Biometrika (2005), **92,** 4, *pp.* 765–778

Printed in Great Britain

Data tracking and the understanding of Bayesian consistency

By STEPHEN G. WALKER

Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NZ, U.K.

s.g.walker@kent.ac.uk

ANTONIO LIJOI and IGOR PRÜNSTER

Dipartimento di Economia Politica e Metodi Quantitativi, Università degli Studi di Pavia, 27100 Pavia, Italy

lijoi@unipv.it igor.pruenster@unipv.it

Summary

We deal with strong consistency for Bayesian density estimation. An awkward consequence of inconsistency is described. It is pointed out that consistency at some density f_0 depends on the prior mass assigned to the ‘pathological’ set of those densities that are close to f_0, in a weak sense, and far apart from f_0, in a Hellinger sense. An analysis of these sets leads to the identification of the notion of ‘data tracking’. Specific examples in which this phenomenon cannot occur are discussed. When it can happen, we show how and where things can go wrong, thus providing more intuition about the sources of inconsistency.

Some key words: Bayesian consistency; Bayesian density estimation; Hellinger distance; Kullback–Leibler divergence; Weak neighbourhood.

1. Introduction

A first formulation of the issue of consistency of Bayesian inferential procedures is given in Doob (1949). It states that, if there exists a consistent sequence of estimators of the unknown parameter, then the posterior estimators are consistent in the sense that the posterior distribution converges to a point mass at the unknown parameter outside a set of prior mass zero. A drawback of such an approach is that the null sets on which convergence fails could be relevant. In this case, the problem can be circumvented by resorting to a ‘frequentist’ notion of consistency which gives rise to the ‘what if’ method adopted by Diaconis & Freedman (1986). The idea consists of generating independent data from a ‘true’ fixed distribution f_0 and checking whether or not the posterior accumulates in suitably-defined neighbourhoods of f_0. This corresponds to requiring the data eventually to swamp the prior.

An early use of the ‘what if’ method can be found in Freedman (1963), where it is shown that weak consistency does not necessarily hold for priors supported by discrete distributions on a countable set of states. However, if the number of states is finite, consistency is achieved and the result extends to the countable case by the introduction of an additional entropy condition. A sufficient condition for weak consistency with more general priors is suggested in Schwartz (1965). This is solely a support condition. Further examples of inconsistency, involving mixtures, are illustrated in Diaconis & Freedman (1986).

When we consider problems of density estimation, it is natural to ask for the strong consistency of Bayesian procedures. An early contribution in this area is an unpublished University of Illinois technical report by A. R. Barron, which is based on uniformly consistent tests. Later developments, combining well-established techniques in the theory of empirical processes with ideas from Barron's report, provide sufficient conditions for strong consistency in terms of metric entropies. For instance, Barron et al. (1999) specify bracketing entropy conditions for strong consistency to hold true and apply their results to a number of commonly used priors in Bayesian nonparametric inference. Following the same lines, Ghosal et al. (1999) provide slightly weaker sufficient conditions for strong consistency in terms of the L_1-metric entropy and deal with mixtures of a Dirichlet process. This approach, reviewed in Wasserman (1998), has also been employed for verifying strong consistency of specific priors in Bayesian nonparametrics; see for example Petrone & Wasserman (2002). New ideas for solving consistency issues are given in Walker (2003, 2004), where a simple sufficient condition for strong consistency is represented by the finiteness of a suitable sum of square roots of prior probabilities.

The present paper aims at providing an understanding of the main issues that arise when dealing with consistency of Bayesian procedures. An argument which motivates the interest in consistency can be based on a notion of merging which differs from the classical one introduced by Blackwell & Dubins (1962). Indeed, we consider the case of two Bayesians sharing the same prior but collecting two independent datasets from the same density f_0. It turns out that if the prior is inconsistent at f_0 then the two Bayesians disagree even if more and more data are collected. This is an unpleasant feature. Given this, it is even more important to determine possible sources of strong inconsistency. In order to develop such an analysis, we still preserve the support condition introduced by Schwartz (1965). We illustrate how consistency at some density f_0 depends on the prior mass assigned to the 'pathological' set of those densities that are close to f_0, in a weak sense, and far apart from f_0, in the L_1-metric. If the prior does not put mass on such sets, then strong consistency is achieved at f_0. Many priors in common use meet such a requirement, and we provide some related illustration. If the prior mass on such sets is positive, one has to take care about densities that track the data, a notion to be made precise later on. In order to get rid of the data-tracking phenomenon, one has to look for sufficient conditions which avoid it. Regarding this aspect, we reconsider and slightly generalise a result of Walker (2004), which is given in terms of a summability condition of prior probabilities. We provide an interpretation and show that this sufficient condition is not necessary. Finally a new sufficient condition is provided. When applied to the prior of the counterexample in Barron et al. (1999) it nicely shows the reason for its inconsistency.

2. Notation and some basic facts

We consider a sequence of observations $(X_n)_{n\geqslant 1}$, each taking values in some complete and separable metric space $\mathbb{X}$ endowed with a σ-algebra $\mathscr{X}$. If $\mathbb{F}$ indicates the space of probability density functions with respect to some measure λ on $\mathbb{X}$, then we define

$$d_H(f,g)=\left[\int_{\mathbb{X}}\{f^{\frac{1}{2}}(x)-g^{\frac{1}{2}}(x)\}^2\lambda(dx)\right]^{\frac{1}{2}},$$

for any f and g in $\mathbb{F}$, and set $\mathscr{F}$ to be the Borel σ-algebra of $\mathbb{F}$. Suppose that Π stands for a prior distribution on $(\mathbb{F}, \mathscr{F})$. In this case, we assume that, given a density f drawn from Π, the observations are independent and identically distributed with common density f; that is

