

On Bayesian consistency

Stephen Walker

University of Bath, UK

and Nils Lid Hjort

University of Oslo, Norway

[Received April 2000. Revised May 2001]

Summary. We consider a sequence of posterior distributions based on a data-dependent prior (which we shall refer to as a pseudoposterior distribution) and establish simple conditions under which the sequence is Hellinger consistent. It is shown how investigations into these pseudoposteriors assist with the understanding of some true posterior distributions, including Pólya trees, the infinite dimensional exponential family and mixture models.

Keywords: Asymptotics; Bayesian sieve; Bayes nonparametrics; Consistency

1. Introduction

Asymptotics play an important role in statistics. In classical density estimation this role is crucial, providing justifications for a wide range of nonparametric estimators such as kernel-based estimators and sieve maximum likelihood estimators (Shen and Wong, 1994; Wong and Shen, 1995) and other nonparametric estimators (see, for example, van de Geer (1993)). Establishing consistency and rates of convergence with respect to a suitable metric, typically the Hellinger distance, is the key point to this area of research.

However, Bayesian nonparametric methods have only recently started to undergo asymptotic studies. Early work was done by Schwartz (1965) who established that a prior which puts positive mass on all Kullback–Leibler neighbourhoods of the true density is weakly consistent. Previously, Freedman (1963), and later Diaconis and Freedman (1986) in a more natural setting, had demonstrated that priors which put positive mass on all weak neighbourhoods of the true distribution function are not necessarily weakly consistent. Recent attention has switched to studying and finding sufficient conditions for strong (Hellinger) consistency.

Suppose that Π is a prior distribution on the set of all probability densities over an interval or region of interest. As data $x_1, x_2, \dots$ accumulate from some unknown underlying density f_0 , will the Bayesian posterior distribution $\Pi^n(\cdot) = \Pi(\cdot | x_1, \dots, x_n)$ concentrate around this f_0 ? Barron *et al.* (1999) presented one such Bayesian nonparametric consistency theorem; the corresponding theorem of Ghosal *et al.* (1999a) is of a similar nature. Barron *et al.* (1999) made two assumptions to prove consistency in the Hellinger metric, i.e.

$$\Pi^n\{f: H(f) > \epsilon\} \rightarrow 0 \quad \text{almost surely}$$

Address for correspondence: Stephen Walker, Department of Mathematical Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, UK.
E-mail: massgw@maths.bath.ac.uk

where

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

is the Hellinger distance. The first is that the prior puts positive mass on all Kullback–Leibler neighbourhoods of the true density and the second is a combined smoothness and tail condition involving a *sieve* and a set of *upper brackets*.

For our purposes, for each positive δ , a sieve is a sequence $\mathcal{F}_n(\delta)$ of sets of densities which are well behaved, in the sense that they are not too ‘wiggly’. This can be made more formal via a set of upper brackets, i.e. for every $f \in \mathcal{F}_n(\delta)$ there is an f_k^U from the set $\{f_1^U, \dots, f_{N_n}^U\}$ such that $\int f_k^U < 1 + \delta$ and $f \leq f_k^U$ (hence upper bracket). The assumptions for posterior Hellinger consistency required by Barron *et al.* (1999) are that the number N_n of brackets for $\mathcal{F}_n(\delta)$ does not increase at a rate greater than $\exp(nc)$ for all but finitely many n , for some $c > 0$, and that the prior mass for the complement set $\mathcal{F}_n'(\delta)$ is exponentially small. There is a need to put low prior mass on the wiggly densities in $\mathcal{F}_n'(\delta)$, which is explained further in Section 2. See also Geman and Hwang (1982) for the use of sieves in nonparametric maximum likelihood estimation problems.

Wasserman (1998) has provided a review of Hellinger consistency. Shen and Wasserman (2001) and Ghosal *et al.* (2000) provide rates of convergence.

