# Asymptotic Properties of Nonparametric Bayesian Procedures

Larry Wasserman

ABSTRACT This chapter provides a brief review of some large sample frequentist properties of nonparametric Bayesian procedures. The review is not comprehensive, but rather, is meant to give a simple, heuristic introduction to some of the main concepts. We mainly focus on consistency but we touch on a few other issues as well.

## 1 Introduction

Nonparametric Bayesian procedures present a paradox in the Bayesian paradigm. On the one hand, they are most useful when we don't have precise information. On the other hand, they require huge amounts of prior information because nonparametric procedures involve high dimensional if not infinite dimensional parameter spaces. The usual hope that the data will dominate the prior was dashed by Freedman (1963, 1965) and then Diaconis and Freedman (1986) who showed that putting mass in weak neighborhoods the true distribution does not guarantee that the posterior accumulates in weak neighborhoods.

Interest in properties like consistency derive from our desire that the posterior not be strongly driven by the prior. Examining the frequentist properties of the posterior is perhaps the only reliable way to understand the import of the many implicit assumptions we make when doing nonparametric Bayesian inference.

This chapter provides a non-technical introduction to the frequentist large sample properties of Bayesian nonparametric procedures. Much of the material in this chapter is derived from or influenced by the aforementioned papers of Diaconis and Freedman, as well as Shen (1995), Tierney (1987), Ghoshal, Ghosh and Ramamoorthi (1997a,b), Cox (1993), Barron (1988), Schwartz (1965) and Barron, Schervish and Wasserman (1997). However, we shall not attempt a comprehensive review of this area. Rather, our hope is merely to give a flavor of some of the main ideas. More thorough reviews include Ghoshal, Ghosh and Ramamoorthi (1997a) and Ghosh and Ramamoorthi (1997). Other relevant references include Diaconis and Freedman (1993) and Freedman and Diaconis (1983).

One of the messages of this chapter is that in some cases, posteriors on infinite dimensional spaces exhibit reasonable large sample behavior. But in other cases, they do not and purely frequentist procedures might be better suited to these problems. Bayesian nonparametric methods – like all statistical methods – have their limitations.

## 2 Frequentist or Bayesian Asymptotics?

In Bayesian inference, there are two styles of asymptotics. Frequentist style asymptotics studies the behavior of the posterior with respect to draws from a fixed sampling distribution. Bayesian style asymptotics studies the behavior of the asymptotics with respect to the marginal distribution obtained by integrating the sampling distribution with respect to the prior.

Diaconis and Freedman (1986) made a compelling case for using frequentist asymptotics. Without delving into too much detail, the case for using frequentist asymptotics can be made using a simple example. Suppose $Y_1, \ldots, Y_n \sim N(\theta, 1)$ and let the prior $\pi$ be a point mass at zero. Clearly, the posterior is also a point mass at zero. Barring the miracle that the true value of $\theta$ is zero, this posterior is inconsistent: the posterior mass never accumulates around the true value.

Yet if we let $A$ denote the event that the posterior eventually accumulates around the true value of $\theta$, then Doob's theorem (Doob 1949) shows that $A$ has a probability one. This would seem to contradict the obvious fact that the posterior is inconsistent. The reason it does not is that Doob's theorem uses the Bayesian's own prior when it claims that $A$ has probability one. Indeed, the posterior is consistent if $\theta$ happens to be 0 and, according to the prior, this is a sure event. Consistency fails on a null set, but the null set is huge. In this example, we don't need fancy techniques to see that the null set is big. But in infinite dimensional spaces, it is not so easy know whether a null set is "big," hence the need to study frequentist consistency seriously.

## 3 Consistency

"Consistency" means that the posterior concentrates around the true distribution. Often, one or two conditions are needed for consistency. The first is a support condition which requires that the prior put positive mass around the true distribution. The second is a smoothness condition which requires that not too much mass be put on rough densities. Consistency requires a balancing act between trying to spread the mass around as much as possible but not spreading it on nasty distributions.

