Statistical Science
2004, Vol. 19, No. 1, 111–117
DOI 10.1214/088342304000000134

Modern Bayesian Asymptotics

Stephen G. Walker

Abstract. A survey of modern Bayesian asymptotics is given. Specific attention is paid to the Hellinger consistency of posterior distributions and the asymptotic study of Bayes factors.

Key words and phrases: Bayes factor, Bayes nonparametrics, consistency, Hellinger distance.

1. INTRODUCTION

The aim of this article is to present recent and new results in Bayesian asymptotics relating to the notion of consistency for density estimation involving independent and identically distributed sampling. For reasons of space I will restrict particular attention to (i) conditions under which a sequence of posterior distributions is consistent and (ii) the asymptotic study of the Bayes factor. A survey of other recent research in Bayesian asymptotics will appear in the last section of this article. As far as possible I will avoid technical details which, while nontrivial, are not essential for the understanding of Bayesian asymptotics.

For the first part (i), the aim is not to make a case for consistency or argue that Bayesian consistency is an obligatory requirement. Previous authors have made such arguments; see Diaconis and Freedman (1986), Wasserman (1998) and Ghosal, Ghosh and Ramamoorthi (1999b), for example. Bayesian consistency has become closely associated with Bayesian nonparametrics, reflecting the realistic assumption that the true distribution function can take any shape. On the other hand, parametric Bayesian inference is based on prior probability 1 being put on density functions having a particular form. If the data suggests otherwise, then what the Bayesian does next can be open to serious criticism. See, for example, Draper (1999). The nonparametric Bayesian avoids such problems by putting prior probability 1 on all density functions. It could also be all distribution functions, but throughout we will assume all relevant unknowns are densities and, for simplicity, are densities with respect to the Lebesgue measure. Consequently we will only consider priors which are supported by such densities. Examples of nonparametric priors include mixtures of Dirichlet processes (Lo, 1984), Pólya trees (Kraft, 1964; Ferguson, 1974; Mauldin, Sudderth and Williams, 1992; Lavine, 1992, 1994) and generalized exponential Gaussian process priors (Lenk, 1988, 1991).

It is prudent here to explain what Bayesian consistency entails. Let Ω be the set of all densities with respect to the Lebesgue measure. A weak neighborhood of the density g is given by

$$A_\varepsilon(g) = \left\{ f \in \Omega : \left| \int \phi_j f - \int \phi_j g \right| < \varepsilon, \quad j = 1, \ldots, k \right\},$$

where the ϕ_j are bounded and continuous functions. See, for example, Ghosal, Ghosh and Ramamoorthi (1999b). A strong neighborhood created by the metric d between densities g and f, $d(f, g)$, is given by

$$A_\varepsilon(g) = \{ f \in \Omega : d(f, g) < \varepsilon \}.$$

Now, if Π_n is the sequence of posterior distributions based on a sample of size n from f_0 (with distribution function F_0), the density function generating the observations, then Bayesian consistency is concerned with conditions to be imposed on the prior Π for which

$$\Pi_n(A_\varepsilon^c(f_0)) \to 0 \quad \text{a.s. } [F_0]$$

for all $\varepsilon > 0$. Here A^c represents the complement set of A; that is, $A^c = \Omega - A$. We write

$$\Pi_n(A) = \frac{\int_A R_n(f) \Pi(df)}{\int R_n(f) \Pi(df)},$$

where

$$R_n(f) = \prod_{i=1}^{n} f(X_i)/f_0(X_i)$$

Stephen G. Walker is Professor, Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (e-mail: s.g.walker@bath.ac.uk).

and $X_1, X_2, \ldots$ are the data. The inclusion of $\prod_{i=1}^{n} f_0(X_i)$ in both numerator and denominator has reasons which will become clear later on. We can term the two types of consistency weak and strong, respectively. For strong consistency, the distance which turns out to be the most useful is the Hellinger distance,

$$d(f,g) = \left\{\int (\sqrt{f} - \sqrt{g})^2\right\}^{1/2} = \left\{2\left(1 - \int \sqrt{fg}\right)\right\}^{1/2}$$

and use will be made of $h(f, f_0) = \frac{1}{2}d^2(f, f_0)$. Reasons for the importance of the Hellinger distance will also become clear later on.

