Available online at www.sciencedirect.com

ScienceDirect

Journal of Multivariate Analysis 98 (2007) 1969 – 1987

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

On posterior consistency in nonparametric regression problems

Taeryon Choi[a,*], Mark J. Schervish[b]

[a]Department of Mathematics and Statistics, University of Maryland, Baltimore County, MD, USA
[b]Department of Statistics, Carnegie Mellon University, USA

Received 26 January 2006
Available online 13 January 2007

Abstract

We provide sufficient conditions to establish posterior consistency in nonparametric regression problems with Gaussian errors when suitable prior distributions are used for the unknown regression function and the noise variance. When the prior under consideration satisfies certain properties, the crucial condition for posterior consistency is to construct tests that separate from the outside of the suitable neighborhoods of the parameter. Under appropriate conditions on the regression function, we show there exist tests, of which the type I error and the type II error probabilities are exponentially small for distinguishing the true parameter from the complements of the suitable neighborhoods of the parameter. These sufficient conditions enable us to establish almost sure consistency based on the appropriate metrics with multi-dimensional covariate values fixed in advance or sampled from a probability distribution. We consider several examples of nonparametric regression problems.

AMS 1991 subject classification: Primary 62G20; 62G08; Secondary 62C10

Keywords: Almost sure consistency; Differentiable functions; Empirical probability measure; Hellinger metric; In probability metric; Sieve

1. Introduction

In this paper, we provide sufficient conditions to verify almost sure consistency for posterior distributions in nonparametric regression problems with additive noise when the regression function and the noise variance are assumed to be unknown. Such problems involve infinite-dimensional

* Corresponding author. Fax: +1 410 455 1066.
E-mail address: tchoi@math.umbc.edu (T. Choi).

0047-259X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2007.01.004

parameters, and consistency of posterior distributions is a much more challenging problem than in the finite-dimensional case. There have been many results giving general conditions under which features of posterior distributions are consistent in infinite-dimensional spaces. For examples, see Schwartz [13]; Barron et al. [2]; Ghosal et al. [6]; Amewou-Atisso et al. [1]; Walker [18]; and Ghosal and Roy [7].

Early results on posterior consistency focused mainly on density estimation, that is on estimating a density function for a random sample without assuming the density belongs to a finite-dimensional parametric family. More recently, attention has turned to posterior consistency in nonparametric regression problems.

One common approach to Bayesian nonparametric regression problems has been made through the orthogonal basis expansion for a regression function. Shen and Wasserman [14] established general results for the rate of convergence of posterior distributions and applied the general results to an example using the orthogonal basis expansion approach to nonparametric regression. Walker [17] provided a general result for establishing posterior consistency for a class of Bayesian regression models. Huang [10] gave a Bayesian convergence rate theorem in the context of regression, assuming the error is normally distributed with known variance and the true regression function is bounded by a known constant. Ghosal and Van der Vaart [8] present general results on convergence rates for non i.i.d. observation which include nonparametric regression problems with known error variance and a suitable distance between two functions. In all of the literature mentioned above, the approaches were based on the assumption that the error variance is assumed to be known for simplicity when the data have normal distribution, and in most cases the assumption that the regression function is uniformly bounded. Further, most of previous results are based on the Hellinger metric assuming that covariate values are sampled from a probability distribution.

However, when the error variance is unknown or covariate values are assumed to be fixed in advance, it is neither obvious that the joint posterior distribution of the regression function and the noise variance is consistent, nor simple to apply existing conditions for posterior consistency based on the Hellinger metric. Compared to the previous approaches, our contributions are the following. First, we assume that the noise variance is unknown and our results cover joint neighborhoods of the regression function and the noise variance. Second, we deal with two types of covariate: randomly sampled from a probability distribution, and fixed in advance. And third, we consider cases in which the regression function is not necessarily bounded by a known constant. However, when proving consistency with respect to the L^1 metric, we still require a boundedness condition. In addition, we also provide results for multi-dimensional covariate values and we prove almost-sure convergence where some previous results prove merely in-probability convergence.

The specific nonparametric regression problem that we present here assumes that the data have normal noise distributions. The error variance is also assumed to be unknown and needs to be estimated. We give two sufficient conditions for posterior consistency and examine them in detail under the nonparametric regression problem. We consider several examples of nonparametric regression problems using appropriate prior distributions and illustrate that posterior distributions are consistent.

The rest of the paper is organized as follows. In Section 2, we state an extension of Schwartz' theorem [13] for independent but non-identically distributed observations, based on the results of Amewou-Atisso et al. [1] and Choudhuri et al. [5]. In Section 3, we describe the model that we are using. In Section 4, we apply the theorem of Section 2 to nonparametric regression problems by introducing two sufficient conditions. In Section 5, we provide posterior consistency theorems in nonparametric regression problems. In Section 6, we give two examples of specific nonparametric regression problems. In Section 7, we discuss some directions on future work. Proofs are given in the Appendix.

2. Consistency theorem for non-i.i.d. observations

Schwartz [13] proved a theorem known as Schwartz' theorem that gave conditions for consistency of posterior distributions of the parameters of distributions of independent and identically distributed random variables. Amewou-Atisso et al. [1] extended Schwartz' theorem to independent non-identically distributed observations and applied it to the semiparametric regression problems. Choudhuri et al. [5] extend Schwartz's theorem to a triangular array of independent non-identically distributed observations for the case of convergence in probability. In this section, we provide another extension of Schwartz's theorem to almost sure convergence for independent but non-identically distributed observations. Our extension is based on both Amewou-Atisso et al. [1] and Choudhuri et al. [5]. We present this extension as Theorem 1 and later we verify the conditions in nonparametric regression problems. The proof of Theorem 1 is given in an Appendix at the end.

Theorem 1. *Let $\{Z_i\}_{i=1}^{\infty}$ be independently distributed with densities $\{f_i(\cdot;\theta)\}_{i=1}^{\infty}$, with respect to a common σ-finite measure, where the parameter θ belongs to an abstract measurable space Θ. The densities $f_i(\cdot;\theta)$ are assumed to be jointly measurable. Let $\theta_0 \in \Theta$ and let P_{θ_0} stand for the joint distribution of $\{Z_i\}_{i=1}^{\infty}$ when θ_0 is the true value of θ. Let $\{U_n\}_{n=1}^{\infty}$ be a sequence of subsets of Θ. Let θ have prior Π on Θ. Define $\Lambda(\theta_0,\theta) = \log \frac{f_i(Z_i;\theta_0)}{f_i(Z_i;\theta)}$, $K_i(\theta_0,\theta) = \mathrm{E}_{\theta_0}(\Lambda(\theta_0,\theta))$ and $V_i(\theta_0,\theta) = \mathrm{Var}_{\theta_0}(\Lambda(\theta_0,\theta))$.*

(A1) *Prior positivity of neighborhoods: Suppose that there exists a set B with $\Pi(B) > 0$ such that:*

(i) $\sum_{i=1}^{\infty} \frac{V_i(\theta_0,\theta)}{i^2} < \infty, \quad \forall \theta \in B$,
(ii) *For all $\epsilon > 0$, $\Pi(B \cap \{\theta : K_i(\theta_0,\theta) < \epsilon$ for all $i\}) > 0$.*

(A2) *Existence of tests: Suppose that there exist test functions $\{\Phi_n\}_{n=1}^{\infty}$, sets $\{\Theta_n\}_{n=1}^{\infty}$ and constants $C_1, C_2, c_1, c_2 > 0$ such that:*

(i) $\sum_{n=1}^{\infty} \mathrm{E}_{\theta_0}\Phi_n < \infty$,
(ii) $\sup_{\theta\in U_n^C \bigcap \Theta_n} \mathrm{E}_\theta(1-\Phi_n) \leqslant C_1 e^{-c_1 n}$,
(iii) $\Pi(\Theta_n^C) \leqslant C_2 e^{-c_2 n}$.

Then

$$\Pi\left(\theta \in U_n^C \,\middle|\, Z_1,\ldots,Z_n\right) \to 0 \quad a.s.\ [P_{\theta_0}]. \tag{1}$$

Fig. 1 is helpful to understand some of the conditions in Theorem 1. The first condition (A1) assumes that there are sets with positive prior probabilities, which could be regarded as neighborhoods of the true parameter θ_0. In Fig. 1, the set B is the neighborhood of the true parameter θ_0 which has a positive prior probability and specifically, it intersects Kullback–Leibler neighborhoods of the true parameter θ_0. The second condition (A2) assumes the existence of certain tests of the hypothesis $\theta = \theta_0$. We construct a test which distinguishes θ_0 from $\Theta_n \cap U_n^C$ as much as possible. Here, Θ_n is a sieve which grows to the parameter space Θ as the sample size increases. We need to consider the sieve Θ_n because the parameter under consideration is infinite dimen-

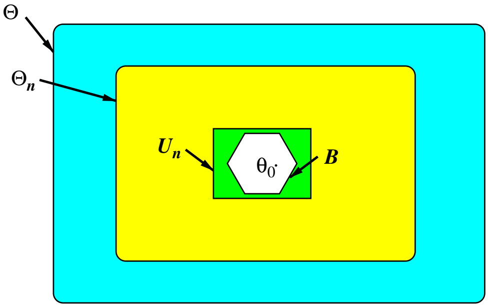

Fig. 1. Graphical display of Theorem 1.

sional. We assume that tests with vanishingly small type I error probability exist. We also assume that these tests have exponentially small type II error probability on part of the complement of a set U_n containing θ_0, namely $\Theta_n \cap U_n^C$. Finally, Θ_n^C is assumed to have exponentially small prior probability.