Let $\mathcal{P}$ be the set of all distributions. We need to introduce some distance measures on $\mathcal{P}$. Let $d_W$ be any distance that metrizes the weak topology i.e. $d_W(P_n, P) \to 0$ if and only if $\int g(x)dP_n(x) \to \int g(x)dP(x)$ for every bounded, continuous function $g$. There are many such metrics and we shall not need to concern ourselves with the details. Let $d_H$ be Hellinger distance:

$$d_H(P, Q)^2 = \int (\sqrt{f}(y) - \sqrt{g}(y))^2 d\lambda(y)$$

where $f$ and $g$ are the densities of $P$ and $Q$ with respect to the dominating measure $\lambda$ Also, let $d_K$ be Kullback-Leibler divergence:

$$d_K(P, Q) = \int dP \log dP/dQ.$$

We shall also write the distances with densities as their arguments when convenient. It is also helpful to define total variation distance $d_1(f, g) = \int |f(y) - g(y)|d\lambda(y)$ which is also called the $L_1$ distance. In what follows, there is little difference between using $d_H$ or $d_1$.

The "support" of a prior $\pi$ is

$$\mathcal{S}(\pi) = \{P;\ \text{for every } \epsilon > 0,\ \pi(N_\epsilon(P)) > 0\} \tag{1.1}$$

where $N_\epsilon(P) = \{Q; d(P, Q) \le \epsilon\}$. The support depends on the choice of distance so we write $\mathcal{S}_W(\pi)$, $\mathcal{S}_H(\pi)$ and $\mathcal{S}_K(\pi)$ for the three choices. It is easy to show that $\mathcal{S}_K(\pi) \subset \mathcal{S}_H(\pi) \subset \mathcal{S}_W(\pi)$.

Say that $\pi$ is consistent at $P_0$ if, for every $\epsilon > 0$, $\pi(N_\epsilon(P_0)|Y^n)$ tends to 1, almost surely with respect to repeated sampling from $P_0$. Let $\mathcal{C}(\pi)$ be the set of all $f_0$ such that $\pi$ is consistent at $f_0$. We call $\mathcal{C}(\pi)$, the "consistency class" of $\pi$. Again, $\mathcal{C}(\pi)$ depends on the choice of distance and we subscript $\mathcal{C}(\pi)$ appropriately as needed. We would like $\mathcal{C}(\pi)$ to be as large as possible.

Now the question is: if we put positive mass in each neighborhood of $P_0$, does it imply that the posterior of every neighborhood of $P_0$ tends to 1? To be precise, the question is whether $\mathcal{S}(\pi) \subset \mathcal{C}(\pi)$.

Freedman (1963) and Diaconis and Freedman (1986) showed that $\mathcal{S}_W(\pi)$ is not a subset of $\mathcal{C}_W(\pi)$. In other words, even if you put positive mass in weak neighborhoods of $P_0$, it does not follow that the posterior concentrates in weak neighborhoods of $P_0$.

On other hand, Schwartz (1965) and Barron (1986) proved that $\mathcal{S}_K(\pi) \subset \mathcal{C}_W(\pi)$. In words: if the prior puts positive mass in Kullback-Leibler neighborhoods of $P_0$, then the posterior concentrates in weak neighborhoods of $P_0$. To state Schwartz's result in more detail, we must first recall the following definition. A sequence of tests $\phi_n(Y_1, \ldots, Y_n)$ is "uniformly consistent" for $H_0 : f = f_0$ versus $H_1 : f \in A^c$ if $E_0(\phi_n) \to 0$ and $\inf_{f \in A^c} E_f \phi_n \to 1$. Here $E_0$ and $E_f$ refer to expectation under repeated sampling from $f_0$ or $f$. Schwartz showed that if $f_0 \in S_K(\pi)$ and if there exists a uniformly consistent test for $H_0 : f = f_0$ versus $H_1 : f \in A^c$, then $\pi(A|Y^n)$ tends to 1 almost surely. If $A$ is a weak neighborhood of $f_0$ then it is known that there does exist a uniformly consistent test and hence consistency holds. As pointed out in Ghoshal, Ghosh and Ramamoorthi (1997), the condition that $f_0 \in S_K(\pi)$ is not a necessary condition. They give the following simple example. Let $Y_1, \ldots, Y_n$ be i.i.d. from a Uniform $(0, \theta)$ with $\theta \in (0, 1]$. The condition fails since the Kullback-Leibler distance from Uniform (0,1) to a Uniform $(0, \theta)$ is infinite. But if a prior has full support on [0,1] then consistency obtains when the true distribution is Uniform (0,1).

