Bernoulli **10**(4), 2004, 651–663

Extending Doob's consistency theorem to nonparametric densities

ANTONIO LIJOI[1*], IGOR PRÜNSTER[1**] and STEPHEN G. WALKER[2]

[1]*Dipartimento di Economia Politica e Metodi Quantitativi, Università degli Studi di Pavia, Via San Felice 5, 27100 Pavia, Italy. E-mail: *lijoi@unipv.it; **igor.pruenster@unipv.it*
[2]*Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK.*
E-mail: S.G.Walker@bath.ac.uk

We extend Doob's well-known result on Bayesian consistency. The extension covers the case where the nonparametric prior is fully supported by densities. However, our use of martingales differs from that of Doob. We also consider rates.

Keywords: consistency; Hellinger distance; martingale; rate of convergence

1. Introduction

In Doob (1949) an application of the martingale convergence theorem to the study of consistency of Bayesian procedures is pointed out. In particular, it is proved that if there exists a consistent estimator of a parameter $\tilde{\theta}$, then the posterior distribution of $\tilde{\theta}$ accumulates in neighbourhoods of $\tilde{\theta}$ almost surely.

In order to highlight connections between the present paper and Doob's result, some notation is introduced. Let $(X_n)_{n\geqslant 1}$ be a sequence of random variables, on some measurable space (Ω, σ), taking values in $(\mathbb{X}, \mathcal{X})$, where $\mathbb{X}$ is a Polish space and $\mathcal{X}$ is the Borel σ-field of sets of $\mathbb{X}$. As usual, $\mathbb{X}^n$ is the n-fold Cartesian product $\mathbb{X} \times \cdots \times \mathbb{X}$ and $\mathbb{X}^\infty$ is the infinite product space, while $\mathcal{X}^n$ and $\mathcal{X}^\infty$ are the corresponding σ-fields. Moreover, denote by Θ the parameter space, assumed to be Polish, and by $\mathcal{A}_\Theta$ a σ-field of sets of Θ. Let $\{P_\theta : \theta \in \Theta\}$ be a family of probability distributions on $(\mathbb{X}, \mathcal{X})$ such that $\theta \mapsto P_\theta(A)$ is $\mathcal{A}_\Theta$-measurable, for every A in $\mathcal{X}$. For our purposes, it is useful to consider $X_1, X_2, \ldots, \tilde{\theta}$ as coordinate random variables defined on $\Omega = \mathbb{X}^\infty \times \Theta$ and to set $\sigma = \mathcal{X}^\infty \otimes \mathcal{A}_\Theta$ as the usual product σ-field. Hence, if Π is a probability distribution on the parameter space $(\Theta, \mathcal{A}_\Theta)$, a probability measure P on the product space $(\mathbb{X}^\infty \times \Theta, \mathcal{X}^\infty \otimes \mathcal{A}_\Theta)$ is uniquely defined by

$$P(A \times B) = \int_B \prod_{i=1}^n P_\theta(A_i)\, \Pi(\mathrm{d}\theta) \tag{1}$$

for any $B \in \mathcal{A}_\Theta$ and $A = A_1 \times \cdots \times A_n \times \mathbb{X}^\infty$, with $A_1, \ldots, A_n \in \mathcal{X}$, for any n. The probability measure Π will be referred to as the prior distribution of $\tilde{\theta}$. The posterior

1350–7265 © 2004 ISI/BS

distribution, given the data $X^{(n)} = (X_1, \ldots, X_n)$, is denoted by Π_n. According to Doob, if $\tilde{\theta}$ is $\mathcal{X}^\infty$-measurable, that is, it is a function of the whole sequence of observations, then

$$\Pi_n(A) \to \mathbb{I}_A(\tilde{\theta}) \text{ a.s.}[P], \tag{2}$$

for any measurable set A, where $\mathbb{I}_B$ is the indicator function of a set B. Schwartz (1965) reconsidered this result and extended it to a decision-theoretic framework.

Denote now by $(\mathbb{F}, H)$ the metric space of densities with respect to some σ-finite measure λ on $\mathbb{X}$, where H is the Hellinger distance defined by

$$H^2(f, g) = \int_{\mathbb{X}} \left\{ f(x)^{1/2} - g(x)^{1/2} \right\}^2 \lambda(\mathrm{d}x)$$

for any pair of densities g and f in $\mathbb{F}$. If λ is the Lebesgue measure on $\mathbb{R}$, the space of absolutely continuous distribution functions is not closed under pointwise limits of sequences of its elements and, thus, hypothesis (A) in Doob (1949) is not met. In other words, if $\tilde{\theta}$ is a nonparametric random density function on $\mathbb{R}$, one needs to prove the consistency result in (2) without resorting to the assumption that $\tilde{\theta}$ is $\mathcal{X}^\infty$-measurable. In this sense we provide an extension of Doob's result and show that the posterior distribution Π_n accumulates in Hellinger neighbourhoods of an essentially unique random density function $\tilde{g}$ and that the X_k are independent and identically distributed (i.i.d.) given $\tilde{g}$.

