Bernoulli **9**(2), 2003, 291–312

Posterior consistency for semi-parametric regression problems

MESSAN AMEWOU-ATISSO[1*], SUBHASHIS GHOSAL[2], JAYANTA K. GHOSH[1**] and R.V. RAMAMOORTHI[3]

[1]*Department of Statistics, Purdue University, 1399 Mathematical Sciences Building, West Lafayette IN 47907, USA. E-mail:* *aamessan@stat.purdue.edu; **ghosh@stat.purdue.edu
[2]*Department of Statistics, North Carolina State University, 220 Patterson Hall, 2501 Founders Drive, Raleigh NC 27695-8203, USA. E-mail: ghosal@stat.ncsu.edu*
[3]*Department of Statistics and Probability, Michigan State University, Wells Hall, East Lansing MI 48824, USA. E-mail: ramamoorthi@stt.msu.edu*

We consider Bayesian inference in the linear regression problem with an unknown error distribution that is symmetric about zero. We show that if the prior for the error distribution assigns positive probabilities to a certain type of neighbourhood of the true distribution, then the posterior distribution is consistent in the weak topology. In particular, this implies that the posterior distribution of the regression parameters is consistent in the Euclidean metric. The result follows from our generalization of a celebrated result of Schwartz to the independent, non-identical case and the existence of exponentially consistent tests of the complement of the neighbourhoods shown here. We then specialize to two important prior distributions, the Polya tree and Dirichlet mixtures, and show that under appropriate conditions these priors satisfy the positivity requirement of the prior probabilities of the neighbourhoods of the true density. We consider the case of both non-stochastic and stochastic regressors. A similar problem of Bayesian inference in a generalized linear model for binary responses with an unknown link is also considered.

Keywords: consistency; Dirichlet mixtures; exponentially consistent test; Kullback–Leibler number; linear regression; Polya tree; posterior distribution

1. Introduction

This paper addresses the consistency of the posterior in regression problems when the unknown distribution of the error variable is endowed with a nonparametric prior. Thus our observations are $Y_1, Y_2, \ldots,$ where

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, 2, \ldots; \tag{1.1}$$

here the errors ϵ_i are independent and identically distributed (i.i.d.) f, with f a density symmetric around 0, and $x_1, x_2, \ldots$ are the values of the covariate X. These may arise as fixed non-random constants or as i.i.d. observations of a random variable X with a known or unknown distribution.

The unknown parameters are f, α, β and formally the parameter space is

1350–7265

$\Theta = \mathcal{F} \times \mathbb{R} \times \mathbb{R}$, where $\mathcal{F}$ is the set of all symmetric densities on $\mathbb{R}$. We start with a prior $\tilde{\Pi}$ for f and, independent of f, a prior μ for (α, β). Let Π stand for the prior $\tilde{\Pi} \times \mu$.

Fix (f_0, α_0, β_0) in Θ. The sequence of posteriors $\Pi(\cdot|Y_1, Y_2, \ldots, Y_n)$ is said to be consistent for (f, α, β) at (f_0, α_0, β_0) if $\Pi(\mathcal{U}|Y_1, \ldots, Y_n)$ converges to 1 almost surely as $n \to \infty$ for any neighbourhood $\mathcal{U}$ of (f_0, α_0, β_0), when the distribution governing $Y_1, Y_2, \ldots$ has the 'true' parameter (f_0, α_0, β_0). An exactly similar definition holds if we want posterior consistency only for the parametric part (α, β) at (f_0, α_0, β_0). It will turn out that the sufficient condition for the latter is weaker than that for the posterior consistency of (f, α, β).

The idea of posterior consistency is due to Freedman (1963), though, in a sense, it goes back to Bayes, Laplace and Von Mises. The relevance of posterior consistency to Bayesians is explained well in Diaconis and Freedman (1986a). Diaconis and Freedman (1986a; 1986b) also provide an example of inconsistency, in a relatively simple setting, for location models with symmetric error distributions. A similar example of inconsistency for the location problem with error distribution having median 0 is given by Doss (1985a; 1985b). The problem of interest then is to identify all or at least a large class of parameter values where consistency obtains. In this paper, although we approach the problem in some generality, it is geared to handling two classes of popular priors on densities – the Polya tree priors and Dirichlet mixtures of a normal kernel.

Recent reviews focusing on general issues of consistency are Ghosal *et al*. (1999a), Ghosh (1998) and Wasserman (1998). In Ghosal *et al*. (1999a; 1999b) and Ghosh (1998) it is argued that a theorem of Schwartz (1965) is the right tool for studying consistency in semi-parametric problems. The same is true of the present paper. However, since the observations are independent but not identically distributed, major changes are needed. We begin with a variant of Schwartz's theorem for independent, non-identically distributed variables. This is discussed in Section 2, while in Sections 3 and 4, we discuss how one can verify the two conditions of this theorem. The lack of i.i.d. structure for the Y_i necessitates assumptions on the x_i to ensure that the exponentially consistent tests required by Schwartz's theorem exist in the present context. Also certain conditions on f_0 are required to verify a condition analogous to Schwartz's on the support of the prior. In Section 4, we relate the properties of the prior on $\mathcal{F}$ to that on the regression parameters and obtain a theorem on consistency. We show in the next section that Polya tree priors of the sort considered in Ghosal *et al*. (1999b) fulfil the requirements. We then turn to Dirichlet mixtures of normal kernel priors. The posterior consistency of these in the context of density estimation was studied in Ghosal *et al*. (1999c). In Section 6 we explore similar problems in the regression setting. In Section 7 we discuss a similar problem of generalized linear models with binary responses and an unknown link function. This may be viewed as a nonparametric generalization of the logistic regression model. A Dirichlet process prior is put on the link distribution function and the consistency of the posterior is briefly discussed. Section 8 indicates the modifications necessary to handle the case of a stochastic regressor.

Although we prove consistency when the covariates are one-dimensional, the arguments easily generalize to more than one dimension. For that we will only need to modify Proposition 3.1 by looking at quadrants under the appropriate modification of Assumption A.

Nonparametric and semi-parametric Bayesian methods are now being used more and more. In view of the example of Diaconis and Freedman (1986a; 1986b), it seems appropriate to see if some validation can be provided through posterior consistency. It will be also interesting to study the rate of convergence of the posterior distribution, as is done in Ghosal *et al*. (2000). In particular, it is of substantial interest to see whether the posterior distribution for the parametric part converges at the classical $\sqrt{n}$ rate. We have not attempted to answer this question here, and will return to it elsewhere.

2. Consistency of posterior

Fix f_0, α_0, β_0. For a density f, let

$$f_{\alpha,\beta,i} = f_{\alpha+\beta x_i}(y) = f(y - (\alpha + \beta x_i)) \tag{2.1}$$

and put $f_{0i} = f_{0,\alpha_0,\beta_0,i}$. For any two densities f and g, let

$$K(f, g) = \int f \log \frac{f}{g}, \qquad V(f, g) = \int f \left(\log_+ \frac{f}{g} \right)^2, \tag{2.2}$$

where $\log_+ x = \max(\log x, 0)$, and put

$$K_i(f, \alpha, \beta) = K(f_{0i}, f_{\alpha,\beta,i}), \qquad V_i(f, \alpha, \beta) = V(f_{0i}, f_{\alpha,\beta,i}). \tag{2.3}$$

As mentioned in the Introduction, the main tool we use is a variant of Schwartz's (1965) theorem. The following theorem is an adaptation to the case when the Y_i are independent but not identically distributed. Here the x_i are non-random. We start with the definition of exponentially consistent tests.

