Bayesian Consistency for Regression Models under a Supremum Distance

Fei Xiang[a,*], Stephen G. Walker[b]

[a]*School of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK*
[b]*School of Mathematics, Statistics and Actuarial Science, Cornwallis Building, University of Kent, Canterbury, Kent CT2 7NF, UK*

Abstract

This paper studies the consistency of Bayesian nonparametric regression models. We concentrate on the use of the sup metric and dealing with non–stochastic, i.e. designed, covariate values. We illustrate our results on a normal mean regression function and demonstrate the usefulness of a model based on piecewise constant functions.

Keywords: Bayesian nonparametric, Consistency, Regression model, Hellinger neighborhood, piecewise constant function

1. Introduction

This paper focuses on the posterior Bayesian consistency of regression models under a supremum based metric and on non–stochastic or design covariates. Observations are of the type $(y_i, x_i)_{i=1}^n$ which arise from some fixed but unknown regression model $f_0(y|x)$. Here, the (x_i) are assumed fixed non-random design points from the covariate space $\mathcal{X} = [0, 1]$. However, to start the theory, we take $\mathcal{X} = (z_1, z_2, z_3, \ldots)$, although examples in later sections consider taking values from $[0, 1]$. In our notation, however, we will simply refer to z_l as l. We use f to denote a conditional density from the Bayesian model and for concreteness $f(y|x)$ is assumed to be a density for all x with respect to the Lebesgue measure. The Bayesian prior is written as $\Pi(df)$ on the function space $\mathcal{F}$ and combines with the data to assign posterior mass to the set of conditional densities A as

$$\Pi^n(A) = \frac{I_A}{I_{\mathcal{F}}} = \frac{\int_A R_n(f)\Pi(df)}{\int_{\mathcal{F}} R_n(f)\Pi(df)},$$

where

$$R_n(f) = \prod_{i=1}^n \frac{f(y_i|x_i)}{f_0(y_i|x_i)}.$$

*Corresponding author

Email addresses: f.xiang@bristol.ac.uk (Fei Xiang), S.G.Walker@kent.ac.uk (Stephen G. Walker)

Our aim is to study consistency with respect to a sup based metric; and hence we are interested in

$$A = A_\epsilon = \{f : \sup_x H(f(\cdot|x), f_0(\cdot|x)) < \epsilon\},$$

where $H(f, f_0)$ is the Hellinger distance between f and f_0, given by

$$H(f, f_0) = \left\{ \int \left(\sqrt{f} - \sqrt{f_0} \right)^2 \right\}^{1/2}.$$

We want to establish sufficient conditions on Π which ensure that

$$\Pi^n(A_\epsilon^c) \to 0$$

in probability.

There are not many results of this type in the literature. One such notable exception is in Shively et al. (2009), where such a consistency result is obtained for a normal mean regression function, when the function is assumed to be monotone on a bounded interval.

Posterior consistency is an important issue in Bayesian nonparametrics. A good review is provided by Choi and Ramamoorthi (2008). Most techniques on demonstrating consistency are developed for the case of independent and identically distributed observations. A popular approach is based on sieves and the existence of uniformly consistent tests; see Barron et al. (1999) and Ghosal et al. (1999) for more details. Another approach was introduced in Walker (2003, 2004) and is based on the summability of square roots of prior masses on the parameter space covering Hellinger balls. It is this latter approach that will form the basis of the current paper.

Recent attention has been given to consistency for regression models. For instance, Amewou-Atisso et al. (2003) studies the weak consistency of semiparametric linear regression models. Also, Ghosal and Roy (2006) investigates the consistency of nonparametric binary regression models with unknown link functions using a Gaussian process prior. On the other hand, Choi (2007); Choi and Schervish (2007); Choi (2005) give alternative methods on the consistency of nonparametric regression models and multiple regression models. However, the assumptions here are that the covariates are generated independent and identically distributed from some distribution function, say Q. Then, taking the integrated Hellinger metric:

$$d(f_0, f) = \int H(f_0(\cdot|x)\, f(\cdot|x))\, dQ(x)$$

the resulting theory is identical to dealing with the independent and identically distributed case involving $(y_i, x_i)_{i=1}^n$. The work about establishing the conditions on Π which provide consistency is typically harder because the model can be high dimensional. A problem with this set–up is that it is not possible to establish consistency for the predictive at a fixed point in the covariate space. Another problem is that it is not always that covariates can be justified as being independent and identically distributed.

When covariates are non–stochastic, the analogous idea is to use the average metric

$$d(f_0, f) = \int H(f_0(\cdot|x)\, f(\cdot|x))\, dQ_n(x)$$

where Q_n is the empirical distribution of the (x_i). Here consistency is established by showing $\Pi^n(A^c_{\epsilon,n}) \to 0$ where

$$A_{\epsilon,n} = \left\{ f : n^{-1} \sum_{i=1}^{n} H\left(f(\cdot|x_i), f_0(\cdot|x_i)\right) < \epsilon \right\}.$$

Consistency here is a rather weak result and still says nothing about the consistency at a particular choice of x.

The layout of the paper is as follows. In Section 2 we describe necessary preliminaries including the basic assumptions. Section 3 contains the consistency results for both the posterior and predictive distributions. In Section 4 we illustrate the findings by establishing consistency for the normal model

$$f_0(y|x) = \mathrm{N}(y|\theta_0(x), \sigma_0^2).$$

Finally, Section 5 contains a discussion and details future work.