Recent studies on consistency rely upon a 'frequentist', or so-called 'what if', method due to Diaconis and Freedman (1986). This approach is usually motivated by the fact that P null sets in (2) can be large. Nonetheless, we believe that Doob-type results are still worth examining in a nonparametric framework for various reasons. Firstly, the objection relating to the size of the P null sets can be circumvented by ensuring that the prior Π has full Hellinger support. This implies that the P null sets are just single densities, meaning that they are isolated points in $\mathbb{F}$ with respect to the L_1 topology. In Section 3 it is proved that many common priors on densities do have full Hellinger support under fairly natural conditions. Secondly, the results presented here are genuinely Bayesian since they refer to the product measure P defined in (1), instead of fixing a 'true' but 'unknown' density function f_0. When Bayesian consistency results are available and P null sets are single densities, one might wonder whether possible inconsistency at a hypothesized f_0 can be thought of as an irrelevant nuisance. Apart from its theoretical relevance, a Doob-type result for random densities is also of practical interest. To clarify this point, here we refer to consistency of decision problems involving sample sizes for which the 'frequentist' approach is meaningless. Let $\mathcal{D} = \{a_1, \ldots, a_N\}$ be a finite set of actions and let $U(a, g)$ be the utility one attains when undertaking action a, where g is a random element describing the 'state of nature' and taking values in a set Θ. Define a_g to be the element in $\mathcal{D}$ that maximizes $a \mapsto U(a, g)$, that is, $U(a_g, g) = \max_{\mathcal{D}} U(a, g)$. A Bayesian would fix a prior Π on Θ and, given a set of n observations, would determine the posterior Π_n. Hence, the corresponding action, denoted by $a(n)$, is such that

$$\int_\Theta U\{a(n), g\}\Pi_n(\mathrm{d}g) = \max_{a \in \mathcal{D}} \int_\Theta U(a, g)\Pi_n(\mathrm{d}g).$$

By definition, the best expected utility is given by

$$\overline{U} := \int_{\Theta} U(a_g, g)\Pi(\mathrm{d}g),$$

and the expected utility $U(n)$ associated with the sample size n is defined as

$$U(n) := \mathrm{E}\left\{\int_{\Theta} U\{a(n), g\}\, \Pi_n(\mathrm{d}g)\right\},$$

where the expectation is computed with respect to the marginal distribution of the observations. At this point, it is natural to require the expected utility $U(n)$ to converge to $\overline{U}$, as the sample size n increases, meaning that the best possible outcome is achieved if 'all the information' is available. This happens if U is bounded and

$$\int_{\Theta} U\{a(n), g\}\, \Pi_n(\mathrm{d}g) \to U(a_{\tilde{g}}, \tilde{g}) \text{ a.s.}[P],$$

where $\tilde{g}$ is distributed according to the prior Π. With further regularity conditions on $U(a, g)$, namely continuity in g, one has the above if Π_n converges weakly to a probability measure with all its mass on $\tilde{g}$. But this is guaranteed if a Doob consistency result holds for Π_n.

In Section 2 the consistency result is stated and proved. In Section 3 some illustrative examples of priors with full Hellinger support are provided. Section 4 provides information concerning rates of convergence, and finally Section 5 contains a brief discussion.

2. Extension of Doob's result

In accordance with notation introduced in the previous section, suppose the parameter space, Θ, coincides with the set, $\mathbb{F}$, of all densities on $(\mathbb{X}, \mathcal{X})$ with respect to some measure λ. Moreover, let $\mathcal{F}$ denote a σ-field of subsets of $\mathbb{F}$. The predictive density function, given n observations $X_1, \ldots, X_n$, is

$$f_n(x) = \int_{\mathbb{F}} f(x)\, \Pi_n(\mathrm{d}f),$$

where

$$\Pi_n(\mathrm{d}f) = \frac{\prod_{i=1}^{n} f(X_i)\Pi(\mathrm{d}f)}{\int_{\mathbb{F}} \prod_{i=1}^{n} f(X_i)\Pi(\mathrm{d}f)}$$

is the posterior distribution. Simple computations lead to the useful equality

$$\frac{\Pi_{n+1}(A)}{\Pi_n(A)} = \frac{f_{n\,A}(X_{n+1})}{f_n(X_{n+1})} \tag{3}$$

for any measurable set of densities A. Here, f_{nA} is the predictive density based on the posterior restricted to the set A, that is,

$$f_{n\,A}(x) = \frac{\displaystyle\int_A f(x)\,\Pi_n(\mathrm{d}f)}{\Pi_n(A)}, \qquad \forall x \in \mathbb{X}.$$

Finally, instead of using the Hellinger distance H, a slight modification of it is considered, that is,

$$h(f,\ g) = 1 - \int_{\mathbb{X}} \sqrt{f(x)\,g(x)}\lambda(\mathrm{d}x), \qquad \forall f,\ g \in \mathbb{F}.$$

In the following, δ_x denotes the Dirac function at the point x. The following theorem can now be proved.

