On Bayesian Consistency

Stephen Walker and Nils Lid Hjort
University of Bath and University of Oslo

ABSTRACT. We consider a sequence of pseudo-posterior distributions [1] and establish simple conditions under which the sequence is Hellinger consistent. It is shown how investigations into these pseudo-posteriors assist with the understanding of some true posteriors, including Pólya-trees, infinite-dimensional exponential family and mixture models.

KEYWORDS: Bayes nonparametrics, Bayesian sieve, asymptotics, consistency.

1. Introduction

Asymptotics play an important role in statistics. In classical density estimation this role is crucial, providing results which justify 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, often the Hellinger distance, are key points to this area of research (see, for example, Shen and Wasserman, 2000).

On the other hand, 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. However, Diaconis and Freedman (1985) 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 Π is a prior distribution on the set of all probability densities over an interval or region of interest. As data $x_1, x_2, \ldots$ accumulate from some unknown underlying density f_0, will the Bayesian posterior distribution $\Pi^n(\cdot) = \Pi(\cdot \mid x_1, \ldots, x_n)$ concentrate around this f_0? The paper of Barron, Schervish and Wasserman (1999), from now on BSW, presents one such

[1] We refer to a posterior based on a data dependent prior as being a pseudo-posterior and a posterior based on a non data dependent prior as being a true posterior.

Bayesian nonparametric consistency theorem; the corresponding theorem of Ghosal, Ghosh and Ramamoorthi (1999) is of a similar nature. BSW make two assumptions to prove consistency in the Hellinger metric. 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. Specifically, for each positive δ a sequence $\mathcal{F}_n(\delta)$ of sets of densities is required to exist such that the prior mass for the complement set $\mathcal{F}_n'(\delta)$ is exponentially small and if $f \in \mathcal{F}_n(\delta)$ then there exists N_n upper brackets $\{f_1^U, \ldots, f_{N_n}^U\}$ such that $\int f_k^U < 1+\delta$ for all k and $f \leq f_k^U$ for some k. One of the assumptions required by BSW is 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 $c > 0$. Wasserman (1998) provides a review of Hellinger consistency and Shen and Wasserman (2000) provide rates of convergence.

Thus, given a nonparametric prior Π, after having ascertained that it satisfies the rather strict requirements of the BSW type, we can generate $\{\Pi^n\}$ knowing that the sequence is, with probability one, Hellinger consistent. Suppose, 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 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 hard to verify or not established, then it is preferable to use $\{Q^n\}$ for inference. This is particularly appropriate in the nonparametric context where construction of Π to incorporate real qualitative information is typically difficult. 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 pseudo-posterior distributions $\{Q^n\}$ related to Π.

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

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

for the Hellinger distance and

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

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

$$A_\varepsilon = \{f\colon H(f) > \varepsilon\}, \quad K_\eta = \{f\colon D(f) \le \eta\}$$

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 over which the densities are defined, for concreteness, although generalisations are easily covered.

In Section 2 we introduce a sequence of pseudo-posteriors which gives rise to a Hellinger consistent sequence of estimators for f. Section 3 considers a number of illustrations where the aim is to show how the pseudo-posteriors assist with the understanding of consistency for true posteriors. The result of BSW is general, covering all priors, and hence might for several special classes of priors 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

$$\Pi^n(\mathrm{d}f) = \frac{R_n(f)\,\Pi(\mathrm{d}f)}{\int R_n(f)\,\Pi(\mathrm{d}f)}.$$

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. BSW present a counter-example in their paper.

Define the pseudo-posterior distribution based on Π as

$$Q^n(\mathrm{d}f) = \frac{R_n^{1/2}(f)\,\Pi(\mathrm{d}f)}{\int R_n^{1/2}(f)\,\Pi(\mathrm{d}f)}.$$

We can view this in one of two ways: we are using the pseudo-likelihood 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

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

Wasserman (2000) used a different pseudo-likelihood function/data-dependent prior to establish statisfactory 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.

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

a consequence of Lemma 1 of BSW, which can also be used to show that $\int R_n^{1/2}(f)\pi(\mathrm{d}f) \neq 0$.

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

$$\Pi_n(\mathrm{d}f) \propto \exp\left\{\tfrac{1}{2} n D_n(f)\right\} \Pi(\mathrm{d}f),$$

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 $D_n(f) < 0$ having 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. BSW impose stronger restrictions on Π to achieve the same effect.

