Consistent semiparametric Bayesian inference about a location parameter

Subhashis Ghosal
Faculty of Mathematics and Computer Science, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Jayanta K. Ghosh[1]
Department of Statistics, Purdue University, 1399 Mathematical Sciences Building, West Lafayette, IN 47907, USA
and
Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, 203 B. T. Road, Calcutta 700035, India

R. V. Ramamoorthi[2]
Department of Statistics and Probability, Michigan State University, Wells Hall, East Lansing, MI 48824, USA

Abstract

We consider the problem of Bayesian inference about the centre of symmetry of a symmetric density on the real line based on independent identically distributed observations. A result of Diaconis and Freedman shows that the posterior distribution of the location parameter may be inconsistent if (symmetrized) Dirichlet process prior is used for the unknown distribution function. We choose a symmetrized Polya tree prior for the unknown density and independently choose θ according to a continuous and positive prior density on the real line. Suppose that the parameters of Polya tree depend only on the level m of the tree and the common values r_m's are such that $\sum_{m=1}^{\infty} r_m^{-1/2} < \infty$. Then it is shown that for a large class of true symmetric densities, including the trimodal distribution of Diaconis and Freedman, the marginal posterior of θ is consistent.

AMS subject classification: Primary 62G20, 62F15.

Key words: Consistency, Kullback-Leibler number, location parameter, Polya tree, Dirichlet process, posterior distribution.

Running head: Bayesian inference about location.

[1]Research supported by NSF grant number 9307727
[2]Research supported by NIH grant number 1 R01 GM49374.

1. Introduction

The starting point of this paper is a result of Diaconis and Freedman (1986a, b). They consider the location problem $X_i = \theta + \epsilon_i$, where the location parameter θ has a prior distribution μ and the ϵ_i's are independent and identically distributed with a symmetric distribution F, and where F itself has a symmetrized Dirichlet prior with base measure α. They then show that, while certain choices of α, for instance when α has a density α' with $\log \alpha'$ convex, ensures the consistency of the posterior at all (θ, F), there are choices of α for which the posterior fails to be consistent at many reasonable "true" values of the parameters. More precisely, when α is Cauchy, they exhibit a pair (θ_0, P_0), where P_0 has a (infinitely differentiable) density and for which, (θ_0, P_0) almost surely, the posterior distribution of θ given $X_1, X_2, \ldots, X_n$ does not converge to θ_0. Similar phenomena was also observed by Doss, who in a series of papers (1984. 1985a,b) carries out a penetrating analysis of the behaviour of the posterior when θ is considered as the median of F, and F, independent of θ has a Dirichlet like prior concentrating on distributions with median 0. Diaconis and Freedman while contending that discreteness of probabilities in the support of the Dirichlet may not be the main issue, construct a class of priors supported by continuous distribution and say " ... Now consider the location problem; we guess this prior is consistent when the expectation is the normal and inconsistent with the Cauchy. The real mathematical issue, it seems to us, is to find computable Bayes procedures and figure out when they are consistent and when they are inconsistent."

In this paper, we study consistency issues in the location problem when the prior on the symmetric distributions is induced by a Polya tree prior. Though the Polya tree prior is different from that constructed by Diaconis and Freedman, we believe that our calculations throws some light on the issues raised by them. Specifically, we consider Polya tree priors that concentrate on symmetric densities. In Theorem 5.1 which is stated informally below, we show that consistency obtains for a large class of true distributions that are supported on the entire real line.

Suppose the relative entropy of the true error distribution with respect to the base measure of the Polya tree is finite and the parameters of the Polya tree $\alpha_{\varepsilon_1 \cdots \varepsilon_m}$ grow like r_m with $\sum_{m=0}^{\infty} r_m^{-1/2} < \infty$. Further, assume that the operation of shifting locations of the true density is continuous in the Kullback-Leibler distance. Then the posterior is consistent.

In Theorem 5.2, we generalize the above result to remove the last hypothesis so that the result is applicable to many more true densities including those considered by Diaconis and Freedman (1986a, b). The main tools in our argument is a theorem of Schwartz and refinement of a theorem of Lavine (1994).

One lesson that emerges from the work of Diaconis and Freedman, and Doss is that the tail free property, which is a natural tool for establishing consistency, is destroyed by the addition of a parameter. The methods of our paper indicates that in semiparametric problems, the Schwartz criterion would be an appropriate tool in

proving consistency.

The results of our paper are stated in the context of location problems though many of the results would carry through to a wider class of semiparametric problems. We do not pursue this aspect.

2. Consistency of the posterior

Our parameter space is $\Theta \times \mathcal{F}^s$ where Θ is the real line and $\mathcal{F}^s$ is the set of all symmetric densities on $\mathbb{R}$. On $\Theta \times \mathcal{F}^s$, we consider a prior $\mu \times \mathbf{P}$ and given (θ, f), $X_1, X_2, \ldots, X_n$ are independent identically distributed with law $P_{\theta,f}$, where $P_{\theta,f}$ is the probability measure corresponding to the density $f(x-\theta)$. We denote by f_θ the density $f(x-\theta)$. Given $X_1, X_2, \ldots, X_n$, we consider the posterior distribution $(\mu \times \mathbf{P})(\cdot|X_1, X_2, \ldots, X_n)$ on $\Theta \times \mathcal{F}^s$ given by the density

$$\frac{\prod f_\theta(X_i)}{\int \prod f_\theta(X_i) d(\mu \times \mathbf{P})(\theta, f)}.$$