Definition 2.1. *Let $\mathcal{W} \subset \mathcal{F} \times \mathbb{R} \times \mathbb{R}$. A sequence of test functions $\Phi_n(Y_1, \ldots, Y_n)$ is said to be* exponentially consistent *for testing*

$$H_0 : (f, \alpha, \beta) = (f_0, \alpha_0, \beta_0) \quad \textit{against} \quad H_1 : (f, \alpha, \beta) \in \mathcal{W} \tag{2.4}$$

if there exist constants C_1, C_2, $C > 0$ such that

(a) $\mathrm{E}_{\Pi_1^n f_{0i}} \Phi_n \leqslant C_1 \mathrm{e}^{-nC}$
(b) $\inf\limits_{(f,\alpha,\beta)\in\mathcal{W}} \mathrm{E}_{\Pi_1^n f_{\alpha,\beta,i}}(\Phi_n) \geqslant 1 - C_2 \mathrm{e}^{-nC}$

Theorem 2.1. *Suppose $\tilde{\Pi}$ is a prior on $\mathcal{F}$ and μ is a prior for (α, β). Let $\mathcal{W} \subset \mathcal{F} \times \mathbb{R} \times \mathbb{R}$. If*

(i) there is an exponentially consistent sequence of tests for

$$H_0 : (f, \alpha, \beta) = (f_0, \alpha_0, \beta_0) \quad \textit{against} \quad H_1 : (f, \alpha, \beta) \in \mathcal{W},$$

(ii) *and for all $\delta > 0$,*

$$\Pi\left\{ (f, \alpha, \beta) : K_i(f, \alpha, \beta) < \delta \textit{ for all } i, \quad \sum_{i=1}^{\infty} \frac{V_i(f, \alpha, \beta)}{i^2} < \infty \right\} > 0,$$

then with $(\prod_{i=1}^{\infty} P_{f_{0i}})$*-probability* 1, *the posterior probability*

$$\Pi(\mathcal{W}|Y_1, \ldots, Y_n) = \frac{\int_{\mathcal{W}} \prod_{i=1}^{n} (f_{\alpha,\beta i}(Y_i)/f_{0i}(Y_i)) \mathrm{d}\Pi(f, \alpha, \beta)}{\int_{\mathcal{F}\times\mathbb{R}\times\mathbb{R}} \prod_{i=1}^{n} (f_{\alpha,\beta i}(Y_i)/f_{0i}(Y_i)) \mathrm{d}\Pi(f, \alpha, \beta)} \to 0. \tag{2.5}$$

Note that $V_i(f, \alpha, \beta)$ bounded above in i is sufficient to ensure the summability of $\sum_{i=1}^{\infty} V_i(f, \alpha, \beta)/i^2$.

The proof of the theorem is similar to that of Schwartz (1965). If we write (2.5) as

$$\Pi(\mathcal{W}|Y_1, \ldots, Y_n) = \frac{I_{1n}(Y_1, \ldots, Y_n)}{I_{2n}(Y_1, \ldots, Y_n)}, \tag{2.6}$$

the proof involves showing, as is done in Schwartz (1965), that condition (i) implies that there exists a $d > 0$ such that $\mathrm{e}^{nd} I_{1n}(Y_1, \ldots, Y_n) \to 0$ a.s., and that condition (ii) implies that for all $d > 0$, $\mathrm{e}^{nd} I_{2n}(Y_1, \ldots, Y_n) \to \infty$ a.s. A sketch of the details is given in the appendix.

It should be noted here that the theorem could have been stated in much more generality, for any semi-parametric problem. Consistency of the posterior holds as long as there is an exponentially consistent test for testing the point null against the complement of the required neighbourhood and (ii) holds. In Section 7 we apply this idea to a binary response regression model with an unknown link.

3. Exponentially consistent tests

Our goal is to establish consistency of the posterior distribution for (f, α, β) or for (α, β) at (f_0, α_0, β_0), and thus the set $\mathcal{W}$ of interest to us is of the type $\mathcal{W} = \mathcal{U}^c$, where $\mathcal{U}$ is a neighbourhood of (f_0, α_0, β_0). In this section we write $\mathcal{W}$ of this type as a finite union of $\mathcal{W}_i$s and show that condition (i) of Theorem 2.1 holds for each of these $\mathcal{W}_i$s. Note that condition (i) does not involve the prior.

We begin with a couple of lemmas.

Lemma 3.1. *For* $i = 1, 2, \ldots$ *let* g_{0i} *and* g_i *be densities on* $\mathbb{R}$. *If for each* i *there exists a function* Φ_i, $0 \leqslant \Phi_i \leqslant 1$, *such that*

$$\mathrm{E}_{g_{0i}}(\Phi_i) = \alpha_i \leqslant \gamma_i = \mathrm{E}_{g_i}(\Phi_i), \tag{3.1}$$

and if

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\gamma_i - \alpha_i) > 0, \tag{3.2}$$

then there exist a constant C, *sets* $B_n \subset \mathbb{R}^n$, $n = 1, 2, \ldots$, *and* n_0 – *all depending only on* (γ_i, α_i) – *such that, for* $n > n_0$,

$$\left[\prod_{i=1}^{n} P_{g_{0i}}\right](B_n) < \mathrm{e}^{-nC}$$

$$\left[\prod_{i=1}^{n} P_{g_i}\right](B_n) > 1 - \mathrm{e}^{-nC}.$$

Proof. Set $B_n = \{\sum_{i=1}^{n}\Phi_i > \sum_{i=1}^{n}(\gamma_i + \alpha_i)/2\}$. Then by Hoeffding's inequality (Dudley 1999, p. 14)

$$\begin{aligned}\left[\prod_{i=1}^{n} P_{g_{0i}}\right](B_n) &\leqslant \left[\prod_{i=1}^{n} P_{g_{0i}}\right]\left\{\sum_{i=1}^{n}(\Phi_i - E_{g_{0i}}(\Phi_i)) > \sum_{i=1}^{n}(\gamma_i - \alpha_i)\right\} \\ &\leqslant \exp\left[-\frac{1}{2n}\left(\sum_{i=1}^{n}(\gamma_i - \alpha_i)\right)^2\right]. \end{aligned} \tag{3.3}$$

On the other hand, applying Hoeffding's inequality to $0 \leqslant 1 - \Phi_i \leqslant 1$,

$$\begin{aligned}\left[\prod_{i=1}^{n} P_{g_i}\right](B_n^c) &\leqslant \left[\prod_{i=1}^{n} P_{g_i}\right]\left\{\sum_{i=1}^{n}((1-\Phi_i) - (1 - \mathrm{E}_{g_i}\Phi)) \leqslant \sum_{i=1}^{n}(\gamma_i - \alpha_i)/2\right\} \\ &\leqslant \exp\left[-\frac{1}{2n}\left(\sum_{i=1}^{n}(\gamma_i - \alpha_i)\right)^2\right].\end{aligned}$$

Taking $C = \frac{1}{4}\liminf_{n\to\infty}((1/n)\sum_{i=1}^{n}(\gamma_i - \alpha_i))^2$, the result follows. □

For a density g and $\theta \in \mathbb{R}$, let g_θ stand for the density $g_\theta(y) = g(y - \theta)$.

Lemma 3.2 *Let g_0 be a continuous symmetric density on $\mathbb{R}$, with $g_0(0) > 0$. Let η be such that* $\inf_{|y|<\eta} g_0(y) = C > 0$.

(i) *For any $\Delta > 0$, there exists a set B_Δ such that*

$$P_{g_0}(B_\Delta) \leqslant \tfrac{1}{2} - C(\Delta \wedge \eta)$$

and, for any symmetric density g,

$$P_{g_\theta}(B_\Delta) \geqslant \tfrac{1}{2}, \qquad \textit{for all } \theta \geqslant \Delta.$$

(ii) *For any $\Delta < 0$, there exists a set $\tilde{B}_\Delta$ such that*

$$P_{g_0}(\tilde{B}_\Delta) \leqslant \tfrac{1}{2} - C(\Delta \wedge \eta)$$

and, for any symmetric density g,

$$P_{g_\theta}(\tilde{B}_\Delta) \geqslant \tfrac{1}{2}, \qquad \text{for all } \theta \leqslant \Delta.$$

Proof. (i) Take $B_\Delta = (\Delta, \infty)$. Since $\theta \geqslant \Delta$ and g_θ is symmetric around θ, $P_{g_\theta}(B_\Delta) \geqslant \frac{1}{2}$.

On the other hand,

$$P_{g_0}(B_\Delta) = \frac{1}{2} - \int_0^\Delta g_0(y)\mathrm{d}y \leqslant \frac{1}{2} - \int_0^{\Delta\wedge\eta} g_0(y)\mathrm{d}y \leqslant \frac{1}{2} - C(\Delta \wedge \eta). \tag{3.5}$$

Similarly, $\tilde{B}_\Delta = (-\infty, \Delta)$ would satisfy (ii). □

Remark 3.1. By considering $I_{B_\Delta}(y - \theta_0)$, it is easy to see that Lemma 3.2 holds if we replace g_0 by g_{0,θ_0} and require $\theta - \theta_0 > \Delta$ or $\theta - \theta_0 < \Delta$.

We return to the regression model.

Assumption A. *There exists* $\varepsilon_0 > 0$ *such that the covariate values* x_i *satisfy*

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n I\{x_i < -\varepsilon_0\} > 0, \qquad \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n I\{x_i > \varepsilon_0\} > 0.$$

Remark 3.2. Assumption A forces the covariate x to take both positive and negative values, that is, values on both sides of 0. However, the point 0 is not special. If the condition is satisfied around any point, then by a simple location shift we can bring that to the present case.