Understanding Bayesian consistency is extremely rewarding. There are two themes. The first is the need to have f_0 in the support of the prior. For consistency it is the Kullback–Leibler support which is important; that is, it is important to have

$$\Pi\{f : d_K(f, f_0) < \delta\} > 0$$

for all $\delta > 0$. Here $d_K(f,g) = \int g \log(g/f)$. We will refer to this as the Kullback–Leibler property for Π if it holds true. Since f_0 is unknown, this Kullback–Leibler property can only be known to hold if $\Pi\{f : d_K(f,g) < \delta\} > 0$ for all $\delta > 0$ and all densities g. In this case the prior must be nonparametric and some examples are given in Barron, Schervish and Wasserman (1999). Alternatively, one could argue that with further conditions on the prior Π, see below, consistency holds for all true f_0 in the Kullback–Leibler support of the prior.

The second theme is the one which causes all the mathematical complexities and is concerned with preventing densities which track the data too closely from dominating the posterior. This can be possible if the prior is supported by all densities.

Summarizing the past very briefly, Schwartz's (1965) result was that a prior which has the Kullback–Leibler property gives rise to a weakly consistent sequence of posterior distributions. On the other hand, Diaconis and Freedman (1986) demonstrated that a prior which puts positive mass on all weak neighborhoods of f_0 does not necessarily give rise to a weakly consistent sequence of posteriors. Recent papers by Barron, Schervish and Wasserman (1999) and Ghosal, Ghosh and Ramamoorthi (1999a) provide sufficient conditions on a prior to give rise to a Hellinger consistent sequence of posterior distributions. Reviews are given in Wasserman (1998), Ghosal, Ghosh and Ramamoorthi (1999b) and Ghosh and Ramamoorthi (2003).

Both the approaches of Barron, Schervish and Wasserman (1999) and Ghosal, Ghosh and Ramamoorthi (1999a) employ the Kullback–Leibler property for Π, which for the moment we can accept as a fundamental property for establishing consistency, though it is not a necessary condition. See Ghosal, Ghosh and Ramamoorthi (1999b) for a counterexample. It is also known that the Kullback–Leibler property for Π is not sufficient in itself to prove consistency. This has been shown in the paper by Barron, Schervish and Wasserman (1999).

For the second part (ii), Bayes factors have been widely studied and used as a Bayesian model selection criterion. See, for example, Bernardo and Smith (1994) for a review. However, to date, asymptotic studies of the Bayes factor have been restricted to situations when one of the competing models is assumed to be correct. This is not a relevant assumption made here.

Section 2 is concerned with Hellinger consistency of posterior distributions, Section 3 with the asymptotics of Bayes factors and Section 4 with other current areas of Bayesian asymptotic research. Proofs of theorems will appear in the Appendix.

2. HELLINGER CONSISTENCY

For Hellinger consistency, current ideas are based on splitting Ω, the set of all densities, into disjoint sets Ω_n and Ω_n^c. Here Ω_n is an increasing sequence of nested sets, a sieve, such that Ω_n^c contains all the densities which track the data too closely and are assigned low prior mass; in fact it is assumed that $\Pi(\Omega_n^c) < \exp(-n\tau)$ for some $\tau > 0$ eventually for all large n. With this, the posterior also assigns exponentially small mass to Ω_n^c. Then, if $A = \{f : h(f, f_0) > \varepsilon\}$,

$$\Pi_n(A) \le \frac{\int_{A\cap\Omega_n} R_n(f)\Pi(df)}{\int R_n(f)\Pi(df)} + \Pi_n(\Omega_n^c).$$

The denominator of the first term on the right-hand side of the inequality can be bounded below by $\exp(-nc)$ eventually for any $c > 0$, if the prior has the Kullback–Leibler property. The task then is to establish an exponential upper bound for the term

$$\int_{A\cap\Omega_n} R_n(f)\Pi(df).$$

This is the tough part, to find a suitable sieve Ω_n. For example, Wong and Shen (1995) provide an entropy condition on Ω_n as being sufficient for

$$\sup_{f\in A\cap\Omega_n} R_n(f)$$

to be exponentially small. This does the job for the Bayesian as the integral will also be bounded appropriately.