3. The model

Let us consider a random response Y corresponding to a covariate vector X taking values in a compact set $T \subset \mathbb{R}^d$. We are interested in estimating the regression function, $\eta(\mathbf{x}) = \mathrm{E}(Y|X = \mathbf{x})$ based on independent observations of (X, Y). We do not assume a parametric form for the regression function, but rather we assume some smoothness conditions. We model the unknown function η as a random function with a suitable prior distribution.

To be specific, the nonparametric regression model we consider here, is the following:

$$\begin{aligned} Y_i &= \eta(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \\ \epsilon_i &\sim N(0, \sigma^2) \text{ given } \sigma, \\ \eta(\cdot) &\sim \Pi_1, \text{ independent of } \sigma \text{ and } (\epsilon_1, \ldots, \epsilon_n), \\ \sigma &\sim \Pi_2, \quad \sigma \in \mathbb{R}^+. \end{aligned} \tag{2}$$

We leave the nature of the sequence $X_1, X_2, \ldots$ of covariates unspecified at present. We will consider two possibilities. In one case, the covariates form an i.i.d. sequence and all of the distributions in (2) are conditional on the sequence of covariates. In the other case, the covariates are all fixed values, known ahead of time.

We assume that the true response function, $\eta_0(\cdot)$ as a function of the covariate X, is continuously differentiable on the compact set T. Without loss of generality, we will assume that $T = [0, 1]^d$ for the remainder of this paper.

4. Sufficient conditions

In this section, we give conditions peculiar to the nonparametric regression problem of (2) under which the conditions of Theorem 1 hold. Some of these conditions can be verified independently of the prior distribution, while the verification of the other conditions is postponed to Section 6, in which we introduce two specific classes of prior distributions.

We use the following notation. The parameter θ in Theorem 1 is (η, σ) with $\theta_0 = (\eta_0, \sigma_0)$. The density $f_i(\cdot; \theta)$ is the normal density with mean $\eta(\mathbf{x}_i)$ and variance σ^2. The parameter space $\boldsymbol{\Theta}$ is a product space of a function space Θ_1 and $\mathbb{R}^+$. Let θ have prior Π, a product measure, $\Pi = \Pi_1 \times \Pi_2$. Finally, we define the joint neighborhoods of the regression function and σ that play the roles of U_n in Theorem 1 and contain θ_0. We consider two different neighborhoods, depending on the nature of the covariate. If the covariate values are sampled from a probability distribution Q, we define the Hellinger neighborhood of (η_0, σ_0), $H_\epsilon = \{(\eta, \sigma) : d_H(f, f_0) < \epsilon\}$. In the definition of H_ϵ, we make use of the fact that a parameter (η, σ) is equivalent to a joint density f for (X, Y) with respect to $\xi = Q \times \lambda$, namely $f(\mathbf{x}, y) = \phi([y - \eta(\mathbf{x})]/\sigma)/\sigma$, where ϕ is the standard normal density, and λ is d-dimensional Lebesgue measure. When the covariate values are fixed in advance, we consider the neighborhood based on the empirical measure of the design points. Let Q_n be the empirical probability measure of the design points, $Q_n(\mathbf{x}) = n^{-1} \sum_{i=1}^n I_{\{\mathbf{x}_i\}}(\mathbf{x})$. Based on Q_n, we define the following neighborhood, $W_{\epsilon,n} = \left\{(\eta, \sigma) : \int |\eta(\mathbf{x}) - \eta_0(\mathbf{x})| \, dQ_n(\mathbf{x}) < \epsilon, \left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon\right\}$.

The marginal neighborhoods of η or σ contain the joint neighborhood that we consider. Thus, if we prove that the posterior probability of joint neighborhoods converge almost surely to 1, then it obviously follows that the posterior probability of marginal neighborhoods converge almost surely to 1.

4.1. Existence of tests

First, we construct test functions that distinguish the true parameter (η_0, σ_0) from $W_{\epsilon,n}^C$ in the sense that the type I error and type II error probabilities of these test functions are exponentially small. However, since we have an infinite dimensional parameter space, it is not always feasible to construct such test functions. Thus, we consider a sieve, Θ_n, which grows eventually to the space of continuously differentiable function on $[0, 1]^d$. For each element of the sieve, we construct a test, as required by condition (A2). Let $\|\cdot\|_\infty$ denote the sup norm and $M_n = \mathcal{O}(n^\alpha)$ for some $\frac{1}{2} < \alpha < 1$ and define $\Theta_n = \Theta_{1n} \times \mathbb{R}^+$, where

$$\Theta_{1n} = \left\{\eta(\cdot) : \|\eta\|_\infty < M_n, \left\|\frac{\partial}{\partial x_i}\eta\right\|_\infty < M_n, \ i = 1, \ldots, d\right\}. \tag{3}$$

Now, the nth test is constructed by combining a collection of tests, one for each of finitely many elements of Θ_n. Those finitely many elements come from a covering of Θ_{1n} by small balls. In addition, it is straightforward from Theorem 2.7.1 of Van der Vaart and Wellner [16] that there exists a constant K' such that the ϵ-covering number $N(\epsilon, \Theta_{1n}, \|\cdot\|_\infty)$ of Θ_{1n} satisfies $\log N(\epsilon, \Theta_{1n}, \|\cdot\|_\infty) \leqslant \frac{K' M_n}{\epsilon^d}$. For each n and each ball in the covering of Θ_{1n}, we find a test with small type I and type II error probabilities. Then we combine the tests and show that they satisfy subconditions (i) and (ii) of (A2).

Theorem 2 gives the existence of tests for the fixed design case based on $W_{\epsilon,n}$. The specific construction, as outlined above, is part of the proof in the Appendix.

Theorem 2. *Suppose that the values of the covariate in* $[0, 1]^d$ *arise according to a fixed design. Then there exist test functions* $\{\Phi_n\}$ *and a constant* $C_5 > 0$ *that satisfy*:

(i) $\sum_{n=1}^\infty \mathrm{E}_{P_0}\Phi_n < \infty$,
(ii) $\sup_{\theta \in W_{\epsilon,n}^C \bigcap \Theta_n} \mathrm{E}_P(1 - \Phi_n) \leqslant \exp(-C_5 n)$,

where P *is the joint distribution of* $\{Y_n\}_{n=1}^\infty$ *assuming that* $\theta = (\eta, \sigma)$.

When the covariate values are sampled from a probability distribution, we introduce the intermediate neighborhoods based on the in-probability metric, defined as

$$d_Q(\eta_1, \eta_2) = \inf\{\epsilon > 0 : Q(\{\mathbf{x} : |\eta_1(\mathbf{x}) - \eta_2(\mathbf{x})| > \epsilon\}) < \epsilon\},$$

and construct test functions in Theorem 3. For the random covariate case, we first establish posterior consistency based on the d_Q metric, and then deduce posterior consistency for the Hellinger metric from the corresponding result for the d_Q metric. Theorem 3 ensures the existence of tests for the random covariate.

Theorem 3. *Let* $U_\epsilon = \left\{(\eta, \sigma) : d_Q(\eta, \eta_0) < \epsilon, \left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon\right\}$. *Suppose that the values of the covariate in* $[0, 1]^d$ *are sampled from a probability distribution Q. Then for the same test function* Φ_n *as in Theorem* 2, *there exists a constant* $C_6 > 0$ *such that*:

(i) $\sum_{n=1}^{\infty} \mathrm{E}_{P_0}\Phi_n < \infty$,
(ii) $\sup_{\theta \in U_\epsilon^C \bigcap \Theta_n} \mathrm{E}_P(1 - \Phi_n) \leqslant \exp(-C_6 n)$.

The proofs of Theorems 2 and 3 are given in the Appendix.

4.2. Conditions for prior distributions

In this section, we state versions of sufficient conditions for the prior distributions of nonparametric regression problems that imply parts of conditions (A1) and (A2) of Theorem 1. We verify these conditions in Section 6 for two specific classes of prior distributions.