$$\mathrm{pr}\{(X_1, \ldots, X_n) \in A\} = \int_A \int_{\mathbb{F}} \left\{ \prod_{i=1}^n f(x_i) \right\} \Pi(df)\lambda(dx_1)\ldots\lambda(dx_n),$$

for each $n \geqslant 1$ and A in $\mathscr{X}^n$. The posterior distribution on $(\mathbb{F}, \mathscr{F})$, given the observations $(X_1, \ldots, X_n)$, coincides with

$$\Pi_n(B) = \frac{\int_B \prod_{i=1}^n f(X_i)\Pi(df)}{\int_{\mathbb{F}} \prod_{i=1}^n f(X_i)\Pi(df)},$$

for all B in $\mathscr{F}$. The frequentist approach to Bayesian consistency is based on the idea of fixing a density f_0 as the 'true' density from which the data are independently sampled and checking whether or not the posterior accumulates in any Hellinger neighbourhood of f_0. We denote by P_0 the probability distribution whose density coincides with f_0 and by P_0^∞ the infinite product measure on $\mathbb{X}^\infty$. Hence Π is 'strongly consistent' or equivalently 'Hellinger-consistent' at f_0 if, for any $\varepsilon > 0$,

$$\Pi_n(A_\varepsilon) \to 1,$$

almost surely with respect to P_0^∞, where $A_\varepsilon = \{f \in \mathbb{F} : d_H(f, f_0) < \varepsilon\}$. In what follows, almost sure convergence will be considered with respect to P_0^∞ even if not explicitly mentioned. An alternative less stringent notion of consistency can be given by referring to the space $\mathbb{P}$ of probability distributions on $(\mathbb{X}, \mathscr{X})$, equipped with the weak topology. A weak neighbourhood of any probability distribution P^* in $\mathbb{P}$ is the set

$$W_\varepsilon = \left\{ P \in \mathbb{P} : \left| \int \phi_i dP - \int \phi_i dP^* \right| < \varepsilon, i = 1, \ldots, k \right\},$$

for a k-tuple of continuous and bounded real-valued functions ϕ_i defined on $\mathbb{X}$. In this case, we say that a prior Π is 'weakly consistent' at f_0 if, for any $\varepsilon > 0$,

$$\Pi_n(W_\varepsilon) \to 1,$$

almost surely, where W_ε stands for a weak neighbourhood of P_0. Recall that the weak topology is coarser than the one induced by d_H, the latter being equivalent to the total variation topology on $\mathbb{P}$.

Another notion we need to consider is that of support of a prior. We say that P_0 is in the support of Π if any neighbourhood of P_0 has positive Π-probability. According to the topology defined on $\mathbb{P}$, we distinguish weak and Hellinger support of Π, which will be denoted by $S_W(\Pi)$ and $S_H(\Pi)$, respectively. One can reasonably think that P_0 being in $S_W(\Pi)$ would imply weak consistency of Π at P_0. Such a guess is wrong, as shown for example by the counterexample in Diaconis & Freedman (1986) for mixtures of Dirichlet processes. Hence, one needs to impose a stronger support condition in order to achieve weak consistency at P_0. To this end, consider two probability distributions P and Q such that P is absolutely continuous with respect to Q, and define the 'Kullback–Leibler divergence' between P and Q as

$$D_K(P, Q) = \int \log(dP/dQ)dP. \qquad (1)$$

If $\mathbb{P}^*$ is a subset of $\mathbb{P}$ formed by all probability distributions dominated by a common σ-finite measure λ, (1) reduces to

$$d_K(f_P, f_Q) = \int f_P \log(f_P/f_Q),$$

where $f_P = dP/d\lambda$ and $f_Q = dQ/d\lambda$ are the densities of P and Q, respectively, with respect to λ for any $P, Q \in \mathbb{P}^*$. Hence, d_K can be seen as a measure of divergence on the corresponding space of densities $\mathbb{F}$. If $K_\varepsilon = \{P \in \mathbb{P}^* : d_K(f_{P_0}, f_P) < \varepsilon\}$ is a neighbourhood of P_0 with respect to d_K, the probability distribution P_0 is in the Kullback–Leibler support $S_K(\Pi)$ of Π if $\Pi(K_\varepsilon) > 0$ for any $\varepsilon > 0$. Note that $S_W(\Pi) \supset S_H(\Pi) \supset S_K(\Pi)$. A fundamental sufficient condition for obtaining weak consistency is due to Schwartz (1965): if P_0 is in $S_K(\Pi)$, then Π is weakly consistent at P_0.

When one is dealing with density estimation, it is more natural to ask for strong consistency and one might hope that a Kullback–Leibler support condition still suffices. However, as has been shown in Barron et al. (1999), this does not happen without any further condition. All the contributions in this area aim at giving simple sufficient conditions for strong consistency and preserve the Kullback–Leibler support condition. In the following sections we attempt to understand the deep reasons for possible strong inconsistencies in cases in which weak consistency holds true.

3. Inconsistency and possible solutions

3·1. *Variation of the problem of merging of opinions*

It is commonly agreed that consistency is an important property of statistical procedures, and this is true in a Bayesian setting as well. Indeed, lack of consistency might yield unpleasant consequences of the type we are going to describe. Before proceeding with the illustration, it is worth recalling that a lot of attention in the literature has focused on the so-called merging of opinions, when two Bayesians assess different priors and one is interested in checking whether or not their posterior inferences tend to coincide as long as more data are collected. Original work on this issue can be found in Blackwell & Dubins (1962), where it is proved that, under a condition of absolute continuity of one prior with respect to the other, merging of opinion occurs in the sense that the L_1-distance between predictive distributions becomes negligible as the sample size increases. Later discussions are provided in Diaconis & Freedman (1986), in Ghosal et al. (1999) and elsewhere. The merging of opinion, or agreement, for large samples, boils down to consistency; that is, posterior distributions accumulate around the same, and correct, density function; see for example Barron et al. (1999).