2. Preliminaries.

This section will describe basic concepts and ideas for the strategies adopted in this paper. To start, we assume that the covariates take values in $\mathcal{X} = \{z_1, z_2, \ldots\}$ and we will typically for simplicity represent these as integers; i.e. $\mathcal{X} = \{1, 2, \ldots\}$. Hence, we are interested in the set

$$A_\epsilon = \{f : H(f(\cdot|l), f_0(\cdot|l)) < \epsilon \ \forall l = 1, 2, ...\}$$

and we will also define

$$A_{\epsilon,l} = \{f : H(f(\cdot|l), f_0(\cdot|l)) < \epsilon\},$$

so that

$$A_\epsilon = \bigcap_l A_{\epsilon,l}.$$

Our aim is to find conditions on Π for which

$$\Pi^n(A^c_\epsilon) \to 0$$

in probability, and hence we become interested in the set

$$A^c_\epsilon = \bigcup_l A^c_{\epsilon,l}.$$

We will be taking finite versions of the model f for each sample size n and will denote the size of the model by N, typically, though not always, suppressing the dependence on n. As $n \to \infty$ it will be that $N \to \infty$, and also $n/N \to \infty$.

Our technique then relies on the idea that if

$$H(f(\cdot|l), f_0(\cdot|l)) > \epsilon$$

for some $l \in \{1, 2, \ldots\}$ then for some $l^* \in \{1, \ldots, N\}$ it is that

$$H(f(\cdot|l^*), f_0(\cdot|l^*)) > \epsilon_N$$

for some $\epsilon_N > 0$. To make this idea more general, let us say for $x \in \mathcal{X}$, now the continuous covariate space, $\mathcal{X} = [0,1]$, if

$$H(f(\cdot|x), f_0(\cdot|x)) > \epsilon$$

for some $x \in \mathcal{X}$ then for some $x^* \in C_N = \{z_1, \ldots, z_N\}$, with each $z_j \in \mathcal{X}$, it is that

$$H(f(\cdot|x^*), f_0(\cdot|x^*)) > \epsilon_N$$

for some $\epsilon_N > 0$. We will write $\epsilon_N = \epsilon\pi_N$ and note that for each N, C_N is a fixed set.

Hence, we can now write, and we are now working under the assumption $\mathcal{X} = [0,1]$, for any N,

$$\begin{aligned} A_\epsilon^c &\subset \bigcup_{1,\ldots,N} \{f : H(f(\cdot|l), f_0(\cdot|l)) > \epsilon_N\} \\ &= \bigcup_{l=1}^{N} A_{\epsilon_N,l}^c. \end{aligned}$$

So, we have

$$\Pi^n(A_\epsilon^c) \le \sum_{l=1}^{N} \Pi^n(A_{\epsilon_N,l}^c).$$

Now each $A_{\epsilon,l}$ defines a set of densities, since the covariate value is the same. Hence, this set of densities can be covered by sets of densities based on Hellinger balls of arbitrary size. This follows from the separability of densities with respect to the Hellinger metric.

Hence, for each $l \in \{1, \ldots, N\}$, there exist disjoint sets

$$\{A_{\epsilon_N,l,j};\, j = 1, 2, \ldots\}$$

such that

$$A_{\epsilon_N,l}^c = \bigcup_j A_{\epsilon_N,l,j}.$$

Each $A_{\epsilon_N,l,j}$ is a subset of a Hellinger ball, centered on $f_{l,j}$, and of size $\frac{1}{2}\epsilon_N$.

Consequently,

$$\Pi^n(A_\epsilon^c) \le \sum_{l=1}^{N} \sum_{j=1}^{\infty} \Pi^n(A_{\epsilon_N,l,j}).$$

The posterior probability for a set A is now written as

$$\Pi^n(A) = \frac{I_{n,A}}{I_n} = \frac{\int_A R_n(f)\,\Pi(df)}{\int_{\mathcal{F}} R_n(f)\,\Pi(df)}$$

where we have again suppressed the dependence on N.

Since we are dealing with N, which will be depending on n, and we need to consider ϵ_N, it follows that we need to proceed as though we are working with rates of convergence ideas. We shall also rewrite the prior as Π_N, as it is the belief on f with respect to each $l = \{1, 2, ..., N\}$. The hard part will be dealing with the numerator and hence here we first consider the denominator. We will consider dealing with the numerator in Section 3.

The following Lemma deals appropriately with the denominator. It is similar to Lemma 8.1 in Ghosal et al. (2000). We need to define the Kullback-Leibler divergence of two densities f and g as

$$K(f,g)=\int f\log\frac{f}{g},$$

and we also define

$$V(f,g)=\int f\left(\log\frac{f}{g}\right)^2.$$

Lemma 1. *Let*

$$K_l(f)=K(f_0(.|z_l),f(.|z_l))\ \textit{and}\ \ V_l(f)=V(f_0(.|z_l),f(.|z_l)).$$

Suppose that for a sequence $\tilde\epsilon_n$ *with* $\tilde\epsilon_n\to 0$ *and* $n\tilde\epsilon_n^2\to\infty$*, and some constant* $C>0$*, the prior satisfies*

$$\Pi_N\Big\{f: K_l(f)<\tilde\epsilon_n^2,\ V_l(f)<\tilde\epsilon_n^2\ \forall l=1,\dots,N\Big\}\geq\exp(-n\tilde\epsilon_n^2C)$$

for all large n*. Then,*

$$\int R_n(f)\,\Pi_N(df)\geq\exp\{-n\tilde\epsilon_n^2(C+1)\}\ \textit{in probability.}$$

PROOF. See Appendix.