4.2.1. Prior positivity

We identify a set B such that the condition (A1) of Theorem 1 holds in the nonparametric regression model, (2). Direct calculations show that

$$K_i(\theta_0; \theta) = \frac{1}{2}\log\frac{\sigma^2}{\sigma_0^2} - \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right) + \frac{1}{2}\frac{[\eta_0(\mathbf{x}_i) - \eta(\mathbf{x}_i)]^2}{\sigma^2}$$

and

$$V_i(\theta_0, \theta) = 2\left[-\frac{1}{2} + \frac{1}{2}\frac{\sigma_0^2}{\sigma^2}\right]^2 + \left[\frac{\sigma_0^2}{\sigma^2}[\eta(\mathbf{x}_i) - \eta_0(\mathbf{x}_i)]\right]^2.$$

For each $\delta > 0$, define

$$B_\delta = \left\{(\eta, \sigma) : \|\eta - \eta_0\|_\infty < \delta, \left|\frac{\sigma}{\sigma_0} - 1\right| < \delta\right\}.$$

Then, from the calculations of $K_i(\theta_0, \theta)$ and $V_i(\theta_0, \theta)$, it is easily shown that (i) for every $\epsilon > 0$, there exists $\delta > 0$ such that $\forall\theta \in B_\delta$, $K_i(\theta_0, \theta) < \epsilon$ for all i and that (ii) $\sum_{i=1}^{\infty} \frac{V_i(\theta_0,\theta)}{i^2} < \infty$, $\forall\theta \in B_\delta$. In addition, because these results hold uniformly in x_i, they hold for both the random and nonrandom design cases.

Hence, if the prior, Π assigns positive probability to B_δ for each $\delta > 0$, then condition (A1) holds.

4.2.2. *Probability of* Θ_n^C

The other condition for prior distributions to satisfy concerns the probability of Θ_n^C. The subcondition (iii) of (A2) requires that there exist constants C_2 and c_2 such that $\Pi(\Theta_n^C) \leqslant C_2 e^{-c_2 n}$. As in the construction of sieves, since $\Pi\left(\Theta_n^C\right) = \Pi_1\left\{\Theta_{1n}^C\right\}$, if the prior distribution for the regression function, Π_1, assigns exponentially small probability to the two sets $\Theta_{1n,0}^C = \{\eta : \|\eta\|_\infty > M_n\}$ and $\Theta_{1n,i}^C = \left\{\eta : \left\|\frac{\partial}{\partial x_i}\eta\right\|_\infty > M_n\right\}$, $i = 1, \ldots, d$, then the condition (iii) of (A2) holds.

5. Posterior consistency in nonparametric regression problems

In Theorems 4 and 6 we provide almost sure consistency results with different conditions on the covariates (random and nonrandom designs) and different topologies (based on empirical measure, L^1, d_Q and Hellinger). In both of these theorems, we assume that the sufficient conditions of Sections 4.1 and 4.2 hold and that η_0 has continuous first-order partial derivatives. (In Section 6, we will give specific examples of priors that do satisfy the conditions of Section 4.2.) Our result for the L^1 topology requires the following additional assumptions:

Assumption NRD_d**.** For each hypercube H in $[0, 1]^d$, let $\lambda(H)$ be its Lebesgue measure. Suppose that there exists a constant K_d, $0 < K_d \leqslant 1$ such that whenever $\lambda(H) \geqslant \frac{1}{K_d n}$, H contains at least one design point.

Assumption B_d**.** Let Π_1^* be a prior distribution for $\eta(\mathbf{x})$ satisfying conditions, specified in Section 4.2. Let V be a constant such that $V > \left\|\frac{\partial}{\partial x_j}\eta_0(x_1, \ldots, x_d)\right\|_\infty$, $j = 1, \ldots, d$. Define

$$\Omega = \left\{\eta : \left\|\frac{\partial}{\partial x_j}\eta(x_1, \ldots, x_d)\right\|_\infty < V, \quad j = 1, \ldots, d\right\}.$$

Assume that $\Pi_1(\cdot) = \Pi_1^*(\cdot \cap \Omega)/\Pi_1^*(\Omega)$ with $\Pi_1^*(\Omega) > 0$.

Theorem 4. *Let* P_0 *denote the joint conditional distribution of* $\{Y_n\}_{n=1}^\infty$ *given the covariate assuming that* η_0 *is the true response function and* σ_0^2 *is the true noise variance. Suppose that the values of the covariate in* $[0, 1]^d$ *are fixed, i.e., known ahead of time.*

(1) *Then for every* $\epsilon > 0$,

$$\Pi\left\{(\eta, \sigma) \in W_{\epsilon,n}^C | Y_1, \ldots, Y_n, x_1, \ldots, x_n\right\} \to 0 \quad a.s.\ [P_0]. \tag{4}$$

(2) *Suppose, in addition, that the values of the covariate in* $[0, 1]^d$ *arise according to a design satisfying Assumption* NRD_d *and the prior satisfies Assumption* B_d*. Let* λ *be d-dimensional Lebesgue measure and* L_ϵ *be the following neighborhood of* (η_0, σ_0) *defined in terms of the* L^1 *metric*:

$$L_\epsilon = \left\{(\eta, \sigma) : \int |\eta(\mathbf{x}) - \eta_0(\mathbf{x})| d\lambda(\mathbf{x}) < \epsilon, \quad \left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon\right\}.$$

Then for every $\epsilon > 0$,

$$\Pi\left\{L_\epsilon^C \middle| Y_1, \ldots, Y_n, x_1, \ldots, x_n\right\} \to 0 \quad a.s.\ [P_0]. \tag{5}$$

Remark 5. Note that Assumption NRD_d is satisfied by the equally spaced design. When the covariate value is one dimensional, Choi [3] shows that (5) holds without Assumption B_d.

Theorem 6. *Suppose that the covariate values are sampled from a probability distribution in* $[0, 1]^d$ *and* P_0 *is the joint distribution of* $\{(X_n, Y_n)\}_{n=1}^{\infty}$ *assuming that* η_0 *is the true response function and* σ_0^2 *is the true noise variance. Then,*

(1) *For every* $\epsilon > 0$,

$$\Pi\left\{(\eta, \sigma) \in U_{\epsilon}^{C} | (X_1, Y_1), \ldots, (X_n, Y_n)\right\} \to 0 \quad a.s.\ [P_0].$$

(2) *For every* $\epsilon > 0$,

$$\Pi\left\{(\eta, \sigma) \in H_{\epsilon}^{C} | (X_1, Y_1), \ldots, (X_n, Y_n)\right\} \to 0 \quad a.s.\ [P_0]. \tag{6}$$

Remark 7. Note that U_ϵ and H_ϵ are closely connected in our problems. By calculating the squared Hellinger distance between two normal density functions, it is observed that Hellinger consistency for H_ϵ follows from consistency for U_ϵ. When σ is known, consistency for H_ϵ follows from consistency for density estimation with the help of some entropy bounds. We could have done the extra work to extend these results to the case of unknown σ. However, since we are proving part (1) of Theorem 6 anyway, we chose to base our proof of part (2) on that result. The detailed calculations and explanations are given at the proof of Theorem 6 in the Appendix.

Remark 8. If the covariate is random and if we assume that the regression function is uniformly bounded, then L^1 consistency follows from Theorem 6 because the equivalence of L^1 convergence and convergence in probability for bounded functions.

6. Examples of nonparametric regression problems

In this section, we present two examples of nonparametric regression priors. One is based on an orthogonal basis expansion, and the other is a version of Gaussian process regression. We give conditions under which each class of examples satisfies the sufficient conditions of Section 4.2, so that the joint posterior distribution of the regression function and σ is consistent. For simplicity, we treat the case of one-dimensional covariates but the examples can easily be extended to the case of multi-dimensional covariates.

6.1. Example 1: orthogonal basis expansion

An orthogonal basis expansion for the regression function, $\eta(x)$, of (2) is a representation of $\eta(x)$ by an infinite sum,

$$\eta(x) = \sum_{j=1}^{\infty} \psi_j \phi_j(x), \tag{7}$$

where $\{\phi_j(x)\}_{j=1}^{\infty}$ is an orthonormal basis for an L^2 space containing η. We will now construct a prior distribution on the set $\Omega = L^2[0, 1] \cap C^1[0, 1]$ that satisfies the sufficient conditions given in Sections 4.2.1 and 4.2.2. We also assume that the true regression function η_0 is in Ω and represented by $\eta_0(x) = \sum_{j=1}^{\infty} \zeta_j \phi_j(x)$. Shen and Wasserman [14] also considered a similar example in Section 5 of their paper to this example with a restricted parameter space for η.