Consequently, given a nonparametric prior Π , after having ascertained that it satisfies the rather strict requirements of the Barron *et al.* (1999) type, we can generate $\{\Pi^n\}$ knowing that the sequence is, with probability 1, Hellinger consistent. Suppose that, with the same prior Π , it is possible to generate another sequence of probability distributions, say $\{Q^n\}$, which is also Hellinger consistent. Moreover, suppose that the conditions on Π for the $\{Q^n\}$ sequence to be consistent are significantly less restrictive than those needed on Π for the $\{\Pi^n\}$ sequence to be consistent. If the extra conditions needed for the $\{\Pi^n\}$ sequence to be consistent are difficult to verify or not established, then it may be preferable to use $\{Q^n\}$ for inference. This is particularly appropriate in the nonparametric context where Π is typically difficult to construct and incorporate real qualitative information. Hence an objective procedure is preferable in such contexts.

In this paper, it is shown that if Π puts positive mass on all Kullback–Leibler neighbourhoods of f_0 then there exists a Hellinger consistent sequence of such pseudoposterior distributions $\{Q^n\}$.

Let $x^n = (x_1, \dots, x_n)$ denote the data of sample size n , i.e. $x_1, x_2, \dots \sim_{\text{IID}} f_0$, where f_0 is the true density, with corresponding probability distribution F_0 . Also, write

$$D(f) = \int f_0 \log(f_0/f)$$

for the Kullback–Leibler divergence from f_0 to f , and let

$$A_\epsilon = \{f: H(f) > \epsilon\},$$

$$K_\eta = \{f: D(f) \leq \eta\}$$

and

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

We take integrals to be with respect to the Lebesgue measure over the interval for which the densities are defined, for concreteness, although generalizations are easily covered.

In Section 2 we introduce a sequence of pseudoposteriors which give rise to a Hellinger consistent sequence of estimators for f . Section 3 considers some illustrations where the aim is to show how the pseudoposteriors assist with the understanding of consistency for the true posteriors. The result of Barron *et al.* (1999) is general, applying to all priors, and hence might for several special classes of prior be requiring more than is actually necessary for consistency.

2. A consistent sequence of distributions

Given a prior Π on a space of probability densities, the true Bayesian posterior distribution is given by

$$\Pi^n(df) = R_n(f) \Pi(df) / \int R_n(f) \Pi(df).$$

Consistency for the sequence $\{\Pi^n\}$ is not guaranteed under the condition, which we now refer to as condition (A), that Π puts positive mass on all Kullback–Leibler neighbourhoods of f_0 . Barron *et al.* (1999) presented a counter-example.

Define the pseudoposterior distribution based on Π as

$$Q^n(df) = R_n^{1/2}(f) \Pi(df) / \int R_n^{1/2}(f) \Pi(df).$$

We can view this in one of two ways: we are using the pseudolikelihood function

$$\mathcal{L}_n(f) \propto \prod_{i=1}^n f^{1/2}(x_i),$$

which is the usual likelihood square rooted; alternatively (our preferred interpretation), we are using the data-dependent prior

$$\Pi_n(df) \propto \Pi(df) / \prod_{i=1}^n f^{1/2}(x_i).$$

Wasserman (2000) used a different pseudolikelihood function–data-dependent prior to establish nice asymptotic properties for mixture models. As with the data-dependent prior of Wasserman (2000), it is the asymptotic properties of the posteriors which justify its use.

It can be shown that Q^n is proper, i.e.

$$\int R_n^{1/2}(f) \Pi(df) < \left\{ \int R_n(f) \Pi(df) \right\}^{1/2} < \infty,$$

a consequence of lemma 1 of Barron *et al.* (1999) which can also be used to show that $\int R_n^{1/2}(f) \pi(df) \neq 0$.

