latter approach in a nonparametric manner. Many of the existing methods are based on assigning priors on the space of densities through countable mixtures of location-scale densities [32] or through generalizing stick-breaking representations [7, 11, 12, 23]. Priors obtained by transforming a Gaussian process [27, 43] and a multivariate generalization of a beta process [44] have also been considered. Ma [31] proposed a generalized Pólya tree, which possesses nice posterior conjugacy properties and hence allows fast computation.

In modern data analysis, often the data may be high-dimensional. Statistical analysis in such a setting is possible only under some sparsity assumption and only if a variable selection procedure is implemented. Many variable selection techniques have been introduced in the frequentist literature, such as discrete subset selection and penalization methods. Popular methods include the least absolute shrinkage and selection operator (lasso) introduced in [41] and the sure independence screening (SIS) proposed in [13]. Under the $p \gg n$ setting, oracle properties of lasso-type estimators have been established for parametric models including linear regression in [22], generalized linear model in [45] and for nonparametric additive models in [26]. For nonparametric (conditional) density estimation problems, however, similar results are only obtained under a fixed p setting in [24, 30].

Bayesian variable selection methods have also gained popularity. For example, stochastic search variable selection (SSVS) adopts an efficient sampling-based method to avoid comparing all possible sub-models [5, 16, 17]. Bayesian model averaging methods incorporate model uncertainty into estimation and predictions [2, 3]. Bayesian variable selection is commonly accomplished by assigning a Bernoulli distribution prior on each covariate [7], whereas an efficient SSVS algorithm is implemented to search the model space and to combine the posterior estimation results from different models. Tokdar *et al.* [43] extended variable selection to dimension reduction by allowing the true sets of covariates determined by a sub-linear space of \mathbf{X} through a projection operator. While these proposed methods show promising numerical results, rates of contraction are largely unknown. Moreover, modern applications often require that we allow the dimension p of the predictor to be also large, possibly much larger than n . So far such results are largely missing from the Bayesian nonparametric literature.

In the linear regression problem, recovery of the regression coefficients requires non-trivial assumptions on the structure of covariates, for example, the restricted isometry property or the compatibility condition to make the underlying problem well posed; see [4], although the corresponding problem of estimating the regression function does not require such conditions, see, for example, [8] for a discussion under a Bayesian framework. In the density regression context, the recovery of the conditional density is analogous to that of the regression function in the linear regression context and hence does not require such conditions.

In the recent years, the literature on Bayesian asymptotics has flourished with many fundamental breakthroughs. General results for posterior contraction rates were established in [18–20, 35, 39, 46]. For density regression models, a consistency result was obtained by Tokdar *et al.* [43] for a logistic Gaussian process prior, by Norets and Pelenis [33] for a kernel stick-breaking process prior and by Pati *et al.* [34] for a probit stick-breaking process prior. Tokdar [42] also obtained the posterior convergence rate for the logistic Gaussian process prior given a fixed p . For high-dimensional Bayesian models, there are very few contraction rates results available. Parametric models have been studied by Jiang [28] for generalized linear model and by Castillo and van der Vaart [6] for Gaussian white noise model. A classification model with categorical predictors was considered by Yang and Dunson [47], who constructed priors using tensor factorizations and obtained a posterior contraction rate allowing p to grow exponentially with n .

In this paper, we consider the Bayesian density regression problem using a finite linear combination of tensor products of B-splines to construct a prior distribution. We obtain the posterior contraction rate under the $p \gg n$ setting and show that the rate is adaptive for both dimension and smoothness in the sense that it agrees with the optimal rate of convergence (up to a logarithmic factor) of the oracle procedure, that uses the knowledge of true predictors and the underlying smoothness of the true conditional density, simultaneously for all smoothness levels and dimension of true predictors. We further extend the result to the anisotropic situation where smoothness can vary in different coordinates of the true conditional density function, and allow both random and deterministic predictors. We also devise an effective computing strategy. Typically, a reversible jump Markov chain Monte Carlo (RJMCMC) introduced by [21] is used for Bayesian computation for models with varying dimension. For high dimensional data, RJMCMC-based methods may be computationally challenging and may give unreliable results due to limited exploration of the model space. We propose a group-wise Dirichlet distribution prior on the coefficients of B-spline functions that leads to a conjugacy-like structure which can be utilized to develop a computing algorithm based on direct sampling without resorting to MCMC techniques. As in the univariate density estimation example in [37], the proposed computing method presents closed form expressions for posterior moments including the mean and the variance.

The paper is organized as follows. In the next section, we describe the model and the prior and discuss some preliminaries on tensor product of B-splines. Posterior contraction rates for both the isotropic and anisotropic cases and for both random and deterministic predictors are obtained in Section 3. Computational strategies are described and simulation results are presented in Section 4. Proofs are presented in Section 5.

2. Bayesian density regression

2.1. Notation

Let $\mathbb{N} = \{1, 2, \dots\}$, $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ and Δ_J be the unit J -dimensional simplex. For any real number x , define $\lfloor x \rfloor$ to be the largest integer less than or equal to x . For a multi-index $\mathbf{l} = (l_1, \dots, l_d) \in \mathbb{N}_0^d$, $d \in \mathbb{N}$, we define the sum $\mathbf{l} = l_1 + \dots + l_d$ and the mixed partial

derivative operator $D^{\mathbf{l}} = \partial^{\mathbf{l}} / \partial x_1^{l_1} \cdots \partial x_d^{l_d}$. For a bounded open connected set $\Omega \subset \mathbb{R}^d$ (e.g., $\Omega = (0, 1)^d$), define the β -Hölder class $\mathcal{C}^\beta(\Omega)$ as the collection of functions f on Ω that has bounded mixed partial derivatives $D^{\mathbf{l}}f$ of all orders up to $\mathbf{l} \leq \beta_0$ and that for every $\mathbf{l} \in \mathbb{N}_0^d$ satisfying $\mathbf{l} = \beta_0$,

$$|D^{\mathbf{l}}f(\mathbf{x}) - D^{\mathbf{l}}f(\mathbf{y})| \leq C \|\mathbf{x} - \mathbf{y}\|_2^{\beta - \beta_0} \quad (1)$$

for some constant $C > 0$, any $\mathbf{x}, \mathbf{y} \in \Omega$ and β_0 as the largest integer strictly smaller than β . Any such function uniquely extends to a continuous function on the closure of Ω .

Let the indicator function be denoted by $\mathbb{1}$. We use “ \lesssim ” to denote an inequality up to a constant multiple. We write $f \asymp g$ if $f \lesssim g \lesssim f$. Let $D(\varepsilon, T, \rho)$ denote the packing number, which is defined as the maximum cardinality of an ε -dispersed subset of T with respect to distance ρ . The symbol \mathbb{P} will stand for a generic probability measure.

2.2. B-spline and its tensor-products

B-spline functions and their tensor-products have been widely used to approximate functions in both mathematics and statistics literature. Here we provide a brief overview of their definitions and approximation properties; see more descriptions in [9]. For natural numbers K and $q \in \mathbb{N}$, let the unit interval $(0, 1)$ be divided into K equally spaced subintervals. A spline of order q with knots at the end points of these intervals is a function f such that the restriction of f in each subinterval is a polynomial of degree less than q and if $q \geq 2$, f is $(q - 2)$ -times continuously differentiable (interpreted as only continuous if $q = 2$). Splines of order q form a linear space of dimension $J = K + q - 1$, a convenient basis of which is given by the set of B-splines B_1, \dots, B_J . In particular, if $q = 1$, then corresponding B-splines form the Haar basis $\{\mathbb{1}_{(j-1/J, j/J]} : 1 \leq j \leq J\}$. The B-splines are nonnegative and add up to one at any x . Each B_j is positive only on an interval of length q/K and at most q many B-splines are nonzero at any given x . Most importantly, splines of order q with knots at $\{0, 1/K, \dots, (K - 1)/K, 1\}$ approximate any function in $C^\alpha(0, 1)$ at the rate $K^{-\alpha}$, or equivalently, any $f \in C^\alpha(0, 1)$ can be approximated by a linear combination of B_1, \dots, B_J up to an error of the order $J^{-\alpha}$.

This idea works in multidimensional case as well. For $(0, 1)^d$ and $d \in \mathbb{N}$, we split $(0, 1)$ into K_i equal intervals and consider corresponding spline functions B_1, \dots, B_{J_i} in the i th direction, where $J_i = q + K_i - 1$. Hence, there are $\prod_{i=1}^d K_i$ equal cubes in total. Define tensor-product B-spline basis functions as the product of univariate basis functions of each direction:

$$B_{\mathbf{j}}(x_1, \dots, x_d) = \prod_{k=1}^d B_{j_k}(x_k), \quad \mathbf{j} = (j_1, \dots, j_d), j_k = 1, \dots, J_k, k = 1, \dots, d. \quad (2)$$

For simplicity, we use \mathbf{B} to denote a column vector of all basis functions and define the total number of basis functions by $J = \prod_{k=1}^d J_k$.

Tensor-products of B-splines maintain a lot of nice properties that the univariate B-splines enjoy. In the following, we list a few of them that will be used in our modeling:

- (i) $0 \leq B_{\mathbf{j}} \leq 1$, for every $\mathbf{j} = (j_1, \dots, j_d) \in \{1, \dots, J_1\} \times \dots \times \{1, \dots, J_d\}$.
- (ii) $\sum_{j_1=1}^{J_1} \dots \sum_{j_d=1}^{J_d} B_{\mathbf{j}}(\mathbf{x}) = 1$, for every $\mathbf{x} \in (0, 1)^d$.
- (iii) For every \mathbf{x} , $B_{\mathbf{j}}(\mathbf{x}) > 0$ only if $\lfloor x_i K \rfloor \leq j_i \leq \lfloor x_i K \rfloor + q - 1$ for every $i = 1, \dots, d$.

We also define the normalized version of a univariate B-spline B by $\bar{B} = B / \int_0^1 B(x) dx$. Like univariate B-splines, the approximation ability of tensor-product B-splines is determined by the smoothness level α of the function to be approximated and J provided that q is chosen to be larger than α . In the following lemma, we state their approximation results. In particular, the result in part (c) suggests that the approximation power remains the same when the coefficients satisfy certain restrictions (positive, adds up to one), which later can help us assign prior distributions.

Lemma 1. (a) For any function $f \in \mathcal{C}^\beta((0, 1)^d)$, $0 < \beta \leq q$, there exists $\boldsymbol{\theta} \in \mathbb{R}^J$ and a constant $C_1 > 0$ such that

$$\left\| f - \sum_{j_1=1}^{J_0} \dots \sum_{j_d=1}^{J_0} \theta_{\mathbf{j}} B_{\mathbf{j}}(\mathbf{x}) \right\|_{\infty} \leq C_1 J_0^{-\beta} \|f^{(\beta)}\|_{\infty},$$

where $\mathbf{j} = (j_1, \dots, j_d)$.

(b) Further, if $f > 0$, then for sufficiently large J_0 , we can choose every element of $\boldsymbol{\theta}$ to be positive.