Let $\{\phi_j(\cdot)\}_{j=1}^{\infty}$ be an orthonormal basis for $L^2[0, 1]$ consisting of differentiable functions. Let $a_j = \sup_{x\in[0,1]} |\phi_j(x)|$ and $b_j = \sup_{x\in[0,1]} |\phi'_j(x)|$. Choose $\{\tau_j\}_{j=1}^{\infty}$ so that $\sum_j a_j\tau_j < \infty$ and $\sum_j b_j\tau_j < \infty$. Let $\{\psi_j\}_{j=1}^{\infty}$ be a sequence of independent random variables with $\psi_j \sim N(0, \tau_j^2)$. The following calculations are similar to those performed in Section 3.3 of Barron et al. [2].

First, note that each $\eta \in \Omega$ has the form (7). For all $x \in [0, 1]$, $\left|\sum_{j=1}^{\infty} \psi_j\phi_j(x)\right| \leqslant \sum_{j=1}^{\infty} \left|\psi_j\right| a_j$. Then for every $t > 0$, by Markov's inequality and Chernoff Bounds, we have

$$\Pr\left\{\sup_{x\in[0,1]}\left|\sum_{j=1}^{\infty}\psi_j\phi_j(x)\right| > M_n\right\} \leqslant \exp\left(-tM_n + \frac{1}{2}t^2\sum_{j=1}^{\infty}a_j^2\tau_j^2\right)\prod_{j=1}^{\infty}2\Phi(a_j\tau_j t),$$

where Φ is the standard normal distribution function, and $M_n = \mathcal{O}(n^{\alpha})$ with $\frac{1}{2} < \alpha < 1$. Note that

$$\log\prod_{j=1}^{\infty}2\Phi(a_j\tau_j t) \leqslant \sum_{j=1}^{\infty}\log\left(1 + \frac{2a_j\tau_j t}{\sqrt{2\pi}}\right) \leqslant \frac{2t}{\sqrt{2\pi}}\sum_{j=1}^{\infty}a_j\tau_j.$$

Then, by taking $t = \frac{M_n}{\sum_{j=1}^{\infty} a_j^2\tau_j^2}$ and assuming n is large enough so that $\frac{4}{\sqrt{2\pi}}\sum_{j=1}^{\infty} a_j\tau_j < \frac{1}{2}M_n$, we have $\Pr\left\{\sup_{x\in[0,1]}\left|\sum_{j=1}^{\infty}\psi_j\phi_j(x)\right| > M_n\right\} \leqslant \exp\left(-\frac{n}{4\sum_{j=1}^{\infty}a_j^2\tau_j^2}\right)$, eventually. Similarly, we have $\Pr\left\{\sup_{x\in[0,1]}\left|\sum_{j=1}^{\infty}\psi_j\phi'_j(x)\right| > M_n\right\} \leqslant \exp\left(-\frac{n}{4\sum_{j=1}^{\infty}b_j^2\tau_j^2}\right)$, by the assumption $\sum_{j=1}^{\infty} b_j\tau_j < \infty$. Hence, there exist constants C_3 and c_3 such that $\Pi(\Theta_n^C) \leqslant C_3\exp(-c_3n)$.

Secondly, the prior positivity condition in Section 4.2.1 is easily verified by assuming $\sum_{j=1}^{\infty} a_j \tau_j < \infty$. In fact, it obviously follows from the last statement in Section 3.3 of [2, pp. 553–554]. Hence, the verification of two conditions on the prior distribution for an orthonormal basis expansion is complete.

For example, consider the following sine–cosine orthonormal basis for $L^2[0, 1] : \phi_1(x) = 1$, and for $n \geqslant 1$, $\phi_{2n}(x) = \sqrt{2}\sin(2n\pi x)$, $\phi_{2n+1}(x) = \sqrt{2}\cos(2n\pi x), \ldots$. It is clear that $a_j \equiv \sup_{x\in[0,1]} |\phi_j(x)| < \sqrt{2}, \forall j$. Moreover, we have $\phi'_{2n}(x) = 2n\pi\sqrt{2}\cos(2n\pi x)$ and $\phi'_{2n+1}(x) = -2n\pi\sqrt{2}\sin(2n\pi x)$. This implies that $b_j \equiv \sup_{x\in[0,1]} |\phi'_j(x)| \leqslant \sqrt{2}\pi j$. Consequently, if we use a prior distribution, Π, that assigns a normal distribution with mean 0 and variance $\tau_j^2 = j^{-2p}$, $p > 2$ to each ψ_j, $j = 1, \ldots$, then we have $\sum_{j=1}^{\infty} a_j\tau_j < \infty$ and $\sum_{j=1}^{\infty} b_j\tau_j < \infty$.

Remark 9. Ghosal and Van der Vaart [8, 7.7.1] and Huang [10] study properties of another type of series, spline series, that can be used to form priors for an unknown regression function. Spline series also have desirable asymptotic properties. For example, Ghosal and Van der Vaart [8, 7.7.1] studied the rate of convergence of posterior distributions for spline series priors.

6.2. Example 2: Gaussian process priors

Another interesting example of Bayesian nonparametric regression is known as Gaussian process regression. Gaussian process regression consists of modeling the unknown one-dimensional regression function, $\eta(x)$ as a Gaussian process a priori. Here, we assume $\eta(\cdot)$ is a Gaussian process with the mean function $\mu_0(\cdot)$ and the covariance function $R(\cdot, \cdot; \beta)$ conditional on the additional parameter, β, which is independent of the noise variance σ^2 and the noises, $(\epsilon_1, \ldots, \epsilon_n)$. Here, we require that the covariance function $R(x, x'; \beta)$ have the form $R_0(\beta|x - x'|)$, where $R_0(x)$ is a positive multiple of a nowhere zero density function and four times continuously differentiable on $\mathbb{R}$ and the mean function $\eta_0(x)$ is continuously differentiable in [0, 1]. The additional parameter β is needed to verify the prior positivity condition of Section 4.2.1. The support of β is $\mathbb{R}^+$ and it is a smoothing parameter that helps determine how far apart x and x' can be while $\eta(x)$ and $\eta(x')$ remain highly correlated. In addition, we assume that for all $n \geqslant 1$, all $\beta > 0$, and all distinct $x_1, \ldots, x_n \in [0, 1]$, the $n \times n$ covariance matrix $((\Sigma_{i,j}))$ with $\Sigma_{i,j} = R(x_i, x_j; \beta)$, is non-singular. Under these assumptions, it can be shown that if η_0 is continuous,

$$P\left(\sup_{x\in[0,1]} |\eta(x) - \eta_0(x)| < \epsilon\right) > 0,$$

for each $\epsilon > 0$. The proof follows from Tokdar and Ghosh [15, Theorem 4.2–4.4] or Ghosal and Roy [7, Theorem 4].

To obtain a suitable probability bound for Θ_n^C as in Section 4.2.2, we need an additional assumption on the prior for β. We assume that there exist $0 < \delta < \frac{1}{2}$ and $b_1, b_2 > 0$ such that $\Pr\left\{\beta > n^\delta\right\} < b_1 \exp(-b_2 n)$, $\forall n \geqslant 1$. Under all of these conditions, if we let $M_n = \mathcal{O}(n^\alpha)$ with α satisfying $\frac{2\delta+1}{2} < \alpha < 1$, we can show that there exist constants C_4 and c_4 such that

$$\Pi(\Theta_n^C) \leqslant C_4 \exp(-c_4 n).$$

The proof follows from Van der Vaart and Wellner [16, Proposition A.2.7], Ghosal and Roy [7, Theorem 5] and Choi [4, Lemmas 3.5.6–3.5.8]. (See [4] for a detailed proof.)

In high dimensions, Gaussian process priors with suitable covariance functions have been investigated and provided successful empirical results. (See for example, [11,12].)

Remark 10. The orthogonal basis expansion in Section 6.1 and the spline series mentioned in Remark 9 belong to the class of Gaussian process priors in Example 6.2. However, when one wishes to specify a particular form of covariance function (such as squared exponential) in order to model a desired form of dependence, applying consistency results for orthogonal basis expansions requires knowing the eigenvalue decomposition of the desired covariance function (analytically, not numerically). Alternatively, one could give up the flexibility of specifying a form of covariance and let the covariance function be determined by the orthogonal basis via Mercer's Theorem or the Karhunen–Loéve expansion. (See for example [9]).

7. Discussion

We began by providing an extension of the posterior consistency theorem of Schwartz [13] for independent but non-identically distributed observations. We then used this theorem to create sufficient conditions for proving almost sure consistency of posterior distributions in nonparametric regression problems. These conditions require the existence of tests with suitable properties and a prior with low probability on large and unsmooth functions and sufficiently high probability near the true regression function. The tests that we have constructed for this purpose were designed independently of the specific prior distribution that we used, and these tests can be used for posterior consistency of general nonparametric regression models with additive noise as long as the prior distribution and the regression function under consideration meet our conditions. We also gave two specific examples of classes of priors for nonparametric regression problems that satisfied our sufficient conditions for consistency.