We now prove that, if Π satisfies condition (A), then $\{Q^n\}$ is, with probability 1, Hellinger consistent. The reason why the prior works from an intuitive point of view is that we can write

$$\Pi_n(df) \propto \exp\{\frac{1}{2}n D_n(f)\} \Pi(df),$$

where $D_n(f) = n^{-1} \sum_{i=1}^n \log\{f_0(x_i)/f(x_i)\}$. The problems of consistency with Π^n can be traced to densities for which $D_n(f) < 0$ have too much weight. Such densities are being assigned low, and sufficiently low, weight in the data-dependent prior that they do not cause a problem in the posterior. In this respect, the prior can be viewed as a Bayesian sieve which downweights sufficiently, rather than removes altogether as in a sieve maximum likelihood estimator, the troublesome densities which make $R_n(f)$ too large, not because they are good densities but rather because they track the data too closely. Our approach is to use the data to downweight the prior. Barron *et al.* (1999) imposed stronger restrictions on Π to achieve the same effect.

Theorem 1. Under assumption (A), with F_0 probability 1, $Q^n(A_\epsilon) \rightarrow 0$ with exponential rate, as $n \rightarrow \infty$ for all sets A_ϵ with $\epsilon > 0$.

Proof. We can write $Q^n(A_\epsilon)$ as

$$Q^n(A_\epsilon) = \int_{A_\epsilon} R_n^{1/2}(f) \Pi(df) / \int R_n^{1/2}(f) \Pi(df).$$

The denominator can be written as

$$L_n = \int \exp\{-\frac{1}{2}n D_n(f)\} \Pi(df).$$

Thus, for any $n > 0$,

$$\exp(n\eta)L_n > \int_{K_{2\eta}} \exp[\frac{1}{2}n\{2\eta - D_n(f)\}] \Pi(df).$$

Arguments laid out by Barron *et al.* (1999) (lemmas 3 and 4), based on Fatou's lemma and condition (A), establish that $L_n > \exp(-n\eta)$ almost surely for large n for all $\eta > 0$.

For the numerator, $U_n = \int_{A_\epsilon} R_n^{1/2}(f) \Pi(df)$,

$$\begin{aligned} \text{pr}_{x^n} \left\{ \int_{A_\epsilon} R_n^{1/2}(f) \Pi(df) > \exp(-nc) \right\} &< \exp(nc) \int_{A_\epsilon} \left\{ \int \sqrt{(ff_0)} \right\}^n \Pi(df) \\ &< \exp(nc) \int_{A_\epsilon} \{1 - \frac{1}{2}H(f)^2\}^n \Pi(df) \\ &< \exp(nc - \frac{1}{2}n\epsilon^2). \end{aligned}$$

Thus, choosing $c < \frac{1}{2}\epsilon^2$, the Borel–Cantelli theorem gives that $U_n < \exp(-nc)$ almost surely for large n for any $c < \frac{1}{2}\epsilon^2$. Consequently, we can choose $\eta < c$ and thus $Q_n(A_\epsilon) < \exp(-n\delta)$ almost surely for large n for any $\delta < \frac{1}{2}\epsilon^2$, completing the proof.

The pseudo-Bayes estimator based on the sequence $\{Q^n\}$ is given by

$$f^n(x) = \int f(x) Q^n(df).$$

Here we establish that $f^n \rightarrow f_0$ almost surely with respect to the Hellinger distance.

Theorem 2. $H(f^n) < \epsilon$ almost surely for large n for any $\epsilon > 0$ and hence $H(f^n) \rightarrow 0$ almost surely.

Proof. Using corollary 1 from Barron *et al.* (1999),

$$H^2(f^n) \leq \int H^2(f) Q^n(df) = \int_{A_\epsilon} H^2(f) Q^n(df) + \int_{A_\epsilon^c} H^2(f) Q^n(df).$$

Now $H^2(\cdot) \leq 2$ so

$$H^2(f^n) \leq 2 Q^n(A_\epsilon) + \epsilon^2,$$

completing the proof.