Theorem 1. *There exists a random element $\tilde{g}$ such that*

$$\Pi_n(A) \to \delta_{\tilde{g}}(A) \text{ a.s.}[P]$$

as n tends to $+\infty$, for any A in $\mathcal{F}$. Moreover, such a $\tilde{g}$ is essentially unique and the X_j are conditionally i.i.d. given $\tilde{g}$.

Proof. If $\sigma_n = \sigma(X_1, \ldots, X_n)$, then $\mathrm{E}(\Pi_{n+1}(A)|\sigma_n) = \Pi_n(A)$ for any $A \in \mathcal{F}$. Hence by the martingale convergence theorem, there exists a random element Π_∞ such that

$$\Pi_n(A) \to \Pi_\infty(A) \text{ a.s.}[P],$$

for all sets $A \in \mathcal{F}$. Moreover, by a result proved in Grey (2001), Π_∞ is a probability measure on a set having P probability 1. By suitably completing the definition of Π_∞ outside such a set, one has that Π_∞ is a random probability measure. Moreover, by virtue of (3), one easily obtains

$$\mathrm{E}\Big\{\Pi_{n+1}^{1/2}(A)|\sigma_n\Big\} = \Pi_n^{1/2}(A)\{1 - h(f_{n\,A},\ f_n)\}.$$

Consider, now, the martingale $(S_N,\ \sigma_{_N})_{N\geqslant 1}$ defined by

$$S_N = \sum_{n=1}^{N}\Big[\Pi_n^{1/2}(A) - \Pi_{n-1}^{1/2}(A)\{1 - h(f_{n\,A},\ f_n)\}\Big] \tag{4}$$

$$= \Pi_N^{1/2}(A) - \Pi^{1/2}(A) + \sum_{n=1}^{N}\Pi_{n-1}^{1/2}(A)h(f_{n-1\,A},\ f_{n-1}).$$

One can prove that

$$\mathrm{E}\Bigg\{\sum_n \Pi_n^{1/2}(A)h(f_{n\,A},\ f_n)\Bigg\} < \infty.$$

Straightforward application of the Borel–Cantelli lemma yields

$$\Pi_n^{1/2}(A)h(f_{n\,A},\ f_n) \to 0 \text{ a.s.}[P],$$

as n tends to $+\infty$. Let Ω_0 be the set on which convergence occurs. Then for any $\omega \in \Omega_0$ there exists a density function g such that $\Pi_\infty(A^\varepsilon) > 0$ for all $\varepsilon > 0$, where $A^\varepsilon = \{f \in \mathbb{F} : h(f,\ g) < \varepsilon\}$, since Π_∞ is a probability measure. Hence, one has

$$h(f_{n\,A^\varepsilon},\ f_n) \to 0 \qquad (n \to +\infty)$$

and, by virtue of the triangular inequality and of the convexity of $h(\cdot,\ \cdot)$, it follows that $h(f_n,\ g) \to 0$. Moreover, it is easy to show that such a g is unique. Thus, let $\tilde{g} : \Omega \to \mathbb{F}$ be a function that associates to each ω a density function g such that $\Pi_\infty(A^\varepsilon) > 0$, for any $\varepsilon > 0$. Such a function is measurable. Indeed, for any $B \in \mathcal{F}$,

$$\tilde{g}^{-1}(B) \subset \{\omega \in \Omega : \Pi_\infty(B) > 0\}.$$

On the other hand, if $\omega \in \Omega$ is such that $\Pi_\infty(B) > 0$, there exists a density g in B such that g is in the support of Π_∞. This means that

$$\tilde{g}^{-1}(B) \supset \{\omega \in \Omega : \Pi_\infty(B) > 0\}$$

and measurability of $\tilde{g}$ follows from the fact that Π_∞ is a random probability measure. Moreover, equality between the two sets above implies

$$\Pi_\infty = \delta_{\tilde{g}}.$$

Finally, knowing that the observations are conditionally i.i.d. given a random density $\tilde{f}$, we wish to prove that $\tilde{f} = \tilde{g}$ a.s.$[P]$. If $P_{\tilde{f}}(B) = \int_B \tilde{f}(x)\lambda(\mathrm{d}x)$ for any B in $\mathcal{X}$, then

$$\begin{aligned}\mathrm{E}\{P_{\tilde{f}}(B)|\sigma_n\} = P(X_{n+1} \in B|\sigma_n) &= \int_{\mathbb{F}} P_f(B)\Pi_n(\mathrm{d}f)\\ &= \int_{\mathbb{F}} P_f(B)\mathrm{E}\{\Pi_\infty(\mathrm{d}f)|\sigma_n\} = \mathrm{E}\left\{\int_{\mathbb{F}} P_f(B)\Pi_\infty(\mathrm{d}f)|\sigma_n\right\}\end{aligned}$$