We have not discussed rates of convergence in this paper. Ghosal and Van der Vaart [8] present general results on convergence rates for non-i.i.d. observations which include nonparametric regression cases. They mention the general results for nonparametric regression based on the Hellinger metric. We would like to extend their results into our model structure. The test functions that we constructed could also be used to compute rates of convergence by replacing the fixed ϵ with a decreasing sequence $\{\epsilon_n\}_{n=1}^{\infty}$. In this case, the challenge is to find the rate at which the prior probability shrinks as we decrease the size of the Kullback–Leibler neighborhood of the true regression function. This problem involves finding lower tail probabilities for the regression function and will depend on the nature of prior distributions under consideration. Finally, we have assumed that the form of the noise distribution is normal. We can easily extend the result on the existence of test functions to the case in which the noise distribution is Laplace (double exponential) instead of normal. (See [4]).

Acknowledgments

The authors thank the referees and an Associate Editor for helpful comments.

Appendix A. Proof of Theorem 1

The posterior probability (1) can be written as

$$
\begin{aligned}
&\Pi(\theta \in U_n^C | Z_1, \dots, Z_n) \\
&\quad = \frac{\int_{U_n^C \cap \Theta_n} \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)}\, d\Pi(\theta) + \int_{U_n^C \cap \Theta_n^C} \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)}\, d\Pi(\theta)}{\int_\Theta \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)}\, d\Pi(\theta)} \\
&\quad \leqslant \Phi_n + \frac{(1-\Phi_n)\int_{U_n^C \cap \Theta_n} \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)} d\Pi(\theta) + \int_{U_n^C \cap \Theta_n^C} \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)}\, d\Pi(\theta)}{\int_\Theta \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)}\, d\Pi(\theta)} \\
&= \Phi_n + \frac{I_{1n}(Z_1, \dots, Z_n) + I_{2n}(Z_1, \dots, Z_n)}{I_{3n}(Z_1, \dots, Z_n)}. \qquad \text{(A.1)}
\end{aligned}
$$

The main proof consists of proving the following results:

$$\Phi_n \to 0 \quad \text{a.s. } [P_{\theta_0}], \tag{A.2}$$

$$e^{\beta_1 n} I_{1n}(Z_1, \dots, Z_n) \to 0 \quad \text{a.s. } [P_{\theta_0}] \text{ for some } \beta_1 > 0, \tag{A.3}$$

$$e^{\beta_2 n} I_{2n}(Z_1, \dots, Z_n) \to 0 \quad \text{a.s. } [P_{\theta_0}] \text{ for some } \beta_2 > 0, \tag{A.4}$$

$$e^{\beta n} I_{3n}(Z_1, \dots, Z_n) \to \infty \quad \text{a.s. } [P_{\theta_0}] \text{ for all}\beta > 0. \tag{A.5}$$

Letting $\beta \leqslant \min\{\beta_1, \beta_2\}$ will imply (1).

(A.2) is obtained from the Markov inequality and Borel–Cantelli lemma. (A.3) and (A.4) are clear from the proof of Choudhuri et al. [5, Theorem 2]. Finally, (A.5) follows from Amewou-Atisso et al. [1, Lemma A.1] and the proof of Amewou-Atisso et al. [1, Theorem 2.1].

Appendix B. Proof of Theorem 2

We will construct the test Φ_n as the maximum of a collection of tests indexed by elements of the covering of a sieve in a manner similar to that of Ghosal and Roy [7]. In what follows, not all details of the proof are given. Readers can refer to Choi [4] for more detailed calculations.

Let $0 < r < \min\{\epsilon/2, 4\sigma_0\sqrt{\epsilon - \epsilon^2}\}$, and $t = \min\{\epsilon/2, r/4\}$. Let N_t be the t covering number of Θ_{1n} in the supremum (L^∞) norm. As mentioned in Section 4.1, $N_t = N(t, \Theta_{1n}, \|\cdot\|_\infty) = \mathcal{O}(M_n)$. Recall that $M_n = \mathcal{O}(n^\alpha)$ where $\frac{1}{2} < \alpha < 1$. Let $\alpha/2 < \tau < 1/2$, and define $c_n = n^\tau$ so that $\log(N_t) = o(c_n^2)$. Let $\eta^1, \ldots, \eta^{N_t} \in \Theta_{1n}$ be such that for each $\eta \in \Theta_{1n}$ there exists i such that $\|\eta - \eta^i\|_\infty < t$. If $\int |\eta - \eta_0| dQ_n > \epsilon$, then $\int |\eta^i - \eta_0| dQ_n > \epsilon/2$.

Next, we construct tests corresponding to each η^i. Let η be a continuous function on $[0, 1]^d$ and let $\delta > 0$. Define $\eta_{*j} = \eta(x_j)$ and $\eta_{0j} = \eta_0(x_j)$ for $j = 1, \ldots, n$. Let $b_j = 1$ if $\eta_{*j} \geqslant \eta_{0j}$ and -1 otherwise. Let $\Psi_{1n}[\eta, \delta]$, $\Psi_{2n}[\delta]$ and $\Psi_{3n}[\eta, \delta]$ be the indicators of the set A_1, A_2 and A_3, respectively, where

$$A_1 = \left\{ \sum_{j=1}^{n} b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) > 2c_n\sqrt{n} \right\},$$

$$A_2 = \left\{ \sum_{j=1}^{n} \frac{(Y_j - \eta_{0j})^2}{\sigma_0^2} > n(1 + \delta) \right\}$$

and

$$A_3 = \left\{ \sum_{j=1}^{n} \frac{(Y_j - \eta_{*j})^2}{\sigma_0^2} \leqslant n(1 - \delta^2) \right\}.$$

Define $\Psi_n[\eta, \delta] = \max\{\Psi_{1n}[\eta, \delta], \Psi_{2,n}[2\delta], \Psi_{3n}[\eta, 2\delta]\}$. Our final test Φ_n is defined by $\Phi_n = \max_{1 \leqslant j \leqslant N_t} \Psi_n[\eta^j, \epsilon/2]$. Next, we bound the type I and type II error probabilities for each $\Psi_n[\eta, \delta]$ and show that these bounds imply the necessary bounds for the error probabilities of Φ_n.

First, consider the type I error, $\mathrm{E}_{P_0}(\Psi_n[\eta, \delta]) \leqslant \mathrm{E}_{P_0}(\Psi_{1n}) + \mathrm{E}_{P_0}(\Psi_{2n}) + \mathrm{E}_{P_0}(\Psi_{3n})$. Note that

(a) By Mill's ratio, we have

$$\mathrm{E}_{P_0}(\Psi_{1n}) \leqslant \frac{1}{2\sqrt{2\pi}} \frac{\exp(-2c_n^2)}{c_n}.$$

(b) By the Markov inequality and Chernoff bounds, we have

$$\mathrm{E}_{P_0}(\Psi_{2n}) \leqslant \exp\left(-n \left[\frac{\delta^2}{4} - \frac{\delta^3}{6} \right] \right).$$

Table B1
Three possible cases of (η, σ) for computing type II error bounds

Case	$(\eta, \sigma) \in D_i, i = 1, 2, 3$
I	$D_1 = \left\{(\eta, \sigma) : \int \lvert\eta(\mathbf{x}) - \eta_0(\mathbf{x})\rvert dQ_n(\mathbf{x}) > \epsilon, \frac{\sigma}{\sigma_0} \leqslant 1 + \epsilon\right\}$
II	$D_2 = \left\{(\eta, \sigma) : \frac{\sigma}{\sigma_0} > 1 + \epsilon\right\}$
III	$D_3 = \left\{(\eta, \sigma) : \frac{\sigma}{\sigma_0} < 1 - \epsilon\right\}$

(c) Let $W \sim \chi_n^2$ and $W' \sim \chi^2(n, \vartheta)$ with $\vartheta = \sum_{j=1}^n \left(\frac{\eta_{*j} - \eta_{0j}}{\sigma}\right)^2$. For all $t < 0$, we have

$$\mathrm{E}_{P_0}(\Psi_{3n}) = P_0 \left\{ \sum_{j=1}^n \left(\frac{Y_j - \eta_{0j}}{\sigma_0} + \frac{\eta_{0j} - \eta_{*j}}{\sigma_0} \right)^2 \leqslant n[1 + \delta](1 - \delta) \right\}$$
$$\leqslant \Pr\left\{ W' \leqslant n\left[1 - \delta^2\right] \right\} \leqslant \Pr\left\{ W \leqslant n\left[1 - \delta^2\right] \right\}.$$

Therefore, by the Chernoff bounds, we get $E_{P_0}(\Psi_{3n}) \leqslant \exp\left(-n\frac{(\delta^*)^2 - (\delta^*)^3}{4(1+\delta^*)}\right)$, where $\delta^* = \frac{\delta^2}{1-\delta^2}$.

