A Note on the Consistency of Bayes Factors for Testing Point Null versus Nonparametric Alternatives

Sarat C. Dass

Michigan State University

Jaeyong Lee

Pennsylvania State University

May 15, 2002

Abstract

When testing a point null hypothesis versus an alternative that is vaguely specified, a Bayesian test usually proceeds by putting a non-parametric prior on the alternative and then computing a Bayes factor based on the observations. This paper addresses the question of consistency, that is, whether the Bayes factor is correctly indicative of the null or the alternative as the sample size increases. We establish several consistency results in the affirmative under fairly general conditions. Consistency of Bayes factors for testing a point null versus a *parametric* alternative has long been known. The results here can also be viewed as the non-parametric extension of the parametric counterpart.

MSC: 62G20; 62C10

Key words: Bayes Factor, Consistency, Dirichlet process, Polya Tree, Infinite Dimensional Exponential Family.

1 Introduction

Non-parametric Bayesian methods have been popular and successful in many estimation problems but their relevance in hypotheses testing situations have become of interest only recently. In particular, the testing of a parametric null versus a non-parametric alternative has received considerable attention from Bayesians, e.g., Berger and Guglielmi (1998), Verdinelli and Wasserman (1998), Carota and Parmigiani (1996), and Florens, Richard and Rolin (1996). Berger and Guglielmi (1998) consider the problem of goodness of fit in the framework of testing a parametric null versus a non-parametric alternative and derive measures of goodness of fit closely related to the Bayes factor. By looking at goodness of fit as a Bayesian test of hypotheses, one can take advantage of many of its attractive features. Bayesian hypothesis testing is not based on asymptotic results, and thus, can be used equally effectively on small or moderate sample sizes. Bayesian hypotheses testing uses Bayes factors to decide between accepting or rejecting the null hypothesis. Thus, as the sample size increases, one can ask if the Bayes factor is correctly indicative of H_0 or H_1 given that the sampling density belongs to one of the two hypotheses. This is the question of consistency.

Even though Bayesian answers in hypothesis testing problems are not operationally based on asymptotics, consistency of the resulting Bayes factor is an important issue that needs to be addressed. In the case of estimation using non-parametric priors, Diaconis and Freedman (1986) show that some posteriors based on n samples need not be consistent, that is, the posterior may not put mass tending to one for sufficiently small neighborhoods of the true parameter value. Thus, inference based on such inconsistent posteriors can be highly misleading.

Analogously, in hypotheses testing, it is important to know if the Bayes procedure based on the Bayes factor actually leads to sensible answers as the sample size increases. Consistency holds for Bayes factors when parametric families are involved in the testing scenario. Even when the sampling distribution does not belong to either H_0 or H_1 in the parametric case, the Bayes factor eventually chooses the hypothesis that is closest to the sampling density in a Kullback Liebler sense. Exact rates of convergence are also well known.

In the case of infinite dimensional parameter spaces, relatively little is known about consistency and rates of convergence of Bayes factors in the case of general non-parametric priors. We establish consistency for the Bayes factor when the null hypothesis is true for any arbitrary non-parametric prior. In the case when the alternative hypothesis is true, we show that the set of all sampling densities under which consistency holds has measure one with respect to the non-parametric prior, regardless of the prior chosen. Our goal is to establish consistency in terms of conditions satisfied by a sampling density in the support of an arbitrary non-parametric prior, and not only on a case by case basis. We only consider non-parametric priors on the space of all probability density functions for reasons explained in Section 3.

The remainder of this paper is organized as follows. Section 2 gives the motivation and definition of consistency pertaining to Bayes factors. Section 3 discusses some well known examples of non-parametric priors on the space of all densities. Sections 4 gives the proofs of theorems in Section 2. We end this paper with a discussion of testing a composite parametric null versus a non-parametric alternative in Section 5.