Other ideas are to be found in Barron, Schervish and Wasserman (1999) and Ghosal, Ghosh and Ramamoorthi (1999a), the latter using the notion of a uniformly consistent sequence of tests, based on original work done by Barron (1988), and the L_1-metric entropy. The result of Ghosal, Ghosh and Ramamoorthi (1999a) is more general than that of Barron, Schervish and Wasserman (1999). In this section, the result of Ghosal, Ghosh and Ramamoorthi (1999a) will be rediscovered in a purely Bayesian way.

The results of Ghosal, Ghosh and Ramamoorthi (1999a) and others give no feel for when a prior is going to give problems. To investigate exactly how a prior would need to be for consistency not to hold, once it had the Kullback–Leibler property, let us consider some predictive densities; so, let us define

$$f_n(x) = \int f(x)\Pi_n(df)$$

to be the predictive density and

$$f_{nA}(x) = \int f(x)\Pi_{nA}(df)$$

to be the predictive density with posterior restricted to the set A; that is, for $\Pi(A) > 0$,

$$\Pi_{nA}(df) = \frac{\mathbf{1}(f \in A)\Pi_n(df)}{\int_A \Pi_n(df)}.$$

Additionally, a prior Π is said to have property Q if the following holds:

- with F_0 probability 1, $h(f_{nA(\varepsilon)}, f_0) > \varepsilon$ for all n and for all $\varepsilon > 0$ when $A(\varepsilon) = \{f : h(f, f_0) > \varepsilon\}$.

The idea behind property Q is that the predictive density based on a posterior restricted to the set A, which does not include any density closer than ε to f_0 in the Hellinger sense, can never itself get closer than a distance ε to f_0. This class of prior would seem to include all reasonable ones; in fact it would be disappointing to find a prior in regular use which did not have property Q.

THEOREM 1. *If Π has the Kullback–Leibler property and property Q, then Π_n is Hellinger consistent.*

Although not required for what follows, the conclusion of the theorem still holds if we replace property Q with the following:

- with F_0 probability 1, $\liminf_n h(f_{nA(\varepsilon)}, f_0) > \varepsilon$ for all $\varepsilon > 0$.

Also, if we replace $A(\varepsilon)$ by the complement set of a weak neighborhood of f_0, then property Q becomes automatic as it is a property of all priors that for all $\varepsilon > 0$ there exists a $\lambda_\varepsilon > 0$ such that $h(f_{nA(\varepsilon)}, f_0) > \lambda_\varepsilon$ for all large n. This can then be used to prove the weak consistency result of Schwartz (1965). See Walker (2003) for further details.

To obtain the consistency result of Ghosal, Ghosh and Ramamoorthi (1999a), first consider the set of densities $B(\eta) = \{f : h(f, f_B) < \eta\}$, where f_B is a fixed density and $h(f_B, f_0) = \delta > \eta$. Then, due to the convexity of h,

$$h(f_{nB(\eta)}, f_B) \le \int h(f, f_B)\Pi_{nB(\eta)}(df) < \eta$$

and so, from the triangular inequality,

$$h(f_{nB(\eta)}, f_0) \ge h(f_B, f_0) - h(f_{nB(\eta)}, f_B) > \delta - \eta > 0$$

and so $\Pi_n(B(\eta)) \to 0$ almost surely (reproduce the proof to Theorem 1).