Weak neighborhoods are large. It would also be nice if we could say that the posterior concentrates in stronger neighborhoods. Barron (1988), Barron, Schervish and Wasserman (1997), Ghoshal, Ghosh and Ramamoorthi (1997a, 1997b) and Shen (1997) discuss this issue. Generally, one can show that $\mathcal{S}_K(\pi) \subset \mathcal{C}_H(\pi)$ if we insist that the prior $\pi$ satisfy a regularity condition. The outline of why this is true, is in the next section.

## 4 Consistency in Hellinger Distance

Here we blend some ideas of Barron (1988), Barron, Schervish and Wasserman (1997), Ghoshal, Ghosh and Ramamoorthi (1997a, 1997b), Schwartz (1963), Shen (1995) and others to give a heuristic explanation of when the posterior concentrates in Hellinger neighborhoods of the true density. Specifically, we will show that $\mathcal{S}_K(\pi) \subset \mathcal{C}_H(\pi)$ under a smoothness condition. To do so, we need some definitions.

Let $\mathcal{G}$ be a set of densities with respect to some underlying measure $\lambda$. A set of pairs of functions $(\ell_1, u_1), \ldots, (\ell_k, u_k)$ is called a an "$\epsilon$-bracketing" of $\mathcal{G}$ if (i) $d_H(\ell_j, u_j) \leq \epsilon$ for $j = 1, \ldots, k$ and (ii) for every $f \in \mathcal{G}$ there is a $j \in \{1, \ldots, k\}$ such that $\ell_j \leq f \leq u_j$ a.e. with respect to a dominating measure. Let $N_{[\,]}(\epsilon, \mathcal{G})$ be the number of brackets in the smallest $\epsilon$ bracketing of $\mathcal{G}$ and let $H_{[\,]}(\epsilon, \mathcal{G}) = \log N_{[\,]}(\epsilon, \mathcal{G})$. The number $H_{[\,]}(\epsilon, \mathcal{G})$ is called the bracketing metric entropy of $\mathcal{G}$. See van der Vaart and Wellner (1996) for more detail about bracketing entropy.

For this section we assume that the prior $\pi$ lives on a set of distributions which all possess densities with respect to some common dominating measure $\lambda$. We shall be somewhat loose in some of our terminology. Thus we shall speak of a density $f$ when we should really talk about an equivalence class of densities. And we shall refer to neighborhoods of a density $f$ to mean the same as a neighborhood around the corresponding measure. The true probability measure generating the data is $P_0$ with density $f_0$. The $n$-fold product measure of $P_0$ is denoted $P_0^n$. The corresponding density is $f_0^n$. Statements like "such and such happens almost surely" refer to repeated sampling from $P_0$. We shall assume all relevant quantities are measurable.

THEOREM. Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. from $f_0$ and suppose that $f_0 \in \mathcal{S}_K(\pi)$. Further, suppose that for every $\epsilon > 0$, there exist a sequence of sets of densities $\mathcal{F}_n$ such that

(i) entropy condition:

$$\int_{\epsilon^2}^{\epsilon} H_{[\,]}^{1/2}(u, \mathcal{F}_n)du \leq \sqrt{n}\epsilon^2 \quad \text{and}$$

(ii) tail condition: there exists a positive constant $r > 0$ such that

$$\pi(\mathcal{F}_n^c) < e^{-nr}$$

for all large $n$. Then, for every $\epsilon > 0$, $\pi(N_\epsilon^c|Y^n)$ tends to 0 exponentially quickly, almost surely, where $N_\epsilon$ is a Hellinger neighborhood around $P_0$ of size $\epsilon$. Thus, $\mathcal{S}_K(\pi) \subset \mathcal{C}_H(\pi)$.