(c) Assume that $f(y|x_1, \dots, x_d)$ is a positive density function in y for every (x_1, \dots, x_d) and as a function of (y, x_1, \dots, x_d) belongs to $\mathcal{C}^\beta((0, 1)^{d+1})$, where $0 < \beta \leq q$. Then for sufficiently large $J = J_0^{d+1}$, there exists $\boldsymbol{\eta} \in (0, 1)^J$ satisfying $\sum_{j_0=1}^{J_0} \eta_{j_0, j_1, \dots, j_d} = 1$ for every fixed $(j_1, \dots, j_d) \in \{1, \dots, J_0\}^d$, and a constant $C_2 > 0$ such that

$$\left\| f(y|x_1, \dots, x_d) - \sum_{j_0=1}^{J_0} \dots \sum_{j_d=1}^{J_0} \eta_{j_0, \dots, j_d} \bar{B}_{j_0}(y) \prod_{k=1}^d B_{j_k}(x_k) \right\|_{\infty} \leq C_2 J_0^{-\beta} = C_2 J^{-\beta/(d+1)}.$$

2.3. Data generating process and the prior

We consider the data generated from n independent and identically distributed pairs of observations $(Y_1, \mathbf{X}_1), \dots, (Y_n, \mathbf{X}_n)$, where $Y_i \in (0, 1)$ and $\mathbf{X}_i \in (0, 1)^p$ for every $i = 1, \dots, n$, and $p \in \mathbb{N}$. It may be noted that the unit intervals appearing in the ranges of these random variables are not special as we can apply an appropriate affine transform on the data otherwise. We assume that Y is related only to d covariates, say $X_{m_1^0}, \dots, X_{m_d^0}$, that is the conditional density of Y given \mathbf{X} is a function of these coordinates only. This is an important sparsity assumption that will allow us to make valid inference about the conditional density even when p is very large, provided that d is small. However, neither d nor these indexes are known. The goal is to estimate the conditional density of Y given \mathbf{X} with accuracy comparable with the oracle procedure which assumes the knowledge of d and m_1^0, \dots, m_d^0 using a Bayesian procedure.

A prior on the conditional density given p covariate values x_1, \dots, x_p can be induced by a finite series expansion in terms of tensor product of B-splines

$$h(y, \mathbf{x} | J_0, \mathbf{J}, \boldsymbol{\eta}) = \sum_{j_0=1}^{J_0} \cdots \sum_{j_p=1}^{J_p} \eta_{j_0, \dots, j_p} \bar{B}_{j_0}(y) \prod_{k=1}^p B_{j_k}(x_k), \quad (3)$$

where $\mathbf{j} = (j_1, \dots, j_p)$ is a p -dimensional index, $\mathbf{J} = (J_1, \dots, J_p)$ and $\boldsymbol{\eta} = (\eta_{1,\mathbf{j}}, \dots, \eta_{J_0,\mathbf{j}})^T$ lies in a J_0 -dimensional simplex for every $\mathbf{j} \leq \mathbf{J}$. Note that η_{j_0, \dots, j_p} does not change for $j_k = 1, \dots, J_k$, if and only if the k th component does not affect the conditional density. In order to incorporate this feature in the prior, we define variable inclusion indicators $\gamma_k = \mathbb{1}$ (the k th variable is in the model). Let $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$. Thus, η_{j_0, \dots, j_p} depends only on j_0 and j_k with $k \in \text{Supp}(\boldsymbol{\gamma}) = \{k: \gamma_k = 1\} = \{m_1, \dots, m_r\}$ for $r = \sum_{k=1}^p \gamma_k$. Thus, the common value of η_{j_0, \dots, j_p} can be denoted by $\theta_{j_0, j_{m_1}, \dots, j_{m_r}}$. Now the conditional density can be written as

$$h(y, \mathbf{x} | J_0, \mathbf{J}, \boldsymbol{\eta}) = \sum_{r=0}^p \sum_{m_1, \dots, m_r} \sum_{j_0=1}^{J_0} \sum_{j_{m_1}=1}^{J_{m_1}} \cdots \sum_{j_{m_r}=1}^{J_{m_r}} \theta_{j_0, j_{m_1}, \dots, j_{m_r}} \bar{B}_{j_0}(y) \prod_{k: \gamma_k=1} B_{j_k}(x_k). \quad (4)$$

By assigning prior distributions on indicator variables γ_k , number of terms J_k and the corresponding coefficients $\boldsymbol{\theta}$, we obtain an induced prior on f . The prior on the model indicator $\boldsymbol{\gamma}$ is constructed by first putting a prior on the total model size r , and then selecting models with size r . More specifically, we construct the prior distribution through the following scheme:

- (A1) Prior on the model size r : Let $r = \sum_{k=1}^p \gamma_k$ be the number of variables included in the model and Π_1 be a fixed, positive prior probability mass function of r . Assume that there exists some constants $c_0, t_0 > 0$, such that for every $r \in \mathbb{N}$,

$$\Pi_1(r) \leq \exp\{-\exp(c_0 r^{t_0})\}. \quad (5)$$

- (A2) Prior on the inclusion variables $(\gamma_1, \dots, \gamma_p)$: Given a value of r , define the support of $\boldsymbol{\gamma}$ by $\{m_1, \dots, m_r\}$. We assume that the probability $\Pi_2(m_1, \dots, m_r | r)$ of each set of variables $1 \leq m_1 < \dots < m_r \leq p$ of size r satisfies

$$c'_1 \frac{1}{\binom{p}{r}} \leq \Pi_2(m_1, \dots, m_r | r) \leq c''_1 \frac{1}{\binom{p}{r}}$$

for some positive constant $c'_1 \leq c''_1$.

- (A3) Prior on the number of terms in the basis expansion: Given the model size r and active predictor indices $1 \leq m_1 < \dots < m_r \leq p$, let the number of terms in the basis expansion in $(r+1)$ -fold tensor products of B-splines be denoted by $(J_0, J_{m_1}, \dots, J_{m_r})$, and let $\Pi_3(\cdot | r; m_1, \dots, m_r)$ stand for their joint prior distribution. We let Π_3 be induced by independently distributed $J_0, J_{m_1}, \dots, J_{m_r}$ with

identical distribution $\tilde{\Pi}_3$, that is,

$$\Pi_3(J_0, J_{m_1}, \dots, J_{m_r} | r; m_1, \dots, m_r) = \tilde{\Pi}_3(J_0 | r) \prod_{k=1}^r \tilde{\Pi}_3(J_{m_k} | r)$$

and that for some fixed constants $c'_2, c''_2 > 0, \kappa' \geq \kappa'' \geq 1$,

$$\exp\{-c'_2 j^{r+1} (\log j)^{\kappa'}\} \leq \tilde{\Pi}_3(j | r) \leq \exp\{-c''_2 j^{r+1} (\log j)^{\kappa''}\}. \quad (6)$$

(A4) Prior on the coefficients: Given the values of r, m_1, \dots, m_r and $J_0, J_{m_1}, \dots, J_{m_r}$, recall that the conditional density of Y given \mathbf{X} is written as

$$\begin{aligned} h(\mathbf{x}, y | r; m_1, \dots, m_r; J_0, J_{m_1}, \dots, J_{m_r}; \boldsymbol{\theta}) \\ = \sum_{j_0=1}^{J_0} \sum_{j_{m_1}=1}^{J_{m_1}} \cdots \sum_{j_{m_r}=1}^{J_{m_r}} \theta_{j_0, j_{m_1}, \dots, j_{m_r}} \bar{B}_{j_0}(y) B_{j_{m_1}}(x_{m_1}) \cdots B_{j_{m_r}}(x_{m_r}), \end{aligned}$$

where $(\theta_{j_0, j_{m_1}, \dots, j_{m_r}} : 1 \leq j_0 \leq J_0) \in \Delta_{J_0}$ for every j_{m_1}, \dots, j_{m_r} . We let every $(\theta_{j_0, j_{m_1}, \dots, j_{m_r}} : 1 \leq j_0 \leq J_0) \in \Delta_{J_0}$ be distributed independently with identical prior distribution $\tilde{\Pi}_4(\cdot | J_0)$ and then denote the induced prior on the coefficients by $\Pi_4(\cdot | r; m_1, \dots, m_r; J_0, J_{m_1}, \dots, J_{m_r})$. In particular, we choose $\tilde{\Pi}_4$ to be a Dirichlet distribution $\text{Dir}(a, \dots, a)$, where a is a fixed positive constant.

Our prior (A1) on the model size includes the truncated binomial prior used in [28] as a special case. Condition (5) implies that $\Pi_1(r > \bar{r}) \leq \exp\{-\exp(c'_0 \bar{r}^{t_0})\}$ for some constant $c'_0 > 0$ and any $\bar{r} \in \mathbb{N}$. Since r should not be greater than p , which changes with n , we may also let Π_1 depend on n . In this case, we assume the decay rate to hold with c_0 and t_0 , which are both free from n , and for any fixed \bar{d} , $\Pi_1(\bar{d})$ is bounded below, or more generally $\Pi_1(\bar{d})$ satisfies $-\log \Pi_d(\bar{d}) = o(n^\delta)$ for all $\delta > 0$.

In (A2), an easy choice is to let $c'_1 = c''_1 = 1$, i.e., assign equal probability for choosing r indices from $\{1, \dots, p\}$. We may also allow c'_1, c''_1 depend on n as long as $\log(1/c'_1) = o(n^\delta)$ and $\log(c''_1) = o(n^\delta)$ for all $\delta > 0$. The posterior contraction rate to be obtained in the next section will remain the same.

For (A3), using the same prior $\tilde{\Pi}_3$ is not necessary, but it is a convenient and appropriate default choice. The independence between components is also not essential. In Section 3.2 when the true density function is anisotropic, we shall have to use a different and more complicated prior, which will be obviously appropriate for this isotropic case as well. Relation (6) is satisfied with $\kappa' = \kappa'' = 1$ if a zero-truncated Poisson distribution is assigned on J^{r+1} in the sense that N is a zero-truncated Poisson and $J = \lfloor N^{1/(r+1)} \rfloor$.

In (A4), the same value of a is not necessary, but we use it as a default choice. In particular, $a = 1$ leads to the uniform prior distribution on the simplex. The same contraction rate will be obtained as long as the parameters are kept in a fixed compact subset of $(0, \infty)$. More generally, we may allow the lower bound approach zero at most polynomially fast in n^{-1} , although the upper bound needs to be fixed.

3. Posterior contraction rates

3.1. Isotropic case

In this section, we establish results on posterior contraction rates for density regression. We allow the total number of covariates p diverge with the increasing sample size n . Let Π be the prior as defined in (A1)–(A4) and denote the posterior distribution based on n pairs of observations $(Y_1, \mathbf{X}_1), \dots, (Y_n, \mathbf{X}_n)$ by $\Pi(\cdot | \mathbf{X}^n, \mathbf{Y}^n)$. Let $\varepsilon_n \rightarrow 0$ be a sequence of positive numbers. Consider a suitable metric on the space of probability densities on $(0, 1)$, such as the Hellinger metric. Let G stand for the common distribution of $\mathbf{X}_1, \dots, \mathbf{X}_n$, which need not be known. We define the root average squared Hellinger distance on the space of conditional densities by

$$\rho^2(f_1, f_2) = \iint \{f_1^{1/2}(y|x_1, \dots, x_p) - f_2^{1/2}(y|x_1, \dots, x_p)\}^2 dy G(dx_1, \dots, dx_p), \quad (7)$$

where f_1 and f_2 stand for generic conditional densities of Y on $(0, 1)$ given \mathbf{X} in $(0, 1)^p$. Let f_0 be a fixed conditional density function for Y on $(0, 1)$ given \mathbf{X} in $(0, 1)^p$, standing for the true conditional density. We say that the posterior distribution of the density regression model based on Π contracts to f_0 at a rate ε_n in the metric ρ if for any $M_n \rightarrow \infty$,

$$\lim_{n \rightarrow \infty} \Pi[\{f: \rho(f_0, f) > M_n \varepsilon_n\} | \mathbf{X}^n, \mathbf{Y}^n] = 0 \quad \text{in probability.} \quad (8)$$

We make the following assumptions.