On $\{f_\theta : (\theta, f) \in \mathcal{F}^s\}$, we assign the topology of weak convergence. It is easy to see that this is equivalent to assigning, on $(\theta, f) \in \mathcal{F}^s$, the product of Euclidean and weak topologies on $\mathbb{R}$ and $\mathcal{F}^s$ respectively. The posterior $(\mu \times \mathbf{P})(\cdot|X_1, X_2, \ldots, X_n)$ is said to be consistent at (θ_0, f_0) if, as $n \to \infty$, $(\mu \times \mathbf{P})(\cdot|X_1, X_2, \ldots, X_n)$ converges weakly to the degenerate measure δ_{θ_0,f_0} almost surely P_{θ_0,f_0}. Clearly, if the posterior is consistent at (θ_0, f_0), the marginal distribution of $(\mu \times \mathbf{P})(\cdot|X_1, X_2, \ldots, X_n)$ on Θ converges to δ_{θ_0} almost surely P_{θ_0,f_0}.

Consistency is also related to robustness with respect to the contamination class of priors of Berger (1994). It is a weaker property in the following sense. Suppose a prior $\mathbf{P}_0$ on the set of probabilities is inconsistent at P_0. Consider a contamination class $\mathcal{P}$ of priors of the form $\{\mathbf{P} : \mathbf{P} = (1-\varepsilon)\mathbf{P}_0 + \varepsilon\delta_P\}$ containing $\mathbf{P}_1 = (1-\varepsilon)\mathbf{P}_0 + \varepsilon\delta_{P_0}$, with respect to which we wish robustness and let ρ be a metric for the weak topology on priors. Letting $\mathbf{P}_0^n$ and $\mathbf{P}_1^n$ stand for the posterior distribution given $X_1, X_2, \ldots, X_n$ under $\mathbf{P}_0$ and $\mathbf{P}_1$ respectively, we have $\rho(\mathbf{P}_1^n, \delta_{P_0}) \to 0$ almost surely by Schwartz's theorem mentioned below whereas $\rho(\mathbf{P}_0^n, \delta_{P_0})$ does not go to 0, by assumption. Clearly $\rho(\mathbf{P}_1^n, \mathbf{P}_0^n)$ cannot tend to 0 as $n \to \infty$.

Our main tool in establishing consistency is a theorem of Schwartz (1965). The relevance of the Schwartz theorem in the present context has been pointed out by Barron (1986). A detailed exposition can be found in Ghosh and Ramamoorthi (1997).

Recall that if f_0 and f_1 are two densities then the Kullback-Leibler divergence measure $K(f_0, f_1)$ is defined by $K(f_0, f_1) = \int_{-\infty}^{\infty} f_0(x) \log(f_0(x)/f_1(x)) dx$. We now state Schwartz's theorem in the form that we need.

Theorem 2.1. *If for all $\delta > 0$,*

$$(\mu \times \mathbf{P})\{(\theta, f) : K(f_{\theta_0}, f_\theta) < \delta\} > 0, \tag{2.1}$$

then the posterior $(\mu \times \mathbf{P})(\cdot|X_1, X_2, \ldots, X_n)$ *is consistent at* (θ_0, f_0).

Remark 2.1. The Kullback-Leibler neighbourhoods arise naturally in the study of general consistency results for the posterior since the posterior is well defined in these neighbourhoods. For instance, in the present context if $\{K(f_{\theta_0}, f_\theta) < \delta\}$ is a Kullback-Leibler neighbourhood of f_{θ_0} then the posterior is $P_{f_{\theta_0}}$-unique in $\{K(f_{\theta_0}, f_\theta) < \delta\}$. On the other hand, when there is no location parameter present, consistency of the posterior can be proved, at least for the standard (but not unique) posteriors for the Dirichlet and Polya tree priors without appealing to the Schwartz theorem.

3. Polya tree priors

Some basic statistical implications of the Polya tree prior can be found in Ferguson (1974), Lavine (1992, 1994) and Mauldin, Sudderth and Williams (1992). In this section we closely follow Lavine (1992, 1994). Let $E = \{0, 1\}$ and E^m be the m-fold Cartesian product $E \times \cdots \times E$ where $E^0 = \emptyset$. Further, set $E^* = \cup_{m=0}^{\infty} E^m$. Let $\pi_0 = \{\mathbb{R}\}$ and for each $m = 1, 2, \ldots$, let $\pi_m = \{B_\varepsilon : \varepsilon \in E^m\}$ be a partition of $\mathbb{R}$ so that sets of π_{m+1} are obtained from a binary split of the sets of π_m and $\cup_{m=0}^{\infty} \pi_m$ is a generator for the Borel sigma-field on $\mathbb{R}$. Let $\Pi = \{\pi_m : m = 0, 1, \ldots\}$.

Definition 3.1. A random probability measure P on $\mathbb{R}$ is said to possess a Polya tree distribution with parameters $(\Pi, \mathcal{A})$, we write $\mathsf{P} \sim \mathrm{PT}(\Pi, \mathcal{A})$, if there exist a collection of nonnegative numbers $\mathcal{A} = \{\alpha_\varepsilon : \varepsilon \in E^*\}$ and a collection $\mathcal{Y} = \{Y_\varepsilon : \varepsilon \in E^*\}$ of random variables such that the following hold:

(i) The collection $\mathcal{Y}$ consists of mutually independent random variables;

(ii) For each $\varepsilon \in E^*$, Y_ε has a beta distribution with parameters $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$;

(iii) The random probability measure P is related to $\mathcal{Y}$ through the relations

$$\mathsf{P}(B_{\varepsilon_1 \cdots \varepsilon_m}) = \left(\prod_{j=1;\varepsilon_j=0}^{m} Y_{\varepsilon_1 \cdots \varepsilon_{j-1}} \right) \left(\prod_{j=1;\varepsilon_j=1}^{m} (1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1}}) \right), \quad m = 1, 2, \ldots,$$

where the factors are $Y_\emptyset$ or $1 - Y_\emptyset$ if $j = 1$.