Here we consider a different set-up which, to our knowledge, has not been investigated before. Suppose that two Bayesians are conducting the same experiment, thereby generating two independent samples $X_1^{(1)}, X_2^{(1)}, \ldots$ and $X_1^{(2)}, X_2^{(2)}, \ldots$, respectively, from the same probability distribution, P_0, which has density with respect to the Lebesgue measure given by f_0. Both Bayesians agree to use the same prior distribution Π on the space of density functions. One would reasonably expect that for large samples the two Bayesians will agree with each other. However, we show that it is possible to construct priors for which agreement is not achieved, even ones which have f_0 in the Kullback–Leibler support of Π, that is ensuring Π to be weakly consistent. Define $g : \mathbb{X}^n \times \mathbb{X}^n \to \mathbb{R}$ as a measurable function of the n-dimensional independent samples $X_1^{(j)}, \ldots, X_n^{(j)}$, for

$j=1,2$. Also, denote by $E_0^{(j)}(g)$ the expectation of g with respect to the jth sample $X_1^{(j)},\ldots,X_n^{(j)}$ keeping $X_1^{(l)},\ldots,X_n^{(l)}$ fixed, where $l \neq j$ and where the $X_i^{(j)}$'s are independent and identically distributed from f_0. Thus, for $j=1$,

$$E_0^{(1)}\{g(X_1^{(1)},\ldots,X_n^{(1)},X_1^{(2)},\ldots,X_n^{(2)})\}$$
$$=\int_{\mathbb{X}^n} g(x_1^{(1)},\ldots,x_n^{(1)},x_1^{(2)},\ldots,x_n^{(2)})\prod_{i=1}^n f_0(x_i^{(1)})\lambda(dx_i^{(1)}).$$

An analogous representation is given for $E_0^{(2)}(g)$.

Theorem 1. *Assume that f_0 is in $S_K(\Pi)$. Then, if Π is not Hellinger-consistent at f_0,*

$$E_0^{(1)}\{D_K(\Pi_n^{(2)},\Pi_n^{(1)})\} > n\delta$$

infinitely often, almost surely, for some $\delta>0$, where $\Pi_n^{(j)}$ denotes the posterior distribution based on the dataset $X_1^{(j)},\ldots,X_n^{(j)}$, for $j=1,2$.

Proofs of the theorems are given in the Appendix.

Such an outcome is certainly startling for two Bayesians using the same prior and sampling from the same density. In particular, there is no merging of information; see for example Barron's technical report. Hence, identification of consistent priors and investigation of possible sources of inconsistency are important issues.

We first consider the latter issue and try to understand why the Kullback–Leibler support condition is sufficient for weak, but not for strong, consistency. It is clear, indeed, that inconsistency at f_0 may be caused by sequences of densities that converge weakly, but not in L_1, to f_0. An example of such behaviour is associated with the sequence of densities $f_n(x)=1+\sin(2\pi nx)$, for x in $[0,1]$. The corresponding sequence of distributions converges weakly to the Un$[0,1]$ distribution, whereas f_n oscillates ever more wildly and does not converge to anything. The oscillating behaviour, together with high peaks at the maxima of the f_n's, causes the undesirable phenomenon of 'tracking the data'. In other terms, data corresponding to these peaks remarkably increase the likelihood and thus may lead to the posterior not swamping mass from the rough densities. In order to understand this phenomenon, first described in Barron et al. (1999), in a more formal setting, one has to focus attention on the set

$$V_{\delta,\varepsilon}=W_\delta\cap A_\varepsilon^c,$$

where W_δ and A_ε denote weak and Hellinger neighbourhoods, respectively, of f_0.

Since by weak consistency the posterior Π_n will accumulate in W_δ, for any $\delta>0$, and $V_{\delta,\varepsilon}$ shrinks as δ goes to 0, the first issue to be faced is whether or not for all small enough δ the prior is prevented from putting mass on $V_{\delta,\varepsilon}$. Intuitively, one can envisage this constraint as being allowed to track the data up to a finite number of observations. In the following subsections we first deal with the case in which $\Pi(V_{\delta,\varepsilon})=0$, for any δ less than some fixed $\delta^*>0$, and we then consider cases in which such a condition is not met.

3·2. *Consistency with* $\Pi(V_{\delta,\varepsilon})=0$

With $V_{\delta,\varepsilon}$ identified as the set that might give rise to inconsistency, one must first look for priors that satisfy $\Pi(V_{\delta,\varepsilon})=0$. Consistency is automatically achieved in this case. Indeed, it turns out that some of the commonly used priors satisfy this condition. Here we provide an illustration by considering some noteworthy examples.

Example 1: *Monotone decreasing densities.* Here we assume that the prior is concentrated on monotone decreasing densities on $\mathbb{R}^+$. Bayesian nonparametric inference with such priors is considered in Brunner & Lo (1989) and further developed in Hansen & Lauritzen (2002). As well as dealing with theoretical and computational issues associated with Bayesian estimation in this setting, they identify consistency as an interesting aspect to be investigated. If F is a probability distribution function corresponding to some monotone decreasing density f, then we have

$$F(x) = \int_{\mathbb{R}^+} F(x;\theta)dG(\theta),$$

where G is a distribution function, $F(x;\theta) = \theta^{-1}\min\{x,\theta\}$ if $\theta > 0$ and $F(x;0)$ is degenerate at 0; see for example Feller (1971, p. 158). Moreover G is uniquely determined by

$$G(\theta) = F(\theta) - \theta f(\theta).$$

In a Bayesian setting G is seen as a random distribution function whose law induces a prior for F. We now verify that Hellinger consistency holds true for any $f_0(x) = \int_x^{+\infty} \theta^{-1} dG_0(\theta)$. Let W_δ be a δ-weak neighbourhood of G_0. We show that $G \in W_\delta$ if and only if f is in an ε-Hellinger neighbourhood, A_ε, of f_0. Assume that the G converges weakly to G_0, so that, for any $x > 0$,

$$\int_x^{+\infty} \frac{1}{\theta} dG(\theta) \to \int_x^{+\infty} \frac{1}{\theta} dG_0(\theta).$$