Convergence rates for $H(f^n)$ can be established by using ideas from Shen and Wasserman (2001). Let t_n be as in lemma 2 and theorem 2 of Shen and Wasserman (2001). Suppose that there is a sequence c_n such that $nc_n \rightarrow \infty$, $c_n \rightarrow 0$ and $c_n \geq 16t_n$. Then

$$H^2(f^n) \leq 2 \exp(-\frac{1}{2}nc_n) + \epsilon_n^2$$

almost surely for large n , for all sequences $\{\epsilon_n\}$ such that

$$\sum_n \exp\{-n(\frac{1}{2}\epsilon_n^2 - c_n)\} < \infty.$$

This result is based on $\int_{A_{\epsilon_n}} R_n^{1/2}(f) \Pi(df) < \exp(-nc_n)$ almost surely for large n . The rate therefore depends on the relative sizes of ϵ_n and t_n .

3. Illustrations

In this section we shall look at several priors and use the consistency of Q^n to help us to establish results for Π^n . The result of Barron *et al.* (1999) is a sufficient condition for Hellinger consistency for any prior. It is not a necessary condition and so the sufficient conditions may be more than is required in some special cases. Here we consider specific priors, including those considered by Barron *et al.* (1999) and Ghosal *et al.* (1999a), and using Q^n establish modified sufficient conditions for the Hellinger consistency of Π^n .

3.1. Pólya trees

We consider Pólya trees on $[0, 1]$ with partition structure the dyadic intervals. This was an example considered by Barron *et al.* (1999). Pólya trees generalize the well-known and widely used Dirichlet process and have seen applications in Lavine (1992) and Walker and Mallick (1997) among others.

For each interval in the dyadic system we allocate a random variable $0 < V_{jk} < 1$, $k = 1, 2, \dots$ and $j = 1, \dots, 2^k$. If j is odd then $V_{j+1k} = 1 - V_{jk}$ and the $\{V_{jk}\}$ for j odd are mutually independent. Define the random probability measure F by

$$F(B_{jk}) = \prod_{l=1}^k V_{l(j)l}$$

and B_{jk} is the j th dyadic interval (from left to right) at level k . Here $B_{l(j)l}$, for $l = 1, \dots, k$, are the unique sequence of dyadic intervals which leads to B_{jk} .

As did Barron *et al.* (1999), we assume that $V_{jk} \sim \text{be}(a_k, a_k)$ for all odd j . Kraft (1964) established that if $\sum_k a_k^{-1} < \infty$ then F is a random probability measure which has a density with respect to the Lebesgue measure on $[0, 1]$. If the Kullback–Leibler divergence between f_0 and the prior predictive is finite and $\sum_k a_k^{-1/2} < \infty$, collectively known as condition (B), then

the Pólya tree prior puts positive mass on all Kullback–Leibler neighbourhoods of f_0 . See Ghosal *et al.* (1999b). Assume that condition (B) holds.

Under our data-dependent prior, for which we shall use a superscript Q , the posterior for the V_{jk}^Q are given, for odd j , by

$$V_{jk}^Q \sim \text{be}(a_k + n_{jk}/2, a_k + n_{j+1k}/2).$$

Hence,

$$E(V_{jk}^Q) = \frac{a_k + n_{jk}/2}{2a_k + n_{jk}/2 + n_{j+1k}/2} = \frac{2a_k + n_{jk}}{4a_k + n_{jk} + n_{j+1k}}$$

which is clearly equal to $E(V_{jk}^\Pi)$, where V_{jk}^Π are obtained as the true posterior based on a Pólya tree prior with parameters $2a_k$.

Consequently, the pseudopredictive density

$$f^n(x) = \lim_{k \rightarrow \infty} \left\{ 2^k \prod_{l=1}^k E(V_{l(x)l}^Q) \right\},$$

i.e.

$$f^n(x) = \lim_{k \rightarrow \infty} \left(2^k \prod_{l=1}^k \frac{2a_l + n_{l(x)l}}{4a_l + n_{l-1(x)l-1}} \right),$$

based on the data-dependent Pólya tree prior with parameters a_k is equivalent to the true predictive density based on a Pólya tree prior with parameters $2a_k$. This indicates that although the posterior distribution of a Pólya tree prior may not be consistent under condition (B) (Barron *et al.* (1999) established $a_k = 8^k$ as being sufficient for this) the predictive density is consistent under condition (B). Note that this requires much less than the $a_k = 8^k$ condition.