We restrict ourselves to partitions $\Pi = \{\pi_m : m = 0, 1, \ldots\}$ that are determined by a strictly positive continuous density α on $\mathbb{R}$ in the following sense: The sets in π_m are intervals of the form $\{x : (k-1)/2^m < \int_{-\infty}^{x} \alpha(t)dt \leq k/2^m\}$, $k = 1, 2, \ldots, 2^m$. We term the measure (corresponding to) α as the base measure because of its role similar to the base measure of Dirichlet process. The above conditions are assumed throughout without explicit mention.

Our next theorem refines Theorem 2 of Lavine (1994) by providing an explicit expression for the parameters.

Theorem 3.1. *Let f_0 be a density and* $\mathbf{P}$ *denote the prior* $\mathrm{PT}(\Pi, \mathcal{A})$*, where* $\alpha_\varepsilon = r_m$ *for all* $\varepsilon \in E^m$ *and* $\sum_{m=1}^{\infty} r_m^{-1/2} < \infty$*. Further assume that* $K(f_0, \alpha) < \infty$*. If* $\mathsf{P} \sim \mathrm{PT}(\Pi, \mathcal{A})$*, then almost surely,* P *has a density* f *and*

$$\mathbf{P}\{\mathsf{P} : K(f_0, f) < \delta\} > 0, \quad \delta > 0. \tag{3.1}$$

Remark 3.1. For any $\delta > 0$, the sequence $r_m = m^{2+\delta}$ suffices for an application of the Theorem 3.1. This sequence grows a little faster than Lavine's choice $r_m = m^2$. Whether consistency obtains under Lavine's choice is still left open. The choice of the parameter sequence and the base measure is likely to play a role in determining the rate of convergence and robustness properties.

Remark 3.2. In a recent article, Ghosal, Ghosh and Ramamoorthi (1998) show that priors arising out of Dirichlet mixtures of normals also satisfy (3.1).

Proof of Theorem 3.1. By the results of Kraft (1964), it follows that the weaker condition $\sum_{m=0}^{\infty} r_m^{-1} < \infty$ implies the existence of a density of the random probability measure P. Considering the transformation $x \mapsto \int_{-\infty}^{x} \alpha(t)dt$, we can without loss of generality assume that f and f_0 are densities on $[0, 1]$. Moreover, Π is then the canonical binary partition. By the martingale convergence theorem, there exist a collection of numbers $\{y_\varepsilon : \varepsilon \in E^*\}$ from $[0, 1]$ such that, with probability one

$$f_0(x) = \lim_{m\to\infty} \left(\prod_{j=1;\varepsilon_j=0}^{m} 2y_{\varepsilon_1\cdots\varepsilon_{j-1}} \right) \left(\prod_{j=1;\varepsilon_j=1}^{m} 2(1 - y_{\varepsilon_1\cdots\varepsilon_{j-1}}) \right), \tag{3.2}$$

where the limit is taken through a sequence $\varepsilon_1\varepsilon_2\cdots$ which corresponds to the dyadic expansion of x. Since the density f of P exists, it similarly follows that

$$f(x) = \lim_{m\to\infty} \left(\prod_{j=1;\varepsilon_j=0}^{m} 2Y_{\varepsilon_1\cdots\varepsilon_{j-1}} \right) \left(\prod_{j=1;\varepsilon_j=1}^{m} 2(1 - Y_{\varepsilon_1\cdots\varepsilon_{j-1}}) \right) \tag{3.3}$$

for almost every realization of f. Now for any $N \geq 1$,

$$K(f_0, f) = M_N + R_{1N} - R_{2N}, \tag{3.4}$$

where

$$M_N = \mathsf{E}\left[\log \left(\prod_{j=1;\varepsilon_j=0}^{N} \left(\frac{y_{\varepsilon_1\cdots\varepsilon_{j-1}}}{Y_{\varepsilon_1\cdots\varepsilon_{j-1}}} \right) \prod_{j=1;\varepsilon_j=1}^{N} \left(\frac{1 - y_{\varepsilon_1\cdots\varepsilon_{j-1}}}{1 - Y_{\varepsilon_1\cdots\varepsilon_{j-1}}} \right) \right) \right], \tag{3.5}$$

$$R_{1N} = \mathsf{E}\left[\log \left(\prod_{j=N+1;\varepsilon_j=0}^{\infty} 2y_{\varepsilon_1\cdots\varepsilon_{j-1}} \prod_{j=N+1;\varepsilon_j=1}^{\infty} 2(1 - y_{\varepsilon_1\cdots\varepsilon_{j-1}}) \right) \right] \tag{3.6}$$

and

$$R_{2N} = \mathsf{E}\left[\log\left(\prod_{j=N+1;\varepsilon_j=0}^{\infty} 2Y_{\varepsilon_1\cdots\varepsilon_{j-1}} \prod_{j=N+1;\varepsilon_j=1}^{\infty} 2(1-Y_{\varepsilon_1\cdots\varepsilon_{j-1}})\right)\right]; \tag{3.7}$$

here E stands for the expectation with respect to the distribution of $(\varepsilon_1, \varepsilon_2, \dots)$ which comes from the binary expansion of x and x is distributed according to the density f_0, for a fixed realization of the Y-values.