PROOF. Let $N$ be an $\epsilon$ Hellinger neighborhood of $f_0$. By Bayes' theorem we have

$$\begin{aligned} \pi(N^c|Y^n) &= \frac{\int_{N^c} \prod_i f(Y_i)d\pi(f)}{\int \prod_i f(Y_i)d\pi(f)} \\ &= \frac{\int_{N^c} R_n(f)d\pi(f)}{\int R_n(f)d\pi(f)} \end{aligned} \tag{1.2}$$

where

$$R_n(f) = \prod_i \frac{f(Y_i)}{f_0(Y_i)}.$$

Now,

$$\int_{N^c} R_n(f)d\pi(f) = \int_{N^c \cap \mathcal{F}_n} R_n(f)d\pi(f) + \int_{N^c \cap \mathcal{F}_n^c} R_n(f)d\pi(f) = A + B, \text{ say.}$$

Property (ii) implies that $B$ is very small. This can be seen using the following trick. Let $Pr$ represent probability statements under $P_0^n$ and let $E$ represent expectation under $P_0^n$. Fix an arbitrary $c > 0$ and, using Markov's inequality and Fubini's theorem:

$$\begin{aligned} Pr(B > c) &\leq c^{-1}E(B) \\ &= c^{-1} \int_{\mathcal{Y}^n} \int_{N^c \cap \mathcal{F}_n^c} R_n(f)d\pi(f)dP_0^n \\ &= c^{-1} \int_{N^c \cap \mathcal{F}_n^c} \int_{\mathcal{Y}^n} R_n(f)dP_0^n d\pi(f) \\ &= c^{-1} \int_{N^c \cap \mathcal{F}_n^c} \int_{\mathcal{Y}^n} R_n(f) f_0^n(Y^n)d\lambda^n d\pi(f) \\ &= c^{-1} \int_{N^c \cap \mathcal{F}_n^c} \int_{\mathcal{Y}^n} \frac{f^n(Y^n)}{f_0^n(Y^n)} f_0^n(Y^n)d\lambda^n d\pi(f) \\ &= c^{-1} \int_{N^c \cap \mathcal{F}_n^c} \int_{\mathcal{Y}^n} f^n(Y^n)d\lambda^n d\pi(f) \end{aligned}$$

$$= c^{-1} \int_{N^c \cap \mathcal{F}_n^c} d\pi(f)$$

$$= c^{-1}\pi(N^c \cap \mathcal{F}_n^c) \le c^{-1}\pi(\mathcal{F}_n^c) \le c^{-1}e^{-nr}$$

by condition (ii). Take $c = e^{-nr/2}$. Conclude that $Pr(B > e^{-nr/2}) \le e^{-nr/2}$. By the first Borel-Cantelli lemma, $B \le e^{-nr/2}$ almost surely, for all large $n$.

Now

$$\int_{N^c \cap \mathcal{F}_n} R_n(f) d\pi(f) \le \sup_{f \in N_c \cap \mathcal{F}_n} R_n(f)\pi(N^c \cap \mathcal{F}_n) \le \sup_{f \in N_c \cap \mathcal{F}_n} R_n(f).$$

Now we use a large deviation inequality which is implied by property (i). Specifically, from Wong and Shen (1995, Theorem 1), (i) implies that, almost surely, for all large $n$,

$$\sup_{f \in N_c \cap \mathcal{F}_n} R_n(f) < e^{-nc} \tag{1.3}$$

for some $c > 0$. We have upper bounded the numerator of (1.2) by $e^{-nc} + e^{-nr/2}$. Now we proceed to lower bound the denominator.

Let $\delta = (1/2)[(r/2) + c] > 0$ and let $K$ be a Kullback-Leibler neighborhood of size $\delta/2$. Let

$$D_n(f) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_0(Y_i)}{f(Y_i)}.$$

Then,

$$e^{n\delta} \int R_n(f) d\pi(f) = e^{n\delta} \int e^{-nD_n(f)} d\pi(f) \ge e^{n\delta} \int_K e^{-nD_n(f)} d\pi(f).$$