3.2. Infinite dimensional exponential family

Here we discuss an application involving the infinite dimensional exponential family on $[0, 1]$. Barron *et al.* (1999) also considered this example in section 3.3 of their paper. Original work on these families was done by Leonard (1978) and Lenk (1988, 1991) who demonstrated their usefulness and application. Let $\Psi = \{\psi_j\}_{j=1}^\infty$ be a set of independent normal random variables with zero means and variances $\{\tau_j^2\}$ and $\{\phi_j\}_{j=1}^\infty$ a set of orthogonal polynomials on $[0, 1]$. Then a random density chosen from the prior $\Pi(\Psi)$ is given by

$$f(x) = \exp \left\{ \sum_{j=1}^\infty \psi_j \phi_j(x) - c(\Psi) \right\}$$

where

$$\exp\{c(\Psi)\} = \int \exp \left\{ \sum_{j=1}^\infty \psi_j \phi_j(x) \right\} dx.$$

Barron *et al.* (1999) established the conditions $\sum_j a_j \tau_j < \infty$ and $\sum_j b_j \tau_j < \infty$, where $a_j = \sup_x |\phi_j(x)|$ and $b_j = \sup_x |\phi_j'(x)|$, as being sufficient for the consistency of $\Pi^n(\Psi)$.

Here we consider the more general version of the prior considered by Lenk (1988, 1991). Let $f \sim \text{LNS}(\mu, \sigma, \xi)$, so $f(x) \propto W_\xi(x)$ where $W_\xi(\cdot)$ is a generalized log-normal process with distribution Λ_ξ characterized by

$$\Lambda_{\xi}(A) \propto \int_A \left\{ \int W(x) dx \right\}^{\xi} d\Lambda(W)$$

and Λ represents a log-normal process, i.e. if $W \sim \Lambda$ then $W(x) = \exp\{Z(x)\}$, $Z(\cdot)$ is a Gaussian process with mean $\mu(x)$, and $\sigma(x, y)$ is the covariance of $Z(x)$ and $Z(y)$. See Lenk (1988) for further details. Then the true posterior for f is given by LNS(μ_n, σ_n, ξ_n) where $\mu_n(x) = \mu(x) + \sum_i \sigma(x, x_i)$, $\sigma_n = \sigma$ and $\xi_n = \xi - n$. The posterior for Q , denoted by Λ_Q^n , is characterized via

$$\Lambda_Q^n(A) \propto \int_A \prod_i W(x_i)^{1/2} \left\{ \int W(x) dx \right\}^{\xi-n/2} d\Lambda(W),$$

and hence it is seen that Λ_Q^n is the true posterior based on the sample-size-dependent prior $f \sim \text{LNS}(\mu, \sigma/2, \xi + n/2)$. This result follows from theorems 1 and 2 of Lenk (1988). By putting $\xi = -n/2$ we obtain the prior of Barron *et al.* (1999), i.e. $f \sim \text{LNS}(\mu, \sigma/2, 0)$. See Lenk (1991) for this connection. Consequently, provided that $\text{LNS}(\mu, \sigma/2, 0)$ satisfies condition (A), the sequence of posterior distributions is Hellinger consistent.