By the definition of a Polya tree, M_N and R_{2N} are independent random variables for all $N \geq 1$. To prove (3.1), it suffices to show that for any $\delta > 0$, there is some $N \geq 1$ such that

$$\mathbf{P}\{M_N < \delta\} > 0, \tag{3.8}$$

$$|R_{1N}| < \delta \tag{3.9}$$

and

$$\mathbf{P}\{|R_{2N}| < \delta\} > 0. \tag{3.10}$$

The set $\{(Y_\varepsilon : \varepsilon \in E^m, m = 0, \dots, N-1) : M_N < \delta\}$ is a nonempty open set in $\mathbb{R}^{2^N-1}$; it is open by the continuity of the relevant map while it is nonempty as $(y_\varepsilon : \varepsilon \in E^m, m = 0, \dots, N-1)$ belongs to this set. Thus (3.8) follows by the nonsingularity of the beta distribution. Relation (3.9) follows from Lemma 2 of Barron (1985). To complete the proof, it remains to show (3.10) for some $N \geq 1$. We shall actually prove the stronger fact

$$\lim_{N\to\infty} \mathbf{P}\{|R_{2N}| \geq \delta\} = 0. \tag{3.11}$$

Let $\mathbf{E}$ stand for the expectation with respect to the prior distribution $\mathbf{P}$ and E, as before, the expectation with respect to the distribution of $(\varepsilon_1, \varepsilon_2, \dots)$. Now

$$\begin{aligned}
&\mathbf{P}\{|R_{2N}| \geq \delta\} \\
&\quad \leq \delta^{-1}\mathbf{E}|R_{2N}| \\
&\quad \leq \delta^{-1}\mathbf{E}\,\mathsf{E}\left[\sum_{j=N+1;\varepsilon_j=0}^{\infty} |\log(2Y_{\varepsilon_1\cdots\varepsilon_{j-1}})| + \sum_{j=N+1;\varepsilon_j=1}^{\infty} |\log(2(1-Y_{\varepsilon_1\cdots\varepsilon_{j-1}}))|\right] \\
&\quad = \delta^{-1}\mathsf{E}\left[\sum_{j=N+1;\varepsilon_j=0}^{\infty} \mathbf{E}|\log(2Y_{\varepsilon_1\cdots\varepsilon_{j-1}})| + \sum_{j=N+1;\varepsilon_j=1}^{\infty} \mathbf{E}|\log(2(1-Y_{\varepsilon_1\cdots\varepsilon_{j-1}}))|\right] \\
&\quad \leq \delta^{-1}\mathsf{E}\left[\sum_{j=N+1}^{\infty} \max\{\mathbf{E}|\log(2Y_{\varepsilon_1\cdots\varepsilon_{j-1}})|, \mathbf{E}|\log(2(1-Y_{\varepsilon_1\cdots\varepsilon_{j-1}}))|\}\right] \\
&\quad \leq \delta^{-1}\sum_{j=N+1}^{\infty} \max_{(\varepsilon_1\cdots\varepsilon_{j-1})\in E^{j-1}} \max\{\mathbf{E}|\log(2Y_{\varepsilon_1\cdots\varepsilon_{j-1}})|, \mathbf{E}|\log(2(1-Y_{\varepsilon_1\cdots\varepsilon_{j-1}}))|\} \\
&\quad = \delta^{-1}\sum_{j=N+1}^{\infty} \eta(r_m),
\end{aligned} \tag{3.12}$$

where $\eta(k) = E|\log(2U_k)|$ with $U_k \sim$Beta(k,k). By Lemma A.1 of appendix, $\eta(k) = O(k^{-1/2})$ as $k \to \infty$. Since $\sum_{m=1}^{\infty} r_m^{-1/2} < \infty$ by assumption, the right hand side (RHS) of (3.12) is the tail of a convergent series. This completes the proof of (3.11) and hence that of the theorem. □

Remark 3.3. A minor modification of the proof shows that the Kullback-Leibler neighbourhoods would continue to have positive measure when the prior is modified as follows: Divide $\mathbb{R}$ into $k+1$ intervals $I_1, I_2, \ldots, I_{k+1}$ and assume that $(P(I_1), P(I_2), \ldots, P(I_k))$ have a joint density which is positive everywhere on the k-dimensional set $\{(a_1, \ldots, a_k) : a_i > 0, j = 1, \ldots, k, \sum_{j=1}^{k} a_i < 1\}$. For each I_j, the conditional distribution given $P(I_j)$ has a Polya tree prior satisfying the assumptions of the Theorem. We point out that the priors are special cases of the priors constructed by Diaconis and Freedman and consequently the consistency results proved later are also valid for the restricted class of Diaconis-Freedman priors. Moreover, it follows from Theorem 1 of Lavine (1994) that such priors can approximate any prior belief upto any desired degree of accuracy in a strong sense.