Consequently, by Scheffé's theorem, one has that $\int |f(x) - f_0(x)| dx \to 0$, or equivalently that $f \to f_0$ in Hellinger distance. To show the converse, we prove that $G \in W_\delta^c$ implies $f \in A_\varepsilon^c$. Define a weak neighbourhood of G_0 as

$$W_\delta = \left\{ G : \left| \int_0^{+\infty} \frac{(\theta - y)\mathbb{I}_{(y,+\infty)}(\theta)}{\theta} dG(\theta) - \int_0^{+\infty} \frac{(\theta - y)\mathbb{I}_{(y,+\infty)}(\theta)}{\theta} dG_0(\theta) \right| < \delta \right\},$$

for a fixed $y > 0$, where $\mathbb{I}_A$ is the indicator function of set A. If $G \notin W_\delta$, then $|F(y) - F_0(y)| > \delta$, which yields $f \in A_\varepsilon^c$ for $\varepsilon < \delta$. Thus, one has that $\Pi(V_{\delta,\varepsilon}) = 0$ and Hellinger consistency holds without any further assumption.

Example 2: *Mixture models.* Consider the mixture model

$$f(x) = \int \phi_h(x - \theta) dQ(\theta),$$

where ϕ_h is the normal density function with mean zero and variance h^2. Moreover, Q has a nonparametric prior and μ is the prior distribution for h. This model is considered by Ghosal et al. (1999). It is assumed that

$$f_0(x) = \int \phi_{h_0}(x - \theta) dQ_0(\theta)$$

is the true density function. Note that h can cause trouble by getting arbitrarily close to 0.

First, let $h_0 > \tau > 0$ and define a neighbourhood of (h_0, Q_0) as the set

$$W_{\delta,\tau} = \{(h, Q) : |h - h_0| < \tau, Q \in W_\delta\}.$$

If $(h, Q) \to (h_0, Q_0)$, then, from

$$\left| \int \phi_h(x-\theta) dQ(\theta) - \int \phi_{h_0}(x-\theta) dQ_0(\theta) \right| \leqslant \int |\phi_h(x-\theta) - \phi_{h_0}(x-\theta)| dQ_0(\theta) + \left| \int \phi_h(x-\theta) d\{Q(\theta) - Q_0(\theta)\} \right|,$$

one has that $f(x) \to f_0(x)$ pointwise for all x, and Scheffé's theorem implies that $f \to f_0$ in the Hellinger distance. This means that $(h, Q) \in W_{\delta,\tau}$ implies that $f \in A_\varepsilon$.

Now consider

$$|P(B) - P_0(B)| = \left| \int \{\Phi_h(B; \theta) - \Phi_{h_0}(B; \theta)\} dQ_0(\theta) + \int \Phi_h(B; \theta) d\{Q(\theta) - Q_0(\theta)\} \right|,$$

where $\Phi_h(B; \theta) = \int_B \phi_h(x-\theta)dx$. Again, for h bounded away from 0, and excluding the case in which $(h, Q) = (h_0, Q_0)$, we can always find a set B to make this positive. Now, consider the case in which h gets arbitrarily close to 0. It is easy to see that $|P(B) - P_0(B)| \to |Q(B) - P_0(B)|$. Hence, consistency can fail when the prior puts positive mass on h in a neighbourhood of 0 and positive mass on Q in Hellinger neighbourhoods of P_0. This renders the quantity $|Q(B) - P_0(B)|$ small with positive probability, for any choice of set B, thus leading to possible problems in the identification of the correct (h_0, Q_0). This can be circumvented by requiring Q and P to have suitable different supports. For example, take P and Q with supports coinciding with the real line and with $[-a, a]$ for some finite and positive a, respectively. This ensures that the prior for Q puts zero mass in Hellinger neighbourhoods of P_0 and $|Q(B) - P_0(B)|$ is away from 0 for some set B.

Example 3: *Finite-dimensional parametric family*. Here we consider sampling models $\{f(x; \theta) : \theta \in \Theta\}$, where Θ is a finite-dimensional parameter space. Provided the support condition is met by the prior, such families lead to consistency. The point is that $\Pi(V_{\delta,\varepsilon}) = 0$ for some $\delta > 0$ and for all $\varepsilon > 0$. For $f \in V_{\delta,\varepsilon}$ for all $\delta > 0$ it is required that the density f be oscillating, the number of oscillations increasing to ∞ as $\delta \downarrow 0$. This just cannot happen if f is based on a finite-dimensional parameter.

To formalise this, we have the following simple conditions which should be easily verifiable for any particular $f(.; \theta)$. If $f(.; \theta_k) \to f(.; \theta_0)$ weakly, that is

$$\int g(x) f(x; \theta_k) dx \to \int g(x) f(x; \theta_0) dx,$$

for all bounded and continuous g, then this implies that $|\theta_k - \theta_0| \to 0$. If $\theta \mapsto f(x; \theta)$ is continuous almost everywhere with respect to the Lebesgue measure, this in turn implies that

$$f(x; \theta_k) \to f(x; \theta_0),$$

pointwise almost everywhere. Then weak neighbourhoods of $f_0(.) \equiv f(.; \theta_0)$ are equivalent to Hellinger neighbourhoods and so for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $\Pi(V_{\delta,\varepsilon}) = 0$.

Example 4: *Discrete model*. Here we assume that observations take values in a countable set, such as $\mathbb{X} = \{1, 2, 3, \ldots\}$. Denote by $f^{(k)}$ the random mass assigned to the integer k by f. Suppose that Π is concentrated on all discrete probability distributions on $\mathbb{X}$.