3.3. Parametric families

Let $\mathcal{F} = \{f(x; \theta); \theta \in \Theta\}$ be a family of densities with respect to Lebesgue measure and suppose that $x_1, \dots, x_n$ are independent and identically distributed from $f(x; \theta_0)$ with $\theta_0 \in \Theta$. We assume that $\hat{f} = f_{\hat{\theta}}$, the maximum likelihood estimator, exists. Let Π be a prior probability on Θ and define

$$L_n(\theta) = \prod_{i=1}^n \frac{f(x_i; \theta)}{f(x_i; \theta_0)}.$$

Theorem 3. If $\Pi(K_{\eta}) > 0$ for all $\eta > 0$ then

$$\Pi^n(A_{\epsilon}) \leq \exp(-nc) L_n^{1/2}(\hat{\theta})$$

almost surely for large n for any $\epsilon > 0$ and $c < \frac{1}{2}\epsilon^2$.

Proof. We define $Q^n(\theta) \propto L_n^{1/2}(\theta) \Pi(\theta)$ and from results established in Section 2 we know that $Q^n(A_{\epsilon}) < \exp(-n\delta)$ almost surely for large n for $\delta < \frac{1}{2}\epsilon^2$. It is easy to see that $\Pi^n(\theta) \propto L_n^{1/2}(\theta) Q^n(\theta)$ and, more precisely,

$$\Pi^n(\theta) = L_n^{1/2}(\theta) Q^n(\theta) \frac{\int L_n^{1/2}(\theta) \Pi(d\theta)}{\int L_n(\theta) \Pi(d\theta)}.$$

Now

$$\frac{\int L_n^{1/2}(\theta) \Pi(d\theta)}{\int L_n(\theta) \Pi(d\theta)} \leq \left\{ \frac{1}{\int L_n(\theta) \Pi(d\theta)} \right\}^{1/2}$$

and the denominator is bounded below almost surely for large n by $\exp(-n\eta)$ for any $\eta > 0$. Thus, using the consistency result for Q^n , we have

$$\Pi^n(A_c) < \exp(\tfrac{1}{2}n\eta) \exp(-n\delta) L_n^{1/2}(\hat{\theta})$$

and hence the result.

Consequently, the consistency of Π^n is guaranteed if the $L_n(\hat{\theta})$ is well behaved. Conditions and special cases for this were studied by van de Geer (1993). If

$$g(\theta) = \sqrt{\{f(x; \theta)/f(x; \theta_0)\}} - 1$$

and

$$\sup_{\theta} |g(\theta) d(F_n - F_{\theta_0})| \rightarrow 0 \quad \text{almost surely}$$

where F_n is the empirical distribution function, then

$$\limsup_n \left[\frac{1}{n} \log \{L_n^{1/2}(\hat{\theta})\} \right] \leq 0 \quad \text{almost surely}$$

and the posterior consistency of Π^n holds. van de Geer (1993), theorem 2.4, provides an entropy condition as being sufficient for the above uniform law of large numbers result to be true.

3.4. Mixture model

In this section we consider the case when $f(x) = \int \phi_h(x - \theta) dP(\theta)$, where ϕ_h is a kernel density with bandwidth h and P is a random probability distribution. This is the model considered by Ghosal *et al.* (1999a) who considered a Dirichlet prior for P and took ϕ_h to be the normal density with standard deviation h . A prior on $(0, \infty)$ is also assigned to h . This is a version of the celebrated mixture of Dirichlet process model, which has been widely used in Bayesian nonparametric problems. See, for example, Escobar and West (1995).

We let Π denote the prior for P and π_n the prior for h , allowing the latter to be sample size dependent. Following Ghosal *et al.* (1999a) we write $f_{h,P}$ to denote a random density $f(x) = \int \phi_h(x - \theta) dP(\theta)$. We shall also use the normal density for ϕ . We define

$$g_{h,P}(x) = \left\{ \int \phi_h^{1/2}(x - \theta) dP(\theta) \right\}^2$$

and note that $g_{h,P} \leq f_{h,P}$. Now let us consider, for any set A ,

$$\Pi(A|x^n) = \frac{\int_A R_n(f_{h,P}) \pi_n(h) dh \Pi(dP)}{\int R_n(f_{h,P}) \pi_n(h) dh \Pi(dP)}.$$