Remark 3.4. It is not necessary that for each m, $\alpha_{\varepsilon_1 \cdots \varepsilon_m}$ be the same for all $(\varepsilon_1, \ldots, \varepsilon_m) \in E^m$. The proof goes through even when only $\alpha_{\varepsilon_1 \cdots \varepsilon_{m-1} 0} = \alpha_{\varepsilon_1 \cdots \varepsilon_{m-1} 1}$ for all $(\varepsilon_1, \ldots, \varepsilon_{m-1}) \in E^{m-1}$, $m \geq 1$, and $r_m := \min\{\alpha_{\varepsilon_1 \cdots \varepsilon_m} : (\varepsilon_1, \ldots, \varepsilon_m) \in E^m\}$ satisfies the condition $\sum_{m=1}^{\infty} r_m^{-1/2} < \infty$.

4. Symmetrization

A prior $\mathbf{P}$ on the set $\mathcal{F}$ of all densities can be used to construct a prior on the set $\mathcal{F}^s$—the space of all symmetric densities. We consider two natural ways of doing this.

Method 1. Let $\mathbf{P}$ be a prior on $\mathcal{F}$. The map $f \mapsto (f(x) + f(-x))/2$ from $\mathcal{F}$ to $\mathcal{F}^s$ induces a measure on $\mathcal{F}^s$.

Method 2. Let $\mathbf{P}$ be a prior on $\mathcal{F}(\mathbb{R}^+)$—the space of densities on $\mathbb{R}^+$. The map $f \mapsto f^*$ where, $f^*(x) = f^*(-x) = f(x)/2$, gives rise to a measure on $\mathcal{F}^s$.

Unlike the Dirichlet process, even if the partitions and α_ε are all symmetric, these two methods yield different probabilities on $\mathcal{F}^s$. However, our consistency results hold under both methods, as the next lemma indicates.

Lemma 4.1. *Let* $\mathbf{P}$ *be a prior on* $\mathcal{F}$ *or on* $\mathcal{F}(\mathbb{R}^+)$ *satisfying* (3.1). *Let* $\mathbf{P}^*$ *be the prior obtained on* $\mathcal{F}^s$ *by method* 1 *or method* 2. *If* $f_0 \in \mathcal{F}^s$, *then*

$$\mathbf{P}^*\{f \in \mathcal{F}^s : K(f_0, f) < \delta\} > 0, \quad \delta > 0 \tag{4.1}$$

Proof. For Method 1, the result follows from Jensen's inequality and the conclusion is immediate for method 2 since, setting $g_0(x) = 2f_0(x)$ and $g(x) = 2f(x)$ for x in $\mathbb{R}^+$, both g_0, g belong to $\mathcal{F}(\mathbb{R}^+)$ and $K(f_0, f) = K(g_0, g)$. □

5. Location parameter problem

As mentioned in Section 1, our parameter space is $\Theta \times \mathcal{F}^s$ and given (θ, f), let $X_1, X_2, \ldots, X_n$ be independent and identically distributed. f_θ.

Definition 5.1. The map $(\theta, f) \mapsto f_\theta$ is said to be *KL-continuous* at $(0, f_0)$ if

$$K(f_0, f_{0,\theta}) = \int_{-\infty}^{\infty} f_0(x) \log(f_0(x)/f_0(x-\theta))dx \to 0 \quad \text{as} \quad \theta \to 0.$$

We would then call $(0, f_0)$ a *KL-continuity point.*

Let $f^*_{0,\theta}$ be the density defined by $f^*_{0,\theta}(x) = (f_{0,\theta}(x) + f_{0,\theta}(-x))/2$, the symmetrization of $f_{0,\theta}$, where $f_{0,\theta}$, as before, stands for $f_0(\cdot - \theta)$.

Theorem 5.1. *Assume that for every sufficiently small $|\theta|$, (4.1) holds with f_0 replaced by $f^*_{0,\theta}$. If μ gives positive mass to all open sets in Θ and if $(0, f_0)$ is KL-continuity point, then the posterior $(\mu \times \mathbf{P}^*)(\cdot|X_1, X_2, \ldots, X_n)$ is consistent at (θ_0, f_0) for all θ_0.*

Proof. It suffices to prove when $\theta_0 = 0$. By Theorem 2.1, it is enough to verify that $\mu \times \mathbf{P}^*$ satisfies the Schwartz condition (2.1), namely $(\mu \times \mathbf{P}^*)\{(\theta, f) : K(f_0, f_\theta) < \delta\} > 0$ for all $\delta > 0$. Now for any θ,

$$\begin{aligned} K(f_0, f_\theta) &= \int_{-\infty}^{\infty} f_0 \log(f_0/f_\theta) \\ &= \int_{-\infty}^{\infty} f_0 \log(f_0/f_{-\theta}) \\ &= \int_{-\infty}^{\infty} f_{0,\theta} \log f_{0,\theta} - \int_{-\infty}^{\infty} f_{0,\theta} \log f. \end{aligned} \tag{5.1}$$

Since

$$\int_{-\infty}^{\infty} f_{0,\theta} \log f^*_{0,\theta} = \int_{-\infty}^{\infty} f^*_{0,\theta} \log f^*_{0,\theta} \tag{5.2}$$

and

$$\int_{-\infty}^{\infty} f_{0,\theta} \log f = \int_{-\infty}^{\infty} f^*_{0,\theta} \log f, \tag{5.3}$$

we have

$$\begin{aligned} K(f_0, f_\theta) &= \int_{-\infty}^{\infty} f_{0,\theta} \log(f_{0,\theta}/f^*_{0,\theta}) + \int_{-\infty}^{\infty} f^*_{0,\theta} \log(f^*_{0,\theta}/f) \\ &\le \frac{1}{2} \int_{-\infty}^{\infty} f_{0,\theta} \log\left(\frac{f_{0,\theta}}{f_{0,\theta}}\right) + \frac{1}{2} \int_{-\infty}^{\infty} f_{0,\theta} \log\left(\frac{f_{0,\theta}}{f_{0,-\theta}}\right) + K(f^*_{0,\theta}, f) \\ &= \frac{1}{2} K(f_0, f_{0,-2\theta}) + K(f^*_{0,\theta}, f). \end{aligned} \tag{5.4}$$