where the last equality follows from the definition of conditional expectation. Since $P_{\tilde{f}}(B)$ and $\int_{\mathbb{F}} P_f(B)\Pi_\infty(\mathrm{d}f)$ are bounded, one has

$$\begin{aligned}\mathrm{E}(P_{\tilde{f}}(B)|\sigma_\infty) &= \mathrm{E}\left\{\int_{\mathbb{F}} P_f(B)\Pi_\infty(\mathrm{d}f)|\sigma_\infty\right\}\\ &= \int_{\mathbb{F}} P_f(B)\Pi_\infty(\mathrm{d}f) = P_{\tilde{g}}(B),\end{aligned}$$

where the last equality follows from the fact that $\Pi_\infty = \delta_{\tilde{g}}$. Moreover, by de Finetti's representation theorem, $P_{\tilde{f}}$ is σ_∞ measurable, and then

$$P_{\tilde{f}}(B) = P_{\tilde{g}}(B) \text{ a.s.}[P],$$

for every B in $\mathcal{X}$. This completes the proof. □

3. Illustrative examples

A few examples involving priors on space of densities are presented. They all have full Hellinger support, under suitable conditions, thus providing evidence of the fact that the only P null sets on which consistency may fail are single densities.

3.1. Mixture models

Let k be a non-negative valued kernel on $(\mathbb{X}\times\mathbb{Y},\ \mathcal{X}\otimes\mathcal{Y})$, $\mathbb{X}$ and $\mathbb{Y}$ being subsets of the real line, such that:

(i) $\int_{\mathbb{X}} k(x,\ y)\,\lambda(\mathrm{d}x) = 1$ for any $y\in\mathbb{Y}$, and for some σ-finite measure λ on $(\mathbb{X},\ \mathcal{X})$;
(ii) $y\mapsto k(x,\ y)$ is bounded, continuous and $\mathcal{Y}$ measurable, for each x in $\mathbb{X}$.

If $\tilde{Q}$ is a random probability measure on $(\mathbb{Y},\ \mathcal{Y})$, the quantity

$$\tilde{f}(x) = \int_{\mathbb{Y}} k(x,\ y)\,\tilde{Q}(\mathrm{d}y) \tag{5}$$

is a random density function. Such a mixture has been investigated by Lo (1984) for Bayesian density estimation when $\tilde{Q}$ is a Dirichlet process, and has recently gained some attention with reference to consistency problems when used as a prior on a space of densities; see, for example, Ghosal *et al.* (1999a) and Petrone and Wasserman (2002).

Here, it is shown that if a kernel satisfies (i)–(ii) and a mild additional condition to be specified later on, the probability distribution of $\tilde{f}$ in (5) puts positive mass on all Hellinger neighbourhoods of $f_0(x) = \int_{\mathbb{Y}} k(x,\ y)\,Q_0(\mathrm{d}y)$ as long as the distribution μ^* of $\tilde{Q}$ has full weak support. In other terms, given any such density f_0, its ε-Hellinger neighbourhoods have positive probability, for any $\varepsilon > 0$.

Let Q_0 be any probability measure on $(\mathbb{Y},\ \mathcal{Y})$ whose δ-weak neighbourhood, $W_\delta^{(0)}$ say, has positive μ^*-probability, for any $\delta > 0$. Given $\varepsilon > 0$, fix a compact set $\mathcal{K}\in\mathcal{X}$ such that

$$\int_{\mathcal{K}^c} f_0(x)\,\lambda(\mathrm{d}x) < \frac{\varepsilon}{8}, \tag{6}$$

with $\lambda(\mathcal{K}) < +\infty$. From now on, Q is any probability measure in $W_\delta^{(0)}$. Then

$$\left|\int_{\mathcal{K}^c}\int_{\mathbb{Y}} k(x,\ y)\,Q(\mathrm{d}y)\,\lambda(\mathrm{d}x) - \int_{\mathcal{K}^c}\int_{\mathbb{Y}} k(x,\ y)\,Q_0(\mathrm{d}y)\lambda(\mathrm{d}x)\right| < \delta$$

since $y\mapsto\int_{\mathcal{K}^c} k(x,\ y)\lambda(\mathrm{d}x)$ is bounded and continuous, and combination with (6) yields

$$\int_{\mathcal{K}^c}\int_{\mathbb{Y}} k(x,\ y)\,Q(\mathrm{d}y)\,\lambda(\mathrm{d}x) = \int_{\mathcal{K}^c} f(x)\,\lambda(\mathrm{d}x) < \frac{\varepsilon}{8} + \delta. \tag{7}$$

Moreover, for any $\rho > 0$, one can set $M_\rho > 0$ such that $Q_0([-M_\rho + \delta,\ M_\rho - \delta]^c) < \rho$ which, in turn, implies $Q([-M_\rho,\ M_\rho]^c) < \rho + \delta$.