2 Consistency of Bayes Factors

The following notations will be used throughout the paper. Let $\mathcal{X}$ be a complete separable metric space (or Polish space), μ be a σ-finite measure on $\mathcal{X}$ and $\mathcal{F}$ be the space of all probability densities with respect to μ with support $\mathcal{X}$. Also, denote by $X_1, X_2, \cdots$, random variables taking values in $\mathcal{X}$, which are independent and identically distributed (iid) with a density $f \in \mathcal{F}$. Consider the following hypothesis testing scenario

$$H_0 : f = f_0 \text{ versus } H_1 : f \neq f_0. \tag{1}$$

Equation (1) is the most general form of testing a point null versus a non-parametric alternative. A Bayesian testing procedure would proceed by first specifying prior probabilities, π_0 and π_1, of the null hypothesis and the alternative, respectively, and a non-parametric prior π on the space of the alternative, H_1. We postpone the discussion of what an appropriate prior should be until Section 3 but for now, assume that a non-parametric prior is given. The Bayes factor for the testing of (1), based on a sample, $\underline{x}_n$, of size n, is the ratio of the marginal under H_0 to the marginal under H_1, and is given by the expression

$$B(\underline{x}_n) = \frac{\prod_{i=1}^n f_0(x_i)}{\int_{H_1} \prod_{i=1}^n f(x_i)\pi(df)}. \tag{2}$$

The Bayes factor in (2) can also be interpreted as the ratio of posterior odds to the prior odds of H_0 to H_1. To see this, define an *overall* prior on $H_0 \cup H_1$ as

$$\pi^*(f) = \pi_0 \cdot I_{f_0}(f) + \pi_1 \cdot I_{H_1}(f) \cdot \pi(f), \tag{3}$$

where $I_A(\cdot)$ stands for the indicator function of the set A, i.e., $I_A(f) = 0$ if $f \notin A$ and $I_A(f) = 1$ if $f \in A$. We use the following notation for generic priors and posteriors, namely, if $g(\cdot)$ is a prior on $H_0 \cup H_1$, then we will denote the posterior derived from g based on a sample, $\underline{x}_n$, of size n, by $g(\,\cdot\,|\underline{x}_n)$. Thus, for the prior π^*, the posterior and prior odds ratio is related to the Bayes factor by

$$\frac{\pi^*(H_0|\underline{x}_n)}{\pi^*(H_1|\underline{x}_n)} = \frac{\pi_0}{\pi_1} \cdot B(\underline{x}_n). \tag{4}$$

For all subsequent discussions, we take the default choice for π_0 and π_1, namely, $\pi_0 = \pi_1 = 1/2$. In this case, the posterior odds ratio is exactly equal to the Bayes factor. Thus, given the observations $x_1, x_2, \ldots, x_n$, large values of B would indicate that there is very strong evidence for H_0 based on the data whereas small values of B would indicate otherwise. As the sample size increases indefinitely, we would expect to get perfect information about the sampling density, say f, and the Bayes factor should also correctly and overwhelmingly be able to decide between H_0 and H_1. This motivates the following definition for the consistency of the Bayes factor.

Let $\mathcal{X}^n$ and $\mathcal{X}^\infty$ be the products of n and infinite copies of $\mathcal{X}$. Also, let P_f^n and P_f^∞ be the n and infinite products of the probability measure P_f, which has density f, on $\mathcal{X}^n$ and $\mathcal{X}^\infty$, respectively.

Definition 1 *The Bayes factor, $B(\underline{x}_n)$, for the testing of (1) is said to be consistent if*

$$\lim_{n\to\infty} B(\underline{x}_n) = \infty, P_{f_0}^\infty - a.s.,$$

and for any $f \neq f_0$,

$$\lim_{n\to\infty} B(\underline{x}_n) = 0, P_f^\infty - a.s..$$

Before we give the theorems establishing consistency of Bayes factors, we need a few more definitions. The Kullback-Leibler divergence, $K(f,g)$, provided it exists, between two densities f and g in $\mathcal{F}$ is defined as

$$K(f,g) = \int f(x) \log \frac{f(x)}{g(x)} \mu(dx). \tag{5}$$

Also let

$$K_\epsilon(f) = \{g \in \mathcal{F} : K(f,g) < \epsilon\}, \text{ for } \epsilon > 0. \tag{6}$$

We say f is in the Kullback-Leibler support of π, if

$$\pi(K_\epsilon(f)) > 0, \text{ for all } \epsilon > 0.$$

With the above definitions, we can now state three theorems establishing consistency of the Bayes factor.