- (B1) The true density f_0 depends only on d predictors $X_{m_1^0}, \dots, X_{m_d^0}$, where d is a fixed number. Further we assume that as a function of y and $x_{m_1^0}, \dots, x_{m_d^0}$, we have $f_0 \in \mathcal{C}^\beta((0, 1)^{d+1})$ for some $0 < \beta \leq q$.
- (B2) The ambient dimension $p \leq \exp(Cn^\alpha)$ for $0 < \alpha < 1$. When $\alpha = 0$, we interpret this condition as $p \leq n^K$ for some constant $K > 0$.
- (B3) The true conditional density f_0 is bounded below by a positive constant \underline{m}_0 .

Theorem 1. *Suppose that we have i.i.d. observations $\mathbf{X}_1, \dots, \mathbf{X}_n$ from a possibly unknown probability distribution G on $(0, 1)^p$. Assume that the true conditional density satisfies conditions (B1)–(B3). If the prior satisfies conditions (A1)–(A4), then the posterior distribution of f contracts at f_0 at the rate*

$$\varepsilon_n = \max\{n^{-(1-\alpha)/2}(\log n)^{1/(2t_0)}, n^{-\beta/(2\beta+d+1)}(\log n)^{\kappa'\beta/(2\beta+d+1)}\} \quad (9)$$

with respect to ρ , where t_0 and κ' are defined in (A1) and (A3).

Theorem 1 establishes ε_n as a bound on the posterior contraction rate at f_0 . It is known that the minimax rate associated with estimating a $(d+1)$ -dimensional density lying in a β -Hölder class is $(n/\log n)^{-\beta/(2\beta+d+1)}$ (see [29]) with respect to the supremum norm.

The minimax rate of convergence for conditional densities with respect to the metric ρ is not known yet, but it is reasonable to expect that the rate $n^{-\beta/(2\beta+d+1)}$ up to a logarithmic factor applies in this situation as well, and can be taken as the oracle rate with which the rates obtained in Theorem 1 can be compared. Thus if p grows polynomially fast in n , then the rate we obtained coincides with the oracle rate up to a logarithmic factor. If p grows exponentially fast, then it makes an impact on the rate. Note that we obtain the optimal rate with the use of the same prior distribution for all values of α and β . Hence our estimation and variable selection procedure is rate-adaptive in the sense that the posterior automatically adapts to the unknown number of covariates d (i.e., the oracle dimension) in the true model and the smoothness level β . Our result also trivially contains the fixed dimensional situation where no variable selection is involved. Note that the contraction at the true density does not necessarily guarantee the convergence of the selected set of predictors to the true set of predictors. The question of recovering the true set of predictors remains open and is beyond the scope of the present paper. However, as contraction rates are regulated by the complexity of the underlying model determined by its dimension, it may be anticipated that the posterior distribution assigns most of its mass to low complexity models relative to the ambient dimension.

Remark 1. Theorem 1 establishes contraction rates for the posterior distribution of the entire conditional density function $f(y|\mathbf{x})$. As a consequence, we can obtain the same posterior contraction rate for other quantities of interest such as conditional quantile functions, conditional moment functions and so on. This rate may not be optimal for the estimation of these quantities because y has been integrated out, that is, we conjecture the optimal rate is $n^{-\beta/(2\beta+d)}$ instead of $n^{-\beta/(2\beta+d+1)}$, up to logarithmic factors.

Remark 2. After examining the proof, we find that condition (5) in (A1) can be relaxed if α is small. For example, if $\alpha = 0$, then we only need $\Pi_1(r) \leq \exp(-c_0 r^{t_0})$.

3.2. Anisotropic case

If predictors are qualitatively different, then it may be interesting to consider the situation where f_0 has different smoothness levels in different directions. In the following, we propose an alternative anisotropic smoothness assumption replacing condition (B1).

For $\beta = (\beta_0, \dots, \beta_d) \in \mathbb{N}^{d+1}$ and $\beta_0, \dots, \beta_d \leq q$, define a tensor Sobolev space $\mathcal{S}^\beta((0, 1)^{d+1})$ of functions f of $(d+1)$ variables by

$$\mathcal{S}^\beta((0, 1)^{d+1}) = \{f: \|D^{\mathbf{l}} f\|_\infty < \infty, \mathbf{l} = (l_0, \dots, l_d), l_k \leq \beta_k, k = 0, \dots, d\}$$

with an associated norm $\|\cdot\|_{\mathcal{S}^\beta}$ defined as

$$\|f\|_{\mathcal{S}^\beta} = \|f\|_\infty + \left\| \frac{d^{\beta_0} f}{dy^{\beta_0}} \right\|_\infty + \sum_{k=1}^d \left\| \frac{d^{\beta_k} f}{dx_k^{\beta_k}} \right\|_\infty.$$

As in Lemma 1, we show that the tensor-product B-splines still have nice approximation abilities within anisotropic function spaces.

Lemma 2. (a) For any function $f \in \mathcal{S}^\beta((0,1)^{d+1})$, where $0 < \beta_0, \dots, \beta_d \leq q$, there exists $\theta \in \mathbb{R}^{\prod_{k=0}^d J_k}$ and a constant $C_1 > 0$ depending only on q, d and β_0, \dots, β_d such that

$$\left\| f - \sum_{j_0=1}^{J_0} \cdots \sum_{j_d=1}^{J_d} \theta_{\mathbf{j}} \mathbf{B}_{\mathbf{j}}(\mathbf{x}) \right\|_{\infty} \leq C_1 \sum_{k=0}^d J_k^{-\beta_k} \left\| \frac{d^{\beta_k} f}{dx_k^{\beta_k}} \right\|_{\infty},$$

where $\mathbf{j} = (j_0, \dots, j_d)$.

(b) Further, if $f > 0$, we can choose every element of θ to be positive.

(c) Assume that $f(y|x_1, \dots, x_d)$ is a positive probability density function in y for every (x_1, \dots, x_d) and as a function of (y, x_1, \dots, x_d) belongs to $\mathcal{S}^\beta((0,1)^{d+1})$, where $\beta = (\beta_0, \dots, \beta_d) \in \mathbb{N}^{d+1}$ satisfying $0 < \beta_0, \dots, \beta_d \leq q$. Then there exists $\eta \in (0,1)^{\prod_{k=0}^d J_k}$ satisfying $\sum_{j_0=1}^{J_0} \eta_{j_0, j_1, \dots, j_d} = 1$ for every fixed $(j_1, \dots, j_d) \in \{1, \dots, J_1\} \times \cdots \times \{1, \dots, J_d\}$ and a constant $C_2 > 0$ such that

$$\left\| f(y|x_1, \dots, x_d) - \sum_{j_0=1}^{J_0} \cdots \sum_{j_d=1}^{J_d} \eta_{j_0, \dots, j_d} \bar{B}_{j_0}(y) \prod_{k=1}^d B_{j_k}(x_k) \right\|_{\infty} \leq C_2 \sum_{k=0}^d J_k^{-\beta_k}.$$

(B4) We assume that the true density f_0 is only related to d predictors with indices $1 \leq m_1^0 < \cdots < m_d^0 \leq p$, where d is a fixed number, and as a function of $(y, x_{m_1^0}, \dots, x_{m_d^0})$ belongs to $\mathcal{S}^\beta((0,1)^{d+1})$.

In order to obtain the adaptive convergence rate, we replace the independent prior distribution on J in condition (A3) by the following joint distribution condition.

(A3') Prior on the number of terms in basis expansion: Given the model size r and active predictor indices $1 \leq m_1 < \cdots < m_r \leq p$, let the number of terms in the basis expansion of $(r+1)$ -fold tensor products of B-splines be denoted by $(J_0, J_{m_1}, \dots, J_{m_r})$, and let $\Pi_3(\cdot|r; m_1, \dots, m_r)$ stand for their joint prior distribution. We assume that for some fixed constants $c'_2, c''_2 > 0, \kappa' \geq \kappa'' \geq 1$,

$$\begin{aligned} & \exp \left\{ -c'_2 J_0 \prod_{k=1}^r J_{m_k} \left(\log J_0 + \sum_{k=1}^r \log J_{m_k} \right)^{\kappa'} \right\} \\ & \leq \Pi_3(J_0, J_{m_1}, \dots, J_{m_r} | r; m_1, \dots, m_r) \\ & \leq \exp \left\{ -c''_2 J_0 \prod_{k=1}^r J_{m_k} \left(\log J_0 + \sum_{k=1}^r \log J_{m_k} \right)^{\kappa''} \right\}. \end{aligned}$$

Then we obtain the posterior convergence rate for anisotropic functions.

Theorem 2. *Suppose that we have i.i.d. observations $\mathbf{X}_1, \dots, \mathbf{X}_n$ from an unknown probability distribution G on $(0, 1)^p$. Assume that the true conditional density satisfies conditions (B2)–(B4). If the prior satisfies conditions (A1), (A2), (A3') and (A4), then the posterior distribution of f contracts at f_0 at the rate*

$$\varepsilon_n = \max\{n^{-(1-\alpha)/2}(\log n)^{1/(2t_0)}, n^{-\beta^*/(2\beta^*+d+1)}(\log n)^{\kappa'\beta^*/(2\beta^*+d+1)}\} \quad (10)$$

with respect to ρ , where $\beta^* = (d+1)(\sum_{k=0}^d \beta_k^{-1})^{-1}$ is the harmonic mean of $\{\beta_0, \dots, \beta_d\}$, t_0 and κ' are defined in (A1) and (A3').

Clearly, the rate reduces to that of the isotropic case when $\beta_0 = \dots = \beta_d$. Thus the rate now can be viewed as the optimal rate (up to a logarithmic factor) of estimating a $(d+1)$ -variate function with smoothness levels β_0, \dots, β_d . Note that the rate is determined by the harmonic mean of smoothness levels in different coordinates, thus the worst smoothness has the most prominent effect on the rate. However, the rate thus obtained is strictly better than that obtained by a naive application of Theorem 1 using the worst smoothness condition in all co-ordinates. Thus additional smoothness in other co-ordinates help improve the rate from the worst case scenario. This result agrees with the minimax rate associated with estimating a $(d+1)$ -dimensional anisotropic density with respect to the Hellinger distance (cf. [1]). Interestingly, the posterior automatically adapts to different vector of smoothness levels. Noticeably, as in the isotropic case, the ambient dimension p does not affect the rate except when it grows exponentially in n . It will be interesting to generalize the result to allow anisotropic Hölder classes with non-integer smoothness levels as in [38]. Since approximation properties of tensor product of B-splines are presently known only for integer smoothness in the anisotropic case, we restrict to smoothness parameters to integer values only.