Let f_0 be any distribution in the support of Π and indicate by $f_0^{(k)}$ the true mass assigned to k. Then P_{f_n} converges weakly to P_{f_0} if and only if $f_n^{(k)} \to f_0^{(k)}$ for all k, which also implies that f_n converges in L_1 to f_0.

3·3. *Consistency with* $\Pi(V_{\delta,\varepsilon}) > 0$

This general case has been the focus of many papers in the literature, wherein conditions on the prior are specified in terms of metric entropy and do not admit an easy interpretation. Our first result allows for a natural identification of the data-tracking behaviour. Indeed, it shows that the posterior mass concentrated on densities that are away from f_0 in a Hellinger sense and do not track the data vanishes as the sample size increases with P_0^∞-probability 1.

Define the data-tracking set as a random set of the type $B_{n,\gamma}^c := \{f : R_n(f) \geqslant e^{n\gamma}\}$, for any $\gamma > 0$ and where $R_n(f) = \prod_{i=1}^n f(X_i)/f_0(X_i)$. This is a clear interpretation of the sets $B_{n,\gamma}^c$, and Theorem 2 will demonstrate that problems of inconsistency arise because the posterior puts sufficient mass into sets of the type $B_{n,\gamma}^c$. First, recall that $A_\varepsilon^c = \{f : d_H(f_0, f) > \varepsilon\}$.

Theorem 2. *Let f_0 be in the Kullback–Leibler support of Π. Then we have*

$$\Pi_n(A_\varepsilon^c \cap B_{n,\gamma}) \to 0,$$

almost surely, for any $\gamma < -2\log(1-\varepsilon)$.

By Theorem 2, problems might arise because of the sets $A_\varepsilon^c \cap B_{n,\gamma}^c$, and we are then interested in finding sufficient conditions for which

$$\Pi_n(A_\varepsilon^c \cap B_{n,\gamma}^c) \to 0, \tag{2}$$

almost surely. We first focus on a prior Π concentrating masses $\Pi_1, \Pi_2, \ldots$ on at most a countable number of densities, with $\sum \Pi_k = 1$. Note that (2) is equivalent to

$$J_n = \sum_{\{k : f_k \in A_\varepsilon^c \cap B_{n,\gamma}^c\}} R_{nk}\Pi_k < \exp(-n\delta),$$

almost surely for all large n for some $\delta > 0$, where we have denoted $R_n(f_k)$ by R_{nk}. Since

$$I_n \geqslant J_n > \Pi(A_\varepsilon^c \cap B_{n,\gamma}^c)\exp(n\gamma),$$

where $I_n := \sum_k R_{nk}\Pi_k < \exp(n\beta)$ almost surely for all large n for any $\beta > 0$, we have that

$$\Pi(A_\varepsilon^c \cap B_{n,\gamma}^c) < \exp(-n\eta),$$

almost surely for all large n for any $\eta < \gamma - \beta$, where we can fix $\beta < \gamma$. Consequently, the Cauchy–Schwarz inequality yields

$$J_n = \sum_{\{k : f_k \in A_\varepsilon^c\}} \mathbb{I}_{B_{n,\gamma}^c}(f_k) R_{nk}\Pi_k \leqslant \sum_{\{k : f_k \in A_\varepsilon^c\}} (R_{nk}^2\Pi_k)^{\frac{1}{2}} \{\Pi(A_\varepsilon^c \cap B_{n,\gamma}^c)\}^{\frac{1}{2}}.$$

Since

$$\sum_{\{k : f_k \in A_\varepsilon^c\}} R_{nk}^2\Pi_k < \left(\sum_k R_{nk}\Pi_k^{\frac{1}{2}}\right)^2$$

a sufficient condition for (2) to hold true is

$$\sum_{\{k:f_k \in A_\varepsilon^c\}} R_{nk} \Pi_k^{\frac{1}{2}} < \exp(n\eta')$$

almost surely for all large n. Recall that $R_{nk} < \exp(n\eta')$ almost surely, for all large n and for all $\eta' > 0$. Thus we can conclude that $\sum_k \Pi_k^{\frac{1}{2}} < \infty$ is sufficient for consistency; see Walker (2004) for different derivations of this result. As a matter of fact, it has been shown in Walker (2004) that a similar condition is sufficient in a more general setting as well.

We now consider a general prior Π, not necessarily discrete. Let f_0 be fixed and take A_ε^c to be the complement of an ε-Hellinger neighbourhood of f_0. By separability of $\mathbb{F}$, such a set can be covered by a countable union of disjoint sets B_j, where $B_j \subseteq B_j^* := \{f : d_H(f, f_j) < \eta\}$, f_j are densities in A_ε^c and η is any number in $(0, \varepsilon)$. If f_0 is in the Kullback–Leibler support of the prior Π and

$$\sum_{j \geqslant 1} \Pi^{\frac{1}{2}}(B_j) < +\infty$$

then Π is Hellinger-consistent at f_0. By virtue of the arguments illustrated at the beginning of the present section, this result can be refined by confining oneself to the determination of a covering of $V_{\delta,\varepsilon} \subseteq A_\varepsilon^c$. Moreover, by mimicking the proof in Walker (2004), one can state that Hellinger-consistency holds true at $f_0 \in S_K(\Pi)$ if, for some $\alpha \in (0, 1)$,

$$\sum_{j \geqslant 1} \Pi^\alpha(V_j) < +\infty, \tag{3}$$

where the sets V_j have diameter $\eta < \varepsilon$ and form a countable partition of $V_{\delta,\varepsilon}$.