Theorem 1 *Under $f_0 \in H_0$,*

$$\lim_{n\to\infty} B(\underline{x}_n) = \infty, P_{f_0}^\infty - a.s.$$

Note that one cannot use Schwartz's criteria for consistency for the prior π^* to obtain Theorem 1. This is because Schwartz's criteria gives consistency only for weak neighborhoods of f_0. We want consistency at $f = f_0$. An involved proof of a weaker result appeared in Verdinelli and Wasserman (1998) in the special case when π is taken to belong to the class of infinite dimensional exponential family priors. Our goal here is to establish strong consistency (almost sure convergence) of the Bayes factor for a general prior π on H_1. Our proof of Theorem 1 follows from a rather simple observation that π^* puts positive mass on H_0 together with (4). Thus, the argument of Doob (1949) is applied without much change. For the completeness, the proof is given in section 4.

Theorem 2 *Let* $\Theta = \{ f \in H_1 : B(\underline{x}_n) \longrightarrow 0, P_f^\infty - a.s..\}$. *Then,* $\pi(\Theta) = 1$.

Theorem 2 states that the Bayes factor is, indeed, consistent for a large set of densities in H_1, namely, a set which has π-probability 1. However, Theorem 2 does not say much about any one *particular* sampling density, f in H_1. To obtain consistency for a particular sampling density, f, we have to further assume that f belongs to the Kullback-Leibler support of the prior, π. This is the result of

Theorem 3 *Suppose $f \in H_1$ is such that f is in the Kullback-Leibler support of the prior π. Then, under f,*

$$\lim_{n\to\infty} B(\underline{x}_n) = 0, P_f^\infty - a.s.$$

We give proofs of the above theorems in Section 4. In the following section, we give examples of non-parametric priors, where the support condition of Theorem 3 has been established for estimation problems. Note that this condition is also sufficient to establish consistency of Bayes factors in hypotheses testing situations by the result of Theorem 3.

3 Examples

3.1 Posterior Consistency of Dirichlet Normal Mixtures

It is well known that the Dirichlet process prior (Ferguson (1973)) puts mass 1 on the space of discrete distributions. However, if one convolutes the Dirichlet prior with an arbitrary number of absolutely continuous density functions, the resulting non-parametric prior would give mass 1 to the space of all probability densities. Using normal kernels gives rise to the Dirichlet mixture of normals. We briefly outline the construction of a Dirichlet normal mixture prior from a Dirichlet process prior. Let $u_1, u_2, \cdots$ be iid random variables from $Beta(1, \alpha(\mathcal{X}))$ and let $Y_1, Y_2, \cdots$ be independent random variables, independent of $u_1, u_2, \cdots$, each distributed according to the probability measure $\alpha_0(\cdot) = \alpha(\cdot)/\alpha(\mathcal{X})$, where $\alpha(\cdot)$ is a finite measure on $\mathcal{X}$. Then, the random P given by

$$P = \sum_{i=1}^{\infty} p_i \delta_{Y_i}, \tag{7}$$

where δ_x is the degenerate probability measure at x, $p_1 = u_1$, and $p_i = u_i \prod_{j=1}^{i-1}(1 - u_j)$ has a Dirichlet process prior. See Sethuraman (1994). The equality in (7) is in the sense of distribution. To obtain Dirichlet normal mixtures from this representation, replace the degenerate probability measure by a normal density with mean Y_i and standard deviation σ. Then, the random density, $g(x)$, of P is of the type

$$g(x) = \sum_{i=1}^{\infty} p_i \frac{1}{\sigma} \phi(\frac{x - Y_i}{\sigma}).$$

In this case, we say P has a Dirichlet normal mixture distribution. There can be various other choices of mixtures based on different choices of the kernel function. The modelling and computational aspects of Dirichlet mixtures were studied, for instance, by MacEachern and Mülller (1998).

Recently, Ghosal, Ghosh and Ramamoorthi (1999b) studied the issue of posterior consistency in the context of density estimation using Dirichlet mixtures. They gave conditions under which the sampling density, f, belongs to the Kullback-Leibler support of the Dirichlet mixture prior.