It is proved that the distribution of $\tilde{f}$ in (5) has full Hellinger support if the following additional condition is met by the kernel k:

(iii) The family $\{k(x, y) : x \in \mathcal{K}\}$ of functions of y, as y varies in $[-M_\rho, M_\rho]$, is uniformly equicontinuous.

By virtue of the Arzelà–Ascoli theorem, (iii) implies that, given $\eta > 0$, there exists a finite partition of $\mathcal{K}$ into sets $A_1, \ldots, A_m$ and points $x_1, \ldots, x_m$ such that

$$\sup_{|y|<M_\rho} |k(x, y) - k(x_j, y)| < \eta$$

for any $x \in A_j$ and for each $j = 1, \ldots, m$. Note that

$$\int_{\mathbb{X}} |f(x) - f_0(x)|\, \lambda(\mathrm{d}x) = \int_{\mathcal{K}} |f(x) - f_0(x)|\, \lambda(\mathrm{d}x) + \int_{\mathcal{K}^c} |f(x) - f_0(x)|\, \lambda(\mathrm{d}x),$$

and observe that the second summand on the right-hand side is bounded by $(\varepsilon/4) + \delta$. As far as the integral over $\mathcal{K}$ is concerned, simple algebra yields

$$\int_{\mathcal{K}} |f(x) - f_0(x)|\, \lambda(\mathrm{d}x) \leqslant I_1 + I_2 + I_3$$

where

$$I_1 = \sum_{j=1}^{m} \int_{A_j} \left| \int_{\mathbb{Y}} k(x, y)\, Q(\mathrm{d}y) - \int_{\mathbb{Y}} k(x_j, y)\, Q(\mathrm{d}y) \right| \lambda(\mathrm{d}x),$$

$$I_2 = \sum_{j=1}^{m} \int_{A_j} \left| \int_{\mathbb{Y}} k(x_j, y)\, Q(\mathrm{d}y) - \int_{\mathbb{Y}} k(x_j, y)\, Q_0(\mathrm{d}y) \right| \lambda(\mathrm{d}x),$$

$$I_3 = \sum_{j=1}^{m} \int_{A_j} \left| \int_{\mathbb{Y}} k(x_j, y)\, Q_0(\mathrm{d}y) - \int_{\mathbb{Y}} k(x, y)\, Q_0(\mathrm{d}y) \right| \lambda(\mathrm{d}x).$$

It is easy to show that $I_2 \leqslant \delta\, \lambda(\mathcal{K})$. When dealing with I_1 and I_3 it is worth considering the partition of $\mathbb{Y}$ into $[-M_\rho, M_\rho]$ and $[-M_\rho, M_\rho]^c$. We confine ourselves to considering just I_1. Set $V := \max_{1 \leqslant j \leqslant m} \sup_y k(x_j, y)$ and notice that it is finite because of (ii). Hence,

$$\begin{aligned} I_1 &\leqslant \sum_{j=1}^{m} \int_{A_j} \int_{|y| \leqslant M_\rho} |k(x, y) - k(x_j, y)|\, Q(\mathrm{d}y)\, \lambda(\mathrm{d}x) \\ &\quad + \sum_{j=1}^{m} \int_{A_j} \int_{|y| > M_\rho} |k(x, y) - k(x_j, y)|\, Q(\mathrm{d}y)\, \lambda(\mathrm{d}x) \\ &\leqslant \eta\, \lambda(\mathcal{K}) + \rho + \rho\, V\, \lambda(\mathcal{K}). \end{aligned}$$

A similar bound is obtained for I_3. Accordingly, if

$$\delta < \frac{\varepsilon}{4(1 + \lambda(\mathcal{K}))}, \qquad \eta < \frac{\varepsilon}{8\, \lambda(\mathcal{K})}, \qquad \rho < \frac{\varepsilon}{8(1 + V\, \lambda(\mathcal{K}))},$$

one has $\int_{\mathbb{X}} |f(x) - f_0(x)| \lambda(\mathrm{d}x) < \varepsilon$.

3.2. Pólya trees

The family of Pólya tree priors has been investigated in depth by Mauldin *et al.* (1992) and Lavine (1992). A brief introductory description is now provided. Let $E = \{0, 1\}$ and $E^m = \{0, 1\}^m$ for $m = 1, 2, \ldots$. Having set $E^0 = \varnothing$, define $E^* = \cup_{m \geqslant 0} E^m$. Introduce the sequence of nested and binary partitions $(\mathcal{P}_m)_{m \geqslant 1}$ of [0, 1] into dyadic intervals such that $\mathcal{P}_m = \{B_\varepsilon : \varepsilon \in E^m\}$. Moreover, let $\mathcal{P}_0 = [0, 1]$ and $\mathcal{P} = \{\mathcal{P}_m : m = 0, 1, 2, \ldots, \}$. Introduce a collection of non-negative numbers $\mathcal{A} = \{\alpha_\varepsilon : \varepsilon \in E^*\}$ and a collection of mutually independent random variables $\mathcal{Y} = \{Y_\varepsilon : \varepsilon \in E^*\}$ with Y_ε distributed according to the beta law of parameters $\alpha_{\varepsilon,0}$ and $\alpha_{\varepsilon,1}$ for each ε in E^*. Hence, a random probability measure $\tilde{P}$ on [0, 1] is said to have a Pólya tree distribution with parameters $(\mathcal{P}, \mathcal{A})$, $\mathrm{PT}(\mathcal{P}, \mathcal{A})$, if