At this stage, one might wonder whether or not (3) is also necessary for consistency to hold true. The answer to such a question is, in general, negative and can be motivated by an argument which shows that violation of (3) does not imply inconsistency. Assume that Π is not consistent at $f_0 \in S_K(\Pi)$ and that, for some $\varepsilon > 0$, $A_\varepsilon \subset \mathrm{S}_H(\Pi)$. Hence, there exists α in $(0, 1)$ such that

$$\sum_{j \geqslant 1} \Pi^\alpha(V_j) = +\infty$$

for any covering of $V_{\delta,\varepsilon}$. Now take $\tilde{f}$ in $A_{\varepsilon/2}$ and denote by $\tilde{V}_j$ the disjoint sets of diameter $\eta < \varepsilon/2$ by means of which $\tilde{V}_{\delta,\varepsilon/2}$ can be covered, where $\tilde{V}_{\delta,\varepsilon/2} = \tilde{W}_\delta \cap \tilde{A}_{\varepsilon/2}$, $\tilde{W}_\delta$ and $\tilde{A}_{\varepsilon/2}$ being, respectively, a weak and a Hellinger neighbourhood of $\tilde{f}$. Note that $\tilde{V}_{\delta,\varepsilon/2} \supseteq V_{\delta,\varepsilon}$ and that any covering of $V_{\delta,\varepsilon}$ can be extended to a covering of $\tilde{V}_{d,\varepsilon/2}$. Thus,

$$\sum_{j \geqslant 1} \Pi^\alpha(\tilde{V}_j) = +\infty$$

must hold. Since $\tilde{f}$ is arbitrary, consistency would fail at each density in $A_{\varepsilon/2}$, thus contradicting Doob's theorem; see Lijoi et al. (2004). Hence, Π cannot be inconsistent at all densities in $A_{\varepsilon/2}$, even if for each such density the series of the Π^α-probabilities diverges, for some α in $(0, 1)$.

Alternatively one can face the issue of establishing the validity of (2), and thus of consistency relying upon the construction of a suitable covering of the random set $B_{n,\gamma}^c$. In the following, $(c_k)_{k \geqslant 1}$ is an increasing sequence of positive numbers such that $0 < \sup_k(c_{k+1} - c_k) = c^* < +\infty$ and, given $\eta > 0$, we set $\delta \geqslant c^* + \eta$. Moreover, let

$$C_{n,k} := \{f : e^{nc_k} \leqslant R_n(f) < e^{nc_{k+1}}\}.$$

Theorem 3. *Let f_0 be in the Kullback–Leibler support of Π. Assume that for all $k \geqslant 1$ there exists a positive integer $n_0 = n_0(k)$ and $\xi_k > 0$, with $\sum \xi_k < +\infty$, such that, for all $n \geqslant n_0$,*

$$P_0^\infty[\Pi(C_{n,k}) < \xi_k \exp\{-n(\delta + c_k)\}] = 1,$$

for some sequence $(c_k)_{k\geqslant 1}$ of the type defined above and $\delta \geqslant c^* + \eta$. Then Π is Hellinger-consistent at f_0.

Special attention is required for Π, based on knowledge of f_0, in order to contradict the assumption of the theorem, bearing in mind that

$$\sum_k \exp(nc_k)\Pi(C_{n,k}) \to 0$$

for all choices of $\{c_k\}$.

It is therefore interesting to investigate why the prior suggested in the counterexample of Barron et al. (1999) does not meet the condition given in the above Theorem 3. Their prior assigns positive masses to single densities for which $R_n(f) = 2^n$, with f_0 being the uniform density on $[0, 1]$, thus explaining the phenomenon of tracking the data. To be more precise, for any sample $X_1, \ldots, X_n$ and $\gamma < \log 2$, one has that $\Pi(B_{n,\gamma}^c) \geqslant \exp(-2)2^{-n}/(2c_0 n)$, where c_0 is some positive constant. For any sequence $(c_k)_{k\geqslant 1}$ defined as above, let $k^* = k^*(n)$ be such that $2^n \in C_{n,k^*}$. This means that $\Pi(C_{n,k^*}) = \Pi(B_{n,\gamma}^c)$ and $\exp(-nc_{k^*+1}) > 2^{-n}$. Hence, for any $\delta \geqslant c^* + \eta$,

$$\exp(-n\delta - nc_{k^*}) < \exp(-n\eta)2^{-n}$$

and, since $\eta > 0$, inequality (4) must hold true for all n large enough:

$$\Pi(C_{n,k^*}) \geqslant \frac{\exp(-2)2^{-n}}{2c_0 n} > \exp(-n\eta)2^{-n} > \xi_{k^*} \exp\{-n(\delta + c_{k^*})\}, \tag{4}$$

for any sample $X_1, \ldots, X_n$.

4. Connecting ideas

The purpose of this section is to bring together the various results for the case in which $\Pi(V_{\delta,\varepsilon}) > 0$ and to understand (3) further, assuming without loss of generality that $\alpha = \frac{1}{2}$. It provides some insight about the sets $V_{\delta,\varepsilon}$.

For (3) to be satisfied with $\alpha = \frac{1}{2}$ we can, when the prior Π does not put mass on single densities, achieve Hellinger neighbourhoods of size no greater than $\varepsilon > 0$ from a dense set $\{f_k\}$, $B_k = N_{e_k}(f_k)$ say, such that $e_k \leqslant \varepsilon$ for all k and $\Pi\{N_{e_k}(f_k)\} \leqslant M/k^{2+r}$ for some finite M and $r > 0$. We can pick the $\{e_k\}$ to make this hold: if $\Pi\{N_\varepsilon(f_k)\} < M/k^{2+r}$ then we take $e_k = \varepsilon$, and otherwise we take $e_k < \varepsilon$ such that $\Pi\{N_{e_k}(f_k)\} = M/k^{2+r}$. Hence, with this we have