The numerator can be written as

$$U_n = \int_A \left\{ \prod_i f_0(x_i)^{-1} \prod_i \int \phi_h(x_i - \theta) dP(\theta) \right\} \pi_n(h) dh \Pi(dP),$$

and using the fact that $\phi_h(z) = \phi_{h/\sqrt{2}}^{1/2}(z) h^{-1/2} \kappa$, where $\kappa^{-1} = \pi^{1/4} \sqrt{2}$, we have

$$U_n \leq \kappa^n \prod_i f_0(x_i)^{-1/2} \int_A R_n^{1/2}(f_{h/\sqrt{2},P}) h^{-n/2} \pi_n(h) dh \Pi(dP).$$

The denominator can be written as

$$L_n = \kappa^n \prod_i f_0(x_i)^{-1/2} \int R_n^{1/2}(g_{h/\sqrt{2}, P}) h^{-n/2} \pi_n(h) dh \Pi(dP).$$

We use the sample-size-dependent prior $\pi_n(h) \propto h^{n/2} \pi(h)$, where $\pi(h)$ is a proper probability density for h , and hence to ensure that $\pi_n(h)$ is a density we require that

$$\int h^\xi \pi(h) dh < \infty \quad \text{for all } \xi > 0.$$

Consequently,

$$\Pi(A|x^n) \leq \frac{\int_A R_n^{1/2}(f_{h/\sqrt{2}, P}) \pi(h) dh \Pi(dP)}{\int R_n^{1/2}(g_{h/\sqrt{2}, P}) \pi(h) dh \Pi(dP)}.$$

If $A_\epsilon = \{(h, P): H(f_{h/\sqrt{2}, P}) > \epsilon\}$ then we know from previous results that the new numerator $U_n \leq \exp(-nc)$ almost surely for large n for any $c < \frac{1}{2}\epsilon^2$. To apply previous results to the denominator, i.e. to ensure that the new denominator $L_n > \exp(-n\delta)$ almost surely for large n for arbitrary $\delta > 0$, we require that

$$\Pi^*\{D(g_{h/\sqrt{2}, P}) < \eta\} > 0$$

for all $\eta > 0$, where $\Pi^*(h, dP) = \pi(h) \Pi(dP)$. Clearly this condition, combined with $\pi_n(h)$ being a proper density for all n , i.e. $\int \pi(h) h^\xi dh < \infty$ for all $\xi > 0$, is a sufficient condition for the Hellinger consistency of posteriors based on the sample-size-dependent prior $\pi_n(h, dP) = \pi_n(h) \Pi(dP)$.

4. Discussion

If the likelihood values are well behaved and maximum likelihood estimators exist then the posterior distributions are consistent, the additional requirement for the Bayesian being condition (A). The problem with models for which maximum likelihood estimators exist is that condition (A) can only be verified for a restricted class of f_0 , i.e. $f_0(\cdot) \in \{f(\cdot; \theta); \theta \in \Theta\}$. Hence, satisfying condition (A) and the non-existence of a maximum likelihood estimator usually go together. A classical solution is the sieve maximum likelihood estimator. Barron *et al.* (1999) presented a Bayesian solution which places extra conditions on Π , the aim being to put low prior mass on densities which can track the data too closely. The solution proposed in this paper uses the data themselves to downweight these troublesome densities in the prior.

If one accepts that Bayesian consistency is important, for which we believe there is increasing support, then prior distributions need to be constructed appropriately. For Hellinger consistency, the current sufficient conditions of the Barron *et al.* (1999) type appear quite difficult to verify. We have shown how to construct a Hellinger consistent sequence of pseudoposterior distributions based on a rather mild regularity condition which avoids the need of a sieve. Moreover, our work, as we have demonstrated in Section 3, sheds light on and simplifies the conditions for Hellinger consistency of standard nonparametric priors, such as those considered by Barron *et al.* (1999) and Ghosal *et al.* (1999a).