Theorem 1. With F_0 probability one, $Q^n(A_\varepsilon) \to 0$ with exponential rate, as $n \to \infty$ for all sets A_ε with $\varepsilon > 0$.

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

$$Q^n(A_\varepsilon) = \frac{\int_{A_\varepsilon} R_n^{1/2}(f)\,\Pi(\mathrm{d}f)}{\int R_n^{1/2}(f)\,\Pi(\mathrm{d}f)}.$$

The denominator can be written as

$$L_n = \int \exp\{-\tfrac{1}{2}nD_n(f)\}\Pi(\mathrm{d}f).$$

Thus, for any $\eta > 0$,

$$\exp(n\eta)L_n > \int_{K_{2\eta}} \exp\left[\tfrac{1}{2}n\left\{2\eta - D_n(f)\right\}\right]\Pi(\mathrm{d}f).$$

Arguments laid out by BSW (Lemmas 3 and 4), based on Fatou's lemma and condition (A), establish that $L_n > \exp(-n\eta)$ a.s. for large n for all $\eta > 0$.

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

$$\begin{aligned}
\mathrm{pr}_{x^n}\left\{\int_{A_\varepsilon} R_n^{1/2}(f)\Pi(\mathrm{d}f) > \exp(-nc)\right\} &< \exp(nc)\int_{A_\varepsilon}\left\{\int\sqrt{f\,f_0}\right\}^n\Pi(\mathrm{d}f)\\
&< \exp(nc)\int_{A_\varepsilon}\{1-\tfrac{1}{2}H(f)^2\}^n\Pi(\mathrm{d}f)\\
&< \exp(nc-\tfrac{1}{2}n\varepsilon^2).
\end{aligned}$$

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

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

$$f^n(x) = \int f(x)\,Q^n(\mathrm{d}f).$$

Here we establish that $f^n \to f_0$ a.s. with respect to the Hellinger distance.

Theorem 2. $H(f^n) < \varepsilon$ a.s. for large n for any $\varepsilon > 0$ and hence $H(f^n) \to 0$ a.s.

Proof. Using Corollary 1 from BSW,

$$H(f^n) \le \int H(f)\,Q^n(\mathrm{d}f) \le \int_{A_\varepsilon} H(f)\,Q^n(\mathrm{d}f) + \int_{A_\varepsilon^c} H(f)\,Q^n(\mathrm{d}f).$$

Now $H(\cdot) \le \sqrt{2}$ so

$$H(f^n) \le \sqrt{2}Q^n(A_\varepsilon) + \varepsilon,$$

completing the proof.

Convergence rates for $H(f^n)$ can be established using ideas from Shen and Wasserman (2000). Let t_n be as in Lemma 2 of Shen and Wasserman (2000), i.e. $\int R_n^{1/2}(f)\Pi(\mathrm{d}f) \ge \exp(-6nt_n)$. Suppose there exists a sequence c_n such that $nc_n \to \infty$, $c_n \to 0$ and $c_n \ge 12t_n$. Then

$$H(f^n) \le \sqrt{2}\exp(-\tfrac{1}{2}nc_n) + \varepsilon_n$$

a.s. for all large n, for all sequences $\{\varepsilon_n\}$ such that

$$\sum_n \exp\left\{-n\left(\tfrac{1}{2}\varepsilon_n^2 - c_n\right)\right\} < \infty.$$

This result is based on

$$\int_{A_{\varepsilon n}} R_n^{1/2}(f)\,\Pi(\mathrm{d}f) < \exp(-nc_n)$$

a.s. for large n. Hence, under simpler conditions than those of BSW, we have a Hellinger-consistent sequence of estimators of f_0, and can also establish rates of convergence.

Remark. The pseudo-Bayes estimator $f^n(x)$ might be hard to compute in its direct form, since it requires the posterior Q^n to be of suitably explicit form, or at least that it should be amenable to simulations. But this is typically difficult as it is for true Bayes estimators. A possible trick is to write the estimator as

$$f^n(x) = \frac{\int f(x)S_n^{1/2}(f)\,\Pi(\mathrm{d}f)}{\int S_n^{1/2}(f)\,\Pi(\mathrm{d}f)},$$

where $S_n(f) = \prod_{i=1}^n f(x_i)/f_1(x_i)$, for a suitable f_1 density taken to secure numerical stability. The point here is that the $f^n(x)$ curve now can be arrived at via simulations from the prior distribution Π alone.