Now consider $\Omega_N = \bigcup_{j=1}^N B_j(\eta)$, where the $B_j(\eta) \subset \{f : h(f, f_j) < \eta\}$ are disjoint and the $\{f_j\}$ are a fixed set of densities, $h(f_j, f_0) \ge \delta$ and $\delta > \eta$. So, from the above, $\Pi_n(\Omega_N) = \sum_{j=1}^N \Pi_n(B_j(\eta)) \to 0$ almost surely as $n \to \infty$.

So, for any N and any $\{B_j\}_{j=1}^N$, for Π_n not to be consistent, densities in $A^* = \Omega_N^c \cap \{f : h(f, f_0) > \varepsilon\}$ must always average together to become close to f_0 in the Hellinger sense. That is,

$$\liminf_n h(f_{nA^*}, f_0) = 0 \quad \text{a.s.}$$

This seems highly unlikely and a prior which allows this must be quite strange. I would suggest that, to find such a prior, knowledge of f_0 would be essential. See the example presented in Barron, Schervish and Wasserman (1999) and also some of the examples presented in Walker and Hjort (2001).

To get the result similar to that of Ghosal, Ghosh and Ramamoorthi (1999a), consider now

$$\Omega_n = \Omega_{N_n} = \bigcup_{j=1}^{N_n} B_j(\eta)$$

and assume that $\Pi(\Omega_n^c) < \exp(-n\tau)$ for some $\tau > 0$ for all but finitely many n. Also, let $A_n = A \cap \Omega_n$, where $A = \{f : h(f, f_0) > \varepsilon\}$ and $\varepsilon > \eta$. Now $\Pi_n(A) \le \Pi_n(A_n) + \Pi_n(\Omega_n^c)$ and $\Pi_n(\Omega_n^c) < \exp(-nd)$ almost surely for all large n and for any $d < \tau$. See Barron, Schervish and Wasserman (1999), Lemma 5, for this

result. Then

$$\Pi_n(A_n) = \sum_{j=1}^{N_n} \Pi_n\big(A \cap B_j(\eta)\big) \le \sum_{j=1}^{N_n} \sqrt{\Pi_n(A \cap B_j(\eta))} = \sum_{j=1}^{N_n} \frac{J_{nj}}{\sqrt{\int R_n(f)\Pi(df)}},$$

where

$$J_{nj} = \sqrt{\int_{A\cap B_j(\eta)} R_n(f)\Pi(df)}.$$

Now,

$$\mathrm{pr}\left\{\sum_{j=1}^{N_n} J_{nj} > \exp(-nd)\right\} \le \exp(nd)\sum_{j=1}^{N_n} \mathrm{E}(J_{nj})$$

and $\mathrm{E}(J_{nj}) \le (1-(\varepsilon-\eta))^n$; see the proof of Theorem 1 from the Appendix. So, to ensure $\Pi_n(A_n) \to 0$ almost surely it is required that $N_n < \exp(n\mu)$ for all but finitely many n, and $\mu < \lambda$, where $\lambda = -\log(1-(\varepsilon-\eta))$. See Walker (2003) for further details. In fact,

$$\mathrm{E}(J_{nj}) \le \big(1-(\varepsilon-\eta)\big)^n \sqrt{\Pi(B_j(\eta))},$$

and this fact has been exploited in Walker (2004) to find alternative sufficient conditions for consistency. See also Section 4.5.

3. BAYES FACTORS

Here we study the asymptotic properties of

$$I_n = \int R_n(f)\Pi(df),$$

which is relevant for the study of Bayes factors. We have already seen what happens to I_n if Π has the Kullback–Leibler property. Namely, $n^{-1}\log I_n \to 0$ almost surely. This is based on $I_n > \exp(-nc)$ almost surely for all large n and for all $c > 0$ and also that $I_n < \exp(nc)$ for all large n and all $c > 0$. For the latter, just consider $\mathrm{pr}\{I_n > \exp(nc)\} < \exp(-nc)\mathrm{E}(I_n) = \exp(-nc)$.