$$\tilde{P}(B_{\varepsilon_1,\ldots,\varepsilon_m}) = \left\{ \prod_{j=1;\, \varepsilon_j=0}^{m} Y_{\varepsilon_1,\ldots,\varepsilon_{j-1}} \right\} \left\{ \prod_{j=1;\, \varepsilon_j=1}^{m} (1 - Y_{\varepsilon_1,\ldots,\varepsilon_{j-1}}) \right\}$$

for $m = 1, 2, \ldots$. In Kraft (1964) it is shown that if Y_ε becomes concentrated around $1/2$ sufficiently rapidly as B_ε shrinks along $\mathcal{P}$, then $\tilde{P}$ will have a density with respect to the Lebesgue measure with probability 1. In Lavine (1994) and Ghosal *et al.* (1999b) properties of the support of $\tilde{P}$ are studied. In particular, Ghosal *et al.* (1999b), by refining a similar result due to Lavine (1994), prove that $\tilde{P}$ has full Kullback–Leibler support provided that α_ε, where $\varepsilon \in E^m$, increases faster than $m^{2+\delta}$, for any $\delta > 0$, as m goes to $+\infty$. Here, it is shown that $\tilde{P}$ enjoys the weaker condition of full Hellinger support without imposing any conditions on the α_ε, apart from those guaranteeing absolute continuity of $\tilde{P}$.

Let f_0 be any density function on [0, 1] and proceed by contradiction, supposing that it does not belong to the Hellinger support of $\mathrm{PT}(\mathcal{P}, \mathcal{A})$, that is, there is a $\delta > 0$ such that

$$\mathrm{PT}\{f \in \mathbb{F} : h(f, f_0) > \delta\} = 1. \tag{8}$$

For any positive integer N, define

$$\tilde{f}_N(x) = \left\{ \prod_{j=1;\varepsilon_j=0}^{N} 2\, Y_{\varepsilon_1,\ldots,\varepsilon_j}(x) \right\} \left\{ \prod_{j=1;\varepsilon_j=1}^{N} 2\left(1 - Y_{\varepsilon_1,\ldots,\varepsilon_j}(x)\right) \right\},$$

where $Y_{\varepsilon_1,\ldots,\varepsilon_j}(x)$ denotes Y indexed by the first j digits in the dyadic expansion of $x \in [0, 1]$. It is proved in Kraft (1964) that $\tilde{f}_N$ converges, in L_1, to the random density function $\tilde{f}$ (almost surely). Given f_0, there exists a collection of numbers $\{y_\varepsilon : \varepsilon \in E^*\}$ such that

$$f_{N,0}(x) = \left\{ \prod_{j=1;\varepsilon_j=0}^{N} 2\, y_{\varepsilon_1,\ldots,\varepsilon_j}(x) \right\} \left\{ \prod_{j=1;\varepsilon_j=1}^{N} 2\left(1 - y_{\varepsilon_1,\ldots,\varepsilon_j}(x)\right) \right\}$$

and $f_0(x) = \lim_N f_{N,0}(x)$ for any x in [0, 1]. By the Scheffé theorem, $f_{N,0}$ converges, in L_1, to f_0. By virtue of the triangular inequality, (8) implies

$$\mathrm{PT}\{f \in \mathbb{F} : h(f, f_N) + h(f_N, f_{N,0}) + h(f_{N,0}, f_0) > \delta\} = 1.$$

The first and third summand in the previous probability statement can be made arbitrarily small. As far as the second is concerned, notice that

$$h(f_N,\, f_{N,0}) = 1 - \int_{[0,1]} \sqrt{f_N(x)\, f_{N,0}(x)}\, \lambda(\mathrm{d}x)$$

is expressed in terms of the product of N independent beta random variables. Hence, nonsingularity of the beta distribution ensures that $h(f_N,\, f_{N,0})$ can be made arbitrarily small with positive probability. Thus, there exists an integer $N' = N'(\delta)$ such that, for any $N \geqslant N'$,

$$\mathrm{PT}\left(\left\{h(f,\, f_N) < \frac{\delta}{3}\right\} \cap \left\{h(f_{N,0},\, f_N) < \frac{\delta}{3}\right\} \cap \left\{h(f_{N,0},\, f_0) < \frac{\delta}{3}\right\}\right) > 0,$$

which contradicts (8).