3. Illustrations

In this section we will look at a number of priors and use the consistency of Q^n to help us establish results for Π^n. The result of BSW is for all priors. Here we consider specific priors, those considered by BSW and Ghosal et al. (1999), and using Q^n establish sufficient conditions for Hellinger consistency.

3.1 Pólya-trees. We consider Pólya-trees on $[0,1]$ with partition structure the dyadic intervals. This was an example considered by BSW. For each interval in the dyadic system we allocate a random variable $0 < V_{jk} < 1$; $k = 1, 2, \ldots$ and $j = 1, \ldots, 2^k$. If j is odd then $V_{j+1\,k} = 1 - V_{jk}$ 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 jth dyadic interval (from left to right) at level k. Here $B_{l(j)\,l}$, for $l = 1, \ldots, k$, make up the unique sequence of dyadic intervals which leads to B_{jk}.

As with BSW, we assume that $V_{jk} \sim \mathrm{be}(a_k, a_k)$ 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, for example, BSW, Section 3.2. Assume condition (B) holds.

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

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

Hence,

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

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

Consequently, the pseudo-predictive density

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

i.e.

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

based on the data-dependent Pólya-tree prior with parameterts a_k is equivalent to the true predictive density based on a Pólya-tree prior with parameters $2a_k$. This indicates that while the posterior distribution of a Pólya-tree prior may not be consistent under condition (B), BSW establish $a_k = 8^k$ 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 $[0,1]$. BSW also consider this example in Section 3.3 of their paper. Original work on these families was done by Leonard (1978), Thorburn (1986) and Lenk (1988, 1991). 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

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

BSW establish 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 $f \sim LNS(\mu, \sigma, \xi)$, so $f(x) \propto W_\xi(x)$ where $W_\xi(\cdot)$ is a generalised lognormal process with distribution Λ_ξ characterised by

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

and Λ represents a lognormal process, i.e. if $W \sim \Lambda$ then $W(x) = \exp\{Z(x)\}$ and $Z(\cdot)$ 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 $LNS(\mu_n, \sigma_n, \xi_n)$ 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 characterised via

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

and hence it is seen that Λ_Q^n is the true posterior based on the sample size dependent prior $f \sim LNS(\mu, \sigma/2, \xi + n/2)$. This result follows from Theorems 1 and 2 of Lenk (1988). By putting $\xi = -n/2$ we obtain the prior of BSW, i.e. $f \sim LNS(\mu, \sigma/2, 0)$. See Lenk (1991) for this connection. Consequently, provided $LNS(\mu, \sigma/2, 0)$ satisfies condition (A), the sequence of posterior distributions are 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 $x_1, \ldots, x_n$ are iid from $f(x; \theta_0)$ with $\theta_0 \in \Theta$. We assume that $\widehat{f} = f_{\hat{\theta}}$, 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

$$\Pi^n(A_\varepsilon) \le \exp(-nc)\, L_n^{1/2}(\widehat{\theta})$$

a.s. for all large n for any $\varepsilon > 0$ and $c < \frac{1}{2}\varepsilon^2$.

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

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

Now

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

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

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

and hence the result.

Consequently, the consistency of Π^n is guaranteed if the $L_n(\widehat{\theta})$ 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

$$\sup_\theta|g(\theta)\,\mathrm{d}(F_n-F_{\theta_0})|\to 0 \;\text{ a.s.}$$

where F_n is the empirical distribution function, then

$$\lim\,\sup_n\left\{\tfrac{1}{n}\log L_n^{1/2}(\widehat{\theta})\right\}\le 0 \;\text{ a.s.}$$