$$\sum_k \Pi^{\frac{1}{2}}(B_k) < +\infty,$$

ensuring that

$$\Pi_n\left\{\bigcup_k B_k \backslash N_\varepsilon(f_0)\right\} \to 0,$$

almost surely. However $\cup_k B_k$ may not cover the space of densities $\mathbb{F}$. To investigate what may be left out, consider $\mathbb{F}^* = \cup_k B_k$ and let $S = \mathbb{F} \backslash \mathbb{F}^*$. We can state immediately that, if it turns out to be that $e_k > \eta > 0$ for some constant η and for all large k, then we have that $S = \varnothing$ and consistency holds. If this is not the case, then S must be closed since $\mathbb{F}^*$ is open. Consequently, S is nowhere dense, since $\mathbb{F} \backslash S$ is dense in $\mathbb{F}$. Therefore S is where the posterior could put mass, a nowhere dense, closed, and thus with empty interior, subset of $\mathbb{F}$. For inconsistency to occur it must be that $S \cap V_{\delta,\varepsilon} \neq \varnothing$ for all $\delta > 0$ and $\varepsilon > 0$; that is S must contain a sequence of densities which converge weakly to f_0 but not in a strong sense.

We use the subscript M to denote the dependence on M, which can be arbitrarily large, and we establish that $\Pi(S_M) \to 0$ as $M \to \infty$. For any $\delta > 0$ and any $\varepsilon > 0$ there exists an $L < +\infty$ such that

$$\Pi\left\{\bigcup_{k=1}^{L} N_\varepsilon(f_k)\right\} > 1 - \delta.$$

If we choose M sufficiently big so that $e_k > \varepsilon$, for all $k \in \{1, \ldots, L\}$, then clearly

$$\Pi\left\{\bigcup_{k} N_{e_k}(f_k)\right\} > 1 - \delta$$

as well. We can do this by taking M such that

$$M/L^{2+r} > \max_{k \in 1,\ldots,L} \Pi\{N_\varepsilon(f_k)\}$$

and noting that

$$\Pi\{N_{e_k}(f_k)\} \geqslant M/L^{2+r},$$

for all $k = 1, \ldots, L$. Therefore, $\Pi(\mathbb{F}^*_M) \to 1$ as $M \to \infty$. In fact, $S_M \downarrow S'$ for some set S' with $\Pi(S') = 0$ and

$$S' = \bigcap_M S_M.$$

Therefore, for inconsistency, Π_n must put mass into $\Delta_M = S_M \backslash S'$, for each M, and $\Delta_M \downarrow \varnothing$.

To summarise, we know that for inconsistency the posterior must put mass into a subset of $V_{\delta,\varepsilon}$. Now we know what this subset is like; it is closed, nowhere dense and, based on the arbitrariness of M, can be made arbitrarily close to the empty set.

Acknowledgement

The authors are grateful to an associate editor and to an anonymous referee for their valuable comments. The research of Stephen G. Walker was funded by an Advanced Research Fellowship from the U.K. Engineering and Physical Sciences Research Council. The research of Antonio Lijoi and Igor Prünster was partially supported by a grant from the Italian Ministry of University and Research. Antonio Lijoi is also affiliated to the Istituto di Matematica Applicata e Tecnologie Informatiche, Milano, Italy. Igor Prünster is also affiliated to the International Center of Economic Research, Torino, Italy.

Appendix

Proofs

Proof of Theorem 1. In order to prove Theorem 1 we need to introduce

$$d_{K,n}(f_0, f) = \frac{1}{n}\sum_{i=1}^{n} \log\{f_0(X_i)/f(X_i)\},$$

the sample Kullback–Leibler divergence between f_0 and f. Since f_0 is in the Kullback–Leibler support of the prior, from Schwartz (1965) one has that

$$\frac{1}{n}\log I_n \to 0,$$

almost surely, and, from A. R. Barron's technical report, that $E_0(n^{-1}\log I_n) \to 0$. Moreover, the identity

$$-\frac{1}{n}\log I_n = \frac{1}{n}D_K(\mu, \Pi) - \frac{1}{n}D_K(\mu, \Pi_n) + \int d_{K,n}(f_0, f)\mu(df) \qquad \text{(A1)}$$

holds true for any measure μ which is absolutely continuous with respect to Π. Indeed,

$$-\log I_n = \int \log\left\{\frac{R_n(f)\Pi(df)}{I_n\Pi(df)}\right\}\mu(df) - \int \log R_n(f)\mu(df)$$

$$= \int \log(d\Pi_n/d\Pi)d\mu + n\int d_{K,n}(f_0, f)\mu(df).$$

Now we let $I_n^{(j)} = \int R_n^{(j)}(f)\Pi(df)$, where

$$R_n^{(j)}(f) = \prod_{i=1}^{n} \frac{f(X_i^{(j)})}{f_0(X_i^{(j)})} \quad (j = 1, 2).$$

If in (A1) we set $\Pi_n = \Pi_n^{(1)}$ and $\mu = \Pi_n^{(2)}$, we have

$$-\frac{1}{n}\log I_n^{(1)} = \frac{1}{n}D_K(\Pi_n^{(2)}, \Pi) - \frac{1}{n}D_K(\Pi_n^{(2)}, \Pi_n^{(1)}) + \int d_{K,n}^{(1)}(f_0, f)\Pi_n^{(2)}(df),$$

where $d_{K,n}^{(1)}(f_0, f) := (1/n)\sum_{i=1}^{n} \log\{f_0(X_i^{(1)})/f(X_i^{(1)})\}$. Take expectations with respect to $X_1^{(1)}, \ldots, X_n^{(1)}$ and keep $X_1^{(2)}, \ldots, X_n^{(2)}$ fixed. Then, using the fact that $E_0^{(1)}(n^{-1}\log I_n^{(1)}) \to 0$, we have that

$$E_0^{(1)}\left\{\frac{1}{n}D_K(\Pi_n^{(2)}, \Pi_n^{(1)})\right\} - \frac{1}{n}D_K(\Pi_n^{(2)}, \Pi) - \int d_K(f_0, f)\Pi_n^{(2)}(df) \to 0,$$