By the KL-continuity assumption there is an ε such that when $|\theta| < \varepsilon$, the first term is less than $\delta/2$. For any θ, since $f^*_{0,\theta}$ is symmetric $\{f : K(f^*_{0,\theta}, f) < \delta/2\}$ has positive $\mathbf{P}^*$ measure. Thus we have, for each $\theta \in [-\varepsilon, \varepsilon]$, $\{f : K(f^*_{0,\theta}, f) < \delta/2\}$ is contained in $\{f : K(f_0, f_\theta) < \delta\}$. This completes the proof. □

The previous theorem establishes the consistency for (θ_0, f_0) when $(0, f_0)$ is a KL-continuity point. This requirement fails when f_0 has support in a finite interval $[-a, a]$. However, the next theorem shows that consistency continues to hold even when f_0 has support in a finite interval, provided f_0 is continuous. We show this by approximating f_0 by a f_1 satisfying conditions of Theorem 5.1. The next lemma indicates the kind of approximation that is needed. The proof is deferred to the appendix.

Lemma 5.1. *Let f_0 and f_1 be densities so that $f_0 \le C f_1$. Then for any f,*

$$K(f_0, f) \le (C+1)\log C + C[K(f_1, f) + \sqrt{K(f_1, f)}].$$

Theorem 5.2. *Assume that for every sufficiently small $|\theta|$, (4.1) holds with f_0 replaced by $f^*_{0,\theta}$ and μ gives positive mass to all open sets in Θ . If f_0 is continuous and has support in a finite interval $[-a, a]$, and $\log \alpha(x)$ is integrable with respect to $N(\mu, \sigma^2)$ for all (μ, σ), then the posterior $\mathbf{P}(\cdot|X_1, X_2, \ldots, X_n)$ is consistent at (θ, f_0) for all θ.*

Proof. We consider two cases.

Case 1. $\inf_{[-a,a]} f_0(x) = \alpha > 0$.

Let

$$f_1(x) = \begin{cases} (1-\eta) f_0(x), & \text{for } -a < x < a, \\ (\eta/2)\phi_{-a,\sigma^2}, & \text{for } x \le -a, \\ (\eta/2)\phi_{a,\sigma^2}, & \text{for } x \ge a, \end{cases} \tag{5.5}$$

where ϕ_{-a,σ^2} and ϕ_{a,σ^2} are respectively the densities of $N(-a, \sigma^2)$ and $N(a, \sigma^2)$ and σ^2 is chosen to ensure that f_1 is continuous at a.

We first show that f_1 is KL-continuous, i.e.,

$$\lim_{\theta \to 0} \int_{-\infty}^{\infty} f_1 \log(f_1/f_{1,\theta}) = \int_{-\infty}^{\infty} \lim_{\theta \to 0} f_1 \log(f_1/f_{1,\theta}) = 0. \tag{5.6}$$

It is enough to establish that for some $\varepsilon > 0$, the family $\{\log(f_1/f_{1,\theta}) : |\theta| < \varepsilon\}$ is uniformly integrable with respect to f_1. This follows since for any M,

$$\sup_{|\theta|<\varepsilon} \sup_{|x|<M} |\log(f_1(x)/f_{1,\theta}(x))| < C_M \quad \text{(say)}$$

and when M is large, for $|x| > M$, $f_{1,\theta}(x) = (\eta/2)(\sigma\sqrt{2\pi})^{-1} \exp[-(x-a-\theta)^2/(2\sigma^2)]$ for all $|\theta| < \varepsilon$, implying

$$\sup_{|\theta|<\varepsilon} \int_{|x|>M} f_1(x) \log(f_1(x)/f_{1,\theta}(x)) dx \to 0 \quad \text{as} M \to \infty.$$

It now follows from Lemma 5.1 that, by setting $C = (1-\eta)^{-1}$ and choosing η close to 1 so that $(C+1)\log C < \delta/2$, we can choose a δ^* such that $K(f_1, f) < \delta^*$ implies $K(f_0, f) < \delta$; consequently $\{(\theta, f) : K(f_1, f_\theta) < \delta^*\} \subset \{(\theta, f) : K(f_0, f_\theta) < \delta\}$. Theorem 5.1 shows that the set on the left hand side has positive $\mu \times \mathbf{P}^*$ measure.

Case 2. $\inf_{[-a,a]} f_0(x) = 0$.

By the continuity of f_0, we can, given any $\eta > 0$, choose a C such that $\int_{-a}^{a}(f_0 \vee C) = 1+\eta$, where $a \vee b = \max(a, b)$. Set $f_1 = (1+\eta)^{-1}(f_0 \vee C)$. Then $f_0 \leq (1+\eta)f_1$ and using Lemma 5.1, we can choose η and δ^* small such that $\{f : K(f_1, f) < \delta^*\} \subset \{f : K(f_0, f) < \delta\}$. Since f_1 is covered by case 1, the theorem follows. □

Remark 5.1. The above consistency theorem notwithstanding, computation of the posterior for θ for the Diaconis-Freedman density shows that convergence for Cauchy base measure is very slow. Even for $n = 500$, one notices the tendency to converge to a wrong value as in the case of the Dirichlet prior with Cauchy base measure. Rapid convergence to the right value does occur in the normal case.