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 $f(x)=\int\phi_h(x-\theta)\,\mathrm{d}P(\theta)$, 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. (1999) 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. We let Π denote the prior for P and π the prior for h. Following Ghosal et al. (1999) we write $f_{h,P}$ to denote a random density $f(x)=\int\phi_h(x-\theta)\,\mathrm{d}P(\theta)$. We will also use the normal density for ϕ. We define $g_{h,P}(x)=\{\int\phi_h^{1/2}(x-\theta)\,\mathrm{d}P(\theta)\}^2$, and note that $g_{h,P}\le f_{h,P}$. Now let us consider, for any set A,

$$\Pi(A|x^n)=\frac{\int_A R_n(f_{h,P})\pi(h)\,\mathrm{d}h\,\Pi(\mathrm{d}P)}{\int R_n(f_{h,P})\pi(h)\,\mathrm{d}h\,\Pi(\mathrm{d}P)}.$$

The numerator can be written as

$$U_n=\int_A\left\{\prod_i f_0(x_i)^{-1}\prod_i\int\phi_h(x_i-\theta)\,\mathrm{d}P(\theta)\right\}\,\pi(h)\,\mathrm{d}h\,\Pi(\mathrm{d}P)$$

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

$$U_n\le\kappa^n\prod_i f_0(x_i)^{-1/2}\int_A R_n^{1/2}(f_{h/\sqrt{2},P})\,h^{-n/2}\pi(h)\,\mathrm{d}h\,\Pi(\mathrm{d}P).$$

The denominator can be written as

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

If we use the sample-size dependent prior $\pi_n(h) \propto h^{n/2}\pi(h)$, which requires $\int h^{\xi}\,\pi(h)\,\mathrm{d}h < \infty$ for all $\xi > 0$, we have

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

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

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

for all $\eta > 0$, where $\Pi(h,\mathrm{d}P) = \pi(h)\,\Pi(\mathrm{d}P)$. Clearly this, combined with $\int \pi(h)\,h^{\xi}\,\mathrm{d}h < \infty$ for all $\xi > 0$, is a sufficient condition for the Hellinger consistency of $\Pi(h,\mathrm{d}P)$.

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. BSW present a Bayesian solution which places extra conditions on Π. The solution proposed in this paper uses the data to downweight troublesome densities in the support of prior.

This procedure, as we have demonstrated in Section 3, sheds much light on the Hellinger consistency of standard nonparametric priors, such as those considered by BSW and Ghosal et al. (1999).

A more general data-modified prior to work with would take the form $\Pi(\mathrm{d}f)/\prod_{i=1}^n f^\alpha(x_i)$, where $0<\alpha<1$; this would also correspond to a pseudo-likelihood $\mathcal{L}_n(f)=\prod_{i=1}^n f^{1-\alpha}(x_i)$. Our choice $\alpha=\frac{1}{2}$ agrees nicely with the Hellinger distance and gives satisfactory results. However, suitably modified arguments lead to a.s. consistency of the posterior with respect to a related metric, say H_α, and similarly to consistency of the pseudo-Bayes estimator $f^n(x)=\int f(x)\,Q_\alpha^n(\mathrm{d}f)$, 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 zero this amounts to a prior and posterior in closer agreement with the real ones.

References

Barron, A., Schervish, M.J. and Wasserman, L. (1999). The consistency of distributions in nonparametric problems. *Annals of Statistics* **27**, 536–561.

Kraft, C.H. (1964). A class of distribution function processes which have derivatives. *Journal of Applied Probability* **1**, 385–388

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

Lenk, P.J. (1988). The logistic normal distribution for Bayesian nonparametric predictive densities. *Journal of the American Statistical Association* **83**, 509–516.

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

Leonard, T. (1978). Density estimation, stochastic processes, and prior information. *Journal of the Royal Statistical Society Series B* **40**, 113–146.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **4**, 10–26.

Shen, X. and Wong, W.H. (1994). Convergence rates of sieve estimators. *Annals of Statistics* **22**, 580–615.

Shen, X. and Wasserman, L. (2000). Rates of convergence of posterior distributions. *Annals of Statistics*, to appear.

Thorburn, D. (1986). A Bayesian approach to density estimation. *Biometrika* **73**, 65–76.

van de Geer, S. (1993). Hellinger consistency of certain nonparametric maximum likelihood estimators. *Annals of Statistics* **21**, 14–44.

WASSERMAN, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Müller and D. Sinha, eds.), 293–304. *Lecture Notes in Statistics*, Springer.

WASSERMAN, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society Series B* **62**, 159–180.

WONG, W.H. AND SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLES. *Annals of Statistics* **23**, 339–362.