It follows that there exists C_3 such that $\mathrm{E}_{P_0}\Psi_n < C_3 \exp(-2c_n^2)$. Since $\log(N_t) = o(c_n^2)$, we have

$$\mathrm{E}_{P_0}\Phi_n \leqslant \sum_{j=1}^{N_t} \mathrm{E}_{P_0}\Psi_n[\eta^j, \epsilon/2] \leqslant C_3 \exp(\log[N_t] - 2c_n^2) \leqslant C_3 \exp(-c_n^2).$$

From this it follows that $\sum_{n=1}^\infty \mathrm{E}_{P_0}\Phi_n < \infty$.

Next we consider the type II error probability. The type II error probability of Φ_n is no larger than the minimum of the type II error probabilites of all of the $\Psi_n[\eta^i, \epsilon/2]$ which, in turn, is no larger than the minimum of the type II error probabilities of the three tests $\Psi_{1n}[\eta^i, \epsilon/2]$, $\Psi_{2n}[\epsilon/2]$ and $\Psi_{3n}[\eta^i, \epsilon/2]$. For each $(\eta, \sigma) \in W_{\epsilon,n}^C \cap \Theta_{1n}$, we need to find only one i and one k such that $\Psi_{kn}[\eta^i, \epsilon/2]$ has type II error probability sufficiently small. We divide the (η, σ) values into three cases, and for each case we calculate one of the type II error probabilities of Ψ_{1n}, Ψ_{2n} or Ψ_{3n}, where the cases are summarized in Table B1. The following notation will be used in all three cases. Let $\eta_{0j} = \eta_0(x_j)$ for $j = 1, \ldots, n$. For arbitrary continuous $\eta \in \Theta_{1n}$, let η^i be such that $\|\eta - \eta^i\|_\infty < t$. Then define $\eta_{1j} = \eta^i(x_j)$ for $j = 1, \ldots, n$.

(a) *Case I*: $\int |\eta(\mathbf{x}) - \eta_0(\mathbf{x})| dQ_n(\mathbf{x}) > \epsilon/2$, $\sigma \leqslant \sigma_0(1 + \epsilon)$.

It follows that $\int |\eta^i(\mathbf{x}) - \eta_0(\mathbf{x})| dQ_n > \epsilon/2$. In this case, we consider the test function Ψ_{1n} and calculate the type II error probability of Ψ_{1n}. For every $r < \epsilon/2$,

$$\sum_{j=1}^n |\eta_{1j} - \eta_{0j}| > rn. \tag{B.1}$$

Assume n is large enough so that $c_n/\sqrt{n} < r/(4\sigma_0)$. Since $\sigma < (1+\epsilon)\sigma_0$,

$$\begin{aligned}
\mathrm{E}_P(1-\Psi_{1n}[\eta^i,\epsilon/2]) &= P\left\{\sum_{j=1}^n b_j\left(\frac{Y_j-\eta_{0j}}{\sigma_0}\right)\leqslant 2c_n\sqrt{n}\right\}\\
&= P\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^n b_j\left(\frac{Y_j-\eta_{*j}}{\sigma}\right)+\frac{1}{\sqrt{n}}\sum_{j=1}^n b_j\left(\frac{\eta_{*j}-\eta_{1j}}{\sigma}\right)\right.\\
&\qquad\left.+\frac{1}{\sqrt{n}}\sum_{j=1}^n\left|\frac{\eta_{1j}-\eta_{0j}}{\sigma}\right|\leqslant 2c_n\frac{\sigma_0}{\sigma}\right\}\\
&\leqslant \frac{4\sigma_0(1+\epsilon)}{r\sqrt{2\pi n}}\exp\left(-\frac{nr^2}{32\sigma_0^2(1+\epsilon)^2}\right),
\end{aligned}$$

where the last inequality is by Mill's ratio.

(b) *Case II*: $\sigma > \sigma_0(1+\epsilon)$.

In this case, we use the type II error probability of the test Ψ_{2n}. Let $W \sim \chi_n^2$ and let W' have a noncentral χ^2 distribution with n degrees of freedom and noncentrality parameter $\frac{\sum_{j=1}^n(\eta_{*j}-\eta_{0j})^2}{\sigma^2}$. Then, since $\sigma > (1+\epsilon)\sigma_0$,

$$\mathrm{E}_P(1-\Psi_{2n}) = \Pr\left(W'\leqslant n\frac{\sigma_0^2}{\sigma^2}[1+\epsilon]\right)\leqslant \Pr\left(W\leqslant\frac{n}{1+\epsilon}\right).$$

By the Chernoff bounds, we get $\mathrm{E}_P(1-\Psi_n[\eta^i,\epsilon/2])\leqslant\exp\left(-n\frac{\epsilon^2-\epsilon^3}{4(1+\epsilon)}\right)$.

(c) *Case III*: $\sigma < \sigma_0(1-\epsilon)$.

For this case, we calculate the type II error probability of Ψ_{3n}.

$$\begin{aligned}
\mathrm{E}_P(1-\Psi_{3n}[\eta^i,\epsilon/2]) &= P\left\{n[1-\epsilon^2/4]\leqslant\sum_{j=1}^n\left(\frac{Y_j-\eta_{1j}}{\sigma_0}\right)^2\right\}\\
&= \Pr\left(n\frac{\sigma_0^2}{\sigma^2}[1-\epsilon^2/4]\leqslant W'\right),
\end{aligned}$$

where W' is a noncentral χ^2 distribution, $\chi^2(n,\vartheta)$, where $\vartheta=\sum_{j=1}^n\left(\frac{\eta_{*j}-\eta_{1j}}{\sigma}\right)^2$. Note that the moment generating function of noncentral $\chi^2(n,\vartheta)$ distribution is given as $\psi(s)=(1-2s)^{-n/2}\exp\left\{-\vartheta/2\left[1-(1-2s)^{-1}\right]\right\}$ for all $s<\frac{1}{2}$. Because $\|\eta^i-\eta\|_\infty < r/4$, we have $|\eta_{1j}-\eta_{*j}| < r/4,\ \forall j=1,\ldots,n$. Therefore, for all $0<s<1/2$,

$$\begin{aligned}
&\Pr\left\{\exp(W's)\geqslant\exp\left(ns[1-\epsilon^2/4]\frac{\sigma_0^2}{\sigma^2}\right)\right\}\\
&\quad\leqslant\exp\left\{\frac{n}{2}\left[-\log(1-2s)-\left(1-\frac{1}{1-2s}\right)\frac{1}{n}\sum_{j=1}^n\left(\frac{\eta_{*j}-\eta_{1j}}{\sigma}\right)^2-2s[1-\epsilon^2/4]\frac{\sigma_0^2}{\sigma^2}\right]\right\}\\
&\quad\leqslant\exp\left\{n\frac{s\sigma_0^2}{\sigma^2}\left[\frac{1}{1-2s}\left((1-\epsilon)^2+\frac{r^2}{16\sigma_0^2}\right)-[1-\epsilon^2/4]\right]\right\}.
\end{aligned}\qquad\text{(B.2)}$$

Because $r^2 < 16\sigma_0^2(\epsilon - \epsilon^2)$, (B.2) is less than $\exp\left\{ns\frac{\sigma_0^2}{\sigma^2}\left[\frac{1}{1-2s}(1-\epsilon) - \left[1-\epsilon^2/4\right]\right]\right\}$. Now, take $s = s^*$ such that $\frac{1}{1-2s^*} = \frac{1-(1+\epsilon)\epsilon^2/4}{1-\epsilon}$ to get

$$\mathrm{E}_P(\Psi_{3n}) \leqslant \exp\left\{-ns^*\frac{\sigma_0^2}{\sigma^2}\frac{\epsilon^3}{4}\right\} \leqslant \exp\left\{-\frac{ns^*}{(1-\epsilon)^2}\frac{\epsilon^3}{4}\right\}.$$

Take C_5 to be the minimum of $r^2/[32\sigma_0^2(1+\epsilon)^2]$, $[\epsilon^2 - \epsilon^3]/[4(1+\epsilon)]$, and $s^*\epsilon^3/[4(1-\epsilon)^2]$ so that $\sup_{\theta \in W_{\epsilon,n}^C \cap \Theta_n} \mathrm{E}_P(1-\Phi_n) \leqslant \exp(-C_5 n)$.

Appendix C. Proof of Theorem 3

As in the proof of Theorem 2, we also consider three different cases of (η, σ) similar to Table B1. Since we deal with d_Q topology for regression functions, the only difference is made for Case I based on d_Q metric, and it follows that cases II and III are independent of η and exactly the same as in Table B1. Theorem 2 and Lemma 11 enable us to use Ψ_{1n} for case I. In addition, for the random design case, Lemma 11 tells us that (B.1) occurs all but finitely often with probability 1. Since there are only finitely many η^j to consider for each n, this suffices to obtain the exponentially small probability bound for Theorem 3.