3.3. Deterministic predictor variables

Our method also applies for the case of deterministic predictors. In order to obtain the posterior convergence rate, we first define the empirical measure $\mathbb{P}_n^{\mathbf{X}} = n^{-1} \sum_{i=1}^n \delta_{\mathbf{X}_i}$, in which $\delta_{\mathbf{X}_i}$ is a point mass probability measure at \mathbf{X}_i , based on the observations $\mathbf{X}_1, \dots, \mathbf{X}_n$. Then we define an empirical Hellinger distance on the space of conditional densities by

$$\rho_n^2(f_1, f_2) = \iint \{f_1^{1/2}(y|x_1, \dots, x_p) - f_2^{1/2}(y|x_1, \dots, x_p)\}^2 dy \mathbb{P}_n^{\mathbf{X}}(dx_1, \dots, dx_p), \quad (11)$$

where f_1 and f_2 are generic conditional densities of Y on $(0, 1)$ given \mathbf{X} in $(0, 1)^p$. We can obtain the same posterior contraction rates as the case of random predictors for both isotropic and anisotropic cases.

Theorem 3. *Suppose that we have deterministic predictors $\mathbf{X}_1, \dots, \mathbf{X}_n$ observed on $(0, 1)^p$. Assume that the prior on the conditional densities satisfies conditions (A1)–(A4). If the true conditional density satisfies (B1)–(B3), then the posterior distribution*

of f contracts at f_0 at the rate ε_n given by (9) with respect to ρ_n . If the true conditional density satisfies (B2)–(B4), and (A3) is replaced by (A3') in the prior specification, then the posterior distribution of f contracts at f_0 at the rate ε_n given by (10) with respect to ρ_n .

4. Numerical results

4.1. Computation

First, we ignore that we have a variable selection issue and pretend that we know which d predictors are relevant, say $\{m_1, \dots, m_d\}$. Thus, we may pretend that we are in the fixed dimensional setting $p = d$ and all predictors are relevant. Then given the observations $(\mathbf{X}_i, Y_i) = (X_{i1}, \dots, X_{id}; Y_i)$, $i = 1, \dots, n$, the conditional likelihood

$$L(\boldsymbol{\eta}, \mathbf{J} | \mathbf{X}, Y) = \prod_{i=1}^n \left\{ \sum_{j_0=1}^{J_0} \sum_{j_{m_1}=1}^{J_{m_1}} \cdots \sum_{j_{m_d}=1}^{J_{m_d}} \eta_{j_0, j_{m_1}, \dots, j_{m_d}} \bar{B}_{j_0}(Y_i) \prod_{k=1}^d B_{j_{m_k}}(X_{i, j_{m_k}}) \right\} \quad (12)$$

expands to $\sum_{\mathbf{s}_1 \in \Lambda_{\mathbf{J}}} \cdots \sum_{\mathbf{s}_n \in \Lambda_{\mathbf{J}}} \prod_{i=1}^n \eta_{\mathbf{s}_i} B_{\mathbf{s}_i}^*(Y_i | \mathbf{X}_i)$, where $\Lambda_{\mathbf{J}} = \{1, \dots, J_0\} \times \cdots \times \{1, \dots, J_{m_d}\}$ and $B_{\mathbf{s}}^*(y | \mathbf{x})$ is defined as $\bar{B}_{s_0}(y) \prod_{k=1}^d B_{s_{m_k}}(x_{m_k})$ for every $(d+1)$ -dimensional index \mathbf{s} taking value in $\Lambda_{\mathbf{J}}$ and $\mathbf{J} = (J_0, J_{m_1}, \dots, J_{m_d}) \in \mathbb{N}^{d+1}$. Since

$$P(\mathbf{J}, \boldsymbol{\eta} | \mathbf{X}, Y) \propto P(\mathbf{X}, Y | \mathbf{J}, \boldsymbol{\eta}) \Pi_4(\boldsymbol{\eta} | \mathbf{J}) \tilde{\Pi}_3(J_0) \prod_{k=1}^d \tilde{\Pi}_3(J_{m_k}),$$

the posterior mean of $f(y | \mathbf{x})$ at point is given by

$$\frac{\sum_{J_0=1}^{\infty} \cdots \sum_{J_{m_d}=1}^{\infty} \tilde{\Pi}_3(J_0) \prod_{k=1}^d \tilde{\Pi}_3(J_{m_k}) \int f(y | \mathbf{x}, \boldsymbol{\eta}, \mathbf{J}) \Pi_4(\boldsymbol{\eta}) L(\boldsymbol{\eta}, \mathbf{J} | \mathbf{X}, Y) d\boldsymbol{\eta}}{\sum_{J_0=1}^{\infty} \cdots \sum_{J_{m_d}=1}^{\infty} \tilde{\Pi}_3(J_0) \prod_{k=1}^d \tilde{\Pi}_3(J_{m_k}) \int \Pi_4(\boldsymbol{\eta}) L(\boldsymbol{\eta}, \mathbf{J} | \mathbf{X}, Y) d\boldsymbol{\eta}}. \quad (13)$$

In view of (12) and the form of $f(y | \mathbf{x})$ given by (4), both numerator and denominator of (13) involve computing integrals of the form $I(\mathbf{s}_1, \dots, \mathbf{s}_n) = \int_{\boldsymbol{\eta}} \Pi(\boldsymbol{\eta}) \prod_{k=1}^n \eta_{\mathbf{s}_k} d\boldsymbol{\eta}$. We collect coefficients $\boldsymbol{\eta}$ with the same index together to form their powers and observe that, by (A4), coefficients whose index differ anywhere except in the zeroth co-ordinate are independent, and the collection of coefficients with the same last d components are Dirichlet distributed. In view of the conjugacy of the Dirichlet functional form with respect to a multinomial function, $I(\mathbf{s}_1, \dots, \mathbf{s}_n)$ can be written down in terms of products of certain Dirichlet forms, and hence can be computed for any given $(\mathbf{s}_1, \dots, \mathbf{s}_n)$. Therefore (13) simplifies to

$$\frac{\sum_{J_0=1}^{\infty} \cdots \sum_{J_d=1}^{\infty} \tilde{\Pi}_3(J_0) \prod_{k=1}^d \tilde{\Pi}_3(J_{m_k}) \sum_{\mathbf{s}_0 \in \Lambda_{\mathbf{J}}} \cdots \sum_{\mathbf{s}_n \in \Lambda_{\mathbf{J}}} I(\mathbf{s}_0, \dots, \mathbf{s}_n) \prod_{i=0}^n B_{\mathbf{s}_i}^*(Y_i | \mathbf{X}_i)}{\sum_{J_0=1}^{\infty} \cdots \sum_{J_d=1}^{\infty} \tilde{\Pi}_3(J_0) \prod_{k=1}^d \tilde{\Pi}_3(J_{m_k}) \sum_{\mathbf{s}_1 \in \Lambda_{\mathbf{J}}} \cdots \sum_{\mathbf{s}_n \in \Lambda_{\mathbf{J}}} I(\mathbf{s}_1, \dots, \mathbf{s}_n) \prod_{i=1}^n B_{\mathbf{s}_i}^*(Y_i | \mathbf{X}_i)}, \quad (14)$$

where (\mathbf{X}_0, Y_0) stands for (\mathbf{x}, y) .

Now, we take the variable selection into consideration. Suppose that the proposed model size is r , which follows the prior distribution Π_1 . Given r , let the covariates X_{m_1}, \dots, X_{m_r} enter the model with probability $\Pi_2(m_1, \dots, m_r | r)$. Define

$$\begin{aligned} W^0(m_1, \dots, m_r | \mathbf{X}, Y) &= \sum_{J_0=1}^{\infty} \sum_{J_{m_1}=1}^{\infty} \cdots \sum_{J_{m_r}=1}^{\infty} \tilde{\Pi}_3(J_0) \prod_{k=1}^r \tilde{\Pi}_3(J_{m_k}) \\ &\quad \times \sum_{\mathbf{s}_0 \in \Lambda_{\mathbf{J}}} \cdots \sum_{\mathbf{s}_n \in \Lambda_{\mathbf{J}}} I(\mathbf{s}_0, \dots, \mathbf{s}_n) \prod_{i=0}^n B_{\mathbf{s}_i}^*(Y_i | \mathbf{X}_i), \\ W^1(m_1, \dots, m_r | \mathbf{X}, Y) &= \sum_{J_0=1}^{\infty} \sum_{J_{m_1}=1}^{\infty} \cdots \sum_{J_{m_r}=1}^{\infty} \tilde{\Pi}_3(J_0) \prod_{k=1}^r \tilde{\Pi}_3(J_{m_k}) \\ &\quad \times \sum_{\mathbf{s}_1 \in \Lambda_{\mathbf{J}}} \cdots \sum_{\mathbf{s}_n \in \Lambda_{\mathbf{J}}} I(\mathbf{s}_1, \dots, \mathbf{s}_n) \prod_{i=1}^n B_{\mathbf{s}_i}^*(Y_i, \mathbf{X}_i). \end{aligned}$$

Then the posterior mean of $f(y | \mathbf{x})$ is given by

$$\frac{\sum_{r=1}^{\bar{r}} \Pi_1(r) \sum_{1 \leq m_1 < \dots < m_r \leq p} \Pi_2(m_1, \dots, m_r | r) W^0(m_1, \dots, m_r | \mathbf{X}, Y)}{\sum_{r=1}^{\bar{r}} \Pi_1(r) \sum_{1 \leq m_1 < \dots < m_r \leq p} \Pi_2(m_1, \dots, m_r | r) W^1(m_1, \dots, m_r | \mathbf{X}, Y)}. \quad (15)$$

Similar expressions can be obtained for other posterior moments, in particular, for the posterior second moment and hence the posterior variance. This estimate can be viewed as a kernel mixture estimator whose kernel is determined jointly by selected covariates and associated tensor product B-splines. Since a B-spline basis function takes nonzero values only at q intervals, the calculation of W^0 for a given r involves $(J_{\max} - J_{\min} + 1)^{r+1} q^{(r+1)(n+1)}$ terms if we restrict J_0 and each J_{m_k} , $k = 1, \dots, r$, to take values between J_{\min} and J_{\max} . Then there will be $\sum_{r=1}^{\bar{r}} \binom{p}{r} (J_{\max} - J_{\min} + 1)^{r+1} q^{(r+1)(n+1)}$ terms in total. Instead of evaluating all terms, we randomly sample a number of terms in both numerator and denominator and take the associated average values. If we choose $q = 1$, then the prior can be viewed as a multivariate random histogram and the number of terms in the expression for the posterior mean will reduce to $\sum_{r=1}^{\bar{r}} \binom{p}{r} (J_{\max} - J_{\min} + 1)^{r+1}$, although the resulting density estimate will be less smooth and the rate adaptation property of the posterior distribution will apply only to smoothness up to order 1. We shall make this choice in our simulation studies to save on computational cost in exchange of sacrificing some smoothness.

4.2. Simulation results

In the following, we provide more details in the prior construction of our model.

(C0) We choose $q = 1$, which leads to histogram basis functions (Haar basis).

- (C1) We assign a uniform prior on the model size ranging from 2 to $\bar{r} = 7$.
- (C2) The prior probability of γ_k follows a Bernoulli distribution with parameter w_k for $0 \leq w_k \leq 1$ and $k = 1, \dots, p$. The values of w_k can depend on the marginal correlation between X_k and Y .
- (C3) Given the model size r chosen, we generate a zero-truncated Poisson random variable K with mean $\lambda = 100$ and then assign the integer part of $K^{1/(r+1)}$ to the number of expansion terms J . We restrict J between 4 and 8, that is, $J_{\min} = 4$ and $J_{\max} = 8$. Then (A3) holds for $\kappa' = \kappa'' = 1$.
- (C4) Given J_0 , we let the vector $(\theta_{j_0, j_{m_1}, \dots, j_{m_r}} : j_0 = 1, \dots, J_0)$ have the uniform distribution over the J_0 -simplex for every feasible value of \mathbf{j} . Then condition (A4) is satisfied for $a = 1$.