3. Consistency results.

For reasons that will become clear later, we will take the first n covariates to be from those N values which we have discussed in Section 2. Therefore, we have $x_i=z_1$ for $i=1,\dots,r_n$, and $x_i=z_2$ for $i=r_n+1,\dots,2r_n$, and so on, to $x_i=z_N$ for $i=(N-1)r_n+1,\dots,Nr_n$. So $Nr_n=n$. We are not implying the order must be this way, we are only suggesting that among the n samples, we take r_n of each covariates $l\in\{1,\dots,N\}$, in any order. For the maths, it is convenient to assign a particular order.

Now define, for $l=1,\dots,N$,

$$L_{n,l,j}=\int_{A_{\epsilon_N,l,j}}R_n(f)\,\Pi(df)$$

from which we can see a standard result, e.g. Walker (2003), that

$$\frac{L_{n,l,j}}{L_{n-1,l,j}}=\frac{f_{n-1,l,j}(y_n|x_n)}{f_0(y_n|x_n)}$$

where $f_{n-1,l,j}(\cdot|x_n)$ is the predictive density for y_n given x_n based on a posterior restricted and normalized to the set $A_{\epsilon_N,l,j}$ and given the sample $(x_i,y_i)_{i=1}^{n-1}$. Therefore,

$$\mathrm{E}\left(\frac{L_{n,l,j}^{1/2}}{L_{n-1,l,j}^{1/2}}\,|\sigma_{n-1}\right)=1-\frac{1}{2}\,H^2(f_{n-1,l,j}(\cdot|x_n),f_0(\cdot|x_n)),$$

where $\sigma_n = \sigma((x_1, y_1), \ldots, (x_n, y_n))$.

Taking z_1 for example, if $x_n \neq z_1$ then we can bound the RHS by 1, i.e.

$$\mathrm{E}\left(\frac{L_{n,1,j}^{1/2}}{L_{n-1,1,j}^{1/2}} \,|\sigma_{n-1} \right) \leq 1.$$

On the other hand, if $x_n = z_1$, then we have

$$H(f_{n-1,1,j}(\cdot|x_n), f_0(\cdot|x_n)) \geq H(f_{1,j}(\cdot|x_n), f_0(\cdot|x_n)) - H(f_{n-1,1,j}(\cdot|x_n), f_{1,j}(\cdot|x_n))$$

where $f_{1,j}$ is the density at the center of $A_{\epsilon_{N,1,j}}$. Hence,

$$H(f_{1,j}(\cdot|x_n), f_0(\cdot|x_n)) > \epsilon_N$$

and due to the convexity of the Hellinger balls,

$$H(f_{n-1,1,j}(\cdot|x_n), f_{1,j}(\cdot|x_n)) \leq \frac{1}{2}\epsilon_N$$

since the Hellinger balls are of size $\frac{1}{2}\epsilon_N$. Thus,

$$\frac{1}{2} H^2(f_{n-1,1,j}(\cdot|x_n), f_0(\cdot|x_n)) \geq \frac{1}{4}\epsilon_N^2$$

and so for $x_n = z_1$ we have

$$\mathrm{E}\left(\frac{L_{n,1,j}^{1/2}}{L_{n-1,1,j}^{1/2}} \,|\sigma_{n-1} \right) \leq 1 - \frac{1}{4}\epsilon_N^2.$$

Putting this together with the covariates selected, we have

$$\mathrm{E}\left(L_{n,1,j}^{1/2}\right) \leq \left(1 - \frac{1}{4}\epsilon_N^2\right)^{r_n} \sqrt{\Pi(A_{\epsilon_N,1,j})}.$$

Now observe that

$$\Pi^n(A_\epsilon^c) \leq \frac{\sum_{l,j} \sqrt{L_{n,l,j}}}{\sqrt{I_n}}.$$

According to the condition appearing in Lemma 1, the denominator has a lower bound, i.e.,

$$\sqrt{I_n} \geq \exp\left\{-\frac{n}{2}\tilde{\epsilon}_n^2(C+1)\right\}$$

in probability.

For the numerator, we see that

$$\mathrm{P}\left(\sum_{l,j} L_{n,l,j}^{1/2} > e^{-n\delta_n^2} \right) \leq e^{n\delta_n^2} e^{-\frac{1}{4}n(\epsilon_N^2/N)} \sum_{l,j} \sqrt{\Pi(A_{\epsilon_N,l,j})}$$

and so if

$$e^{n\delta_n^2} e^{-\frac{1}{4}n(\epsilon_N^2/N)} \sum_{l,j} \sqrt{\Pi(A_{\epsilon_N,l,j})} \to 0$$

then

$$\sum_{l,j} L_{n,l,j}^{1/2} < e^{-n\delta_n^2}$$

in probability. We are obviously interested in the case when $\delta_n^2 = C_1\tilde{\epsilon}_n^2$ for some constant C_1, and hence we are interested in determining conditions on the prior for which

$$e^{-n[-C_1\tilde{\epsilon}_n^2+\frac{1}{4}(\epsilon_N^2/N)]}\sum_{l,j}\sqrt{\Pi(A_{\epsilon_N,l,j})} \to 0.$$