Suppose now Π has the property that $\Pi\{f : d_K(f, f_0) < c\} > 0$ only for, and for all, $c > \delta$ for some $\delta \ge 0$. We will refer to this as the Kullback–Leibler (δ) property for Π. Then it is not difficult to show (an obvious modification of Lemma 4 from Barron, Schervish and Wasserman, 1999) that

$$\liminf_n n^{-1}\log I_n \ge -\delta \quad \text{a.s.}$$

We now introduce another "reasonable" property for Π similar to property Q. We say the prior has property Q^* if the following holds:

- $\liminf_n d_K(f_{nA(\varepsilon)}, f_0) \ge \varepsilon$ for all $\varepsilon > 0$ when $A(\varepsilon) = \{f : d_K(f, f_0) > \varepsilon\}$.

As with property Q, it would be disappointing if a prior in regular use did not possess property Q^*.

THEOREM 2. *If Π has property Q^*, the Kullback–Leibler (δ) property and*

$$\sum_n n^{-2}\,\mathrm{Var}\{\log(I_n/I_{n-1})\} < \infty,$$

then $n^{-1}\log I_n \to -\delta$ almost surely.

An illustration of this theorem is given at the end of this section. The condition $\sum_n n^{-2}\,\mathrm{Var}\{\log(I_{n+1}/I_n)\} < \infty$ is a rather weak condition and holds, for example, if $\sup_n \mathrm{Var}\{\log(I_{n+1}/I_n)\} < \infty$.

We now have access to the asymptotic properties of a Bayes factor for comparing two Bayesian models, say Π_1 and Π_2. If Π_1 has the Kullback–Leibler (δ_1) property and Π_2 has the Kullback–Leibler (δ_2) property and assuming both satisfy property Q^* and

$$\sum_n n^{-2}\,\mathrm{Var}\{\log(I_{jn}/I_{jn-1})\} < \infty,$$

where $I_{jn} = \int R_n(f)\Pi_j(df)$, then

$$n^{-1}\log B_n \to \delta_2 - \delta_1$$

almost surely. Here $B_n = I_{1n}/I_{2n}$ is the Bayes factor, and so asymptotically the Bayes factor prefers the model with the smallest δ value. This motivates priors which have $\delta = 0$ as a property, and the paper of Barron, Schervish and Wasserman (1999) contains examples of such (nonparametric) priors.

Here we provide a simple illustration of Theorem 2. Consider the model $f(x;\theta) = \theta\exp(-x\theta)$ with prior $\pi(\theta) = \exp(-\theta)$. Also assume $f_0(x) = \Gamma(a)^{-1}x^{a-1}\exp(-x)$. Then

$$d_K(f_\theta) = -\log\Gamma(a) + (a-1)\int \log x\, f_0(x)\,dx - a - \log\theta + a\theta$$

and so is minimized when $\theta = 1/a$ and hence

$$\delta = -\log\Gamma(a) + (a-1)\int \log x\, f_0(x)\,dx - a + \log a + 1.$$

Note when $a = 1$ then $\delta = 0$. Now

$$I_n = \frac{n!\exp(nS_n)\Gamma(a)^n}{(1+nS_n)^{1+n}\prod_{i=1}^n X_i^{a-1}}$$

and so

$$n^{-1}\log I_n = n^{-1}\log n! + S_n - (1+1/n)\log(1+nS_n) + \log\Gamma(a) - (a-1)n^{-1}\sum_{i=1}^{n}\log X_i.$$

Here $S_n = n^{-1}\sum_{i=1}^{n} X_i$. Now $S_n \to a$ almost surely and $n^{-1}\log n! - \log(1+n) \to -1$ and so $n^{-1}\log I_n \to -\delta$ almost surely.

It is not possible usually to write I_n and δ in the nonparametric cases. However, simulation studies can demonstrate convergence and have been undertaken in Walker, Damien and Lenk (2004).