For each $f \in K$, by the law of large numbers, $D_n(f)$ converges almost surely to $D(f) \equiv d_K(f_0, f) < \delta$. Thus, $\lim_n n(\delta - D_n(f)) = \infty$ almost surely. This convergence is almost sure on the sample space, for each fixed $f$. The null set can depend on $f$. However, by Fubini's theorem, we can choose a single set of probability one on the sample space such that $D_n(f)$ converges to $D(f)$ expect possibly for a set of $f's$ of $\pi$ measure 0. So, by Fatou's lemma,

$$\text{liminf}_n e^{n\delta} \int R_n(f) d\pi(f) \ge \int_K \text{liminf}_n e^{n[\delta - D_n(f)]} d\pi(f) = \infty$$

as long as $\pi(K) > 0$, which holds by property (i). Thus, for large $n$, $e^{n\delta} \int e^{-nD_n(f)} \ge 1$, i.e. $\int R_n(f) d\pi(f) \ge e^{-n\delta}$ almost surely for large $n$.

Having bounded the numerator and denominator, we can now see that

$$\pi(N^c | Y^n) \le \frac{e^{-nc} + e^{-nr/2}}{e^{-n\delta}} = e^{-(n/2)[c + (r/2)]} \tag{1.4}$$

almost surely, for large $n$. ♣

The conditions of the Theorem are stronger than needed. Ghoshal, Ghosh and Ramamoorthi (1997b) have a consistency result that uses $L_1$ entropy which is weaker than the bracketing entropy condition. Their result is as follows. Suppose that $f_0 \in \mathcal{S}_K(\pi)$ and that for every $\epsilon > 0$ there is a $\delta < \epsilon$, $c_1, c_2 > 0$, $\beta < \epsilon^2/2$ and $\mathcal{F}_n$ such that (i) $\pi(\mathcal{F}_n^c) < c_1 e^{-nc_2}$ and (ii) $J(\delta, \mathcal{F}_n) < n\beta$ where $J(\beta, \mathcal{F}_n)$ is the logarithm of the cardinality of the smallest number $m$ of densities $f_1, \ldots, f_m$ such that $\mathcal{F}_n \subset \cup_{j=1}^m \{f; d_1(f, f_j) < \delta\}$. Then Hellinger (and total variation) consistency obtains.

A more general result is contained in Barron (1986) which is also summarized nicely in Ghoshal, Ghosh and Ramamoorthi (1997a). Suppose again that $f_0 \in \mathcal{S}_K(\pi)$. Then Barron shows that the following two statements are equivalent:

(1) For any measurable set $U$, there exists $\beta_0 > 0$ such that

$$P_0(\pi(U^c|Y^n) > e^{-n\beta_0}, i.o.) = 0;$$

(2) There exist subsets $V_n$ and $W_n$, positive numbers $c_1, c_2, \beta_1, \beta_2$ and tests $\{\phi_n\}$ such that (i) $U^c = V_n \cup W_n$, (ii) $\pi(W_n) \leq c_1 e^{-n\beta_1}$, (iii) $Pr(\phi_n > 0, i.o.) = 0$ and $\inf_{f \in V_n} E_f(\phi_n) \geq 1 - c_2 e^{-n\beta_2}$.

The next question is: do commonly used methods satisfy the conditions of the theorem? Fortunately, the answer is yes. For example, Barron, Schervish and Wasserman (1997) show that random histograms, Polya trees and infinite dimensional exponential families all satisfy the conditions of the theorem. For example, Polya trees with $Beta(a_k, a_k)$ distributions at the $k^{th}$ level are shown to be consistent if $a_k = 8^k$. Ghoshal, Ghosh and Ramamoorthi (1997) show that Dirichlet process mixtures are consistent. This result requires technical conditions on the tail of the true density as well as some conditions on the base measure of the Dirichlet.

## 5 Other Asymptotic Properties

Consistency is a very weak property. We would like the posterior to possess other properties as well. If we study how the posterior concentrates in neighborhoods $N_{\epsilon_n}$ where $\epsilon_n$ tends to zero, then we can assess the rate of convergence of the posterior. Shen (1995), for example, has results on rates of convergence. Some of the results are negative: in certain cases the rate of convergence of the posterior is slower than the best obtainable rates by other means.