3.3. Infinite-dimensional exponential families

This particular family of priors on the space of densities was studied by Leonard (1978) and Lenk (1988; 1991). Let $\Psi = (\psi_n)_{n\geqslant 1}$ be a sequence of independent Gaussian random variables with ψ_n having zero mean and variance equal to τ_n^2. Moreover, introduce a sequence $\Phi = (\phi_n)_{n\geqslant 1}$ of orthogonal polynomials on [0, 1] and choose the τ_n in such a way that $\sum_j \sup_{0\leqslant x\leqslant 1} |\phi_j(x)|\, \tau_j < +\infty$. Hence

$$f(x) = \exp\left\{\sum_{n=1}^{\infty} \psi_n \phi_n(x) - C(\Psi)\right\}$$

is a random probability density function on [0, 1], with respect to the Lebesgue measure λ. The quantity $C(\Psi)$ is the normalizing constant. In Barron *et al.* (1999) it is proved that the distribution of f has full Kullback–Leibler support among densities f_0 for which $\int f_0 \log f_0 < \infty$. Thus, *a fortiori*, it has full Hellinger support.

4. Rates of convergence

As pointed out in Section 2, consistency is mainly based on the almost sure convergence of the sequence of predictive densities to the random density $\tilde{g}$, in the Hellinger distance. This also raises the issue of determining the rates at which the sequence $(f_n)_{n\geqslant 1}$ converges to $\tilde{g}$. This result is new, since previous contributions provide rates within the 'frequentist' approach; see, for example, Shen and Wasserman (2001) and Ghosal *et al.* (2000). It should be pointed out that it is not possible to compare the results contained in such papers with ours. Indeed, we consider rates of convergence to the random density $\tilde{g}$ while they study, in a different setting, convergence to some fixed f_0.

Here we determine rates of convergence of the cumulative average $N^{-1}\sum_{n=1}^{N} h(f_n,\, \tilde{g})$, where $(f_n)_{n\geqslant 1}$ is the sequence of predictive densities. Taking into account that h is the square of H, the following result essentially means that the rate for the cumulative averages of Hellinger distances is $N^{-1/4}$.

Theorem 2. *Let* $(X_n)_{n\geqslant 1}$ *be a sequence of random variables which are conditionally i.i.d. given a random density function* $\tilde{g}$. *Denote by* $(f_n)_{n\geqslant 1}$ *the sequence of predictive density functions. Then, for any* $\eta > 0$, *there exists a positive constant* k_η *and an integer* N_0 *such that*

$$P\left\{\frac{1}{N}\sum_{n=1}^{N} h(f_n,\ \tilde{g}) < \frac{k_\eta}{N^{1/2-r}}\right\} \geqslant 1-\eta,$$

for any $N > N_0$ *and for any* $r > 0$.

Proof. According to Theorem 1, we denote by Ω_0 a set such that $P(\Omega_0) = 1$ and $h(f_n,\ \tilde{g}) \to 0$ as $n \to +\infty$ for any $\omega \in \Omega_0$. Define a function that associates with each $\omega \in \Omega$ a decreasing sequence $(B_n)_{n\geqslant 1}$ of measurable sets of densities such that:

(a) $\tilde{g}(\omega) \in \cap_{n\geqslant 1} B_n(\omega)$ for any ω.
(b) $\mathrm{diam}_h(B_n(\omega)) < \rho_n\eta/2$ for any $n \geqslant 1$ and for each ω, where $(\rho_n)_{n\geqslant 1}$ is a sequence of positive numbers decreasing to 0 and $\mathrm{diam}_h(B)$ indicates the diameter of the set of densities B with respect to distance h.
(c) $\omega \mapsto B_n(\omega)$ is σ_n-measurable, for any $n \geqslant 1$.

For notational simplicity, from now on we suppress the dependence of B_n on ω. By virtue of the Markov inequality and of the convexity of h one can easily show that, for any $N \geqslant 1$,

$$\begin{aligned} P\left\{\frac{1}{N}\sum_{n=1}^{N} h(f_{n\,B_n},\ \tilde{g}) \geqslant \rho_N\right\} &\leqslant \frac{1}{N\rho_N}\sum_{n=1}^{N} \mathrm{E}[h(f_{n\,B_n},\ \tilde{g})] \\ &\leqslant \frac{1}{N\rho_N}\sum_{n=1}^{N}\int_\Omega\int_{B_n(\omega)} h(f,\ \tilde{g}(\omega))\,\frac{\Pi_n(\mathrm{d}f)}{\Pi_n(B_n)}\,P(\mathrm{d}\omega) \\ &\leqslant \eta/2. \end{aligned} \tag{9}$$