4. OTHER BAYESIAN ASYMPTOTICS

4.1 Predictive Densities

There does not as yet appear to be a refined set of conditions for the Hellinger consistency of the predictive density $f_n = \int f\Pi_n(df)$ [i.e., $h(f_n, f_0) \to 0$ almost surely]. It is merely stated in the literature (see, e.g., Barron, Schervish and Wasserman, 1999), that the Hellinger consistency of Π_n implies the Hellinger consistency of f_n. The proof, taken from Barron, Schervish and Wasserman (1999), is as follows:

$$\begin{aligned} h(f_n, f_0) &\le \int h(f, f_0)\Pi_n(df) \\ &\le \int_{A(\varepsilon)} h(f, f_0)\Pi_n(df) \\ &\quad + \int_{A(\varepsilon)^c} h(f, f_0)\Pi_n(df), \end{aligned}$$

and $\int_{A(\varepsilon)} h(f, f_0)\Pi_n(df) < \Pi_n(A(\varepsilon))$, which goes to zero almost surely, and also $\int_{A(\varepsilon)^c} h(f, f_0)\Pi_n(df) < \varepsilon$. Since ε is arbitrary it must be that $h(f_n, f_0) \to 0$ almost surely.

The Kullback–Leibler property for Π is, however, tantalizingly close to being able to establish the Hellinger consistency of f_n. It is shown in Walker (2003) that if Π has the Kullback–Leibler property, then $h(f^N, f_0) \to 0$ almost surely, where

$$f^N = N^{-1}\sum_{n=1}^{N} f_n.$$

Again, it is evident that if f^N is Hellinger consistent, and yet f_n is not, then Π should be somewhat strange.

4.2 Rates of Convergence

This is a fast-moving area of current research. The idea is to find bounds for $\Pi_n(A_n^c)$, where A_n is a shrinking neighborhood of f_0. Two recent articles written on this subject are by Ghosal, Ghosh and van der Vaart (2000) and Shen and Wasserman (2001). At the moment the story for rates of convergence is mixed. According to Shen and Wasserman (2001), "Although it is too early to draw general conclusions, it appears that the choice of prior in an infinite dimensional problem is more difficult if one wants to achieve good rates."

4.3 Nonparametric and Semiparametric Regression Models

The most visible successes for Bayesian asymptotics and consistency have been with independent and identically distributed data. Few semiparametric regression models have been studied. Exceptions to this are the binary regression model considered by Diaconis and Freedman (1993, 1995), $\mathrm{E}(X_i) = \eta(\xi_i)$, where η is the unknown function to be estimated. With the prior for η considered by Diaconis and Freedman, consistency is the rule, though $\eta_0 \equiv \frac{1}{2}$ needs special attention.

Diaconis and Freedman (1998) also studied a normal regression model, as did Shen and Wasserman (2001). The model of Shen and Wasserman (2001) is given by

$$X_i = \eta(\xi_i) + \varepsilon_i,$$

where the ξ_i are independent uniform random variables from $[0, 1]$ and the ε_i are independent and identically distributed as standard normal.

A semiparametric regression model was recently considered by Shen (2002). The model assumes

$$X_i = \theta\xi_i + \eta(\zeta_i) + \varepsilon_i,$$

where now (ξ_i, ζ_i) are independently uniformly distributed over $[0, 1]^2$. Shen regards η as a (nonparametric) nuisance parameter and studies the asymptotic marginal posterior distribution of θ.

4.4 Priors on Distribution Functions

The most well-known prior on distribution functions is the Dirichlet process (Ferguson, 1973). This is a special case of both the neutral priors of Doksum (1974) and Pólya trees (Kraft, 1964; Ferguson, 1974). Pólya trees are *tailfree*. A prior is tailfree with respect to a nested sequence of partitions $C_k = \{B_\varepsilon; \varepsilon \in \{0, 1\}^k\}$, that is, B_ε splits into $B_{\varepsilon 0}$ and $B_{\varepsilon 1}$, if $F(B_0)$, $F(B_{00}|B_0)$, $F(B_{10}|B_1), \ldots$ are all independent. Here

$F(B)$ is the random mass allocated to the set B for a random distribution F chosen from the prior. For Pólya trees, $F(B_{\varepsilon 0}|B_\varepsilon)$ are beta distributions. It is known (see, e.g., Ghosal, Ghosh and Ramamoorthi, 1999b) that tailfree priors give rise to a weakly consistent sequence of posterior distributions.