Proposition 3.1. *If Assumption A holds,* f_0 *is continuous at* 0 *and* $f_0(0) > 0$*, then there is an exponentially consistent sequence of tests for*

$$H_0 : (f, \alpha, \beta) = (f_0, \alpha_0, \beta_0) \quad \text{against} \quad H_1 : (f, \alpha, \beta) \in \mathcal{W}$$

in each of the following cases:

(i) $\mathcal{W} = \{(f, \alpha, \beta) : \alpha > \alpha_0, \beta - \beta_0 > \Delta\}$,
(ii) $\mathcal{W} = \{(f, \alpha, \beta) : \alpha < \alpha_0, \beta - \beta_0 > \Delta\}$,
(iii) $\mathcal{W} = \{(f, \alpha, \beta) : \alpha > \alpha_0, \beta - \beta_0 < -\Delta\}$,
(iv) $\mathcal{W} = \{(f, \alpha, \beta) : \alpha < \alpha_0, \beta - \beta_0 < -\Delta\}$.

Proof. (i) Let $K_n = \{i : 1 \leqslant i \leqslant n, x_i > \varepsilon_0\}$ and $\#K_n$ stand for the cardinality of K_n. We will construct a test using only those Y_i for which the corresponding i is in K_n.

If $i \in K_n$, then $(\alpha + \beta x_i) - (\alpha_0 + \beta_0 x_i) > \Delta x_i$, and by Lemma 3.2, for each $i \in K_n$, there exists a set A_i such that

$$\alpha_i := P_{f_{0i}}(A_i) < \tfrac{1}{2} - C(\eta \wedge \Delta x_i)$$

and

$$\gamma_i := \inf_{(f,\alpha,\beta)\in\mathcal{W}} P_{f_{\alpha,\beta,i}}(A_i) \geqslant \tfrac{1}{2},$$

where ':=' denotes equality by definition.

If $i \leqslant n$ and $i \notin K_n$, set $A_i = \mathbb{R}$, so that $\alpha_i = \gamma_i = 1$. Thus

$$\begin{aligned}\liminf_{n\to\infty}\left(n^{-1}\sum_{i=1}^{n}(\gamma_i-\alpha_i)\right) &\geqslant \liminf_{n\to\infty}\left(n^{-1}\sum_{i\in K_n}C(\eta\wedge\Delta x_i)\right)\\ &\geqslant C(\eta\wedge\Delta\varepsilon_0)\liminf_{n\to\infty}\#K_n/n>0. \end{aligned} \tag{3.6}$$

With $\Phi_i = I_{A_i}$, the result follows from Lemma 3.1.

(ii) In this case we construct tests using Y_i such that $i \in M_n := \{1 \leqslant i \leqslant n : x_i < -\varepsilon_0\}$. If $i \in M_n$, then

$$(\alpha+\beta x_i)-(\alpha_0+\beta_0 x_i) < \Delta x_i < -\Delta\varepsilon_0.$$

Now using (ii) of Lemma 3.2, we obtain sets $\tilde{B}_i$ and then obtain exponentially consistent tests using Lemma 3.1 as in part (i).

The other two cases follow similarly. □

The union of the $\mathcal{W}$s in Proposition 3.1 is the set $\{(f, \alpha, \beta) : |\beta-\beta_0| > \Delta\}$. The next proposition takes care of $\{(f, \alpha, \beta) : |\alpha-\alpha_0| > \Delta\}$. The proof is along the same lines and is omitted.

Proposition 3.2. *Under the assumptions of Proposition* 3.1, *there exists an exponentially consistent sequence of tests for testing*

$$H_0 : (f, \alpha, \beta) = (f_0, \alpha_0, \beta_0) \quad \textit{against} \quad H_1 : (f, \alpha, \beta) \in \mathcal{W}$$

when $\mathcal{W}$ *is*

(i) $\{(f, \alpha, \beta) : \alpha-\alpha_0 > \Delta, \beta > \beta_0\}$,
(ii) $\{(f, \alpha, \beta) : \alpha-\alpha_0 > \Delta, \beta < \beta_0\}$,
(iii) $\{(f, \alpha, \beta) : \alpha-\alpha_0 < -\Delta, \beta > \beta_0\}$,
(iv) $\{(f, \alpha, \beta) : \alpha-\alpha_0 < -\Delta, \beta < \beta_0\}$.

Remark 3.3. If random fs are not symmetrized around zero, α is not identifiable. So the posterior distribution for α will not be consistent. Consistency for β will hold under appropriate conditions. To prove the existence of uniformly consistent tests for β, we pair Y_is and consider the difference $Y_i - Y_j$, which has a density that is symmetric around $\beta(x_i - x_j)$. We can now handle the problem in essentially the same way as in Proposition 3.1 to construct strictly unbiased tests. A result analogous to Proposition 3.2 then follows immediately. The verification of the other conditions in Sections 4, 5 and 6 is along exactly similar lines.

The next proposition considers neighbourhoods of f_0 to obtain posterior consistency for the true density rather than only the parametric part. We need an additional assumption.

Assumption B. *For some* L, $|x_i| < L$ *for all* i.

In practice, the range of interest of the regressor is often a bounded interval, since the linearity of the regression function can only be expected on a range of values. Therefore, the assumption may not be very restrictive from a practical point of view.

Proposition 3.3. *Suppose that Assumption B holds. Let* $\mathcal{U}$ *be a weak neighbourhood of* f_0 *and let* $\mathcal{W} = \mathcal{U}^c \times \{(\alpha, \beta) : |\alpha - \alpha_0| < \Delta, \ |\beta - \beta_0| < \Delta\}$. *Then there exists an exponentially consistent sequence of tests for testing*

$$H_0 : (f, \alpha, \beta) = (f_0, \alpha_0, \beta_0) \quad \textit{against} \quad H_1 : (f, \alpha, \beta) \in \mathcal{W}.$$

Proof. Without loss of generality take

$$\mathcal{U} = \left\{ f : \int \Phi(y) f(y) - \int \Phi(y) f_0(y) < \varepsilon \right\}, \tag{3.7}$$

where $0 \leqslant \Phi \leqslant 1$ and Φ is uniformly continuous.

Since Φ is uniformly continuous, given $\varepsilon > 0$, there exists $\delta > 0$ such that $|y_1 - y_2| < \delta$ implies $|\Phi(y_1) - \Phi(y_2)| < \varepsilon/2$.

Let Δ be such that

$$|(\alpha - \alpha_0) + (\beta - \beta_0)x_i| < \delta$$

for $\alpha, \beta \in \mathcal{W}$. Set $\tilde{\Phi}_i(y) = \Phi(y - (\alpha_0 + \beta_0 x_i))$. Then

$$\mathrm{E}_{f_{0i}} \tilde{\Phi}_i = \mathrm{E}_{f_0} \Phi, \qquad \mathrm{E}_{f_{\alpha,\beta,i}} \tilde{\Phi}_i = \mathrm{E}_{f_{(\alpha-\alpha_0),(\beta-\beta_0),i}} \Phi. \tag{3.8}$$

Noting that

$$\int \Phi(y - ((\alpha - \alpha_0) + (\beta - \beta_0)x_i)) f_{(\alpha-\alpha_0)+(\beta-\beta_0)x_i}(y) \mathrm{d}y = \int \Phi(y) f(y) \mathrm{d}y,$$

we have, by the uniform continuity of Φ,

$$\begin{aligned} \int \tilde{\Phi}_i(y) f_{\alpha,\beta,i}(y) \mathrm{d}y &\geqslant \int \Phi(y) f(y) \mathrm{d}y - \int |\Phi(y) - \Phi(y - ((\alpha - \alpha_0) + (\beta - \beta_0)x_i))| \\ &\quad \times f_{(\alpha-\alpha_0)+(\beta-\beta_0)x_i}(y) \mathrm{d}y \\ &\geqslant \int \Phi(y) f(y) \mathrm{d}y - \frac{\varepsilon}{2} \\ &\geqslant \mathrm{E}_{f_0} \Phi + \frac{\varepsilon}{2} \end{aligned}$$

for any $f \in \mathcal{U}^c$. An application of Lemma 3.1 completes the proof. □

4. Prior positivity of neighbourhoods

In this section we develop sufficient conditions to verify condition (ii) of Theorem 2.1. A similar problem in the location-parameter context was studied in Ghosal *et al*. (1999b).