Remark 5.2. While we have discussed consistency issues, it would be interesting to explore how the robustness calculations in Section 4 of Lavine (1994) can be made in the context of a location parameter.

Remark 5.3. Lemma 5.1 and the Schwartz theorem can be used to yield an analogue of Theorem 5.1 for general semiparametric models. Let $(\theta, f) \mapsto \phi(\theta, f)$, where $\phi(\theta, f)$ is a density on $\mathbb{R}$. Suppose a prior $\mu \times \mathbf{P}$ on $(\Theta, \mathcal{F})$ satisfies

(i) μ gives positive mass to every neighbourhood of θ_0

(ii) For all sufficiently small $|\theta - \theta_0|$, and all $\varepsilon > 0$,

$$\mathbf{P}\{f : K(\phi(\theta, f_0), \phi(\theta, f)) < \varepsilon\} > 0.$$

Then if (θ_0, f_0) is a point such that

(a) $\dfrac{\phi(\theta_0, f_0)}{\phi(\theta, f_0)} \leq C(\theta)$, where $C(\theta) \to 1$ as $\theta \to \theta_0$,

(b) $\lim_{\theta \to \theta_0} K(\phi(\theta, f_0), \phi(\theta, f)) = K(\phi(\theta_0, f_0), \phi(\theta_0, f))$ for all f,

then the posterior is consistent at (θ_0, f_0).

For a proof, take $\phi(\theta_0, f_0)$ and $\phi(\theta, f_0)$ as f_0 and f_1 respectively in Lemma 5.1. Then for each θ close to θ_0, $\{f : K(\phi(\theta_0, f_0), \phi(\theta, f)) < \varepsilon\}$ will contain a set of the form $\{f : K(\phi(\theta, f_0), \phi(\theta, f)) < \varepsilon'\}$, and this set has positive measure by assumptions (i), (ii) and (b) above.

Appendix

Lemma A.2. *If* $U_k \sim$Beta(k, k), *then* $E|\log(2U_k)| = O(k^{-1/2})$ *as* $k \to \infty$.

Proof. The proof uses Laplace's method. Let $\eta_k = E|\log(2U_k)|$. In other words

$$\eta_k = \frac{1}{B(k,k)} \int_0^1 |\log(2u)| u^{k-1}(1-u)^{k-1} du, \tag{A.1}$$

implying that

$$\eta_k = \frac{1}{B(k,k)} \int_0^1 |\log(2(1-u))| u^{k-1}(1-u)^{k-1} du. \tag{A.2}$$

Adding (A.1) and (A.2) and observing that $\log(2u)$ and $\log(2(1-u))$ are always of the opposite sign, we obtain

$$2\eta_k = \frac{1}{B(k,k)} \int_0^1 |\log(u/(1-u))| u^{k-1}(1-u)^{k-1} du. \tag{A.3}$$

This implies by Jensen's inequality that

$$\begin{aligned} 4\eta_k^2 &\leq \frac{1}{B(k,k)} \int_0^1 (\log(u/(1-u)))^2 u^{k-1}(1-u)^{k-1} du \\ &= \frac{1}{B(k,k)} \int_0^1 \{1 + (\log(u/(1-u)))^2\} u^{k-1}(1-u)^{k-1} du - 1. \end{aligned} \tag{A.4}$$

Now

$$\{1 + (\log(u/(1-u)))^2\} u^{k-1}(1-u)^{k-1} = \exp(g_k(u)), \tag{A.5}$$

where

$$g_k(u) = (k-1)\log u + (k-1)\log(1-u) + h(u)$$

and

$$h(u) = \log\{1 + (\log(u/(1-u)))^2\}.$$

It is easily observed that $g_k(1/2) = -2(k-1)\log 2$, $g_k'(1/2) = 0$ and $g_k'(u)$ is decreasing in u so that $g_k(u)$ has a unique maximum at $1/2$. Fix $\delta > 0$ and let $\lambda = \sup\{h''(u) : |u - 1/2| < \delta\}$. Thus on $u \in (\frac{1}{2} - \delta, \frac{1}{2} + \delta)$, we have

$$g_k(u) \leq -2(k-1)\log 2 - \frac{(u - \frac{1}{2})^2}{2}(8(k-1) - \lambda). \tag{A.6}$$

Thus

$$\begin{aligned} &4\eta_k^2 \\ &\leq \frac{1}{B(k,k)} \int_{1/2-\delta}^{1/2+\delta} \exp\left[-2(k-1)\log 2 - 4(k-1)\left(1 - \frac{\lambda}{8(k-1)}\right)\left(u - \frac{1}{2}\right)^2\right] du \\ &\quad + \frac{1}{B(k,k)} \int_{|u-\frac{1}{2}|>\delta} \{1 + (\log(u/(1-u)))^2\} u^{k-1}(1-u)^{k-1} du - 1 \\ &\leq \frac{\Gamma(2k)}{(\Gamma(k))^2} 2^{-2(k-1)} \int_{-\infty}^{\infty} \exp\left[-4(k-1)\left(1 - \frac{\lambda}{8(k-1)}\right)\left(u - \frac{1}{2}\right)^2\right] du \\ &\quad + \frac{1}{B(k,k)} \int_{|u-\frac{1}{2}|>\delta} \{1 + (\log(u/(1-u)))^2\} u^{k-1}(1-u)^{k-1} du - 1. \end{aligned} \tag{A.7}$$