Neutral priors are not tailfree and do not admit densities and so other techniques to demonstrate consistency are required for these priors. Such work has been done by Kim and Lee (2001). They use the Hjort (1990) parameterization of neutral to the right processes,

$$A(t) = \int_0^t \frac{dF(s)}{1 - F(s-)},$$

where F is a neutral to the right process, and find sufficient conditions for which A given the data, written as A_n, converges weakly to A_0 in $D[0, \tau]$. This is achieved by finding sufficient conditions under which $\mathrm{E}(A_n(t)) \to A_0(t)$ for all $t \in [0, \tau]$ and $\mathrm{Var}(A_n(t)) \to 0$ for all $t \in [0, \tau]$.

4.5 Further Developments

A different approach to Hellinger consistency has recently been found by Walker (2004). The Kullback–Leibler property for the prior is retained. The further condition requires the separability of the space of densities with respect to the Hellinger metric. Let $\{B_j\}$ be disjoint subsets of Hellinger balls around the densities $\{f_j\}$ and $\Omega = \bigcup_j B_j$. If

$$\sum_j \sqrt{\Pi(B_j)} < \infty,$$

then Π_n is Hellinger consistent. The proof is elementary and the condition is often straightforward to use when applied to particular priors.

4.6 Discussion

As is evident from the difficulties in understanding and dealing with consistency issues of posteriors associated with independent and identically distributed data and the type of models discussed in Section 4.3, obtaining general Bayesian consistency theorems for semi-parametric and nonparametric regression models is going to be very hard. Rates of convergence are also going to be challenging.

With respect to the Hellinger consistency of posterior distributions, it can be regarded that the Kullback–Leibler property for the prior is practically sufficient, provided a sensible prior is being used (with property Q). As for Bayes factors, a prior with the Kullback–Leibler property will beat all other models. The fact that this is an asymptotic result is not relevant in this case. A decision maker would if allowed compute a Bayes factor with as large a dataset as possible. So if he or she knew that, for large datasets, a particular model would always come out on top, then surely he or she would select this model. This motivates the use of priors with the Kullback–Leibler property. See Walker, Damien and Lenk (2004) for further discussion and examples of this point.

APPENDIX

PROOF OF THEOREM 1. The proof is not complicated at all. Let $A = A(\varepsilon) = \{f : h(f, f_0) > \varepsilon\}$. A key identity is given by

$$\int_A R_{n+1}(f)\Pi(df) = \frac{f_{nA}(X_{n+1})}{f_0(X_{n+1})} \int_A R_n(f)\Pi(df).$$

We then define

$$J_n = \sqrt{\int_A R_n(f)\Pi(df)},$$

so that

$$\begin{aligned}\mathrm{E}(J_{n+1}|\mathcal{F}_n) &= J_n \int \sqrt{f_0\, f_{nA}} \\ &= J_n\{1 - h(f_{nA}, f_0)\} \le J_n(1 - \varepsilon),\end{aligned}$$

where $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. The numerator for $\Pi_n(A)$ is J_n^2 and the above gives $J_n < \exp(-nd)$ almost surely for all large n and for all $d < -\log(1 - \varepsilon)$. The denominator of $\Pi_n(A)$ is $I_n = \int R_n(f)\Pi(df)$, which with the Kullback–Leibler property is bounded below by $\exp(-nc)$ almost surely for all large n and for all $c > 0$. Then pick $c < d$. □