Lemma 11. *Assume the covariate values are sampled from a probability distribution Q. Let η be a function such that $d_Q(\eta, \eta_0) > \epsilon$. Let $0 < r < \epsilon^2$, and define*

$$A_n = \left\{\sum_{i=1}^{n} |\eta(X_i) - \eta_0(X_i)| \geqslant rn\right\}.$$

Then there exists $C_{11} > 0$ such that $\Pr(A_n^C) \leqslant \exp(-C_{11}n)$ for all n and A_n occurs all but finitely often with probability 1. *The same C_{11} works for all η such that $d_Q(\eta, \eta_0) > \epsilon$.*

Proof. Let $B = \{x | \eta(x) - \eta_0(x)| > \epsilon\}$, so that $Q(B) > \epsilon$. Let $Z = n - \sum_{i=1}^n I_B(X_i)$, and notice that Z has a binomial distribution with parameters n and $1 - Q(B)$. Let $q = r/\epsilon < \epsilon$, and let Z' have a binomial distribution with parameters n and $1-\epsilon$ so that Z' stochastically dominates Z. Then

$$\Pr(A_n^C) \leqslant \Pr(Z > n[1-q]) \leqslant \Pr(Z' > n[1-q]).$$

Note that for all $t > 0$, $\Pr(Z' > n[1-q]) \leqslant [\epsilon + [1-\epsilon]\exp(t)]^n \exp(-tn[1-q])$. Let

$$t = \log\left(\frac{\epsilon(1-q)}{q(1-\epsilon)}\right) > 0, \quad C_{11} = q\log\left(\frac{q}{\epsilon}\right) + (1-q)\log\left(\frac{1-q}{1-\epsilon}\right) > 0.$$

Then $\Pr(A_n^C) \leqslant \exp(-C_{11}n)$, and C_{11} does not depend on the particular η. The probability one claim follows from the first Borel–Cantelli lemma.

To complete the proof, consider the same t, N_t and $\eta^1, \ldots, \eta^{N_t} \in \Theta_{1n}$ as in the proof of Theorem 2. If $d_Q(\eta, \eta_0) > \epsilon$, then $d_Q(\eta^j, \eta_0) > \epsilon/2$. Consider the same test, Φ_n as in Theorem 2. It is clear that we have

$$\text{(i)} \sum_{n=1}^{\infty} \mathrm{E}_{P_0}\Phi_n < \infty \quad \text{and} \quad \text{(ii)} \sup_{\theta \in U_\epsilon^C \cap \Theta_n} \mathrm{E}_P(1-\Phi_n) \leqslant \exp(-C_6 n). \qquad \Box$$

Appendix D. Proof of Theorem 4

The proof of Theorem 4 (1) obviously follows from Theorems 1 and 2. We state the proof of Theorem 4 (2) in detail as follows.

The test constructions count on the existence of the quantity in (B.1) under L^1 topology of $\eta(\mathbf{x})$. Thus, we need to prove first that if $\int |\eta_1(\mathbf{x}) - \eta_0(\mathbf{x})| dQ > \epsilon$, then there exists a suitable constant r such that (B.1) holds as follows:

$$\sum_{j=1}^{n} |\eta_{1j} - \eta_{0j}| > rn.$$

When the dimension of a covariate grows, this problem becomes more complicated in particular when we use a L^1 metric compared to the metric based on empirical measure in Theorem 2. The following two lemmas show that under Assumptions B_d and NRD$_d$, there exist enough number of design points to acquire (B.1) in d dimensions.

Lemma 12. *Assume Assumption NRD$_d$. Let λ be Lebesgue measure in $\mathbb{R}^d$. Let K_d be the constant mentioned in Assumption NRD$_d$ and V be a positive constant. Let A_V be the set of all continuous functions γ such that $\forall \mathbf{x}_1, \mathbf{x}_2 \in [0,1]^d$, $|\gamma(\mathbf{x}_1) - \gamma(\mathbf{x}_2)| \leqslant V\|\mathbf{x}_1 - \mathbf{x}_2\|$. For each such function γ and $\epsilon > 0$, define $B_{\epsilon,\gamma} = \{x : |\gamma(x)| > \epsilon\}$. Then for each $\epsilon > 0$ there exists an integer N such that, for all $n \geqslant N$ and all $\gamma \in A_V$,*

$$\sum_{i=1}^{n} |\gamma(x_i)| \geqslant nK_d \lambda(B_{\epsilon,\gamma}) \frac{\epsilon}{2}. \tag{D.1}$$

Proof. Let $\epsilon > 0$ and let N be large enough so that $\frac{\sqrt{d}V}{(K_d n)^{1/d}} < \epsilon/2$ for all $n \geqslant N$. Let $\gamma \in A_V$. Since $B_{\epsilon,\gamma}$ is an open set in $\mathbb{R}^d$, it can be approximated by a finite collection of disjoint and equally-sized hypercubes, $\left\{H^k_{\epsilon,\gamma}\right\}_{k=1}^{N_{\epsilon,\gamma}}$. Specifically,

$$\lambda(H^k_{\epsilon,\gamma}) = \frac{1}{K_d n}, \quad H^k_{\epsilon,\gamma} \cap H^j_{\epsilon,\gamma} = \emptyset (j \neq k), \quad H^k_{\epsilon,\gamma} \cap B_{\epsilon,\gamma} \neq \emptyset \quad \text{and} \quad B_{\epsilon,\gamma} \subset \bigcup_{k=1}^{N_{\epsilon,\gamma}} H^k_{\epsilon,\gamma}.$$

By Assumption NRD$_d$, each hypercube of $\left\{H^k_{\epsilon,\gamma}\right\}_{k=1}^{N_{\epsilon,\gamma}}$ contains at least one design point. Let x^*_k be a design point in $H^k_{\epsilon,\gamma}$. If x^*_k is also in $B_{\epsilon,\gamma}$, then it is obvious that $|\gamma(x^*_k)| > \epsilon/2$. If $x^*_k \in H^k_{\epsilon,\gamma} \cap B^c_{\epsilon,\gamma}$, then it is still true that $|\gamma(x^*_k)| > \epsilon/2$ as follows: Let x^* be a point in $H^k_{\epsilon,\gamma} \cap B_{\epsilon,\gamma}$. Since x^*_k is in $H^k_{\epsilon,\gamma}$, the distance between x^*_k and x^* is less than $\frac{\sqrt{d}}{(K_d n)^{1/d}}$, i.e. $\|x^*_k - x^*\|_d \leqslant \frac{\sqrt{d}}{(K_d n)^{1/d}}$, where $\|\cdot\|_d$ is the Euclidean distance. In addition, by the assumption on A_V, it is clear that

$$\gamma(x^*_k) \geqslant \gamma(x^*) - \frac{\sqrt{d}V}{(K_d n)^{1/d}} > \epsilon - \frac{\sqrt{d}V}{(K_d n)^{1/d}} > \frac{\epsilon}{2}.$$

Thus, there are $N_{\epsilon,\gamma}$ design points, $\{x^*_k\}_{k=1}^{N_{\epsilon,\gamma}}$ such that $|\gamma(x^*_k)| > \epsilon/2$. Therefore, $\sum_{k=1}^{N_{\epsilon,\gamma}} |\gamma(x^*_k)| > N_{\epsilon,\gamma}\frac{\epsilon}{2}$. Furthermore, since $\lambda\left(\bigcup_{k=1}^{N_{\epsilon,\gamma}} H^k_{\epsilon,\gamma}\right) = \frac{1}{K_d n} N_{\epsilon,\gamma} \geqslant \lambda\left(B_{\epsilon,\gamma}\right)$, we have $\sum_{i=1}^{n} |\gamma(x_i)| \geqslant \{K_d n \lambda \left(B_{\epsilon,\gamma}\right)\} \frac{\epsilon}{2}, \forall d \geqslant 2.$ □

Lemma 13. *Assume Assumption NRD$_d$. Let K_d and V be constants defined in the previous Lemma. Let A_V be the set of all continuous functions η such that $\|\eta\|_\infty < M_n$ and $\|\frac{\partial}{\partial x_j}\eta(x_1, \ldots, x_d)\|_\infty < V$, $j = 1, \ldots, d$. Then for each $\epsilon > 0$ there exist an integer N and $r > 0$ such that, for all $n \geqslant N$ and all $\eta \in A_V$ such that $\|\eta - \eta_0\|_1 > \epsilon$, $\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| \geqslant rn$.*

Proof. Let V_0 be an upper bound of η_0 and let V be an upper bound on all partial derivatives of η_0. Let $0 < \delta < \epsilon$. Let $r = K_d(\epsilon - \delta)/4$ and $D_i = \{x : (i-1)\delta < |\eta(x) - \eta_0(x)| < i\delta\}$.