We apply the MCMC-free calculation method described in Section 4.1 on the following two examples,

$$Y|\mathbf{X} \sim \text{Beta}(4X_1 + 3X_2^2, 10X_2), \quad (16)$$

$$Y|\mathbf{X} \sim \text{Beta}(5X_2 \exp(2X_1), 5X_3^2 + 3X_4). \quad (17)$$

For each example, we generate p covariates X_1, \dots, X_p uniformly from $[0.05, 0.95]$. In the computation of (15), we randomly draw $N^* = 100$ or 500 terms in the sums of W^1 and W^0 for every fixed choice of m_1, \dots, m_r and r . We compare our method (rsp) with least-squares kernel conditional density estimation (L-S) developed by [40], where they use L_1 -regularization to select variables. Prediction errors under the L_2 -loss and their maximum standard errors associated with 10 Monte Carlo replications are summarized in Tables 1 and 2.

Compared with the least-squares method, our approach has a better performance in most cases. Since we are directly sampling a fixed number of terms from the sums in (15), our prediction error does not change too much with p , which makes the proposed method outperform L-S when p is large. Comparing the prediction errors under the choices of $N^* = 100$ and 500, we find that their performances are quite close to each other. Hence direct sampling does not introduce too much variability. We also carry out

Table 1. Simulation example 1: true density generated by (16)

	$n = 100$			$n = 500$		
	L-S	rsp($N^* = 100$)	rsp($N^* = 500$)	L-S	rsp($N^* = 100$)	rsp($N^* = 500$)
$p = 5$	0.88	0.76	0.86	0.81	1.01	1.01
$p = 10$	1.04	0.78	0.81	1.10	1.10	1.12
$p = 50$	0.69	0.66	0.68	0.73	1.05	1.04
$p = 100$	0.69	0.72	0.70	0.63	0.76	0.96
$p = 500$	0.96	0.78	0.67	0.86	0.76	0.80
$p = 1000$	1.15	0.59	0.63	1.26	0.77	0.95
max s.e.	0.07	0.06	0.06	0.08	0.10	0.10

Table 2. Simulation example 2: true density generated by (17)

	$n = 100$			$n = 500$		
	L-S	rsp($N^* = 100$)	rsp($N^* = 500$)	L-S	rsp($N^* = 100$)	rsp($N^* = 500$)
$p = 5$	0.76	0.61	0.64	0.86	0.74	0.71
$p = 10$	0.97	0.66	0.61	1.00	0.70	0.71
$p = 50$	0.69	0.64	0.61	0.72	0.71	0.74
$p = 100$	0.72	0.67	0.64	0.74	0.73	0.72
$p = 500$	0.95	0.61	0.71	1.00	0.73	0.78
$p = 1000$	1.26	0.68	0.66	1.25	0.69	0.73
max s.e.	0.08	0.04	0.05	0.08	0.06	0.06

a sensitivity analysis by using different parameter values in the prior distribution, for example, $\bar{r} = 6$, $\lambda = 50$ and $J_{\min} = 5$, $J_{\max} = 10$. Similar results are obtained. In practice, one may choose \bar{r} as a constant multiple (e.g., twice) of the possible maximal model size to let all important covariates be included in the considered model with a high probability. The Poisson mean parameter λ in (C3) shall be modified according to the choice of \bar{r} to ensure that $\lambda^{1/r}$ falls into an appropriate range, say, between 4 and 20.

5. Proofs

Proof of Theorem 1. Note that the conditional density $f(y|\mathbf{x})$ is the same as the joint density of (\mathbf{X}, Y) at (\mathbf{x}, y) with respect to the dominating measure μ equal to the product of G and the Lebesgue measure. Further, the distance ρ on the space of conditional densities is equivalent to the Hellinger distance on the space of joint densities with respect to μ . Hence, in order to derive contraction rate of the posterior distribution of the conditional density at a true density $f_0(\cdot|\cdot)$, we need only to apply the standard result on posterior convergence rate for (joint) densities given by Theorem 1 of [18]. The required conditions characterizing the posterior contraction rate $\varepsilon_n \rightarrow 0$ can therefore be rewritten in the present context as follows: there exists a sequence of subsets \mathcal{F}_n of the space of conditional densities, called sieves, such that

$$\Pi(\mathcal{F}_n^c) \lesssim \exp\{-8n\varepsilon_n^2\}, \quad (18)$$

$$\log D(\varepsilon_n, \mathcal{F}_n, \rho) \lesssim n\varepsilon_n^2, \quad (19)$$

$$\Pi(f: K(f_0, f) \leq \varepsilon_n^2, V(f_0, f) \leq \varepsilon_n^2) \gtrsim \exp\{-n\varepsilon_n^2\}, \quad (20)$$

where $K(f_0, f) = \iint f_0(y|\mathbf{x}) \log(f_0(y|\mathbf{x})/f(y|\mathbf{x})) dy dG(\mathbf{x})$ is the Kullback–Leibler divergence and $V(f_0, f) = \iint f_0(y|\mathbf{x}) \log^2(f_0(y|\mathbf{x})/f(y|\mathbf{x})) dy dG(\mathbf{x})$ is the Kullback–Leibler variation. We define a sieve in the following way:

$$\mathcal{F}_n = \{h(\mathbf{x}, y|r; m_1, \dots, m_r; J_0, J_{m_1}, \dots, J_{m_r}; \boldsymbol{\theta}): r \leq \bar{r}_n,$$

$$1 \leq m_1 < \dots < m_r \leq p; J_0, J_{m_1}, \dots, J_{m_r} \leq \tilde{J}_n, \boldsymbol{\theta} \in (\Delta_{J_0})^{\prod_{k=1}^r J_{m_k}}, \quad (21)$$

where $\tilde{J}_n = \lfloor (LJ_n^*)^{(d+1)/(r+1)} (\log n)^{\kappa/(r+1)} \rfloor$, J_n^* and \bar{r}_n are two sequences of number going to infinity, L and κ are some fixed positive constants. We shall choose the values of these numbers later.

We first verify (18). Note that $\Pi(\mathcal{F}_n^c)$ is bounded by

$$\begin{aligned} & \Pi_1(r > \bar{r}_n) + \sum_{r=1}^{\bar{r}_n} \sum_{1 \leq m_1 < \dots < m_r \leq p} \Pi_3(J_{m_k} > \tilde{J}_n \text{ for some } k = 1, \dots, r | r, m_1, \dots, m_r) \\ & \leq \exp\{-\exp(c_0 \bar{r}_n^{t_0})\} + \sum_{r=1}^{\bar{r}_n} \binom{p}{r} r \tilde{\Pi}_3(J > \tilde{J}_n) \\ & \leq \exp\{-\exp(c_0 \bar{r}_n^{t_0})\} + \bar{r}_n p^{\bar{r}_n} \sum_{r=1}^{\bar{r}_n} \exp\{-c_2'' \tilde{J}_n^{r+1} (\log \tilde{J}_n)^{\kappa''}\} \\ & \leq \exp\{-\exp(c_0 \bar{r}_n^{t_0})\} + \bar{r}_n^2 p^{\bar{r}_n} \exp\{-c_3 L^{d+1} (J_n^*)^{d+1} (\log \tilde{J}_n)^{\kappa''+\kappa}\} \\ & = \exp\{-\exp(c_0 \bar{r}_n^{t_0})\} + \exp\{2 \log \bar{r}_n + \bar{r}_n \log p - c_3 L^{d+1} (J_n^*)^{d+1} (\log \tilde{J}_n)^{\kappa''+\kappa}\} \\ & \leq \exp(-b n \varepsilon_n^2) \end{aligned}$$

for any $b > 0$ and some constant $c_3 > 0$ provided L is chosen sufficiently large and the following relations hold

$$\begin{aligned} \log \bar{r}_n &\lesssim n \varepsilon_n^2, & \bar{r}_n \log p &\lesssim n \varepsilon_n^2, \\ (J_n^*)^{d+1} (\log n)^{\kappa+\kappa''} &\gtrsim n \varepsilon_n^2, & \exp(c_0 \bar{r}_n^{t_0}) &\gtrsim n \varepsilon_n^2. \end{aligned} \quad (22)$$

Now we bound the covering number $D(\varepsilon_n, \mathcal{F}_n, \rho)$ using the relation $D(\varepsilon_n, \mathcal{F}_n, \rho) \leq D(\varepsilon_n^2, \mathcal{F}_n, \|\cdot\|_1)$, where $\|\cdot\|_1$ stand for the L_1 -distance on the space of conditional densities given by

$$\|f_1 - f_2\|_1 = \iint |f_1(y|\mathbf{x}) - f_2(y|\mathbf{x})| dy dG(\mathbf{x}) \leq \sup_{\mathbf{x}} \int |f_1(y|\mathbf{x}) - f_2(y|\mathbf{x})| dy.$$

We split \mathcal{F}_n in layers corresponding to different r , different m_1, \dots, m_r and different $J_0, J_{m_1}, \dots, J_{m_r}$:

$$\mathcal{F}_n = \bigcup_{r=1}^{\bar{r}_n} \bigcup_{1 \leq m_1 < \dots < m_r \leq p, 1 \leq J_0, J_{m_1}, \dots, J_{m_r} \leq J_n^*} \mathcal{F}_{n;r;m_1, \dots, m_r; J_0, J_{m_1}, \dots, J_{m_r}}.$$

For any given $r, m_1, \dots, m_r, J_0, J_{m_1}, \dots, J_{m_r}$, consider $\boldsymbol{\theta}, \boldsymbol{\theta}' \in (\Delta_{J_0})^{\prod_{k=1}^r J_{m_k}}$. We can write $\boldsymbol{\theta} = (\boldsymbol{\theta}_{j_{m_1}, \dots, j_{m_r}} : 1 \leq j_{m_1}, \dots, j_{m_r} \leq J_n^*)$, $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_{j_{m_1}, \dots, j_{m_r}} : 1 \leq j_{m_1}, \dots, j_{m_r} \leq J_n^*)$

where $\boldsymbol{\theta}_{j_{m_1}, \dots, j_{m_r}} = (\theta_{j_0, j_{m_1}, \dots, j_{m_r}} : 1 \leq j_0 \leq J_n^*)$ and $\boldsymbol{\theta}'_{j_{m_1}, \dots, j_{m_r}} = (\theta'_{j_0, j_{m_1}, \dots, j_{m_r}} : 1 \leq j_0 \leq J_n^*)$. Then

$$\begin{aligned} & \|h(\mathbf{x}, y|r; m_1, \dots, m_r; J_0, J_{m_1}, \dots, J_{m_r}; \boldsymbol{\theta}) - h(\mathbf{x}, y|r; m_1, \dots, m_r; J_0, J_{m_1}, \dots, J_{m_r}; \boldsymbol{\theta}')\|_1 \\ & \leq \sup_{\mathbf{x}} \sum_{j_0=1}^{J_0} \sum_{j_{m_1}=1}^{J_{m_1}} \cdots \sum_{j_{m_r}=1}^{J_{m_r}} |\theta_{j_0, j_{m_1}, \dots, j_{m_r}} - \theta'_{j_0, j_{m_1}, \dots, j_{m_r}}| B_{j_{m_1}}(x_{m_1}) \cdots B_{j_{m_r}}(x_{m_r}) \\ & \leq \max_{j_{m_1}, \dots, j_{m_r}} \|\boldsymbol{\theta}_{j_{m_1}, \dots, j_{m_r}} - \boldsymbol{\theta}'_{j_{m_1}, \dots, j_{m_r}}\|_1 \end{aligned}$$

since the collection $B_j(x)$ s add up to 1 for any x . Using the fact that $D(\varepsilon, \Delta_d, \|\cdot\|_1) \leq (3/\varepsilon)^d$, we obtain

$$\begin{aligned} D(\varepsilon_n^2, \mathcal{F}_{n;r;m_1, \dots, m_r; J_0, J_{m_1}, \dots, J_{m_r}}, \|\cdot\|_1) & \leq \prod_{1 \leq J_{m_1}, \dots, J_{m_r} \leq \bar{J}_n} D(\varepsilon_n^2, \Delta_{J_0}, \|\cdot\|_1) \\ & \leq \prod_{1 \leq J_{m_1}, \dots, J_{m_r} \leq \bar{J}_n} \left(\frac{3}{\varepsilon_n^2} \right)^{J_0} \\ & = \left(\frac{3}{\varepsilon_n^2} \right)^{\bar{J}_n^{r+1}} \\ & \leq \left(\frac{3}{\varepsilon_n^2} \right)^{L^{d+1} (J_n^*)^{d+1} (\log n)^\kappa}. \end{aligned} \tag{23}$$