We can put this into a Theorem:

Theorem 1. *Since*

$$H(f_N(\cdot|x), f_0(\cdot|x)) > \epsilon$$

for some $x \in \mathcal{X}$ implies

$$H(f_N(\cdot|l), f_0(\cdot|l)) > \epsilon_N$$

for some $l \in \{1,\dots,N\}$, if the prior Π_N satisfies the condition of Lemma 1 and

$$e^{-n[-C_1\tilde{\epsilon}_n^2+\frac{1}{4}(\epsilon_N^2/N)]}\sum_{l,j}\sqrt{\Pi_N(A_{\epsilon_N,l,j})} \to 0,$$

then

$$\Pi^n(A_\epsilon^c) \to 0$$

in probability when the covariates are taken according to $x_i = z_j$ for $(j-1)r_n + 1 \le i \le jr_n$, for $j = 1,\dots,N$ and $r_n = n/N$.

We notice that it is desirable for $\epsilon_N = \pi\epsilon$, or specifically that π_N does not go to 0. This will be clear in the example that follows in Section 4.

The predictive based on the sequence of posterior Π^n and any new fixed x is given by

$$\widehat{f}(y|x) = \int f(y|x)\Pi^n(df).$$

The next theorem shows this specific predictive is also consistent.

Theorem 2. *If the posterior is consistent, i.e.*

$$\Pi^n(A_\epsilon^c) \to 0$$

in probability, then

$$H(\widehat{f}(\cdot|x), f_0(\cdot|x)) \to 0$$

in probability.

Proof. By Jensen's inequality and the convexity of the squared Hellinger distance we have

$$\begin{aligned}
H(\widehat{f}(\cdot|x), f_0(\cdot|x)) &= H\Big(\int f(\cdot|x)\Pi^n(df), f_0(\cdot|x)\Big) \\
&< \int H(f(\cdot|x), f_0(\cdot|x))\Pi^n(df) \\
&= \int_{A_\epsilon^c} H(f(\cdot|x), f_0(\cdot|x))\Pi^n(df) + \int_{A_\epsilon} H(f(\cdot|x), f_0(\cdot|x))\Pi^n(df) \\
&< \Pi^n(A_\epsilon^c) + \epsilon.
\end{aligned}$$

Thus the proof is completed by the fact that ϵ is arbitrary and $\Pi^n(A_\epsilon^c) \to 0$ in probability.

4. Example

This section will consider a normal example, assuming that y is normally distributed with mean function $\theta(x)$ and variance σ^2. So prior distributions will be constructed for θ and σ. We also assume that the covariate space is $\mathcal{X} = [0,1]$. Following Shively et al. (2009) we need to make assumptions on the true function $\theta_0(x)$ in order to obtain results for the sup metric. The assumption here is less restrictive than that of Shively et al. (2009).

We will also discuss here the choice of θ_N. From Sections 2 and 3 it turns out to be quite important that we have $\epsilon_N = \pi_N\epsilon$ and π_N stays away from 0. For example, using a polynomial of order N for modeling θ_N results in a π_N which goes to 0 too fast. This actually ensures there is no consistency result as we can not deal with the numerator and denominator to get the desired consistency. A form of function for θ_N that does allow $\pi_N \to 1$ is a piecewise constant function. So we define

$$\theta_N(x) = \beta_j \quad \text{for } x \in (\,(j-1)/N,\, j/N\,] \tag{1}$$

for $j = 1, \ldots, N$. For us, each β_j will be allocated an independent normal prior distribution with zero mean and variance t_j^2.

Lemma 2. Assume that θ_0 is Lipschitz continuous on $[0,1]$ such that for any $x_1, x_2 \in \mathcal{X}$,

$$|\theta_0(x_1) - \theta_0(x_2)| \le C|x_1 - x_2|,$$

for some constant $C < +\infty$. If $|\theta_N(x) - \theta_0(x)| > \epsilon$ for some $x \in [0,1]$, then for some $j \in \{1, \ldots, N\}$ it is that $|\theta_N(j/N) - \theta_0(j/N)| > \epsilon_N$, where $\epsilon_N = \epsilon - \phi/N$, and ϕ depends only on θ_0.

Proof. We have that, for $x \in ((j-1)/N, j/N]$,

$$|\theta_N(j/N) - \theta_0(j/N)| > |\theta_N(j/N) - \theta_0(x)| - |\theta_0(x) - \theta_0(j/N)|.$$

The first term on the right side is lower bounded by ϵ and the second term has an upper bound of ϕ/N due to the Lipschitz condition. This completes the proof.

4.1. Denominator

We first deal with the denominator and therefore we need to consider $K_l(\theta,\sigma) = K(f_{\theta_0,\sigma_0}(\cdot|l), f_{\theta,\sigma}(\cdot|l))$ and $V_l(\theta,\sigma) = V(f_{\theta_0,\sigma_0}(\cdot|l), f_{\theta,\sigma}(\cdot|l))$ for $l \in \{1,\ldots,N\}$. The next result introduces a way to calculate the prior mass on the Kullback-Leibeler neighborhood of f_0.