If we take the interpretation of our pseudoposteriors as a standard non-data-dependent

prior combined with a square-rooted likelihood, then we can understand the behaviour of Q^n for finite samples. Square-rooting has the consequence of flattening the likelihood, effectively downweighting the influence of the data in the posterior. From this perspective the Q^n can be thought of as *robustifying* posterior distributions. With Π^n the posterior distributions will be sharper but, without having verified the conditions for consistency, they could be sitting in the wrong place.

A more general data-modified prior to work with would take the form

$$\Pi(df) \Big/ \prod_{i=1}^n f^\alpha(x_i),$$

where $0 < \alpha < 1$; this would also correspond to a pseudolikelihood

$$\mathcal{L}_n(f) = \prod_{i=1}^n f^{1-\alpha}(x_i).$$

Our choice $\alpha = \frac{1}{2}$ agrees nicely with the Hellinger distance and gives satisfactory results. However, suitably modified arguments lead to almost sure consistency of the posterior with respect to a related metric, say H_α , and similarly to consistency of the pseudo-Bayes estimator

$$f^n(x) = \int f(x) Q_\alpha^n(df),$$

say. Specifically, arguments used suggest using the distance function $H_\alpha^2(f) = 1 - \int f_0^\alpha f^{1-\alpha}$. With an α closer to 0 this amounts to a prior and posterior in closer agreement with the real ones.

Acknowledgements

We thank two referees for their comments on an earlier draft of the paper.

References

- Barron, A., Schervish, M. J. and Wasserman, L. (1999) The consistency of distributions in nonparametric problems. *Ann. Statist.*, **27**, 536–561.
- Diaconis, P. and Freedman, D. (1986) On the consistency of Bayes estimates. *Ann. Statist.*, **14**, 1–26.
- Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Statist. Ass.*, **90**, 577–588.
- Freedman, D. (1963) On the asymptotic behaviour of Bayes estimates in the discrete case. *Ann. Math. Statist.*, **34**, 1386–1403.
- van de Geer, S. (1993) Hellinger consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, **21**, 14–44.
- Geman, S. and Hwang, C.-R. (1982) Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.*, **10**, 401–414.
- Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999a) Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.*, **27**, 143–158.
- (1999b) Consistent semiparametric Bayesian inference about a location parameter. *J. Statist. Planning Inf.*, **77**, 181–193.
- Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000) Convergence rates of posterior distributions. *Ann. Statist.*, **28**, 500–531.
- Kraft, C. H. (1964) A class of distribution function processes which have derivatives. *J. Appl. Probab.*, **1**, 385–388.
- Lavine, M. (1992) Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, **20**, 1222–1235.

- Lenk, P. J. (1988) The logistic normal distribution for Bayesian nonparametric predictive densities. *J. Am. Statist. Ass.*, **83**, 509–516.
- (1991) Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, **78**, 531–543.
- Leonard, T. (1978) Density estimation, stochastic processes and prior information (with discussion). *J. R. Statist. Soc. B*, **40**, 113–146.
- Schwartz, L. (1965) On Bayes procedures. *Z. Wahrsch. Ver. Geb.*, **4**, 10–26.
- Shen, X. and Wasserman, L. (2001) Rates of convergence of posterior distributions. *Ann. Statist.*, **29**, in the press.
- Shen, X. and Wong, W. H. (1994) Convergence rates of sieve estimators. *Ann. Statist.*, **22**, 580–615.
- Walker, S. G. and Mallick, B. K. (1997) Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *J. R. Statist. Soc. B*, **59**, 845–860.
- Wasserman, L. (1998) Asymptotic properties of nonparametric Bayesian procedures. *Lect. Notes Statist.*, 293–304.
- (2000) Asymptotic inference for mixture models using data-dependent priors. *J. R. Statist. Soc. B*, **62**, 159–180.
- Wong, W. H. and Shen, X. (1995) Probability inequalities for likelihood ratios and convergence rates of sieve MLES. *Ann. Statist.*, **23**, 339–362.