Note that the function $u(1-u)\{1+(\log(u/(1-u))^2\}$ is bounded on $(0,1)$ by M (say). Hence the second term on the RHS of (A.7) is dominated by

$$\begin{aligned}
&\frac{M}{B(k,k)}\int_{|u-1/2|>\delta} u^{k-2}(1-u)^{k-2}du \\
&\quad = M\frac{(2k-1)(2k-2)}{(k-1)^2}P\left\{\left|U_{k-1}-\frac{1}{2}\right|>\delta\right\} \\
&\quad \le \frac{M}{\delta^2}\frac{(2k-1)(2k-2)}{(k-1)^2}E|U_{k-1}-\frac{1}{2}|^2 \\
&\quad = O(k^{-1}). \qquad \text{(A.8)}
\end{aligned}$$

The first term on the RHS of (A.7) is

$$\frac{\Gamma(2k)}{(\Gamma(k))^2}2^{-2k+2}(2\pi)^{1/2}(8(k-1)-\lambda)^{-1/2}, \qquad \text{(A.9)}$$

which, by an application of Stirling's inequalities [Whittaker and Watson (1928), p. 253], can be dominated by

$$\begin{aligned}
&\frac{(2k)^{2k-1/2}e^{-2k}(2\pi)^{1/2}\exp[(24k)^{-1}]}{(k^{k-1/2}e^{-k}(2\pi)^{1/2})^2}2^{-2k+2}(2\pi)^{1/2} \\
&\qquad \times 2^{-3/2}(k-1)^{-1/2}\left(1-\frac{\lambda}{8(k-1)}\right)^{-1/2} \\
&\quad = \left(\frac{k}{k-1}\right)^{1/2}\exp[(24k)^{-1}]\left(1-\frac{\lambda}{8(k-1)}\right)^{-1/2} \\
&\quad = 1+O(k^{-1}). \qquad \text{(A.10)}
\end{aligned}$$

Thus $\eta_k^2 = O(k^{-1})$, completing the proof. □

Proof of Lemma 5.1. We begin with the following inequality which is found in Hannan (1960). If f_0 and f_1 are densities

$$\int f_0[\log(f_0/f_1)]^- = \int f_0[\log(f_1/f_0)]^+ \le \int f_0\left(\frac{f_1}{f_0}-1\right)^+ = \frac{\|f_0-f_1\|}{2}. \qquad \text{(A.11)}$$

Hence if $f_0 \le Cf_1$,

$$\int f_0\log(f_0/f) = \int_{f\le f_1} f_0\log(f_0/f) + \int_{f>f_1} f_0\log(f_0/f) = \text{(I)+(II) (say)}, \qquad \text{(A.12)}$$

where we have

$$\begin{aligned}(\mathrm{I}) &\leq C\int f_1 \log C + C\int_{f\leq f_1} f_1 \log(f_1/f) \\ &\leq C\log C + C[K(f_1,f) - \int_{f>f_1} f_1\log(f_1/f) \\ &= C\log C + C\left[K(f_1,f) + \frac{\|f_0 - f_1\|}{2}\right] \qquad (A.13)\end{aligned}$$

and

$$(\mathrm{II}) \leq \int f_0 \log(Cf_0/f_0) \leq \log C; \qquad (A.14)$$

consequently

$$K(f_0,f) \leq C\log C + \log C + C[K(f_1,f) + \frac{1}{2}\|f_1 - f\|]. \qquad (A.15)$$

Since [Hannan (1960)] $K(f_1,f) \geq \|f_1 - f\|^2/4$, the lemma follows. □

Acknowledgement

Research of the first author was carried out at the Indian Statistical Institute, Calcutta, and was supported by a post doctoral grant from the National Board of Higher Mathematics, Bombay, India.

References

Barron, A. R. (1985). The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem. *Ann. Probab.* **13**, 1292–1303.

Barron, A. R. (1986). Discussion of "On the consistency of Bayes estimates" by P. Diaconis and D. Freedman. *Ann. Statist.* **14**, 26–30.

Berger, J. O. (1994). An overview of robust Bayesian analysis. *Test* **3**, 5–124.

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

Diaconis, P. and Freedman, D. (1986b). On inconsistent Bayes estimates of location. *Ann. Statist.* **14**, 68–87.

Doss, H. (1984). Bayesian estimation in the symmetric location problem *Z. Wahrsch. Verw. Gebiete* **68**, 127–147

Doss, H. (1985a). Bayesian nonparametric estimation of the median; Part I : Computation of the estimates *Ann. Statist.* **13**, 1432–1444.

Doss, H. (1985b). Bayesian nonparametric estimation of the median; Part II: Asymptotic properties of the estimates *Ann. Statist.* **13**, 1445–1464.

Ghosal, S., Ghosh, J. K. and ramamoorthi, R. V. (1998). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* (to appear).

Ghosh, J. K. and Ramamoorthi R. V. (1997). *Lecture notes on Bayesian asymptotics.* Under preparation.

Hannan, J. (1960). Consistency of maximum likelihood estimation of discrete distributions. In *Contributions to Probability and Statistics: Essays in Honour of Harold Hotelling.* (I. Olkin *et al.*, eds.) **1**, 249–257. Stanford Univ. Press, Stanford, CA.

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 Polya tree distributions for statistical modelling. *Ann. Statist.* **20**, 1222–1235.

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

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

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

Whittaker, E. T. and Watson, G. N. (1927). *A Course of Modern Analysis, Fourth Edition.* Oxford University Press.