There, the authors managed with Kullback–Leibler continuity of f_0 at θ_0 – the true value of the location parameter – and the requirement that $\Pi\{K(f^*_{0,\theta}, f) < \delta\} > 0$ for all θ in a neighbourhood of θ_0 and for $f^*_{0,\theta}$ close to but different from f_0. However, this approach does not carry over to the regression context since, even though the true parameter remains (α_0, β_0), for each i we encounter parameters $\theta_i = \alpha_0 + \beta_0 x_i$. Here we take a different approach. Since we have no assumptions on the structure of the random condition f, the assumption on f_0 is somewhat strong. This condition is weakened in Section 6, where we consider the Dirichlet mixture of normals. In that case, the random f is better behaved.

Lemma 4.1. *Suppose $f_0 \in \mathcal{F}$ satisfies the condition that there exist $\eta > 0$, C_η and a symmetric density g_η such that, for $|\eta'| < \eta$,*

$$f_0(y - \eta') < C_\eta g_\eta(y), \qquad \textit{for all } y. \tag{4.1}$$

Then,

(a) *for any $f \in \mathcal{F}$ and $|\theta| < \eta$,*

$$K(f_0, f_\theta) \leqslant (C_\eta + 1)\log C_\eta + C_\eta\left[K(g_\eta, f) + \sqrt{K(g_\eta, f)}\right];$$

(b) *if, in addition, $V(g_\eta, f) < \infty$, then*

$$\sup_{|\theta|<\eta} V(f_0, f_\theta) < \infty.$$

Proof. Part (a) is an immediate consequence of Lemma 5.1 of Ghosal *et al*. (1999a) and the fact that $K(f_{0,\theta}, f) = K(f_0, f_\theta)$, which follows from the symmetry of f_0 and f.

For (b), note that

$$\int f_0\left[\log_+ \frac{f_0}{f_\theta}\right]^2 = \int f_{0,\theta}\left[\log_+ \frac{f_{0,\theta}}{f}\right]^2 \leqslant C_\eta \int g_\eta\left[\log_+ \frac{C_\eta g_\eta}{f}\right]^2, \tag{4.2}$$

which is finite under the assumed condition. □

We write the assumption of last lemma as follows:

Assumption C. For $\eta > 0$, *sufficiently small, there exist $g_\eta \in \mathcal{F}$ and constant $C_\eta > 0$ such that for $|\eta'| < \eta$,*

$$f_0(y - \eta') < C_\eta g_\eta(y) \quad \text{for all } y$$

and

$$C_\eta \to 1 \quad \text{as } \eta \to 0.$$

Proposition 4.1. *Suppose that Assumptions B and C hold. Let $\tilde{\Pi}$ be a prior for f, and μ be a prior for (α, β). If (α_0, β_0) is in the support of μ and if, for all η sufficiently small and for all $\delta > 0$,*

$$\tilde{\Pi}\{K(g_\eta, f) < \delta,\ V(g_\eta, f) < \infty\} > 0, \tag{4.3}$$

then, for all $\delta > 0$ *and some* $M > 0$*,*

$$(\tilde{\Pi} \times \mu)\{(f, \alpha, \beta) : K_i(f, \alpha, \beta) < \delta,\ V_i(f, \alpha, \beta) < M \text{ for all } i\} > 0. \tag{4.4}$$

Proof. Choose η, δ_0 such that (4.3) holds with $\delta = \delta_0$ and

$$(C_\eta + 1)\log C_\eta + C_\eta[\delta_0 + \sqrt{\delta_0}] < \delta.$$

Let

$$V = \left\{(\alpha, \beta) : |\alpha - \alpha_0| < \frac{\eta}{2}, \quad |\beta - \beta_0| < \frac{\eta}{2L}\right\}.$$

Note that

$$K_i(f_0, \alpha, \beta) = K(f_0, f_{(\alpha-\alpha_0)+(\beta-\beta_0)x_i})$$

and

$$V_i(f_0, \alpha, \beta) = V(f_0, f_{(\alpha-\alpha_0)+(\beta-\beta_0)x_i}),$$

and $(\alpha, \beta) \in V$ implies that $|(\alpha - \alpha_0) + (\beta - \beta_0)x_i| < \eta$ for all x_i. An application of Lemma 4.1 immediately gives the result. □

Theorem 4.1. *Suppose that:*

(i) *the covariates* $x_1, x_2, \ldots$ *satisfy Assumptions A and B;*
(ii) f_0 *is continuous,* $f_0(0) > 0$ *and* f_0 *satisfies Assumption C;*
(iii) *for all sufficiently small* η *and for all* $\delta > 0$*,*

$$\tilde{\Pi}\{K(g_\eta, f) < \delta,\ V(g_\eta, f) < \infty\} > 0,$$

where g_η *is as in Assumption C.*

Then for any weak neighbourhood $\mathcal{U}$ *of* f_0*,*

$$\Pi\{(f, \alpha, \beta) : f \in \mathcal{U},\ |\alpha - \alpha_0| < \delta,\ |\beta - \beta_0| < \delta | Y_1, Y_2, \ldots, Y_n\} \to 1 \tag{4.5}$$

a.s. $\prod_{i=1}^{\infty} P_{f_{0i}}$. *In other words, the posterior distribution is weakly consistent at* (f_0, α_0, β_0).

Proof. Note that

$$\{(f, \alpha, \beta) : f \in \mathcal{U},\ |\alpha - \alpha_0| < \delta,\ |\beta - \beta_0| < \delta\}^c \tag{4.6}$$

is the union of sets considered in Propositions 3.1, 3.2 and 3.3. The required exponentially consistent test therefore exists. Proposition 4.1 shows that condition (ii) of Theorem 2.1 holds and hence (4.5) follows. □

Remark 4.1. Assumption (ii) of Theorem 4.1 is satisfied if f_0 is Cauchy or normal. If f_0 is Cauchy, then $g_\eta = f_0$ satisfies Assumption C. If f_0 is normal, then Assumption C holds with

$$g_\eta = f^s_{0,\eta^*} = \tfrac{1}{2}\{f_0(y-\eta^*) + f_0(-y-\eta^*)\}, \tag{4.7}$$

where $\eta^* \to 0$ as $\eta \to 0$ but $\eta^*/\eta \to \infty$.

Remark 4.2. Assumption B is used in two places: Propositions 3.3 and 4.1. For specific f_0s one may be able to obtain the conclusion of Proposition 4.1 without Assumption B. In such cases one would be able to obtain consistency at (α_0, β_0) without having to establish consistency at (f_0, α_0, β_0).

Remark 4.3. In order to strengthen Theorem 4.1 to variation neighbourhoods $\mathcal{U}$ of f_0, one also needs to find, for all $\varepsilon > 0$, a sequence of subsets $\mathcal{F}_n \subset \mathcal{F}$ with $\tilde{\Pi}(\mathcal{F}_n^c)$ exponentially small such that, for some $\delta < \varepsilon/2$ and $\beta < \varepsilon^2/8$, the L_1-metric entropy $J(\delta, \mathcal{F}_n) < n\beta$. See Theorem 2 of Ghosal *et al*. (1999c) for details.

5. Polya tree priors

In this section we show that Polya tree priors, with a suitable choice of parameters, satisfy condition (iii) of Theorem 4.1 and hence the posterior distribution is weakly consistent. To obtain a prior on symmetric densities, we consider Polya tree priors on densities f on the positive half-line and then consider the symmetrization $f^s(y) = \frac{1}{2}f(|y|)$. Since $K(f, g) = K(f^s, g^s)$ and $V(f, g) = V(f^s, g^s)$, this symmetrization presents no problems.

We briefly recall Polya tree priors; for more details the reader should refer to Lavine (1992; 1994) and Mauldin *et al*. (1992).

Let $E = \{0, 1\}$, $E^m = \{0, 1\}^m$ and $E^* = \bigcup_{m=1}^{\infty} E^m$. For each m, $\{B_{\underline{\epsilon}} : \underline{\epsilon} \in E^m\}$ is a partition of $\mathbb{R}^+$, and for each $\underline{\epsilon}$, $\{B_{\underline{\epsilon}0}, B_{\underline{\epsilon}1}\}$ is a partition of $B_{\underline{\epsilon}}$. Furthermore, $\{B_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$ generates the Borel σ-algebra.