Therefore,

$$\begin{aligned} D(\varepsilon_n, \mathcal{F}_n, \rho) & \leq D(\varepsilon_n^2, \mathcal{F}_n, \|\cdot\|_1) \\ & \leq \sum_{r=1}^{\bar{r}_n} \sum_{1 \leq m_1 < \dots < m_r \leq p, 1 \leq J_0, J_{m_1}, \dots, J_{m_r} \leq \bar{J}_n} \left(\frac{3}{\varepsilon_n^2} \right)^{L^{d+1} (J_n^*)^{d+1} (\log n)^\kappa} \\ & \leq \sum_{r=1}^{\bar{r}_n} \binom{p}{r} \bar{J}_n^{r+1} \left(\frac{3}{\varepsilon_n^2} \right)^{L^{d+1} (J_n^*)^{d+1} (\log n)^\kappa} \\ & \leq \sum_{r=1}^{\bar{r}_n} p^r L^{d+1} (J_n^*)^{d+1} (\log n)^\kappa \left(\frac{3}{\varepsilon_n^2} \right)^{L^{d+1} (J_n^*)^{d+1} (\log n)^\kappa} \\ & \lesssim \bar{r}_n p^{\bar{r}_n} (J_n^*)^{d+1} (\log n)^\kappa \exp\{L^{d+1} (J_n^*)^{d+1} (\log n)^\kappa \log(3/\varepsilon_n^2)\} \\ & = \exp\{\log \bar{r}_n + \bar{r}_n \log p + (d+1) \log J_n^* + \kappa \log(\log n) \\ & \quad + L^{d+1} (J_n^*)^{d+1} (\log n)^\kappa \log(3/\varepsilon_n^2)\} \\ & \leq \exp\{c_4 (J_n^*)^{d+1} (\log n)^{\kappa+1} + \bar{r}_n \log p\} \end{aligned}$$

for some $c_4 > 0$. Thus it suffices to have the following relations

$$(J_n^*)^{d+1} (\log n)^{\kappa+1} \lesssim n\varepsilon_n^2, \quad \bar{r}_n \log p \lesssim n\varepsilon_n^2. \quad (24)$$

For (20), in order to lower bound the prior concentration probability around $f_0(y|\mathbf{x})$, we shall restrict to the oracle model consisting of d true covariates $X_{m_1^0}, \dots, X_{m_d^0}$. By Lemma 1, there exists $\boldsymbol{\theta}_0 = (\theta_{j_0, j_{m_1^0}, \dots, j_{m_d^0}}^0 : 1 \leq j_0, j_{m_1^0}, \dots, j_{m_d^0} \leq J_n^*)$ such that

$$\sup_{\mathbf{x}, y} |f_0(y|\mathbf{x}) - h(\mathbf{x}, y|d; m_1^0, \dots, m_d^0; J_n^*, \dots, J_n^*; \boldsymbol{\theta}_0)| \lesssim (J_n^*)^{-\beta} \leq \varepsilon_n. \quad (25)$$

Now for every $(j_{m_1^0}, \dots, j_{m_d^0})$, we define $\boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}}^0 = (\theta_{j_0, j_{m_1^0}, \dots, j_{m_d^0}}^0 : 1 \leq j_0 \leq J_n^*) \in \Delta_{J_n^*}$. Then $\boldsymbol{\theta}_0$ can be written by $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}}^0 : 1 \leq j_{m_1^0}, \dots, j_{m_d^0} \leq J_n^*)$. We consider $\boldsymbol{\theta} = (\boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}} : 1 \leq j_{m_1^0}, \dots, j_{m_d^0} \leq J_n^*)$ and $\boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}} = (\theta_{j_0, j_{m_1^0}, \dots, j_{m_d^0}} : 1 \leq j_0 \leq J_n^*) \in \Delta_{J_n^*}$. If

$$\max_{1 \leq j_{m_1^0}, \dots, j_{m_d^0} \leq J_n^*} \|\boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}} - \boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}}^0\|_1 \leq \varepsilon, \quad (26)$$

then

$$\begin{aligned} & |h(\mathbf{x}, y|d; m_1^0, \dots, m_d^0; J_n^*, \dots, J_n^*; \boldsymbol{\theta}) - h(\mathbf{x}, y|d; m_1^0, \dots, m_d^0; J_n^*, \dots, J_n^*; \boldsymbol{\theta}_0)| \\ & \leq \sum_{j_0=1}^{J_n^*} \sum_{j_{m_1^0}=1}^{J_n^*} \cdots \sum_{j_{m_d^0}=1}^{J_n^*} |\theta_{j_0, j_{m_1^0}, \dots, j_{m_d^0}} - \theta_{j_0, j_{m_1^0}, \dots, j_{m_d^0}}^0| \bar{B}_{j_0}(y) B_{j_{m_1^0}}(x_{m_1^0}) \cdots B_{j_{m_d^0}}(x_{m_d^0}). \end{aligned}$$

Since $0 \leq B_j(x) \leq 1$ and $0 \leq \bar{B}_j(y) \leq J_n^*$ for any j , we have

$$\begin{aligned} & \sup_{\mathbf{x}, y} |h(\mathbf{x}, y|d; m_1^0, \dots, m_d^0; J_n^*, \dots, J_n^*; \boldsymbol{\theta}) - h(\mathbf{x}, y|d; m_1^0, \dots, m_d^0; J_n^*, \dots, J_n^*; \boldsymbol{\theta}_0)| \\ & \leq (J_n^*)^{d+1} \max_{1 \leq j_{m_1^0}, \dots, j_{m_d^0} \leq J_n^*} \|\boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}} - \boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}}^0\|_1 \leq \varepsilon_n \end{aligned} \quad (27)$$

provided that

$$\|\boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}} - \boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}}^0\|_1 \leq (J_n^*)^{-(d+1)} \varepsilon_n \quad \text{for all } j_{m_1^0}, \dots, j_{m_d^0}. \quad (28)$$

To simplify the notation, we denote $h(\mathbf{x}, y|d; m_1^0, \dots, m_d^0; J_n^*, \dots, J_n^*; \boldsymbol{\theta})$ by $f_{\boldsymbol{\theta}}$. Combining (25) and (27), we have the desired approximation $\sup_{\mathbf{x}, y} |f_0(y|\mathbf{x}) - f_{\boldsymbol{\theta}}(y|\mathbf{x})| \leq 2\varepsilon_n$.

Using condition (B3), $\inf f_{\boldsymbol{\theta}} \geq \inf f_0 - \|f_0 - f_{\boldsymbol{\theta}}\|_{\infty} \geq \underline{m}_0/2$ given that ε_n is sufficiently small. This implies that $\|f_0/f_{\boldsymbol{\theta}}\|_{\infty} \leq 2\|f_0\|_{\infty}/\underline{m}_0 < \infty$ since f_0 can be regarded as a fixed continuous function on the compact set $[0, 1]^{d+1}$. Hence, for every $f_{\boldsymbol{\theta}}$ satisfying $\|f_{\boldsymbol{\theta}} - f_0\|_{\infty} \leq 2\varepsilon_n$,

$$\rho^2(f_0, f_{\boldsymbol{\theta}}) = \int \frac{|f_0(y|\mathbf{x}) - f_{\boldsymbol{\theta}}(y|\mathbf{x})|^2}{(f_0^{1/2}(y|\mathbf{x}) + f_{\boldsymbol{\theta}}^{1/2}(y|\mathbf{x}))^2} d\mathbf{y} dG(\mathbf{x}) \leq \frac{1}{\underline{m}_0} \|f_0 - f_{\boldsymbol{\theta}}\|_{\infty}^2 \lesssim \varepsilon_n^2. \quad (29)$$

Therefore, in view of Lemma 8 of [20], we obtain

$$\begin{aligned} K(f_0, f_{\theta}) &\leq 2\rho^2(f_0, f_{\theta}) \left\| \frac{f_0}{f_{\theta}} \right\|_{\infty} \lesssim \varepsilon_n^2, \\ V(f_0, f_{\theta}) &\lesssim \rho^2(f_0, f_{\theta}) \left(1 + \left\| \frac{f_0}{f_{\theta}} \right\|_{\infty} \right)^2 \lesssim \varepsilon_n^2. \end{aligned} \quad (30)$$

Thus, it suffices to lower bound the prior probability of the event in (28), which is

$$\begin{aligned} &\Pi_1(d) \times \Pi_2(\{m_1^0, \dots, m_d^0\} | r = d) \times \{\tilde{\Pi}_3(J_n^*)\}^{d+1} \\ &\quad \times \prod_{1 \leq j_{m_1^0}, \dots, j_{m_d^0} \leq J_n^*} \tilde{\Pi}_4(\|\theta_{j_{m_1^0}, \dots, j_{m_d^0}} - \theta_{j_{m_1^0}^0, \dots, j_{m_d^0}^0}\|_1 \leq (J_n^*)^{-(d+1)} \varepsilon_n) \\ &\gtrsim \frac{1}{\binom{p}{d}} \exp\{-(d+1)c_2'(J_n^*)^{d+1}(\log J_n^*)^{\kappa'}\} \times \exp\left\{-(J_n^*)^d c_5 J_n^* \log \frac{(J_n^*)^{d+1}}{\varepsilon_n}\right\} \end{aligned}$$

for some constant $c_5 > 0$ by the small ball probability estimates of a Dirichlet distribution in Lemma 6.1 of [18]. As long as J_n^* and ε_n^{-1} are powers of n within slowly varying factors, the last expression can be bounded below by $\exp\{-d \log p - c_6(J_n^*)^{d+1}(\log n)^{\kappa'}\}$ for some $c_6 > 0$. Hence in order to obtain (20), it suffices to have the following relationships:

$$(J_n^*)^{-\beta} \lesssim \varepsilon_n, \quad \log p \lesssim n\varepsilon_n^2, \quad (J_n^*)^{d+1}(\log n)^{\kappa'} \lesssim n\varepsilon_n^2. \quad (31)$$

We can determine the rate ε_n as the smallest sequence of numbers that satisfies (22), (24) and (31), that is,

$$\begin{aligned} \varepsilon_n &= \max\{n^{-(1-\alpha)/2}(\log n)^{1/(2t_0)}, n^{-\beta/(2\beta+d+1)}(\log n)^{\kappa'\beta/(2\beta+d+1)}\}, \\ J_n^* &= \lfloor (n\varepsilon_n^2)^{1/(d+1)}(\log n)^{1/(d+1)} \rfloor + 1, \end{aligned} \quad (32)$$

and $\kappa = \kappa' - \kappa''$, $\bar{r}_n = L'(\log n)^{1/t_0}$ for some sufficiently large L' provided that the condition $\exp\{\exp(c_0 \bar{r}_n^{t_0})\} \gg n\varepsilon_n^2$ is satisfied. \square

Proof of Theorem 2. The proof essentially follows the outline given in Theorem 1 except for two main differences. First, we shall need to use different $J_0, J_{m_1^0}, \dots, J_{m_d^0}$ due to the approximation result by Lemma 2. In particular, we need $J_0^{-\beta_0} \asymp J_{m_1^0}^{-\beta_1} \asymp \dots \asymp J_{m_d^0}^{-\beta_d}$. This will slightly change our definition of the sieve and the calculation of the prior concentration. The second difference is that we now have a dependent prior distribution in (A3'), which will change the calculation of the prior concentration rate.