Consider the martingale $(T_N,\ \sigma_N)_{N\geqslant 1}$ defined by

$$T_N = \sum_{n=1}^{N-1}\left\{\Pi_{n+1}^{1/2}(B_n) - \Pi_n^{1/2}(B_n)(1 - h(f_{n\,B_n},\ f_n))\right\},$$

and use the monotonicity of the sequence $(B_n)_{n\geqslant 1}$ to show that

$$\begin{aligned} T_N &\geqslant \sum_{n=1}^{N-1}\left\{\Pi_{n+1}^{1/2}(B_{n+1}) - \Pi_n^{1/2}(B_n)(1 - h(f_{n\,B_n},\ f_n))\right\} \\ &= \Pi_N^{1/2}(B_N) - \Pi_1^{1/2}(B_1) + \sum_{n=1}^{N-1}\Pi_n^{1/2}(B_n)h(f_{n\,B_n},\ f_n) =: Z_N. \end{aligned}$$

Moreover, if $(b_n)_{n\geqslant 1}$ is a sequence defined in such a way that $b_n = n^{1/2+r}$, for each $n \geqslant 1$, the stability theorem yields

$$\frac{T_N}{b_N} \to 0 \text{ a.s.}[P]$$

which, in turn, implies

$$\frac{N}{b_N}\left\{\frac{1}{N}(\Pi_N^{1/2}(B_N) - \Pi_1^{1/2}(B_1)) + \frac{1}{N}\sum_{n=1}^{N-1}\Pi_n^{1/2}(B_n)h(f_{n\,B_n}, f_n)\right\} \to 0$$

almost surely. Consequently, for any $\eta > 0$, one can determine a $k_\eta > 0$ and a positive integer $n_0 = n_0(\eta)$ such that, for any $N \geqslant n_0$,

$$P\left\{\frac{Z_N}{N} < \frac{k_\eta}{N^{1/2-r}}\right\} \geqslant 1 - \frac{\eta}{2}.$$

Now, condition (a) for the sequence $(B_n)_{n\geqslant 1}$ and $\Pi_\infty = \delta_{\tilde{g}}$ give

$$P\left\{\liminf_n \Pi_n(B_n) = 1\right\} = 1.$$

Hence, $P\{\Pi_N(B_N)^{1/2} - \Pi_1(B_1)^{1/2} \geqslant 0\} = 1$ for all N greater than some $\bar{n}_0$ and

$$P\left\{\frac{1}{N}\sum_{n=1}^{N-1} h(f_{n\,B_n}, f_n) < \frac{k_\eta}{N^{1/2-r}}\right\} > 1 - \frac{\eta}{2}, \qquad \forall N \geqslant n_0 \vee \bar{n}_0.$$

Since we are interested in convergence rates for $h(f_n, \tilde{g})$, choose $\rho_N = k_\eta\, b_N/N$ in (9) and use the triangular inequality to obtain

$$\begin{aligned}
&P\left\{\frac{1}{N}\sum_{n=1}^{N-1} h(f_n, \tilde{g}) < \frac{k_\eta}{N^{1/2-r}}\right\} \\
&\geqslant P\left(\left\{\frac{1}{N}\sum_{n=1}^{N-1} h(f_{n\,B_n}, f_n) < \frac{k_\eta}{N^{1/2-r}}\right\} \cap \left\{\frac{1}{N}\sum_{n=1}^{N-1} h(f_{n\,B_n}, \tilde{g}) < \frac{k_\eta}{N^{1/2-r}}\right\}\right) \\
&\geqslant 1 - P\left(\frac{1}{N}\sum_{n=1}^{N-1} h(f_{n\,B_n}, f_n) > \frac{k_\eta}{N^{1/2-r}}\right) - P\left(\frac{1}{N}\sum_{n=1}^{N-1} h(f_{n\,B_n}, \tilde{g}) > \frac{k_\eta}{N^{1/2-r}}\right) \\
&\geqslant 1 - \eta,
\end{aligned}$$

thus completing the proof. □

5. Discussion

If the prior puts positive mass on all Hellinger neighbourhoods of all densities then Bayesian consistency holds almost surely with respect to the prior, in the sense that the posterior accumulates in Hellinger neighbourhoods of an essentially unique random density function, conditional on which the data are i.i.d. In this case the null sets on which consistency can fail are single densities. To our knowledge no work has been done on this version of Bayesian consistency since Doob (1949) and Schwartz (1965).

Acknowledgements

A. Lijoi was partially supported by the Italian Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR), grants 2002015111 and 2001138888. I. Prünster gratefully acknowledges support from MIUR, grant 2002015111, and the Progetto Giovani Ricercatori. The research of S.G. Walker was funded by an EPSRC Advanced Research Fellowship.

The authors are grateful to Patrizia Berti and Pietro Rigo for providing some insights which improved a part of the manuscript. Special thanks are also due to a referee for comments on an earlier version of this paper.

References

Barron, A., Schervish, M.J. and Wasserman, L. (1999) The consistency of posterior 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.

Doob, J.L. (1949) Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique 13, pp. 23–27. Paris: CNRS.

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

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

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

Grey, D.R. (2001) A note on convergence of probability measures. *J. Appl. Probab.*, **38**, 1055–1058.

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.

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

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

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

Leonard, T. (1978) Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B*, **40**, 113–146.

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

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

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

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

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

Received January 2003 and revised September 2003