PROOF OF THEOREM 2. The key to the proof is the martingale sequence

$$S_N = \sum_{n=1}^N \{\log(I_n/I_{n-1}) + d_K(f_{n-1}, f_0)\},$$

which is a martingale by virtue of $\mathrm{E}\{\log(I_n/I_{n-1})|\mathcal{F}_{n-1}\} = -d_K(f_{n-1}, f_0)$. For such martingales it is known that if

$$\sum_n n^{-2}\,\mathrm{Var}\{\log(I_n/I_{n-1})\} < \infty,$$

then $S_N/N \to 0$ almost surely. Therefore,

$$N^{-1}\log I_N + N^{-1}\sum_{n=1}^N d_K(f_{n-1}, f_0) \to 0$$

almost surely. With property Q^* for Π it follows that

$$\liminf_N N^{-1} \sum_{n=1}^{N} d_K(f_{n-1}, f_0) \geq \delta$$

almost surely and hence $\limsup_N N^{-1} \log I_N \leq -\delta$ almost surely With the Kullback–Leibler (δ) property for Π we have $\liminf_N N^{-1} \log I_N \geq -\delta$ almost surely and hence

$$N^{-1} \log I_N \to -\delta \quad \text{a.s.} \qquad \square$$

ACKNOWLEDGMENT

The research of the author is funded by an Engineering and Physical Sciences Research Council Advanced Research Fellowship.

REFERENCES

BARRON, A. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Unpublished manuscript.

BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561.

BERNARDO, J. M. and SMITH, A. F. M. (1994). *Bayesian Theory*. Wiley, New York.

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

DIACONIS, P. and FREEDMAN, D. (1993). Nonparametric binary regression: A Bayesian approach. *Ann. Statist.* **21** 2108–2137.

DIACONIS, P. and FREEDMAN, D. (1995). Nonparametric binary regression with random covariates. *Probab. Math. Statist.* **15** 243–273.

DIACONIS, P. and FREEDMAN, D. (1998). Consistency of Bayes estimates for nonparametric regression: Normal theory. *Bernoulli* **4** 411–444.

DOKSUM, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2** 183–201.

DRAPER, D. (1999). Discussion of "Bayesian nonparametric inference for random distributions and related functions," by S. G. Walker, P. Damien, P. Laud and A. F. M. Smith. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 510–513.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629.

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999a). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143–158.

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999b). Consistency issues in Bayesian nonparametrics. In *Asymptotics, Nonparametrics and Time Series*: *A Tribute to Madan Lal Puri* (S. Ghosh, ed.) 639–667. Dekker, New York.

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. and RAMAMOORTHI, R. V. (2003). *Bayesian Nonparametrics.* Springer, New York.

HJORT, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18** 1259–1294.

KIM, Y. and LEE, J. (2001). On posterior consistency of survival models. *Ann. Statist.* **29** 666–686.

KRAFT, C. (1964). A class of distribution function processes which have derivatives. *J. Appl. Probability* **1** 385–388.

LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20** 1222–1235.

LAVINE, M. (1994). More aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **22** 1161–1176.

LENK, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric predictive densities. *J. Amer. Statist. Assoc.* **83** 509–516.

LENK, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika* **78** 531–543.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357.

MAULDIN, R. D., SUDDERTH, W. D. and WILLIAMS, S. C. (1992). Pólya trees and random distributions. *Ann. Statist.* **20** 1203–1221.

SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26.

SHEN, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.* **97** 222–235.

SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714.

WALKER, S. G. (2003). On sufficient conditions for Bayesian consistency. *Biometrika* **90** 482–488.

WALKER, S. G. (2004). New approaches to Bayesian consistency. *Ann. Statist*. To appear.

WALKER, S. G., DAMIEN, P. and LENK, P. J. (2004). On priors with a Kullback–Leibler property. *J. Amer. Statist. Assoc*. **99** 404–408.

WALKER, S. G. and HJORT, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B Stat. Methodol*. **63** 811–821.

WASSERMAN, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statist.* **133** 293–304. Springer, New York.

WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362.