Theorem 3. *For all $\delta > 0$, there exist constants $c_1 > 0$ and $c_2 > 0$, independent of δ, such that if*

$$\sup_x |\theta(x) - \theta_0(x)| < \delta \quad \textit{and} \quad |(\sigma_0/\sigma) - 1| < \delta$$

then it is that

$$\sup_x K_x(\theta,\sigma) < c_1\delta \quad \textit{and} \quad \sup_x V_x(\theta,\sigma) < c_2\delta.$$

Proof. See Appendix.

Hence, we focus on finding a lower bound for

$$\Pi\left\{(\theta,\sigma) : \sup_{j\in\{1,\ldots,N\}} |\theta_N(j/N) - \theta_0(j/N)| < \delta,\ |(\sigma_0/\sigma) - 1| < \delta\right\}.$$

If we model θ and σ independently then we separate

$$\Pi_\theta\left\{\theta : \sup_j |\theta_N(j/N) - \theta_0(j/N)| < \delta\right\} \quad \text{and} \quad \Pi_\sigma\left\{|(\sigma_0/\sigma) - 1| < \delta\right\}.$$

For the σ probability, it is easy to show that

$$\Pi_\sigma\left\{|(\sigma_0/\sigma) - 1| < \delta\right\} \sim 2\pi_\sigma(\sigma_0)\delta\sigma_0$$

for small δ, where π_σ is the prior density for σ and we assume Π_σ put positive mass around σ_0, which is easy to achieve in practice by having the prior to be something like a gamma density, for example. To confirm this result, it is of interest to compute

$$\begin{aligned}\Pi\left\{\frac{\sigma_0}{1+\delta} < \sigma < \frac{\sigma_0}{1-\delta}\right\} &= \int_{\frac{\sigma_0}{1+\delta}}^{\frac{\sigma_0}{1-\delta}} \pi(u)du \\ &\approx 2\pi(\sigma_0)\frac{\delta\sigma_0}{(1-\delta)(1+\delta)} \\ &\approx 2\pi(\sigma_0)\delta\sigma_0,\end{aligned}$$

given the δ is very small.

For the θ probability, we note that

$$\Pi_\theta\left\{\theta : \sup_j |\theta_N(j/N) - \theta_0(j/N)| < \delta\right\} = \Pi_\beta\left\{\beta : \sup_j |\beta_j - \theta_0(j/N)| < \delta\right\}.$$

This is given by

$$\prod_{j=1}^N \Pi_{\beta_j}\left\{\beta_j : |\beta_j - \theta_0(j/N)| < \delta\right\}$$

which is given by

$$\prod_{j=1}^{N}\frac{1}{t_j\sqrt{2\pi}}\int_{\theta_0(j/N)-\delta}^{\theta_0(j/N)+\delta} e^{-0.5s^2/t_j^2}\,ds$$

which is for small δ asymptotically equivalent to

$$\left(\frac{2\delta}{\sqrt{2\pi}}\right)^N\prod_{j=1}^{N}t_j^{-1}\exp\left\{-\theta_0^2(j/N)/t_j^2\right\}.$$

Since θ_0^2 is bounded above and the (t_j) are constant, then for some $\psi<1$ it is that

$$\Pi\left\{(\theta,\sigma):\sup_{j\in\{1,\ldots,N\}}|\theta_N(j/N)-\theta_0(j/N)|<\delta,\ |(\sigma_0/\sigma)-1|<\delta\right\}>\psi^N\delta^{N+1}.$$

Therefore, to apply Lemma 1, we would be looking for $\tilde{\epsilon}_n^2$ such that

$$-N\log\tilde{\epsilon}_n^2-N\log\psi<n\tilde{\epsilon}_n^2$$

which is satisfied for

$$\tilde{\epsilon}_n^2=(N/n)^{1-\alpha}$$

for any $\alpha>0$.

4.2. Numerator

We now look at the numerator. First, we establish that

$$\frac{1}{2}H^2(f(\cdot|x),f_0(\cdot|x))=1-\sqrt{\tau(\sigma,\sigma_0)}\exp\left\{-\frac{1}{4}\frac{(\theta(x)-\theta_0(x))^2}{\sigma^2+\sigma_0^2}\right\},$$

where

$$\tau(\sigma,\sigma_0)=\frac{2\sigma\sigma_0}{\sigma^2+\sigma_0^2}.$$

So, if $H(f(\cdot|x),f_0(\cdot|x))>\epsilon^*$ for some ϵ^*, which depends on ϵ, then either

$$|(\sigma_0/\sigma)-1|>\epsilon\quad\text{or}\quad|\theta_0(x)-\theta(x)|>\epsilon.$$

Thus

$$A_\epsilon\subset\bigcup_{l=1}^{N}\{\theta:|\theta(l)-\theta_0(l)|>\epsilon_N\}\bigcup\{\sigma:|(\sigma_0/\sigma)-1|>\epsilon\}.$$

The final term of the union can be dealt with straightforwardly; it is quite easy to show that

$$\int_{B_\epsilon}R_n(f)\,\Pi(df)<e^{-nc_\epsilon}$$

a.s. for all large n for some $c_\epsilon>0$, where $B_\epsilon=\{\sigma:|(\sigma_0/\sigma)-1|>\epsilon\}$. The union of the remaining terms can be dealt with using Theorem 1. So, since we have

$$\tilde{\epsilon}_n^2=(N/n)^{1-\alpha}$$

for any $\alpha > 0$, we need to determine N_n such $N_n \tilde{\epsilon}_n^2 \to 0$. This happens when we take N so that

$$N^{2-\alpha}/n^{1-\alpha} \to 0.$$

Therefore, $N = n^{\frac{1}{2}-\alpha}$ for any $\alpha > 0$ is sufficient. Now we are only left with showing that

$$\sum_{l,j} \sqrt{\Pi(A_{\epsilon_N,l,j})} < M_N$$

where $M_N < \infty$ and that $M_{N_n} e^{-\epsilon n/N_n} \to 0$.