Zhao (1993, 1997) studied rates of convergence of the Bayes estimator in the following model. Suppose that

$$Y_i = \beta_i + \epsilon_i \tag{1.5}$$

for $i = 1, 2, \ldots$ where the $\epsilon_i$'s are independent, identically distributed random variables with mean 0 and variance $\sigma_n^2 = 1/n$ and $\sum_i \beta_i^2 < \infty$. Take the prior to make the $\beta_i^\prime s$ Normal with mean 0 and variance $i^{-2p}$. Suppose that $\beta$ lies in $\Omega = \{\beta; \sum_i i^{2q}\beta_i^2 \leq B\}$ where $q$ and $B$ are known. The optimal minimax rate for squared error loss is known to be $n^{2q/(2q+1)}$. In other words

$$0 < \lim_{n\to\infty} \inf_{\hat{\beta}} \sup_{\theta\in\Omega} n^{\frac{2q}{2q+1}} R(\hat{\beta}, \beta) < \infty \qquad (1.6)$$

where $R(\hat{\beta}, \beta) = E(\sum_i (\hat{\beta}_i - \beta_i)^2)$.

Zhao found that the Bayes estimate achieves the optimal rate when $p = (2q+1)/2$. But in this case, the prior and posterior have measure 0 on $\Omega$! And she shows that any independent Gaussian prior with support on $\Omega$ cannot attain the optimal rate. On a positive note, she shows the following result. Suppose we take the prior to be $\pi = \sum_j w_j \pi_j$ where $w_j \sim j^{-2}$ and $\pi_j$ is Gaussian on the first $j$ coordinates and sets all other $\beta_j^\prime s$ to 0. Then this prior has positive mass on $\Omega$ and the Bayes estimator achieves the optimal rate.

It is also natural to inquire about asymptotic normality of the posterior. The results are few and mixed. Shen (1995) gives conditions which ensure that the posterior of functions of the distribution has, asymptotically, a normal distribution. He also investigates whether the posterior is efficient, i.e. the asymptotic variance is as small as possible. The conditions needed to guarantee to normality are complicated. Diaconis and Freedman (1997) studied the following problem; a related investigation is in Cox (1993). Here we summarize the results of Diaconis and Freedman.

Suppose we again have the model above used by Zhao where now the prior makes the $\beta_i^\prime s$ Normal with mean 0 and variance $\tau_i^2 = A/i^\alpha$, $1 < \alpha < \infty$. Let $T_n = \sum_{i=1}^\infty (\beta_i - \hat{\beta}_i)^2$ where $\hat{\beta}_i = E(\beta_i|Y) = \tau_i^2 Y_i/(\sigma_n^2 + \tau_i^2)$. They prove that, with respect to the posterior,

$$T_n \approx n^{-1+\frac{1}{\alpha}} C + n^{-1+\frac{1}{2\alpha}} \sqrt{D} Z_n \qquad (1.7)$$

where $C$ and $D$ are numbers and $Z_n$ converges in law to a $N(0, 1)$. On the other hand, with respect to the sampling distribution,

$$T_n \approx n^{-1+\frac{1}{\alpha}} C + n^{-1+\frac{1}{2\alpha}} \sqrt{F} U_n(\beta) + n^{-1+\frac{1}{2\alpha}} \sqrt{G} V_n(\beta). \qquad (1.8)$$

The last equation holds for almost all $\beta$ drawn from the prior. That is, it is a frequentist result holding for fixed $\beta$, and the $\beta$'s for which it holds are a set of measure 1 with respect to the prior. Moreover, $G < D$, $U_n(\beta)$ and $V_n(\beta)$ both converge in law to $N(0, 1)$ random variables and

$$\text{liminf}_n U_n(\beta) = -\infty \quad \text{and} \quad \text{limsup}_n U_n(\beta) = \infty.$$

Thus, the posterior behaves quite differently from the usual Normal limiting behavior we would expect in finite dimensional models. However, the

Bayesian and frequentist limiting behavior can be made equivalent if the prior variances of the $\beta_j' s$ are chosen to decay exponentially quickly.