A random probability measure P on $\mathbb{R}^+$ is said to be distributed as a Polya tree with parameters $(\Pi, \mathcal{A})$, where Π is a sequence of partitions as described in the previous paragraph, and $\mathcal{A} = \{\alpha_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$ is a collection of non-negative numbers, if there exists a collection $\{Y_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$ of mutually independent random variables such that:

(i) each $Y_{\underline{\epsilon}}$ has a beta distribution with parameters $\alpha_{\underline{\epsilon}0}$ and $\alpha_{\underline{\epsilon}1}$;
(ii) the random measure P is given by

$$P(B_{\epsilon_1\cdots\epsilon_m}) = \left[\prod_{j=1,\epsilon_j=0}^{m} Y_{\epsilon_1\cdots\epsilon_{j-1}}\right]\left[\prod_{j=1,\epsilon_j=1}^{m} (1 - Y_{\epsilon_1\cdots\epsilon_{j-1}})\right].$$

We restrict ourselves to partitions $\Pi = \{\Pi_m : m = 0, 1, \ldots\}$ that are determined by a strictly positive, continuous density α on $\mathbb{R}^+$ in the following sense: the sets in Π_m are intervals of the form

$$\left\{y : \frac{k-1}{2^m} < \int_{-\infty}^{y} \alpha(t)\mathrm{d}t \leqslant \frac{k}{2^m}\right\}.$$

Theorem 5.1. *Let $\tilde{\Pi}$ be a Polya tree prior on densities on $\mathbb{R}^+$ with $\alpha_{\underline{\epsilon}} = r_m$ for all $\underline{\epsilon} \in E^m$. If $\sum_{m=1}^{\infty} r_m^{-1/2} < \infty$, then for any density g such that $K(g, \alpha) < \infty$ and $\mathrm{E}_g(\log g)^2 < \infty$, we have, for all $\delta > 0$,*

$$\lim_{M\to\infty} \tilde{\Pi}\{f : K(g, f) < \delta, V(g, f) < M\} > 0. \tag{5.1}$$

Proof. We will show that

$$\lim_{M\to\infty} \tilde{\Pi}\{f : V(g, f) < M\} \to 1. \tag{5.2}$$

This, together with Theorem 3.1 of Ghosal *et al*. (1999b), where it is shown that $\tilde{\Pi}\{f : K(f, g) < \delta\} > 0$ when $\sum_{m=1}^{\infty} r_m^{-1/2} < \infty$, would then prove the theorem.

Since

$$V(g, f) \leqslant \mathrm{E}_g(\log f)^2 + \mathrm{E}_g(\log g)^2 + 2\sqrt{\mathrm{E}_g(\log f)^2\mathrm{E}_g(\log g)^2}, \tag{5.3}$$

it is enough to show that, as $M \to \infty$,

$$\tilde{\Pi}\{\mathrm{E}_g(\log f)^2 > M\} \to 0. \tag{5.4}$$

If y has the binary expansion $\underline{\epsilon} = \epsilon_1\epsilon_2\cdots$, then, for almost all y,

$$f(y) = \lim_{m\to\infty}\left[\prod_{j=1,\epsilon_j=0}^{m} 2Y_{\epsilon_1\cdots\epsilon_{j-1}}\right]\left[\prod_{j=1,\epsilon_j=1}^{m} 2(1 - Y_{\epsilon_1\cdots\epsilon_{j-1}})\right], \tag{5.5}$$

so that

$$\mathrm{E}_g(\log f)^2 = \mathrm{E}_g\left[\sum_{j=1,\epsilon_j=0}^{\infty}\log(2Y_{\epsilon_1\cdots\epsilon_{j-1}}) + \sum_{j=1,\epsilon_j=1}^{\infty}\log(2(1 - Y_{\epsilon_1\cdots\epsilon_{j-1}}))\right]^2, \tag{5.6}$$

where E_g now stands for the expectation over $\underline{\epsilon}$ when y has density g.

Now letting $\mathcal{E}$ stand for the expectation with respect to $\tilde{\Pi}$, we have, by Chebyshev's inequality,

$$\tilde{\Pi}[\mathrm{E}_g(\log f)^2 > M] \leqslant M^{-1}\mathcal{E}\mathrm{E}_g\left[\sum_{j=1,\epsilon_j=0}^{\infty}\log(2Y_{\epsilon_1\cdots\epsilon_{j-1}}) + \sum_{j=1,\epsilon_j=1}^{\infty}\log(2(1 - Y_{\epsilon_1\cdots\epsilon_{j-1}}))\right]^2. \tag{5.7}$$

Interchanging the order of expectations and exploiting independence, the right-hand side of (5.7) can further be bounded by

$$2M^{-1}\mathrm{E}_g\left[\sum_{j=1,\epsilon_j=0}^{\infty}\mathcal{E}(\log(2Y_{\epsilon_1\cdots\epsilon_{j-1}}))^2+\left(\sum_{j=1,\epsilon_j=0}^{\infty}\mathcal{E}(\log(2Y_{\epsilon_1\cdots\epsilon_{j-1}}))\right)^2\right.$$

$$\left.+\sum_{j=1,\epsilon_j=1}^{\infty}\mathcal{E}(\log(2(1-\mathrm{Y}_{\epsilon_1\cdots\epsilon_{j-1}})))^2+\left(\sum_{j=1,\epsilon_j=1}^{\infty}\mathcal{E}(\log(2(1-Y_{\epsilon_1\cdots\epsilon_{j-1}})))\right)^2\right].$$

Since $Y_{\epsilon_1\cdots\epsilon_{j-1}}$ and $1-Y_{\epsilon_1\cdots\epsilon_{j-1}}$ have the same distribution, the last expression is equal to

$$2M^{-1}\mathrm{E}_g\left[\sum_{j=1}^{\infty}\mathcal{E}(\log(2Y_{\epsilon_1\cdots\epsilon_{j-1}}))^2+\left(\sum_{j=1}^{\infty}\mathcal{E}(\log(2Y_{\epsilon_1\cdots\epsilon_{j-1}}))\right)^2\right].$$

Note that the terms inside E_g do not involve the particular sequence $\underline{\epsilon}$. Letting $\varphi(k)=\mathrm{E}|\log(2U_k)|$ and $\psi(k)=\mathrm{E}(\log(2U_k))^2$, where $U_k\sim\mathrm{Beta}(k,k)$, the last expression can be written as

$$2M^{-1}\left[\sum_{m=1}^{\infty}\psi(r_m)+\left(\sum_{m=1}^{\infty}\varphi(r_m)\right)^2\right].$$

It is shown in the Appendix that $\varphi(k)$ and $\psi(k)$ are respectively $O(k^{-1})$ and $O(k^{-1/2})$. Since $\sum_{m=1}^{\infty}r_m^{-1/2}<\infty$, both infinite series are summable and hence the last expression goes to 0 as $M\to\infty$. □

Although Polya trees give rise to naturally interpretable priors on densities and lead to consistent posterior, sample paths of Polya trees are very rough, having discontinuities everywhere. Such a drawback can easily be overcome by considering a mixture of Polya trees. Posterior consistency continues to hold in this case since, by Fubini's theorem, prior positivity holds under mild uniformity conditions.

6. Dirichlet mixture of normals

In this section, we look at random densities that arise as mixtures of normal densities. Let ϕ_h denote the normal density with mean 0 and standard deviation h. For any probability P on $\mathbb{R}$, $f_{h,P}$ will stand for the density

$$f_{h,P}(y)=\int\phi_h(y-t)\mathrm{d}P(t). \tag{6.1}$$

Our model consists of a prior μ for h and a prior $\tilde{\Pi}$ for P. Consistency issues related to these priors, in the context of density estimation, are explained in Ghosal *et al.* (1999c). Here we look at similar issues when the error density f in the regression model is endowed with these priors.

To ensure that the prior sits on symmetric densities, we let P be a random probability on $\mathbb{R}^+$ and set

$$f_{h,P}(y) = \frac{1}{2}\int \phi_h(y-t)\mathrm{d}P(t) + \frac{1}{2}\int \phi_h(y+t)\mathrm{d}P(t). \tag{6.2}$$

We will denote by $\tilde{\Pi}$ both the prior for P and the prior for $f_{h,P}$.

The following lemma shows that the random f generated by the prior under consideration is more regular than those generated by Polya tree priors, and hence the conditions on f_0 are more transparent than those in Section 5 or those in Ghosal *et al*. (1999b).

Lemma 6.1. *Let f_0 be a density such that*

$$\int y^4 f_0(y)\mathrm{d}y < \infty \quad \textit{and} \quad \int f_0(y)|\log f_0(y)|^2\,\mathrm{d}y < \infty. \tag{6.3}$$

If $f(y) = \int \phi_h(y-t)\mathrm{d}P(t)$ and $\int t^2\,\mathrm{d}P(t) < \infty$, then

(i) $\lim_{\theta\to 0} K(f_0, f_\theta) = K(f_0, f)$;
(ii) $\lim_{\theta\to 0} V(f_0, f_\theta) = V(f_0, f)$.