We define a new sieve as follows

$$\mathcal{F}_n = \left\{ h(\mathbf{x}, y | r; m_1, \dots, m_r; J_0, J_{m_1}, \dots, J_{m_r}; \theta) : r \leq \bar{r}_n, \right.$$

$$1 \leq m_1 < \dots < m_r \leq p; J_0 \prod_{k=1}^r J_{m_k} \leq \tilde{J}_n^{r+1}, \theta \in \Delta_{J_0}^{\prod_{k=1}^r J_{m_k}} \Big\}, \quad (33)$$

where $\tilde{J}_n = \lfloor (LJ_n^*)^{(d+1)/(r+1)} (\log n)^{\kappa/(r+1)} \rfloor$, J_n^* and \bar{r}_n are two sequences of number going to infinity, and L and κ are some fixed positive constants. We shall choose the values of these numbers later.

We first verify (18). It follows that

$$\Pi(\mathcal{F}_n^c) \leq \Pi_1(r > \bar{r}_n) + \sum_{r=1}^{\bar{r}_n} \sum_{1 \leq m_1 < \dots < m_r \leq p} \Pi_3 \left(J_0 \prod_{k=1}^r J_{m_k} > \tilde{J}_n^{d+1} \right). \quad (34)$$

Note that the joint distribution of $(J_0, J_{m_1}, \dots, J_{m_r})$ depends only on the value of their product $J_0 \prod_{k=1}^r J_{m_k}$. Let t be a given integer and let N_t stand for the number of ways one can choose $\{J_0, J_{m_1}, \dots, J_{m_r}\}$ such that $J_0 \prod_{k=1}^r J_{m_k} = t$. Then

$$\Pi_3 \left(J_0 \prod_{k=1}^r J_{m_k} = t \right) \leq N_t \exp\{-c_2''' t (\log t)^{\kappa''}\}$$

for some $c_2''' > 0$. Clearly $N_t \leq t^{r+1}$. Thus,

$$\Pi_3 \left(J_0 \prod_{k=1}^r J_{m_k} = t \right) \leq \exp\{(r+1) \log t - c_2''' t (\log t)^{\kappa''}\} \leq \exp\{-c_7 t (\log t)^{\kappa''}\}$$

for some $c_7 > 0$ provided that $t (\log t)^{\kappa''} \gg (r+1) \log t$, which is satisfied if $t \gg r$ since $\kappa'' \geq 1$. Note that the distribution of the product $J_0 \prod_{k=1}^r J_{m_k}$ has a better-than-geometric tail starting from a large multiple of r , and hence

$$\Pi_3 \left(J_0 \prod_{k=1}^r J_{m_k} \geq t \right) \leq \exp(-c_8 t \log t)$$

for some $c_8 > 0$. In the sieve, we choose the cut-off \tilde{J}_n^{r+1} which is clearly of order higher than r and hence the requirement is met. As a result, the second term in (34) is bounded by

$$\begin{aligned} & \sum_{r=1}^{\bar{r}_n} p^r \exp\{-c_8 \tilde{J}_n^{r+1} \log \tilde{J}_n^{r+1}\} \\ & \leq \sum_{r=1}^{\bar{r}_n} \exp\{r \log p - c_8' L^{d+1} (J_n^*)^{d+1} (\log n)^{\kappa+\kappa''}\} \\ & \leq \exp\{\log \bar{r}_n + \bar{r}_n \log p - c_8' L^{d+1} (J_n^*)^{d+1} (\log n)^{\kappa+\kappa''}\} \end{aligned}$$

for some $c'_8 > 0$, which is of the same form of the corresponding bound for the isotropic case, and that L can be chosen sufficiently large. Thus, relation (22) is obtained.

The calculation of entropy proceeds in the same way as in the isotropic case. We split \mathcal{F}_n into layers following the same definition. Then

$$D(\varepsilon_n^2, \mathcal{F}_{n;r;m_1,\dots,m_r;J_0,J_{m_1},\dots,J_{m_r}}, \|\cdot\|_1) \leq \left(\frac{3}{\varepsilon_n^2}\right)^{J_0 \prod_{k=1}^r J_{m_k}} \leq \left(\frac{3}{\varepsilon_n^2}\right)^{\tilde{J}_n^{r+1}}$$

and the remaining calculations are identical, which give entropy estimates of the sieve as in the isotropic case and hence relation (24) is obtained.

Now we estimate the prior concentration rate. Consider the oracle model given by $(d; m_1^0, \dots, m_d^0; J_{n,0}^*, J_{n,1}^*, \dots, J_{n,d}^*)$, where

$$(J_{n,0}^*)^{-\beta_0} \asymp (J_{n,1}^*)^{-\beta_1} \asymp \dots \asymp (J_{n,d}^*)^{-\beta_d} \leq \varepsilon_n.$$

By Lemma 2, there exists $\theta_0 = (\theta_{j_0, j_{m_1^0}, \dots, j_{m_d^0}} : 1 \leq j_0 \leq J_{n,0}^*, 1 \leq j_{m_k^0} \leq J_{n,k}^*, k = 1, \dots, d)$ such that

$$\sup_{\mathbf{x}, y} |f_0(y|\mathbf{x}) - h(\mathbf{x}, y|d; m_1^0, \dots, m_d^0; J_{n,0}^*, \dots, J_{n,d}^*; \theta_0)| \lesssim \sum_{k=0}^d (J_{n,k}^*)^{-\beta_k} \lesssim \varepsilon_n.$$

Given $j_{m_1^0}, \dots, j_{m_d^0}$, define $\theta_{j_{m_1^0}, \dots, j_{m_d^0}}^0 = (\theta_{j_0, j_{m_1^0}, \dots, j_{m_d^0}}^0 : 1 \leq j_0 \leq J_{n,0}^*) \in \Delta_{J_{n,0}^*}$. Then $\theta_0 \in (\Delta_{J_{n,0}^*})^{\prod_{k=1}^d J_{n,k}^*}$. Let $\theta \in (\Delta_{J_{n,0}^*})^{\prod_{k=1}^d J_{n,k}^*}$ and be represented by $(\theta_{j_{m_1^0}, \dots, j_{m_d^0}} : 1 \leq j_{m_k^0} \leq J_{n,k}^*, k = 1, \dots, d)$. Then as before,

$$\begin{aligned} & |h(\mathbf{x}, y|d; m_1^0, \dots, m_d^0; J_{n,0}^*, \dots, J_{n,d}^*; \theta) - h(\mathbf{x}, y|d; m_1^0, \dots, m_d^0; J_{n,0}^*, \dots, J_{n,d}^*; \theta_0)| \\ & \leq \sum_{j_0=1}^{J_{n,0}^*} \sum_{j_{m_1^0}=1}^{J_{n,1}^*} \dots \sum_{j_{m_d^0}=1}^{J_{n,d}^*} |\theta_{j_0, j_{m_1^0}, \dots, j_{m_d^0}} - \theta_{j_0, j_{m_1^0}, \dots, j_{m_d^0}}^0| \bar{B}_{j_0}(y) B_{j_{m_1^0}}^*(x_{m_1^0}) \dots B_{j_{m_d^0}}^*(x_{m_d^0}) \\ & \leq J_{n,0}^* \prod_{k=1}^d J_{n,k}^* \max_{\substack{1 \leq j_{m_k^0} \leq J_{n,k}^*, \\ k=1, \dots, d}} \|\theta_{j_{m_1^0}, \dots, j_{m_d^0}} - \theta_{j_{m_1^0}, \dots, j_{m_d^0}}^0\|_1 \\ & \leq (J_{n,0}^*)^{d+1} \max_{\substack{1 \leq j_{m_k^0} \leq J_{n,k}^*, \\ k=1, \dots, d}} \|\theta_{j_{m_1^0}, \dots, j_{m_d^0}} - \theta_{j_{m_1^0}, \dots, j_{m_d^0}}^0\|_1, \end{aligned}$$

where $J_n^* = \lfloor (\prod_{k=0}^d J_{n,k}^*)^{1/(d+1)} \rfloor + 1$ is the smallest integer greater than the geometric mean of $J_{n,0}^*, \dots, J_{n,d}^*$. Thus it suffices to lower bound

$$\Pi_1(d) \times \Pi_2(\{m_1^0, \dots, m_d^0\} | r = d) \times \Pi_3(J_{n,0}^*, \dots, J_{n,d}^*)$$

$$\times \prod_{1 \leq j_0 \leq J_{n,0}^*, 1 \leq j_{m_k^0} \leq J_{n,k}^*, k=1, \dots, d} \tilde{\Pi}_4(\|\boldsymbol{\theta}_{j_{m_1^0}, \dots, j_{m_d^0}} - \boldsymbol{\theta}_{j_{m_1^0}^0, \dots, j_{m_d^0}^0}\| \leq (J_n^*)^{-(d+1)} \varepsilon_n).$$

Since the other factors are as before, it suffices to look at the third factor only, whose lower bound is given by

$$\exp \left\{ -c_2'' \left(\prod_{k=0}^d J_{n,k}^* \right) \left(\sum_{k=0}^d \log J_{n,k}^* \right)^\kappa \right\} \gtrsim \exp \{ -c_9 (J_n^*)^{d+1} (\log J_n^*)^\kappa \}$$

for some constant $c_9 > 0$, which is identical with the corresponding expression for the isotropic case. Thus we need

$$(J_{n,k}^*)^{-\beta_k} \lesssim \varepsilon_n, \quad \log p \lesssim n \varepsilon_n^2, \quad (J_n^*)^{d+1} (\log n)^{\kappa'} \lesssim n \varepsilon_n^2, \quad (35)$$

$$J_n^* \asymp \left(\prod_{k=0}^d J_{n,k}^* \right)^{1/(d+1)}.$$

Combining (22), (24) and (35), we can choose $\kappa = \kappa' - \kappa''$, \bar{r}_n as a large multiple of $(\log n)^{1/t_0}$, $J_{n,k}^* = \varepsilon_n^{-1/\beta_k}$ and

$$\varepsilon_n = \max \{ n^{-\beta^*/(2\beta^*+d+1)} (\log n)^{\kappa' \beta^*/(2\beta^*+d+1)}, n^{-(1-\alpha)/2} (\log n)^{1/(2t_0)} \},$$

where $\beta^* = (d+1)(\sum_{k=0}^d \beta_k^{-1})^{-1}$ is the harmonic mean of $(\beta_0, \beta_1, \dots, \beta_d)$. \square

Proof of Theorem 3. Note that the distance ρ_n on the space of conditional densities mathematically can be expressed as the Hellinger distance on the space of joint densities with respect to the dominating measure μ , which is the product of $\mathbb{P}_n^{\mathbf{X}}$ and the Lebesgue measure. This is notwithstanding the fact that the predictor variables are actually deterministic. We only need to replace $G(\cdot)$ by $\mathbb{P}_n^{\mathbf{X}}$ in the definitions of Kullback–Leibler divergence and Kullback–Leibler variation in (20). The rest of arguments proceed exactly in the same way as in Theorems 1 and 2 for the isotropic and the anisotropic cases respectively. \square

Proof of Lemma 1. Part (a) is a well-known approximation result for tensor product splines, see Theorem 12.7 of [36] or Lemma 2.1 of [10], for example. Part (b) is a direct multivariate generalization of Lemma 1, part (b) of [37].