Before we state the conditions for the posterior consistency of Dirichlet normal mixtures, we introduce some notation and bounds for the tail probabilities of a random probability measure from the Dirichlet process prior. These bounds are first shown by Doss and Sellke (1982) and used by Ghosal, Ghosh and Ramamoorthi (1999b) in their result.

Let ϕ_h be the normal density with mean 0 and standard deviation h and $f_\sigma = \phi_\sigma * f$, i.e., $f_\sigma(x) = \int \phi_\sigma(x-y)f(y)dy$. Suppose P has a Dirichlet process prior with parameter α, where α is a finite measure on $\mathcal{X}$,i.e., $P \sim DP(\alpha)$. Then, there exist $k > 0$ and x_0 such that, for all P in a set of $DP(\alpha)$-probability 1,

$$\begin{array}{lll} P(x,\infty) \geq l_1(x), & P(x + k\log x, \infty) \leq u_1(x), & \text{for } x > x_0, \text{ and} \\ P(-\infty, x) \geq l_2(x), & P(-\infty, x - k\log|x|) \leq u_2(x), & \text{for } x < -x_0, \end{array}$$

where

$$\begin{aligned} l_1 &= \exp[-2\log|\log\alpha_0(x,\infty)|/\alpha_0(x,\infty)], \\ l_2 &= \exp[-2\log|\log\alpha_0(-\infty,x)|/\alpha_0(-\infty,x)], \\ u_1 &= \exp[-\frac{1}{\alpha_0(x+k\log x,\infty)|\log\alpha_0(x-k\log x,\infty)|^2}], \\ u_2 &= \exp[-\frac{1}{\alpha_0(-\infty,x-k\log|x|)|\log\alpha_0(-\infty,x-k\log|x|)|^2}]. \end{aligned}$$

Define

$$L_h(x) = \begin{cases} \phi_h(k\log x)(l_1(x) - u_1(x)), & \text{if } x > 0, \\ \phi_h(k\log|x|)(l_2(x) - u_2(x)), & \text{if } x < 0. \end{cases}$$

We give the conditions of Ghosal, Ghosh and Ramamoorthi (1999b) below.

Theorem 4 *(Ghosal, Ghosh and Ramamoorthi 1999) Suppose* 0 *is in the support of the prior on* σ *and* f *is in the support of* $DP(\alpha)$*. If*

$$\lim_{\sigma\to 0}\int f(x)\log(\frac{f(x)}{f_\sigma(x)})dx = 0;$$

for all $\sigma > 0$,

$$\lim_{a\to\infty}\int f(x)\log(\frac{f_\sigma(x)}{\int_{-a}^{a}\phi_\sigma(x-\theta)f(\theta)d\theta})dx = 0;$$

for all $h > 0$,

$$\lim_{M\to\infty}\int_{|x|>M} f(x)\log(\frac{f_h(x)}{L_h(x)})dx = 0,$$

then $\pi(K_\epsilon(f)) > 0$ for all $\epsilon > 0$.

Note that the above conditions are also sufficient for the consistency of Bayes factors for the hypotheses testing situation by Theorem 3.

3.2 Posterior Consistency of Polyá Tree Priors

We quote two theorems from Ghosal, Ghosh and Ramamoorthi (1999a). Here, $\{B_{\epsilon_1,\epsilon_2,\cdots,\epsilon_k}\}$ represents a usual hierarchical partition of the real line associated with the construction of a Polyá Tree Prior, and the conditional probabilities $P(B_{\epsilon_1,\epsilon_2,\cdots,\epsilon_k}|B_{\epsilon_1,\epsilon_2,\cdots,\epsilon_{k-1}})$ are distributed according to $Beta(\alpha_{\epsilon_1,\epsilon_2,\cdots,\epsilon_k}, 1-\alpha_{\epsilon_1,\epsilon_2,\cdots,\epsilon_k})$ for some constants $0 < \alpha_{\epsilon_1,\epsilon_2,\cdots,\epsilon_k} < 1$. The first theorem gives conditions whereby the Polyá tree prior puts mass one to the class of all absolutely continuous distributions, i.e., distribution with densities.