Theorem 4. *The sum of square rooted prior mass on Hellinger covering balls is bounded by*

$$\sum_{l,j} \sqrt{\Pi(A_{\epsilon_N,l,j})} < M_N$$

and

$$e^{-n/N} M_N \to 0.$$

Proof. See Appendix.

In summary, we are showing that the numerator is bounded above by

$$e^{-n\epsilon/N} M_N$$

for some constant $\epsilon > 0$, where $N_n = n^{\frac{1}{2}-\alpha}$ for any $\alpha > 0$, and $M_N e^{-\epsilon(n/N)} \to 0$. On the other hand, the denominator is bounded below by $e^{-n(N/n)^{1-\alpha}}$ for any $\alpha > 0$. Putting these together yields the desired consistency.

5. Discussion

In this paper we have concentrated on establishing consistency for Bayesian regression models with non–stochastic covariates and with respect to a sup Hellinger metric. The example has illuminated the strategy of constructing the prior based on the sample size. We would advocate such a procedure. Bayesian infinite mixture models with each mixture representing a nonparametric regression model would appear to be too unidentifiable and with a finite data set are clearly over–parameterized. Allowing the model to increase with the data makes perfect sense. In the mean regression function we advocate a piecewise constant model for the function with effectively $\sqrt{n}$ pieces.

We would next extend the theory to mixture models and it is our anticipation that we would work with N_n mixtures for a sample of size n where N_n would need to be determined to present the sup Hellinger consistency. Work on this is ongoing.

We are not pursuing the standard Bayesian consistency ideas for regression for a number of reasons. The theory, once the covariates are assumed to be i.i.d., is just that for i.i.d. data and the toughness is due to verifying the conditions for the i.i.d. case which now are needed to be worked out in higher dimensions. Additionally covariates are rarely i.i.d. and are often designed. Our results give an indication of how to take covariates to ensure consistency. Finally, we are able to establish consistency for the predictive at a particular choice of covariate value, not possible if they are assumed to be i.i.d.

6. Appendix

Proof of lemma 1. The $f(\cdot|z_l)$ is a density indexed by z_l and $(x_i)_{i=1}^n$ take values from $(z_1,\ldots,z_N)$. Therefore, $K_l(f) < \tilde{\epsilon}_n^2$ and $V_l(f) < \tilde{\epsilon}_n^2$ for $l = 1, 2, ...$ are equivalent to

$$K_i(f) = K(f_0(.|x_i), f(.|x_i)) < \tilde{\epsilon}_n^2, \quad \forall i = 1,\ldots,n$$

and

$$V_i(f) = V(f_0(.|x_i), f(.|x_i)) < \tilde{\epsilon}_n^2, \quad \forall i = 1,\ldots,n.$$

Hence the condition on the prior is equivalent to

$$\Pi\left\{f : K_i(f) < \tilde{\epsilon}_n^2 \;\; V_i(f) < \tilde{\epsilon}_n^2 \;\; \forall i = 1,\ldots,n\right\} \geq \exp(-n\tilde{\epsilon}_n^2 C)$$

for all large n.

The proof follows similarly to Ghosal et al. (2000). By Jensen's inequality,

$$\log \int \prod_{i=1}^n \frac{f}{f_0}(y_i|x_i)\Pi(df) \geq \sum_{i=1}^n \int \log \frac{f}{f_0}(y_i|x_i)\Pi(df).$$

Thus

$$\begin{aligned}
& \mathrm{P}\left(\int \prod_{i=1}^n \frac{f}{f_0}(y_i|x_i)\Pi(df) \leq \exp(-n\tilde{\epsilon}_n^2(C+1))\right) \\
= \; & \mathrm{P}\left(\log \int \prod_{i=1}^n \frac{f}{f_0}(y_i|x_i)\Pi(df) \leq -n\tilde{\epsilon}_n^2(C+1)\right) \\
\leq \; & \mathrm{P}(\sum_{i=1}^n \int \log \frac{f}{f_0}(y_i|x_i)\Pi(df) \leq -n\tilde{\epsilon}_n^2(C+1)) \\
= \; & \mathrm{P}\left(\frac{\sqrt{n}}{n}\sum_{i=1}^n \int \log \frac{f}{f_0}(y_i|x_i)\Pi(df) \leq -\sqrt{n}\tilde{\epsilon}_n^2(C+1)\right) \\
= \; & \mathrm{P}\left(\frac{\sqrt{n}}{n}\sum_{i=1}^n \int \log \frac{f}{f_0}(y_i|x_i)\Pi(df) - P \leq -\sqrt{n}\tilde{\epsilon}_n^2(C+1) - P\right)
\end{aligned}$$

where

$$P = \frac{\sqrt{n}}{n}\sum_{i=1}^n \int f_0(y_i|x_i) \int \log \frac{f}{f_0}(y_i|x_i)\Pi(df)\, dy_i$$

so $P > -\sqrt{n}\tilde{\epsilon}_n^2$.