Let λ be Lebesgue measure. Then, $\|\eta - \eta_0\|_1$ can be bounded as follows:

$$\sum_i i\delta\lambda(D_i) \geqslant \|\eta - \eta_0\|_1 > \epsilon. \tag{D.2}$$

Let $\zeta(x) = |\eta(x) - \eta_0(x)|$ and $\zeta_m(x) = \min\{m\delta, \zeta(x)\}$, for $m = 0, \ldots, n$. Note that $\zeta_{\lceil M_n + V_0/\delta\rceil}(x)$ is the same as $\zeta(x)$.

For $m = 1, \ldots, n$, define $B_m \equiv \{x : \zeta_m(x) > (2m-1)\delta/2\}$. Then, for all $x \in B_m$, $\zeta_m(x) - \zeta_{m-1}(x) > \delta/2$. Thus, Lemma 12 (with $\gamma = \zeta_m - \zeta_{m-1}$ and $\epsilon = \delta/2$) implies $\sum_{i=1}^n \left(\zeta_m(x_i) - \zeta_{m-1}(x_i)\right) \geqslant \{nK_d\lambda(B_m)\}\frac{\delta}{4}$.

Now, write

$$\begin{aligned}
\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| &= \sum_{i=1}^n \zeta_{\lceil (M_n+V_0)/\delta\rceil}(x_i) \\
&= \sum_{i=1}^n \sum_{m=1}^{\lceil (M_n+V_0)/\delta\rceil} \left\{\zeta_m(x_i) - \zeta_{m-1}(x_i)\right\} \\
&= \sum_{m=1}^{\lceil (M_n+V_0)/\delta\rceil} \sum_{i=1}^n \left\{\zeta_m(x_i) - \zeta_{m-1}(x_i)\right\}, \\
&\geqslant \sum_{m=1}^{\lceil (M_n+V_0)/\delta\rceil} \left[\lambda(B_m)n\right] K_d\frac{\delta}{4}.
\end{aligned} \tag{D.3}$$

Also, for $m = 1, \ldots, n$, $B_m \supset \bigcup_{i=m+1}^{\lceil (M_n+V_0)/\delta\rceil} D_i$. Therefore, it follows that

$$\sum_{m=1}^{\lceil (M_n+V_0)/\delta\rceil} \delta\lambda(B_m) \geqslant \sum_{i=2}^{\lceil (M_n+V_0)/\delta\rceil} (i-1)\delta\lambda(D_i) \geqslant \left\{\sum_{i=2}^{\lceil (M_n+V_0)/\delta\rceil} i\delta\lambda(D_i)\right\} - \delta \geqslant \epsilon - \delta,$$

where the last inequality follows from (D.2). Combining this with (D.3) gives

$$\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| \geqslant n(\epsilon - \delta)\frac{K_d}{4} \quad \text{for all } n \geqslant N. \qquad \square$$

Hence, the proof obviously follows from Lemmas 12 and 13 and Theorems 1 and 2.

Appendix E. Proof of Theorem 6

The proof of Theorem 6 (1) obviously follows from Theorems 1 and 3 and we describe the proof of Theorem 6 (2) in detail as follows.

In order to obtain almost sure consistency based on Hellinger metric, we need to calculate the Hellinger distance between two density functions, $d_H(f, f_0)$. For simplified calculation, we consider the quantity $h(f, f_0)$ defined as $h(f, f_0) = \frac{1}{2}d_H^2(f, f_0)$, and $h(f, f_0)$ is calculated as follows:

$$\begin{aligned} h(f, f_0) &= 1 - \frac{1}{\sqrt{2\pi\sigma\sigma_0}} \iint \exp\left\{-\frac{1}{4\sigma^2}\left[y - \eta(\mathbf{x})\right]^2 - \frac{1}{4\sigma_0^2}\left[y - \eta_0(\mathbf{x})\right]^2\right\} dy\, dQ \\ &= 1 - \int \sqrt{\frac{2\sigma\sigma_0}{\sigma^2 + \sigma_0^2}} \exp\left\{-\frac{1}{16\sigma^2\sigma_0^2}\left[\eta(\mathbf{x}) - \eta_0(\mathbf{x})\right]^2 \Bigg/ \left(\frac{1}{4\sigma^2} + \frac{1}{4\sigma_0^2}\right)\right\} dQ. \end{aligned} \tag{E.1}$$

The integral in (E.1) is of the form $\int c_1 \exp(-c_2[\eta(\mathbf{x}) - \eta_0(\mathbf{x})]^2)\, dQ(\mathbf{x})$, where c_1 can be made arbitrarily close to 1 by choosing $|\sigma/\sigma_0 - 1|$ small enough and c_2 is bounded when σ is close to σ_0. In addition, since $c_1 \exp(-c_2[\eta(\mathbf{x}) - \eta_0(\mathbf{x})]^2)$ is bounded by c_1, we have

$$\begin{aligned} &\int 1 - c_1 \exp(-c_2[\eta(\mathbf{x}) - \eta_0(\mathbf{x})]^2)\, dQ(\mathbf{x}) \\ &\quad = \int_{|\eta - \eta_0| \leqslant \delta} 1 - c_1 \exp(-c_2[\eta(\mathbf{x}) - \eta_0(\mathbf{x})]^2)\, dQ(\mathbf{x}) \\ &\qquad + \int_{|\eta - \eta_0| > \delta} 1 - c_1 \exp(-c_2[\eta(\mathbf{x}) - \eta_0(\mathbf{x})]^2)\, dQ(\mathbf{x}) \\ &\quad \leqslant \left\{1 - c_1 \exp(-c_2\delta^2)\right\} + (1 + c_1) Q(\mathbf{x} : |\eta(\mathbf{x}) - \eta_0(\mathbf{x})| > \delta). \end{aligned}$$

Therefore, it follows that for each ϵ there exists a δ such that (E.1) will be less than ϵ whenever $|\sigma/\sigma_0 - 1| < \delta$ and $d_Q(\eta, \eta_0) < \delta$. Hence, it is proven.

References

[1] M. Amewou-Atisso, S. Ghosal, J.K. Ghosh, R.V. Ramamoorthi, Posterior consistency for semiparametric regression problems, Bernoulli 9 (2003) 291–312.

[2] A. Barron, M.J. Schervish, L. Wasserman, The consistency of posterior distributions in nonparametric problems, Ann. Statist. 27 (1999) 536–561.

[3] T. Choi, Alternative posterior consistency results in nonparametric binary regression using Gaussian process priors, J. Statist. Plann. Inference, 2007, to appear.

[4] T. Choi, Posterior consistency in nonparametric regression problems under Gaussian process priors, Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, 2005.

[5] N. Choudhuri, S. Ghosal, A. Roy, Bayesian estimation of the spectral density of time series, J. Amer. Statist. Assoc. 99 (2004) 1050–1059.

[6] S. Ghosal, J.K. Ghosh, R.V. Ramamoorthi, Posterior consistency of Dirichlet mixtures in density estimation, Ann. Statist. 27 (1999) 143–158.

[7] S. Ghosal, A. Roy, Posterior consistency of Gaussian process prior for nonparametric binary regression, Ann. Statist. 34 (2006) 2413–2429.

[8] S. Ghosal, A.W. Van der Vaart, Convergence rates of posterior distributions for noniid observations, Ann. Statist. 35 (2007) in press.

[9] U. Grenander, Abstract Inference, Wiley, NY, 1981.

[10] T.M. Huang, Convergence rates for posterior distributions and adaptive estimation, Ann. Statist. 32 (2004) 1556–1593.

[11] C. Paciorek, M.J. Schervish, Spatial modelling using a new class of nonstationary covariance functions, Environmetrics 17 (2006) 483–506.

[12] C.J. Paciorek, M.J. Schervish, Nonstationary covariance functions for Gaussian process regression, Adv. Neural Inform. Process. Syst. 16 (2004) 273–280.

[13] L. Schwartz, On Bayes procedures, Z. Wahr. Verw. Gebiete 4 (1965) 10–26.

[14] X. Shen, L. Wasserman, Rates of convergence of posterior distributions, Ann. Statist. 29 (2001) 666–686.

[15] S. Tokdar, J.K. Ghosh, Posterior consistency of Gaussian process priors in density estimation, J. Statist. Plann. Inference 137 (2007) 34–42.

[16] A.W. Van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer, New York, 1996.

[17] S.G. Walker, Bayesian consistency for a class of regression problems, South African Statist. J 37 (2003) 149–167.

[18] S.G. Walker, New approaches to Bayesian consistency, Ann. Statist. 32 (2004) 2028–2043.