Proof. Clearly $f(y)$ is positive and continuous, and

$$|\log f_\theta(y)| \leqslant |\log \sqrt{2\pi}h| + \left|\log \int \mathrm{e}^{-(y-\theta-t)^2/(2h^2)}\,\mathrm{d}P(t)\right|. \tag{6.4}$$

Since $\log\int \mathrm{e}^{-(y-\theta-t)^2/(2h^2)}\mathrm{d}P(t) < 0$, by Jensen's inequality applied to $-\log x$, the last expression is bounded by

$$|\log \sqrt{2\pi}h| + \int \frac{(y-\theta-t)^2}{h^2}\,\mathrm{d}P(t).$$

The dominated convergence theorem now applies. □

We now return to the regression model.

Theorem 6.1. *Suppose $\tilde{\Pi}$ is a normal mixture prior for f. If*

(i) *Assumptions A and B hold,*
(ii) $\tilde{\Pi}\{f : K(f_0, f) < \delta,\ V(f_0, f) < \infty\} > 0$ *for all* $\delta > 0$,
(iii) $\mathrm{E}_{f_0}(y^2) < \infty,\ \mathrm{E}_{f_0}(\log f_0)^2 < \infty$,
(iv) $\int\int t^2\,\mathrm{d}P(t)\mathrm{d}\tilde{\Pi}(P) < \infty$,

then the posterior distribution $\Pi(\cdot|Y_1, \ldots, Y_n)$ is weakly consistent for (f, α, β) at (f_0, α_0, β_0) provided (α_0, β_0) is in the support of the prior for (α, β).

Proof. By (iv), $\{P : \int t^2\,\mathrm{d}P(t) < \infty\}$ has $\tilde{\Pi}$-probability 1. So we may assume that

$$\tilde{\Pi}\{\mathcal{U}\} > 0, \tag{6.5}$$

where $\mathcal{U} = \{f : f = f_P, \text{(ii) holds}, \int t^2 \, dP(t) < \infty\}$.

For every $f \in \mathcal{U}$, using Lemma 6.1, choose δ_f such that, for $\theta < \delta_f$,

$$K(f_0, f) < \delta, \qquad V(f_0, f) < \delta. \tag{6.6}$$

Now choose ε_f such that $|\alpha - \alpha_0 + (\beta - \beta_0)x_i| < \delta_f$ whenever $|\alpha - \alpha_0| < \varepsilon_f$, $|\beta - \beta_0| < \varepsilon_f / L$.

Clearly if $f \in \mathcal{U}$ and $|\alpha - \alpha_0| < \varepsilon_f$ and $|\beta - \beta_0| < \varepsilon_f / L$, we have

$$K_i(f, \alpha, \beta) < 2\delta, \quad V_i(f, \alpha, \beta) < V(f_0, f) + \delta. \tag{6.7}$$

Since

$$\tilde{\Pi}\{(f, \alpha, \beta) : f \in \mathcal{U}, |\alpha - \alpha_0| < \varepsilon_f, |\beta - \beta_0| < \varepsilon_f / L\} > 0, \tag{6.8}$$

we have

$$\Pi\left\{(f, \alpha, \beta) : K_i(f_0, \alpha, \beta) < \delta \text{ for all } i, \sum_{i=1}^{\infty} \frac{V_i(f, \alpha, \beta)}{i^2} < \infty\right\} > 0. \tag{6.9}$$

An application of Theorem 2.1 completes the proof. □

It is shown in Ghosal *et al*. (1999c) that if f_0 has compact support or if $f_0 = f_P$ with P having compact support, then $\tilde{\Pi}\{f : K(f_0, f) < \delta\} > 0$ for all $\delta > 0$. The argument given there also shows that in these cases condition (ii) of Theorem 6.1 holds when $\tilde{\Pi}$ is Dirichlet with base measure γ. Ghosal *et al*. (1999c) also describe f_0s whose tail behaviour is related to that of γ such that $\tilde{\Pi}\{f : K(f_0, f) < \delta\} > 0$. In the case when the prior is Dirichlet, the double integral in (iv) is finite if and only if $\int t^2 \, d\gamma(t) < \infty$. While normal f_0 is covered by these results the case of Cauchy f_0 cannot be resolved by the methods in that paper. However, Dirichlet location and scale mixtures of normal should be able to handle Cauchy f_0 which is a normal scale mixture. This scale mixing measure does not have a compact support so the results of Ghosal *et al*. (1999c) still do not apply.

7. Binary response regression with unknown link

A distinguishing feature of the regression problem considered in this paper is the change in the parameter value with i. A similar situation arises in other models such as the regression of the Bernoulli parameter with an unknown link function. This may be viewed as a nonparametric version of logistic regression problems and the methods developed here can be used to handle these problems too. We give an indication of how this can be done without going into much detail.

Consider k levels of a drug on a suitable scale, say $x_1, \ldots, x_k$, with probability of a response (which may be death or some other specified event) p_i, $i = 1, \ldots, k$. To study the effects at different levels, n subjects are treated with the drug. The ith level of the drug is given to n_i subjects and the number of responses r_i noted. We thus obtain k independent binomial variables with parameters n_i and p_i, where $n = n_1 + \ldots + n_k$. The object usually is to find x such that $p = 0.5$. Often, p_i is modelled as

$$p_i = F(\alpha + \beta x_i) = H(x_i), \tag{7.1}$$

say, where F is a response distribution and α and β are parameters. Here p_i may be estimated by r_i/n_i, but if the n_i are small, the estimates will have large variances. The model provides a way of combining all the data. In logistic regression, F is taken as logistic function. Other link functions such as the normal distribution function are also used. The choice of the functional form of the link function is somewhat arbitrary, and this may substantially influence inference, particularly at the two ends where the data is sparse. In recent years, there has been a lot of interest in link functions with unknown functional form. In nonparametric problems of this kind, one puts a prior on F or H. Such an approach was taken by Albert and Chib (1993), Chen and Dey (1998), Basu and Mukhopadhyay (1998, 2000) among others. If one puts a prior on F, one has to put conditions on F such as specifying the values of two quantiles to make (F, α, β) identifiable. In this case, one can develop sufficient conditions for posterior consistency at (F_0, α_0, β_0) using our variant of Schwartz's theorem. However, in practice, one usually puts a Dirichlet process or some other prior on F and, independently of this, a prior on (α, β). Due to the discreteness of Dirichlet selections, many authors actually prefer the use of other priors such as Dirichlet scale mixtures of normals; see Basu and Mukhopadhyay (1998, 2000) and the references therein. Because of the lack of identifiability, the posterior for (α, β) is not consistent. This will show up in simulations as flat, rather than peaked, posteriors. On the other hand, a Dirichlet process prior and a prior on (α, β) provide a prior on H and one can ask for posterior consistency of $H^{-1}(\frac{1}{2})$ at, say, $H_0^{-1}(\frac{1}{2})$. This problem can be solved by Theorem 2.1 as follows.

Without loss of generality, one may take $n_i = 1$ for all i, and hence $k = n$. To verify condition (ii) of Theorem 2.1, consider

$$Z_i = \log \frac{(H_0(x_i))^{r_i}(1 - H_0(x_i))^{1-r_i}}{(H(x_i))^{r_i}(1 - H(x_i))^{1-r_i}}, \tag{7.2}$$

where r_i is 1 or 0 with probability $H(x_i)$ and $1 - H(x_i)$ respectively, and the true H is denoted by H_0. Then

$$\mathrm{E}_{H_0}(Z_i) = H_0(x_i)\log\frac{H_0(x_i)}{H(x_i)} + (1 - H_0(x_i))\log\frac{1 - H_0(x_i)}{1 - H(x_i)} \tag{7.3}$$

and

$$\mathrm{E}_{H_0}(Z_i^2) \leqslant 2H_0(x_i)\left(\log\frac{H_0(x_i)}{H(x_i)}\right)^2 + 2(1 - H_0(x_i))\log\left(\frac{1 - H_0(x_i)}{1 - H(x_i)}\right)^2. \tag{7.4}$$

Assume that the x_i lie in a bounded interval containing $H_0^{-1}(\frac{1}{2})$, and the support of H_0 contains a bigger interval. Since the range of the x_i is bounded, the sequence of formal empirical distributions $n^{-1}\sum_{i=1}^{n}\delta_{x_i}$ of $x_1, \ldots, x_n$ is relatively compact. Assume that all subsequential limits converge to distributions which give positive measure to all non-degenerate intervals, provided the intervals are contained in a certain interval containing $H_0^{-1}(\frac{1}{2})$. Therefore, a positive fraction of the x_i lie in an interval of positive length if the interval is close to the the point $H_0^{-1}(\frac{1}{2})$. Also assume that H_0 is continuous and the support of the prior for H contains H_0. For instance, if the prior is Dirichlet with a base measure whose support contains the support of H_0, then the above condition is satisfied. Mixture priors often have large supports too. For instance, the Dirichlet scale mixture of normal prior used by Basu and Mukhopadhyay (1998; 2000) will have this property if the true link function is also a scale mixture of normal cumulative distribution functions.