Therefore,

$$\begin{aligned}
& \mathrm{P}\left(\int \prod_{i=1}^n \frac{f}{f_0}(y_i|x_i)\Pi(df) \leq \exp(-n\tilde{\epsilon}_n^2(C+1))\right) \\
\leq \; & \mathrm{P}\left(\frac{\sqrt{n}}{n}\sum_{i=1}^n \int \log \frac{f}{f_0}(y_i|x_i)\Pi(df) - P \leq -\sqrt{n}\tilde{\epsilon}_n^2 C\right).
\end{aligned}$$

Since

$$\mathrm{E}\left(\int \log \frac{f}{f_0}(y_i|x_i)\Pi(df)\right) = \int f_0(y_i|x_i) \int \log \frac{f}{f_0}(y_i|x_i)\Pi(df)\, dy_i,$$

applying Chebyshev's inequality, the previous probability is bounded above by

$$\begin{aligned}
& \frac{\mathrm{Var}[\frac{\sqrt{n}}{n}\sum_{i=1}^n \int \log \frac{f}{f_0}(y_i|x_i)\Pi(df)]}{n\tilde{\epsilon}_n^4 C^2} \\
= & \frac{\frac{1}{n}\sum_{i=1}^n \mathrm{Var}[\int \log \frac{f}{f_0}(y_i|x_i)\Pi(df)]}{n\tilde{\epsilon}_n^4 C^2} \\
\leq & \frac{\frac{1}{n}\sum_{i=1}^n \int f_0[\int \log \frac{f}{f_0}(y_i|x_i)\Pi(df)]^2 dy_i}{n\tilde{\epsilon}_n^4 C^2} \\
\leq & \frac{\frac{1}{n}\sum_{i=1}^n \int f_0 \int [\log \frac{f}{f_0}(y_i|x_i)]^2 \Pi(df) dy_i}{n\tilde{\epsilon}_n^4 C^2} \quad \text{(by Jensen's inequality)} \\
\leq & \frac{\tilde{\epsilon}_n^2}{n\tilde{\epsilon}_n^4 C^2} \\
= & \frac{1}{n\tilde{\epsilon}_n^2 C^2} \to 0.
\end{aligned}$$

Therefore,

$$\int \prod_{i=1}^n \frac{f}{f_0}(y_i|x_i)\Pi(df) \geq \exp(-n\tilde{\epsilon}_n^2(C+1)) \text{ in probability}$$

completing the proof.

Proof of Theorem 3. If

$$\sup_x |\theta(x) - \theta_0(x)| < \delta, \quad \text{and} \quad \left|\frac{\sigma_0}{\sigma} - 1\right| < \delta,$$

then

$$\sup_x |\theta(x) - \theta_0(x)|^2 < \delta^2 \quad \text{and} \quad \left|\frac{\sigma^2}{\sigma_0^2} - 1\right| < \frac{2\delta - \delta^2}{(1-\delta)^2}.$$

Therefore, let $\psi = 2\delta/(1-\delta)^2 > (2\delta - \delta^2)/(1-\delta)^2$, so it is also true that

$$\sup_x(\theta(x) - \theta_0(x))^2 < \psi \quad \text{and} \quad \left|\frac{\sigma^2}{\sigma_0^2} - 1\right| < \psi.$$

Now

$$\begin{aligned}
K_x(\theta, \sigma) &= \frac{1}{2}\log\frac{\sigma^2}{\sigma_0^2} - \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right) + \frac{1}{2}\frac{[\theta(x) - \theta_0(x)]^2}{\sigma^2} \\
&< \frac{1}{2}\psi\left(\frac{1}{1-\psi} + \frac{1}{1-\psi} + \frac{1}{(1-\psi)\sigma_0^2}\right).
\end{aligned}$$

Hence

$$K_x(\theta, \sigma) < \left(2 + \sigma_0^{-2}\right)\frac{\psi}{1-\psi}.$$

In addition,

$$\begin{aligned}
V_x(\theta,\sigma) &\leq 2\left[-\frac{1}{2}+\frac{1}{2}\frac{\sigma_0^2}{\sigma^2}\right]^2+\left[\frac{\sigma_0^2}{\sigma^2}\{\theta(x)-\theta_0(x)\}\right]^2 \\
&< \frac{3}{2}\frac{\psi^2}{(1-\psi)^2}.
\end{aligned}$$

Since

$$\frac{\psi}{1-\psi}=\frac{2\delta}{1-4\delta+\delta^2}<4\delta,$$

for $0<\delta<\frac{1}{8}$, the proof can be completed by choosing $c_1=4(2+\sigma_0^{-2})$ and $c_2=24$.

Proof of Theorem 4. The covering Hellinger balls of the function space is constructed as follows and is based on the squared Hellinger distance between two parameter $\eta_1=(\theta_1(x),\sigma_1)$ and $\eta_2=(\theta_2(x),\sigma_2)$ being

$$H^2[f_{\eta_1}(\cdot|x),f_{\eta_2}(\cdot|x)]=2-2\exp\left\{-\frac{[\theta_1(x)-\theta_2(x)]^2}{4(\sigma_1^2+\sigma_2^2)}\right\}\sqrt{\frac{2}{\frac{\sigma_1}{\sigma_2}+\frac{\sigma_2}{\sigma_1}}}.$$

Then $H[f_{\eta_1},f_{\eta_2}]<\delta$ is equivalent to

$$\sup_x|\theta_1(x)-\theta_2(x)|<\delta^*,$$

given the information of the variance. Recall the $\theta(x)$ is piecewise constant β_k on $((k-1)/N,k/N]$. Thus the partition of θ space is the partition of β space given σ. That is,

$$A_{km_k}=\{\beta_k: m_k\epsilon\sigma^2<\beta_k<(m_k+1)\epsilon\sigma^2\},$$

where the m_k are integers from $(-\infty,+\infty)$.