Theorem 5 *Suppose λ is a continuous probability measure on $\mathcal{R}$ with $\lambda(B_{\epsilon_1,\epsilon_2,\cdots,\epsilon_k}) = 2^{-k}$ for all k and further $\alpha_{\epsilon_1,\epsilon_2,\cdots,\epsilon_k} = a_k$. If $\sum_k a_k^{-1} < \infty$, then the resulting Polyá tree gives mass 1 to the set of all distributions that are absolutely continuous with respect to λ.*

The next theorem gives conditions for a density to be in the Kullback-Leibler support of a Polyá tree prior.

Theorem 6 *Suppose that λ is a continuous probability measure with $\lambda(B_{\epsilon_1,\epsilon_2,\cdots,\epsilon_k}) = 2^{-k}$ for all k and further $\alpha_{\epsilon_1,\epsilon_2,\cdots,\epsilon_k} = a_k$. If $\sum_k a_k^{-1/2} < \infty$, then any density f with respect to λ with $\int f\log f d\lambda < \infty$ belongs to the Kullback-Leibler support of the Polyá tree.*

Thus, when a Polyá tree prior is chosen as the prior for the non-parametric alternative in (1), Theorem 6 provides conditions ensuring consistency of the resulting Bayes factor.

3.3 Infinite Dimensional Exponential Family Priors

Verdinelli and Wasserman (1998) discuss the use of the infinite dimensional exponential family priors for testing goodness of fit. They cast the testing problem (1) into the testing problem of

$$\begin{aligned} H_0: &\quad F_0(X_1), F_0(X_2), \cdots, F_0(X_n) \overset{\text{iid}}{\sim} Uniform(0,1) \\ &\quad \text{versus} \\ H_1: &\quad F_0(X_1), F_0(X_2), \cdots, F_0(X_n) \overset{\text{iid}}{\nsim} Uniform(0,1), \end{aligned}$$

where F_0 is the cumulative distribution function (cdf) of the density f_0 in (1). The infinite dimensional exponential family is constructed for distributions with support on the unit interval $[0,1]$. They use a sequence of Legendre polynomials, $\{\xi_j(\cdot),\ j = 1, 2, \ldots\}$, defined by

$$\xi_j(x) = \frac{1}{2^j j!}\frac{d^j}{dx^j}(x^2-1)^j,$$

and use the Legendre polynomials together with other coefficients, $\phi = (\phi_1, \phi_2, \ldots)$, to define infinite exponential densities of the form

$$g(u|\phi) = \exp(\sum_{j=1}^{\infty} \phi_j \xi_j(u) - c(\phi)),$$

where $c(\phi) = \log \int_0^1 \exp(\sum_j \phi_j \xi_j(u))du$ is the normalizing constant. In order to get random densities, Verdinelli and Wasserman (1998) put priors on the coefficients, ϕ, given by

$$\phi_j \sim \text{ independent } N(0, \tau^2/c_j^2), \tag{8}$$

where τ and c_j's are constants. To establish the Kullback-Leibler support of a sampling density, f, they quote a theorem from Barron (1988), namely,

Theorem 7 *(Barron (1988))*
If $K(f_0, f) < \infty$ and π is the infinite dimensional exponential family prior with $c_j = j^k$ in (8) where $k > 8$ and $\tau > 0$, then f is in the Kullback-Leibler support of π.

While Verdinelli and Wasserman (1998) showed that under an $f \in H_1$, the Bayes factor converges to 0 in probability, Theorem 3 gives stronger convergence (almost sure) of the Bayes factor for the same set of assumptions.

4 Proof of Theorems

In this section, we give the proofs of Theorems 1, 2 and 3. The reader is referred to Section 2 for the notation used here. We use the weak topology on $\mathcal{F}$ and the usual topology on $\mathcal{X}^\infty$. The product topology on the space $\mathcal{F} \times \mathcal{X}^\infty$ is generated in the usual way.