If H_ν is a sequence converging weakly to H_0, then, by Polya's theorem, the convergence is uniform. Note that the functions $p\log(p/q) + (1-p)\log((1-p)/(1-q))$ and $p(\log(p/q))^2 + (1-p)(\log((1-p)/(1-q)))^2$ in q converge to 0 as $q \to p$, uniformly in p lying in a compact subinterval of (0, 1). Thus given $\delta > 0$, we can choose a weak neighbourhood $\mathcal{U}$ of H_0 such that if $H \in \mathcal{U}$, then $\mathrm{E}_{H_0}(Z_i) < \delta$ and the $\mathrm{E}_{H_0}(Z_i^2)$ are bounded. By the assumption on the support of the prior, condition (ii) of Theorem 2.1 holds.

For the existence of exponentially consistent tests in condition (i) of Theorem 2.1, consider, without loss of generality, testing $H^{-1}(\frac{1}{2}) = H_0^{-1}(\frac{1}{2})$ against $H^{-1}(\frac{1}{2}) > H_0^{-1}(\frac{1}{2}) + \varepsilon$ for small $\varepsilon > 0$. Let

$$K_n = \{i : H_0^{-1}(1/2) + \varepsilon/2 \leqslant x_i \leqslant H_0^{-1}(1/2) + \varepsilon\}.$$

Since

$$\mathrm{E}_H(r_i) = H(x_i) \leqslant H(H_0^{-1}(1/2) + \varepsilon) \leqslant \tfrac{1}{2} \tag{7.5}$$

and

$$\mathrm{E}_{H_0}(r_i) = H_0(x_i) \geqslant H_0(H_0^{-1}(1/2) + \varepsilon/2) > \tfrac{1}{2}, \tag{7.6}$$

the test

$$\frac{1}{\#K_n} \sum_{i \in K_n} r_i < \frac{1}{2} + \eta \tag{7.7}$$

for $\eta = (H_0(H_0^{-1}(\frac{1}{2}) + \varepsilon/2) - \frac{1}{2})/2$ is exponentially consistent by Hoeffding's inequality and the fact that $\#K_n/n$ converge to positive limits along subsequences. Therefore Theorem 2.1 applies and the posterior distribution of $H^{-1}(\frac{1}{2})$ is consistent at $H_0^{-1}(\frac{1}{2})$.

8. Stochastic regressor

In this section, we consider the case where the independent variable X is stochastic. We assume that the X observations $X_1, X_2, \ldots$ are i.i.d. with a probability density function $g(x)$ and are independent of the errors $\epsilon_1, \epsilon_2, \ldots$. We argue below that all the results on consistency hold under appropriate conditions.

Let $G(x) = \int_{-\infty}^{x} g(u)\mathrm{d}u$, denote the cumulative distribution function of X. We shall assume that the following condition holds:

Assumption D. *The independent variable* X *is compactly supported and* $0 < G(0-) \leqslant G(0) < 1$.

Under these assumptions, results will follow from a conditionality argument and the corresponding results for the non-stochastic case, conditioned on a sequence $x_1, x_2, \ldots$ such that Assumptions A and B hold. Note that if g satisfies Assumption D, then P_g^∞-almost all sequences $x_1, x_2, \ldots$ satisfy Assumptions A and B.

Observe that for a stochastic $x_1, x_2, \ldots$ with a known density g, the expressions for the posterior probabilities are still given by (2.6), as the factor $\Pi_{i=1}^n g(x_i)$ is cancelled in the numerator and the denominator. As g has no role, we need no knowledge of it provided that it is a priori independent of the other parameters. We need not specify a prior distribution for g, but assume that the sampled gs are compactly supported and satisfy Assumption D. If f_0 and the prior $\tilde{\Pi}$ satisfy conditions (ii) and (iii) of Theorem 4.1, it then follows that, for any neighbourhood $\mathcal{U}$ of f_0,

$$\tilde{\Pi}\{(f, \alpha, \beta) : f \in \mathcal{U}, |\alpha - \alpha_0| < \delta, |\beta - \beta_0| < \delta | (X_1, Y_1), \ldots, (X_n, Y_n)\} \to 1$$

a.s. $P^\infty_{f_0, g_0, \alpha_0, \beta_0}$, where $P_{f_0, g_0, \alpha_0, \beta_0}$ is the distribution of (X, Y), X has density g_0, $Y = \alpha_0 + \beta_0 X + \epsilon$, X is independent of ϵ and ϵ has density f_0.

Thus if X is stochastic and Assumption D replaces Assumptions A and B in Theorems 5.1 and 6.1, posterior consistency holds.

Appendix

Lemma A.1. *Let* f *and* g *be probability densities. Let* $\|f - g\|_1 = \int |f - g|$ *stand for the* L_1*-distance and let* $K^+(f, g) = \int f \log_+(f/g)$ *and* $K^-(f, g) = \int f \log_-(f/g)$, where $\log_- x = \max(-\log x, 0)$. *Then*

$$K^-(f, g) \leqslant \tfrac{1}{2}\|f - g\| \leqslant \sqrt{K(f, g)/2} \tag{A.1}$$

and

$$K^+(f, g) \leqslant \tfrac{1}{2}\|f - g\| + K(f, g) \leqslant K(f, g) + \sqrt{K(f, g)/2}. \tag{A.2}$$

Proof. Using $\log x \leqslant x - 1$, as in Hannan (1960), we obtain $K^-(f, g) = \int_{g>f} f \log(g/f) \leqslant \int_{g>f}(g - f) = \|f - g\|_1/2$. The second part of (A.1) follows from Kemperman's inequality (Kemperman 1969, Theorem 6.1). Relation (A.2) follows because $K^+ = K + K^-$. □

Remark A.1. Using the inequality $\log x \leqslant 2(\sqrt{x} - 1)$, the following alternative bound can be derived:

$$K^-(f, g) \leqslant \|f - g_1\| - H^2(f, g), \tag{A.3}$$

where $H^2(f, g) = \int (f^{1/2} - g^{1/2})^2$ is the squared Hellinger distance.

Proof of Theorem 2.1. The proof proceeds along the same lines as in Theorem 6.1 of Schwartz (1965). Here is a sketch of the argument.

Write the posterior probability in (2.5) as

$$\frac{I_{1n}}{I_{2n}} \leqslant \Phi_n + \frac{(1-\Phi_n)I_{1n}}{I_{2n}}, \tag{A.4}$$

where I_{1n} and I_{2n} are as in (2.6).

Clearly, in view of the Borel–Cantelli lemma, condition (a) in Definition 2.1 implies that $\Phi_n \to 0$ a.s. $\prod_{i=1}^{\infty} P_{f_{0i}}$.

Note that

$$\begin{aligned} \mathrm{E}_{\Pi_1^n f_{0i}}((1-\Phi_n)I_{1n}) &= \int (1-\Phi_n)\int_{\mathcal{W}^c} \prod_{i=1}^{n} \frac{f_{\alpha,\beta,i}(y_i)}{f_{0i}(y_i)} \mathrm{d}\Pi(f,\alpha,\beta) \prod_{i=1}^{n} f_{0i}(y_i)\mathrm{d}y_i \\ &= \int_{\mathcal{W}^c}\int (1-\Phi_\mathrm{n}) \prod_{i=1}^{n} f_{\alpha,\beta,i}(y_i)\mathrm{d}y_i\, \mathrm{d}\Pi(f,\alpha,\beta) \\ &\leqslant \sup_{\mathcal{W}^c} \mathrm{E}_{\Pi_1^n f_{\alpha,\beta,i}}(1-\Phi_n) \\ &\leqslant C_2 \mathrm{e}^{-nC}. \end{aligned}$$

Therefore,

$$\mathrm{e}^{nC/2}(1-\Phi_n)I_{1n} \to 0 \tag{A.5}$$

a.s. $\prod_{i=1}^{\infty} P_{f_{0i}}$.