The Hellinger distance between f_{η_1} and f_{η_2} bounded by δ, with respect to the variance, is equivalent to $|\sigma_1/\sigma_2|<\delta_2^*$. Thus the partition of the half real line for σ is

$$A_{m^*}=\{\sigma: m^*\delta_2^*<\log\sigma<(m^*+1)\delta_2^*\},$$

where m^* is also integer from $(-\infty,+\infty)$. Hence, the probability of interest is given by

$$\begin{aligned}
&\sum_j\sqrt{\Pi(A_{\epsilon_N,l,j})} \\
= &\sum_{m^*=-\infty}^{\infty}\sqrt{\mathrm{P}[\log\sigma\in\{m^*\delta_2^*,(m^*+1)\delta_2^*\}]}\times\left\{2\prod_{k=1}^{N}\sum_{m=0}^{\infty}\sqrt{\mathrm{P}[\beta_k\in\{m_k\epsilon\sigma^2,(m_k+1)\epsilon\sigma^2|\sigma\}]}\right\} \\
< &\ 4\sum_{m^*=0}^{\infty}\left\{\sqrt{\mathrm{P}[\log\sigma\in\{m^*\delta_2^*,(m^*+1)\delta_2^*\}]}\prod_{k=1}^{N}[1+4(2\pi)^{-\frac{1}{4}}(\frac{t_k}{\epsilon\sigma^2})^{3/2}]\right\}.
\end{aligned}$$

The calculation is similar to that of the infinite-dimensional exponential families of Walker (2004). Put $t_k = k^{-s}$ with $s > 1$, we have, for some constant $\lambda = 4(2\pi)^{-\frac{1}{4}}/\epsilon > 0$

$$\begin{aligned}
&\sum_j \sqrt{\Pi(A_{\epsilon_N,l,j})} \\
< \quad & 4\sum_{m^*=0}^{\infty}\left\{\sqrt{\mathrm{P}[\log\sigma \in \{m^*\delta_2^*, (m^*+1)\delta_2^*\}]}(1+\lambda e^{-3m^*\delta_2^*})^N\right\} \\
< \quad & 4\sum_{m^*=0}^{\infty}\left\{\sqrt{\frac{\delta_2^*}{\sqrt{2\pi}t}}\exp\left(-\frac{{m^*}^2{\delta_2^*}^2}{4t^2}\right)(1+\lambda)^N\right\}.
\end{aligned}$$

Thus, for some constant Q

$$\sum_{l,j}\sqrt{\Pi(A_{\epsilon_N,l,j})} < QN(1+\lambda)^N.$$

That is, we can take $M_N = N\exp(N\tau)$ for some $\tau = \log(1+\lambda) > 0$.

Now $e^{-\epsilon n/N}$ is $\exp(-\epsilon n^{\frac{1}{2}+\alpha})$ and $M_n = N\exp(\tau n^{\frac{1}{2}-\alpha})$, and so

$$e^{-n\epsilon/N}M_N \to 0$$

as required.

Acknowledgements. The authors would like to thank two anonymous referees for their comments and suggestions which have allowed to improve the paper.

References

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., Ramamoorthi, R. V., 2003. Posterior consistency for semi-parametric regression problems. Bernoulli 9, 291–312.

Barron, A., Schervish, M. J., Wasserman, L., 1999. The consistency of posterior distributions in nonparametric problems. Ann. Statistics 27, 536–561.

Choi, T., 2005. Posterior consistency in nonparametric regression problems under gaussian process priors. Ph.D. thesis, Carnegie Mellon Univ., Pittsburgh, PA.

Choi, T., 2007. Alternative posterior consistency results in nonparametric binary regression using gaussian process priors. J. Statist. Plann. Inference 137, 2975–2983.

Choi, T., Ramamoorthi, R. V., 2008. Remarks on consistency of posterior distributions. IMS Collection 3, 170–186.

Choi, T., Schervish, M. J., 2007. On posterior consistency in nonparametric regression problems. J. Multivariate Anal 98, 1969–1987.

Ghosal, S., Ghosh, J. K., Ramamoorthi, R. V., 1999. Posterior consistency of dirichlet mixtures in density estimation. Ann. Statistics 27, 143–158.

Ghosal, S., Ghosh, J. K., Van Der Vaart, A. W., 2000. Convergence rates of posterior distributions. Ann. Statistics 28, 500–531.

Ghosal, S., Roy, A., 2006. Posterior consistency of gaussian process prior for nonparametric binary regression. Ann. Statistics 34, 2413–2429.

Shively, T. S., Sager, T. W., Walker, S. G., 2009. A bayesian approach to non-paramtric monotone function estimation. J. R. Stat. Soc. Ser. B Stat. Methodol. 71.

Walker, S. G., 2003. On sufficient conditions for bayesian consistency. Biometrika 90, 482–488.

Walker, S. G., 2004. New approaches to bayesian consistency. Ann. Statistics 32, 2028–2043.