For part (c), note that by part (b) we have $\boldsymbol{\theta} \in (0, 1)^J$ such that

$$\left\| f(y|x_1, \dots, x_d) - \sum_{j_0=1}^{J_0} \cdots \sum_{j_d=1}^{J_0} \theta_{j_0, \dots, j_d} B_{j_0}(y) \prod_{k=1}^d B_{j_k}(x_k) \right\|_\infty \leq C_1 J^{-\beta/(d+1)}$$

for constant $C_1 = C \|f^{(\beta)}\|_\infty$. Define $\boldsymbol{\xi}$ as the column vector of $\xi_{j_0, \dots, j_d} = \theta_{j_0, \dots, j_d} \int_0^1 B_{j_0}(y) dy$ and \mathbf{B}^* as the column vector of $B_j^*(y, \mathbf{x}) = \bar{B}_{j_0}(y) \prod_{k=1}^d B_{j_k}(x_k)$.

Then

$$\|f(y|x_1, \dots, x_d) - \boldsymbol{\xi}^T \mathbf{B}^*(y, \mathbf{x})\|_\infty \leq C_1 J^{-\beta/(d+1)}. \quad (36)$$

In particular, since $\|f\|_\infty < \infty$, it follows that $\|\boldsymbol{\xi}^T \mathbf{B}^*\|_\infty$ is uniformly bounded.

By integration, and using the fact that B-splines add to 1, it follows that

$$\begin{aligned} & \left\| \sum_{j_1=1}^{J_0} \cdots \sum_{j_d=1}^{J_0} \left(1 - \sum_{j_0=1}^{J_0} \xi_{j_0, \dots, j_d} \right) \prod_{k=1}^d B_{j_k}(x_k) \right\|_\infty \\ &= \left\| 1 - \sum_{j_0=1}^{J_0} \cdots \sum_{j_d=1}^{J_0} \xi_{j_0, \dots, j_d} \prod_{k=1}^d B_{j_k}(x_k) \right\|_\infty \leq C_1 J^{-\beta/(d+1)} \end{aligned}$$

for any $\mathbf{x} \in (0, 1)^d$. Applying a multivariate analog of Theorem 4.38 of [36] for tensor product of B-splines, we can bound the maximum norm of coefficients in a tensor product B-spline expansion by a constant multiple of the supremum norm of the function formed by corresponding linear combination. This is possible by forming a dual basis consisting of tensor product of functions in a dual basis for univariate B-splines and by noting that the supremum norms of the elements of the dual basis can be taken to be uniformly bounded (see Theorem 4.41 of [36]). This leads to the relation

$$\left| 1 - \sum_{j_0=1}^{J_0} \xi_{j_0, \dots, j_d} \right| \leq C'_1 J^{-\beta/(d+1)} \quad (37)$$

for any $(j_1, \dots, j_d) \in \{1, \dots, J_0\}^d$ and some constant $C'_1 > 0$.

Define $\boldsymbol{\eta}$ by the relations $\eta_{j_0, \dots, j_d} = \xi_{j_0, \dots, j_d} / \sum_{m=1}^{J_0} \xi_{m, j_1, \dots, j_d}$. Thus $\boldsymbol{\eta} \in \Delta_{J_0}^{J_0^d}$. Then using (37) and the boundedness of $\|\boldsymbol{\xi}^T \mathbf{B}^*\|_\infty$, we obtain

$$\begin{aligned} & \|\boldsymbol{\xi}^T \mathbf{B}^* - \boldsymbol{\eta}^T \mathbf{B}^*\|_\infty \\ &= \sup_{\mathbf{x}, y} \left| \sum_{j_0=1}^{J_0} \cdots \sum_{j_d=1}^{J_0} \xi_{j_0, \dots, j_d} B_{j_0}(y) \prod_{k=1}^d B_{j_k}(x_k) \left[\left(\sum_{m=1}^{J_0} \xi_{m, j_1, \dots, j_d} \right)^{-1} - 1 \right] \right| \\ &\lesssim \max_{j_1, \dots, j_d} \left| 1 - \sum_{m=1}^{J_0} \xi_{m, j_1, \dots, j_d} \right| \|\boldsymbol{\xi}^T \mathbf{B}^*\|_\infty \\ &\leq C_2 J^{-\beta/(d+1)} \end{aligned}$$

for some positive constant C_2 . Combining with (36), the result now follows. \square

Proof of Lemma 2. Part (a) is a well-known approximation result for tensor Sobolev space, see Theorem 12.7 of [36], for example. The proof of (b) and (c) proceed exactly as in Lemma 1. \square

Acknowledgements

The authors would like to thank the Associate Editor and two referees for their helpful comments that greatly improves the quality of the paper.

References

- [1] BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields* **71** 271–291. [MR0816706](#)
- [2] BROWN, P.J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 627–641. [MR1626005](#)
- [3] BROWN, P.J., VANNUCCI, M. and FEARN, T. (2002). Bayes model averaging with selection of regressors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 519–536. [MR1924304](#)
- [4] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Heidelberg: Springer. [MR2807761](#)
- [5] CARLIN, B.P. and CHIB, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 473–484.
- [6] CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. [MR3059077](#)
- [7] CHUNG, Y. and DUNSON, D.B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Amer. Statist. Assoc.* **104** 1646–1660. [MR2750582](#)
- [8] DALALYAN, A.S. and TSYBAKOV, A.B. (2012). Mirror averaging with sparsity priors. *Bernoulli* **18** 914–944. [MR2948907](#)
- [9] DE BOOR, C. (2001). *A Practical Guide to Splines*, Revised ed. *Applied Mathematical Sciences* **27**. New York: Springer. [MR1900298](#)
- [10] DE JONGE, R. and VAN ZANTEN, J.H. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electron. J. Stat.* **6** 1984–2001. [MR3020254](#)
- [11] DUNSON, D.B. and PARK, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95** 307–323. [MR2521586](#)
- [12] DUNSON, D.B., PILLAI, N. and PARK, J.-H. (2007). Bayesian density regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 163–183. [MR2325270](#)
- [13] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. [MR2530322](#)
- [14] FAN, J., YAO, Q. and TONG, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83** 189–206. [MR1399164](#)
- [15] FAN, J. and YIM, T.H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika* **91** 819–834. [MR2126035](#)
- [16] GEORGE, E.I. and MCCULLOCH, R.E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.
- [17] GEORGE, E.I. and MCCULLOCH, R.E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.
- [18] GHOSAL, S., GHOSH, J.K. and VAN DER VAART, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. [MR1790007](#)
- [19] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. [MR2332274](#)

- [20] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. [MR2336864](#)
- [21] GREEN, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. [MR1380810](#)
- [22] GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. [MR2108039](#)
- [23] GRIFFIN, J.E. and STEEL, M.F.J. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 179–194. [MR2268037](#)
- [24] HALL, P., RACINE, J. and LI, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc.* **99** 1015–1026. [MR2109491](#)
- [25] HALL, P., WOLFF, R.C.L. and YAO, Q. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* **94** 154–163. [MR1689221](#)
- [26] HUANG, J., HOROWITZ, J.L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38** 2282–2313. [MR2676890](#)
- [27] JARA, A. and HANSON, T.E. (2011). A class of mixtures of dependent tail-free processes. *Biometrika* **98** 553–566. [MR2836406](#)
- [28] JIANG, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *Ann. Statist.* **35** 1487–1511. [MR2351094](#)
- [29] KHASMINSKII, R.Z. (1978). A lower bound for risks of nonparametric density estimates in the uniform metric. *Theory Probab. Appl.* **23** 794–798.
- [30] LEPSKI, O. (2013). Multivariate density estimation under sup-norm loss: Oracle approach, adaptation and independence structure. *Ann. Statist.* **41** 1005–1034. [MR3099129](#)
- [31] MA, L. (2012). Recursive partitioning and Bayesian inference on conditional distributions. Technical report. Duke Univ.
- [32] MÜLLER, P., ERKANLI, A. and WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83** 67–79. [MR1399156](#)
- [33] NORETS, A. and PELENIS, J. (2014). Posterior consistency in conditional density estimation by covariate dependent mixtures. *Econometric Theory* **30** 606–646. [MR3205608](#)
- [34] PATI, D., DUNSON, D.B. and TOKDAR, S.T. (2013). Posterior consistency in conditional distribution estimation. *J. Multivariate Anal.* **116** 456–472. [MR3049916](#)
- [35] RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.* **7** 311–333. [MR2934953](#)
- [36] SCHUMAKER, L.L. (2007). *Spline Functions: Basic Theory*, 3rd ed. *Cambridge Mathematical Library*. Cambridge: Cambridge Univ. Press. [MR2348176](#)
- [37] SHEN, W. and GHOSAL, S. (2012). Adaptive Bayesian procedures using random series prior. Technical report. Available at [arXiv:1403.0625](#).
- [38] SHEN, W., TOKDAR, S.T. and GHOSAL, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. [MR3094441](#)
- [39] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. [MR1865337](#)
- [40] SUGIYAMA, M., TAKEUCHI, I., KANAMORI, T., SUZUKI, T., HACHIYA, H. and OKANOHARA, D. (2010). Least-squares conditional density estimation. *IEICE Trans. Inf. Syst.* **E93-D** 583–594.
- [41] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- [42] TOKDAR, S.T. (2011). Dimension adaptability of Gaussian process models with variable selection and projection. Technical report. Available at [arXiv:1112.0716](#).

- [43] TOKDAR, S.T., ZHU, Y.M. and GHOSH, J.K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal.* **5** 319–344. [MR2719655](#)
- [44] TRIPPA, L., MÜLLER, P. and JOHNSON, W. (2011). The multivariate beta process and an extension of the Polya tree model. *Biometrika* **98** 17–34. [MR2804207](#)
- [45] VAN DE GEER, S.A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. [MR2396809](#)
- [46] VAN DER VAART, A.W. and VAN ZANTEN, J.H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. [MR2418663](#)
- [47] YANG, Y. and DUNSON, D. (2012). Bayesian conditional tensor factorizations for high-dimensional classification. Technical report. Available at [arXiv:1301.4950](#).

Received July 2013 and revised June 2014