as $n \to +\infty$. Finally, the inconsistency of Π, applied to $\Pi_n^{(2)}$, yields

$$\limsup_n \int d_K(f_0, f)\Pi_n^{(2)}(df) > 0,$$

almost surely, so that

$$\limsup_n E_0^{(1)}\left\{\frac{1}{n}D_K(\Pi_n^{(2)}, \Pi_n^{(1)})\right\} > 0,$$

almost surely, and the result follows. □

Proof of Theorem 2. Note that, for any f in $B_{n,\gamma}$, one has $R_n(f)e^{-n\gamma} < 1$, which yields

$$e^{-n\gamma}\int_{A_\varepsilon^c \cap B_{n,\gamma}} R_n(f)\Pi(df) < e^{-n\gamma/2}\int_{A_\varepsilon^c \cap B_{n,\gamma}} R_n^{1/2}(f)\Pi(df). \qquad \text{(A2)}$$

If we observe that

$$E_0\left\{\int_{A_\varepsilon^c\cap B_{n,\gamma}} R_n^{1/2}(f)\Pi(df)\right\} < (1-\varepsilon)^n\Pi(A_\varepsilon^c),$$

where E_0 denotes the expected value with respect to P_0^∞, and apply the Markov inequality, then

$$P_0^\infty\left\{\int_{A_\varepsilon^c\cap B_{n,\gamma}} R_n^{1/2}(f)\Pi(df) > e^{-n\delta}\right\} \leqslant \Pi(A_\varepsilon^c)e^{-n\{-\log(1-\varepsilon)-\delta\}},$$

where $\delta > 0$ is chosen in such a way that $\gamma/2 < \delta < -\log(1-\varepsilon)$. Hence, the Borel–Cantelli lemma leads to

$$P_0^\infty\left[\bigcup_{N\geqslant 1}\bigcap_{n\geqslant N}\left\{\int_{A_\varepsilon^c\cap B_{n,\gamma}} R_n^{1/2}(f)\Pi(df) \leqslant e^{-n\delta}\right\}\right] = 1,$$

which, combined with (A2), implies that, for all but a finite number of n's,

$$\int_{A_\varepsilon^c\cap B_{n,\gamma}} R_n(f)\Pi(df) < \exp\{-n(\delta-\gamma/2)\},$$

almost surely. Since $f_0 \in S_K(\Pi)$, one has that, for any $\beta > 0$ and for all but a finite number of n's,

$$I_n = \int_{\mathbb{F}} R_n(f)\Pi(df) > e^{-n\beta},$$

almost surely. If we fix $\beta < \delta - \gamma/2$, then $\Pi_n(A_\varepsilon^c \cap B_{n,\gamma}) \to 0$. □

Proof of Theorem 3. Let

$$\frac{\int_{A_\varepsilon^c} R_n(f)\Pi(df)}{\int_{\mathbb{F}} R_n(f)\Pi(df)} = \Pi_n(A_\varepsilon^c) = \Pi_n(A_\varepsilon^c\cap B_{n,\gamma}) + \Pi_n(A_\varepsilon^c\cap B_{n,\gamma}^c).$$

By Theorem 2, the first summand on the right-hand side tends to 0, almost surely with respect to P_0^∞. Thus, we focus attention on $\Pi_n(A_\varepsilon^c\cap B_{n,\gamma}^c)$. If $(c_k)_{k\geqslant 1}$ is the sequence described above, with $c_1 = \gamma$, then

$$\Pi_n(A_\varepsilon^c\cap B_{n,\gamma}^c) = \frac{J_n}{I_n} = \frac{\sum_k \int_{C_{n,k}\cap A_\varepsilon^c} R_n(f)\Pi(df)}{\int_{\mathbb{F}} R_n(f)\Pi(df)},$$

$$I_n > J_n > \sum_k e^{nc_k}\Pi(C_{n,k}).$$

According to Lemma 1 in Barron et al. (1999), the fact that f_0 is in $S_K(\Pi)$ implies that, for each n, $I_n \notin \{0, \infty\}$ almost surely. By virtue of the hypothesis on Π, one has

$$J_n < \sum_k e^{nc_{k+1}}\Pi(C_{n,k}) < \sum_k \xi_k e^{-n(\delta - c_{k+1}+c_k)} < e^{-n\eta}\sum_k \xi_k,$$

almost surely. Hence, since $I_n < \exp(n\beta)$ for any $\beta > 0$, choosing $\beta < \eta$ leads to $\Pi_n(A_\varepsilon^c\cap B_{n,\gamma}^c) \to 0$, almost surely with respect to P_0^∞. □

References

Barron, A., Schervish, M. J. & Wasserman, L. (1999). The consistency of distributions in nonparametric problems. *Ann. Statist.* **27**, 536–61.

Blackwell, D. & Dubins, L. (1962). Merging of opinions with increasing information. *Ann. Math. Statist.* **33**, 882–6.

Brunner, L. J. & Lo, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.* **17**, 1550–66.

Diaconis, P. & Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–26.

Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, pp. 23–27. Paris: Colloques Internationaux du Centre National de la Recherche Scientifique.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. II, 2nd ed. New York: Wiley.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386–403.

Ghosal, S., Ghosh, J. K. & Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143–58.

Hansen, M. B. & Lauritzen, S. L. (2002). Nonparametric Bayes inference for concave distribution functions. *Statist. Neer.* **56**, 110–27.

Lijoi, A., Prünster, I. & Walker, S. G. (2004). Extending Doob's consistency theorem to nonparametric densities. *Bernoulli* **10**, 651–63.

Petrone, S. & Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *J. R. Statist. Soc.* B **64**, 79–100.

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

Walker, S. G. (2003). On sufficient conditions for Bayesian consistency. *Biometrika* **90**, 482–8.

Walker, S. G. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32**, 2028–43.

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

[*Received October* 2004. *Revised May* 2005]