Let $\mathcal{V}$ be the set displayed in condition (ii) of the theorem. Note that with $W_i = \log_+(f_{0i}/f_{\alpha,\beta,i})(Y_i)$, we have $\mathrm{var}(W_i) \leqslant V_i(f,\alpha,\beta)$, and hence $\sum_{i=1}^{\infty} \mathrm{var}(W_i)/i^2 < \infty$ for all $f \in \mathcal{V}$. Applying Kolmogorov's strong law of large numbers for independent non-identical variables to the sequence $W_i - \mathrm{E}(W_i)$, it follows from Lemma A.1 that, for each $f \in \mathcal{V}$,

$$\begin{aligned} \liminf_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_{\alpha,\beta,i}(Y_i)}{f_{0i}(Y_i)}\right) &\geqslant -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} \log_+ \frac{f_{0i}(Y_i)}{f_{\alpha,\beta,i}(Y_i)}\right) \\ &= -\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} K_i^+(f,\alpha,\beta) \\ &\geqslant -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} K_i(f,\alpha,\beta) + \frac{1}{n}\sum_{i=1}^{n}\sqrt{K_i(f,\alpha,\beta)/2}\right) \\ &\geqslant -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} K_i(f,\alpha,\beta) + \sqrt{\frac{1}{n}\sum_{i=1}^{n} K_i(f,\alpha,\beta)/2}\right). \end{aligned} \tag{A.6}$$

a.s. $\prod_{i=1}^{\infty} P_{f_{0i}}$. Since, for $f \in \mathcal{V}$, $n^{-1}\sum_{i=1}^{n} K_i(f,\alpha,\beta) < \delta$, we have, for each $f \in \mathcal{V}$,

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_{\alpha,\beta,i}(Y_i)}{f_{0i}(Y_i)} \geqslant -(\delta + \sqrt{\delta/2}). \tag{A.7}$$

Choosing δ so that $\delta + \sqrt{\delta/2} \leqslant C/8$ and noting that

$$I_{2n} \geqslant \int_{\mathcal{V}} \prod_{i=1}^{n} \frac{f_{\alpha,\beta,i}(Y_i)}{f_{0i}(Y_i)}\, \mathrm{d}\Pi(f, \alpha, \beta),$$

it follows from Fatou's lemma that

$$\mathrm{e}^{nC/4} I_{2n} \to \infty \tag{A.8}$$

a.s. $\prod_{i=1}^{\infty} P_{f_{0i}}$. Combining this with (A.4) and (A.5), we obtain (2.5). Indeed, the convergence is exponentially fast. This proves the theorem. □

Remark A.2. Condition (ii) of the theorem can be weakened. It can be seen from the proof that if the prior assigns positive probability to the set

$$\left\{\frac{1}{n}\sum_{i=1}^{n} K_i(f, \alpha, \beta) < \delta \text{ for all } n, \sum_{i=1}^{\infty} \frac{V_i(f, \alpha, \beta) + K_i^2(f, \alpha, \beta)}{i^2} < \infty\right\},$$

then the posterior is also consistent.

We state Lemma 5.1 from Ghosal *et al*. (1999b) for easy reference.

Lemma A.2. *If $f_0 \leqslant Cf_1$, where f_0 and f_1 are densities, then, for any f,*

$$K(f_0, f) \leqslant (C+1)\log C + C[K(f_1, f) + \sqrt{K(f_1, f)}]. \tag{A.9}$$

Lemma A.3. *If $U_k \sim \mathrm{Beta}(k, k)$, then*

$$\mathrm{E}(\log(2U_k))^2 = O(k^{-1}). \tag{A.10}$$

Proof. Let

$$I_k = \mathrm{E}(\log(2U_k))^2 = \frac{1}{B(k, k)}\int_0^1 (\log(2u))^2 u^{k-1}(1-u)^{k-1}\, \mathrm{d}u, \tag{A.11}$$

where $B(k, k) = \int_0^1 u^{k-1}(1-u)^{k-1}\, \mathrm{d}u$ is the beta function.

By a change of variable,

$$I_k = \frac{1}{B(k, k)}\int_0^1 (\log 2(1-u))^2 u^{k-1}(1-u)^{k-1}\, \mathrm{d}u. \tag{A.12}$$

Note that $\log(2u)$ and $\log(2(1-u))$ are always of opposite sign for $0 < u < 1$. Therefore,

$$
\begin{aligned}
2I_k &= \frac{1}{B(k,\,k)}\int_0^1 \{(\log(2u))^2 + (\log(2(1-u)))^2\} u^{k-1}(1-u)^{k-1}\,\mathrm{d}u \\
&= \frac{1}{B(k,\,k)}\int_0^1 \{\log(2u) - \log(2(1-u))\}^2 u^{k-1}(1-u)^{k-1}\,\mathrm{d}u \\
&\quad + \frac{1}{B(k,\,k)}\int_0^1 2(\log(2u))(\log(2(1-u)))u^{k-1}(1-u)^{k-1}\,\mathrm{d}u \\
&\leqslant \frac{1}{B(k,\,k)}\int_0^1 \left(\log\frac{u}{1-u}\right)^2 u^{k-1}(1-u)^{k-1}\,\mathrm{d}u. \qquad \text{(A.13)}
\end{aligned}
$$

Using the Laplace approximation, it has been shown in the proof of Lemma A.1 of Ghosal *et al*. (1999b) that the right-hand side of (A.13) is $O(k^{-1})$. This completes the proof. □

Acknowledgement

This paper is dedicated to Professor J. Pfanzagl for his outstanding contribution to asymptotics.

Most of the work for this paper was done while the second named author was at the University of Minnesota.

References

Albert, J.H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, **88**, 669–679.

Basu, S. and Mukhopadhyay, S. (1998) Binary response regression with normal scale mixtures links. In D.K. Dey, S.K. Ghosh and B.K. Mallick (eds), *Generalized Linear Models: A Bayesian Perspective*, pp. 231–241. New York: Marcel Dekker.

Basu, S. and Mukhopadhyay, S. (2000) Bayesian analysis of binary regression using symmetric and asymmetric links. *Sankhyā Ser. B,* **62**, 372–387.

Chen, M.-H. and Dey, D.K. (1998) Bayesian modeling of correlated binary responses via scale mixtures of multivariate normal link functions. *Sankhyā Ser. A,* **60**, 322–343.

Diaconis, P. and Freedman, D. (1986a) On the consistency of Bayes estimates (with discussion). *Ann. Statist.*, **14**, 1–67.

Diaconis, P. and Freedman, D. (1986b) On inconsistent Bayes estimates. *Ann. Statist.*, **14**, 68–87.

Doss, H. (1985a) Bayesian nonparametric estimation of the median. I. Computation of the estimates. *Ann. Statist.*, **13**, 1432–1444.

Doss, H. (1985b) Bayesian nonparametric estimation of the median. II. Asymptotic properties of the estimates. *Ann. Statist.*, **13**, 1445–1464.

Dudley, R.M. (1999) *Uniform Central Limit Theorems.* Cambridge: Cambridge University Press.

Freedman, D. (1963) On the asymptotic distribution of Bayes estimates in the discrete case I. *Ann. Math. Statist.*, **34**, 1386–1403.

Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999a) Consistency issues in Bayesian nonparametrics. In S. Ghosh (ed.), *Asymptotics, Nonparametrics, and Time Series: A Tribute to Madan Lal Puri*, pp. 639–668. New York: Marcel Dekker.

Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999b) Consistent semiparametric Bayesian inference about a location parameter. *J. Statist. Plann. Inference*, **77**, 181–193.

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

Ghosal, S., Ghosh, J.K. and van der Vaart, A.W. (2000) Convergence rates of posterior distributions. *Ann. Statist.*, **28**, 500–531.

Ghosh, J.K. (1998) Bayesian density estimation. *Doc. Math. J. DMV, Extra Vol. ICM*, **III**, 237–243 (electronic).

Hannan, J. (1960) Consistency of maximum likelihood estimation of discrete distributions. In I. Olkin, S.G. Churye, W. Hoeffding, W.G. Madow and H.B. Mann (eds), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pp. 249–257. Stanford, CA: Stanford University Press.

Kemperman, J.H.B. (1969) On the optimum rate of tramsmitting information. *Ann. Math. Statist.*, **40**, 2156–2177.

Lavine, M. (1992) Some aspects of Polya tree distributions for statistical modeling. *Ann. Statist.*, **20**, 1222–1235.

Lavine, M. (1994) More aspects of Polya tree distributions for statistical modeling. *Ann. Statist.*, **22**, 1161–1176.

Mauldin, R.D., Sudderth, W.D. and Williams, S.C. (1992) Polya trees and random distributions. *Ann. Statist.*, **20**, 1203–1221.

Schwartz, L. (1965) On Bayes procedures. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **4**, 10–26.

Wasserman, L. (1998) Asymptotic properties of nonparametric Bayesian procedures. In D. Dey, P. Müller and D. Sinha (eds), *Practical Nonparametric and Semiparametric Bayesian Statistics*, Lecture Notes in Statist. 133, pp. 293–304. New York: Springer-Verlag.

Received April 2001 and revised March 2002