Another desirable property of the posterior is that sets with posterior probability $1-\alpha$ should have frequentist coverage nearly equal to $1-\alpha$. There is a long history of such results for parametric models; see Kass and Wasserman (1996). The correspondence breaks down even in complicated finite dimensional models; For example, Wasserman (1998) showed that for certain finite dimensional mixture models, accurate coverage can only be obtained by data dependent priors. Less is known in the nonparametric case. Cox (1993) showed that in certain Gaussian process models, the correspondence between posterior probability and coverage breaks down. The results of Diaconis and Freedman above also show a breakdown of the coverage properties in infinite dimensions. Suppose we choose $c_n$ such that $I_n = \{\beta; T_n \leq c_n\}$ has posterior probability $1-\alpha$. Then, it is easy to see from their results that there almost certainly will be subsequences along which the coverage of the $I_n$ is arbitrarily small.

## 6 The Robins-Ritov Paradox

Here we summarize an intriguing example, due to Robins and Ritov (1997) which shows the limitation of nonparametric Bayesian methods. Consider independent, identically distributed data $W_1, \ldots, W_n$ where $W_i = (X_i, R_i, Y_i)$. The $Y_i' s$ are binary and our goal is simply to estimate $\psi = E(Y_1)$. The random variable $X_i$ lives in $\mathcal{R}^p$ where $p$ is large. $R_i$ is another binary random variable which indicates whether or not we observe $Y_i$. We are also given a known function $g : \mathcal{R}^p \to [0,1]$ such that $\inf_x g(x) > 0$.

The data are obtained as follows. First we observe $X_i$. Then, with probability $\pi_i = g(X_i)$ we get to observe $Y_i$ i.e. $R_i = 1$ and with probability $1-\pi_i$ we do not observe $Y_i$ i.e. $R_i = 0$. To make the problem easier, we will assume that the marginal distribution of $X_1$ is known. (This is not essential to the problem.) The likelihood is

$$\begin{aligned}
\mathcal{L} &= \prod_i f_X(X_i) f_{R|X}(R_i|X_i)[f_{Y|X}(Y_i|X_i)]^{R_i} && (1.9)\\
&= \prod_i f_X(X_i) \pi_i^{R_i}(1-\pi_i)^{1-R_i} h(X_i)^{R_i Y_i}(1-h(X_i))^{R_i(1-Y_i)} && (1.10)\\
&\propto \prod_i h(X_i)^{R_i Y_i}(1-h(X_i))^{R_i(1-Y_i)} && (1.11)
\end{aligned}$$

where $h(X) = Pr(Y=1|X)$. The only unknown "parameter" in this problem is the function $h(\cdot)$ which is a binary regression function over a $p$-dimensional covariate.

A Bayesian would put a prior on $h$, then compute the posterior of $h$. We will assume that the prior is chosen independently of the function $g$. The posterior of $h$ induces a posterior on $\psi = E(Y) = \int E(Y|X = x)f_X(x)dx = \int h(x)f_X(x)dx$. For the sake of argument, let us suppose $p = 20$, not at all unrealistic. Performing this 20 dimensional nonparametric regression problem is no small feat. Due to the curse of dimensionality, we cannot realistically expect to do well in such a high dimensional problem. Indeed, Robins and Ritov show that the Bayes estimate of $\psi$ cannot be $\sqrt{n}$ consistent and, moreover, cannot be uniformly consistent, unless one imposes severe smoothness restrictions.

This would not be so disturbing except that there is a simple frequentist estimator which behaves quite well. The estimator is the well-known Horvitz-Thompson estimator

$$\hat{\psi} = \frac{1}{n}\sum_i \frac{Y_i R_i}{\pi_i} \tag{1.12}$$

which is $\sqrt{n}$ consistent regardless of the joint law of $(X, R, Y)$ (assuming only that $g$ is bounded away from 0). In addition to its appealing asymptotic properties, this estimator is notable for its simplicity compared to the Bayesian nonparametric approach. The appearance of the $\pi_i's$ is disturbing from the Bayesian point of view. The $X_i's$ are ancillary and the usual Bayesian methods will not depend on the probability of observing the $X_i's$. Still, the alternative of trying to estimate $h$ nonparametrically is even less appealing. Imposing a parametric model on $h$ is also unappealing since any such model is bound to be wrong. If we do wish to impose such a model, Robins and Ritov show that the Horvitz-Thompson estimator can be modified in such a way that the resulting estimator retains $\sqrt{n}$ consistency if the model is wrong. But if the model is right, then their estimator attains the smallest possible asymptotic variance (among estimators that are $\sqrt{n}$ consistent when the model is wrong).

There are many other examples that compare Bayesian and frequentist methods, some favoring one or the other. The point is simply that we should not assume that Bayesian nonparametric methods are a panacea.

## 7 Conclusion

The frequentist properties of Bayesian methods are well understood for parametric models. This is less so for nonparametric models. Fortunately, this is now an active area and we can expect in the coming years to have a much deeper understanding of the issues. Of course, most of the work pertains to large sample properties. Little seems to be known about the finite sample properties of posteriors.

Nonparametric Bayesian methods are valuable in many situations. But the properties of the posterior in nonparametric problems are only now becoming well understood. There is still much to learn but, from what is now known, we see that there is reason for caution.

**Acknowledgments.** I would like to thank Steve MacEachern, Peter Mueller R.V. Ramamoorthi, Jamie Robins and a referee for reading a draft of this chapter and providing valuable feedback. Research supported by NIH grant RO1-CA54852 and NSF grants DMS-9303557 and DMS-9357646.

## References

Barron, A. (1988). “The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions.” Technical Report 7, Department of Statistics, University of Illinois, Champaign, IL.

Barron, A., Schervish, M. and Wasserman, L. (1997). “The consistency of posterior distributions in nonparametric problems.” Technical report, Department of Statistics, Carnegie Mellon.

Cox, D. (1993). “An analysis of Bayesian inference for nonparametric regression.” *The Annals of Statistics*, **21**, 903-923.

Diaconis, P. and Freedman, D. (1997). “On the Bernstein-von Mises Theorem with Infinite Dimensional Parameters,” unpublished manuscript.

Diaconis, P. and Freedman, D. (1986). “On the consistency Bayes estimates.” *The Annals of Statistics*, **14**, 1-26.

Diaconis, P. and Freedman, D. (1993). “Nonparametric binary regression: A Bayesian approach.” *The Annals of Statistics*, **21**, 2108-2137.

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

Freedman, D. (1963), “On the asymptotic behavior of Bayes’ estimates in the discrete case.” *The Annals of Mathematical Statistics*, **34**, 1386-1403.

Freedman, D. (1965), “On the asymptotic behavior of Bayes’ estimates in the discrete case, II.” *The Annals of Mathematical Statistics*, **36**, 454-456.

Freedman, D., Diaconis, P. (1983). “On inconsistent Bayes estimates in the discrete case.” *The Annals of Statistics*, **11**, 1109-1118.

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

Ghoshal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1997a). "Consistency issues in Bayesian nonparametrics," unpublished manuscript.

Ghoshal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1997b). "Posterior consistency of Dirichlet mixtures in density estimation," unpublished manuscript.

Kass, R.E. and Wasserman, L. (1996). "The selection of prior distributions by formal rules." *Journal of the American Statistical Association,* **91**, 1343-1370.

Robins, J. and Ritov, Y. (1997). "Toward a curse of dimensionality appropriate asymptotic theory for semiparametric models." To appear: *Statistics and Medicine.*

Schwartz, L. (1960). "Consistency of Bayes procedures." Ph.D. dissertation, University of California.

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

Shen, X. (1995). "On the properties of Bayes procedures in general parameter spaces." Unpublished manuscript.

Tierney, L. (1987). "Asymptotic nonparametric posterior distributions," Technical report 497, Department of Statistics, University of Minnesota.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer: New York.

Wasserman, L. (1997). "Asymptotic inference for mixture models using data dependent priors," In preparation.

Wong, W.H. and Shen, X. (1995). "Probability inequalities for likelihood ratios and convergence rates of sieve MLE's." *Annals of Statistics*, **23**, 339-362.

Zhao, L. (1993). "Frequentist and Bayesian aspects of some nonparametric estimation." Ph.D. Thesis, Cornell University.

Zhao, L. (1997). "Bayesian Aspects of Some Nonparametric Problems." Technical report, University of Pennsylvania.