Let ν be a probability measure on $\mathcal{F}$. Given an f sampled from ν, the observations $X_1, X_2, \ldots$ are independent and identically distributed according to P_f, the probability measure corresponding to f. The probability measure P_f is uniquely determined by f and vice versa upto an equivalence class resulting from the equivalence relation, $\sim$, defined by

$$f \sim g \quad \text{if and only if} \quad f = g \text{ a.e. } \mu \tag{9}$$

The notation f will now stand for the equivalence class that it generates. For the prior probability ν on $\mathcal{F}$, write Q_ν for the probability measure on $\mathcal{F} \times \mathcal{X}^\infty$ defined by

$$Q_\nu(A \times B) = \int_A P_f^\infty(B)\, \nu(df), \tag{10}$$

where A is Borel in $\mathcal{F}$ and B is Borel in $\mathcal{X}^\infty$.

Proof of Theorem 1. Assume, first, that there is a consistent estimate for f_0, i.e., $f_0 \in \sigma(X_1, X_2, \cdots)$, the sigma algebra generated by $X_1, X_2, \ldots$. Let π^* be as before, namely, $\pi^*(f) = \frac{1}{2} \cdot I_{f_0}(f) + \frac{1}{2} \cdot I_{H_1}(f) \cdot \pi(f)$. Since

$$\begin{aligned} \pi^*(\,\{f_0\}\,|\underline{x}_n\,) &= \frac{\int_{\{f_0\}} \prod_{i=1}^n f(x_i)\pi^*(df)}{\int_{\mathcal{F}} \prod_{i=1}^n f(x_i)\pi^*(df)} \\ &= \frac{\prod_{i=1}^n f_0(x_i)}{\prod_{i=1}^n f_0(x_i) + \int_{H_1} \prod_{i=1}^n f(x_i)\pi(df)} \\ &= \frac{B}{B+1}, \end{aligned}$$

it suffices to show that

$$\pi^*(\{f_0\}\,|\underline{x}_n) \to 1, P_{f_0}^\infty - \text{a.s..}$$

By the martingale convergence theorem,

$$\lim_{n\to\infty} E(I_{f_0}(f)|X_1,\cdots,X_n) = E(I_{f_0}(f)|X_1,X_2,\cdots), \quad Q_{\pi^*} - a.s..$$

Since $f_0 \in \sigma(X_1, X_2, \cdots)$, $I_{f_0}(\cdot)$ is measurable with respect to $\sigma(X_1, X_2, \cdots)$; hence

$$E(I_{f_0}(f)|X_1,X_2,\cdots) = I_{f_0}(f), \quad Q_{\pi^*} - a.s..$$

Let

$$\Omega_0 = \{(f, x_1, x_2, \ldots) : E(I_{f_0}(f)|X_1, X_2, \cdots, X_n) \to I_{f_0}(f)\}.$$

Then, we have shown that

$$Q_{\pi^*}(\Omega_0) = 1.$$

For an $f \in \mathcal{F}$, define $\Omega_f = \{(x_1, x_2, \ldots) : (f, x_1, x_2, \ldots) \in \Omega_0\}$. Then,

$$1 = Q_{\pi^*}(\Omega_0) = \int_{\mathcal{F}} P_f^\infty(\Omega_f)\pi^*(df).$$

Hence, we get

$$P_f^\infty(\Omega_f) = 1, \pi^* - a.s.$$

Since $\pi^*(\{f_0\}) = 1/2 > 0$, it follows that

$$P_{f_0}^\infty(\Omega_{f_0}) = 1.$$

To show that f_0 is measurable with respect to $\sigma(X_1, X_2, \ldots)$, we need only to compute P_{f_0} from $X_1, X_2, \ldots$, since P_{f_0} uniquely determines f_0. Thus, we need only compute $\int g\,dP_{f_0}$ for any bounded, continuous function g. But,

$$\int g\,dP_{f_0} = \lim_{n\to\infty} \frac{1}{n}[g(X_1) + g(X_2) + \ldots + g(X_n)]$$

by the law of large numbers. □

To prove Theorem 2, we need a few lemmas.

Lemma 1 *The posterior,* $\pi^*(\cdot|\underline{x}_n) \longrightarrow \delta_f(\cdot)$ *weakly,* $Q_{\pi^*} - a.s.$*, where* $\delta_f(\cdot)$ *is the degenerate probability at* f*.*

Proof. The reader is referred to Diaconis and Freedman (1986) for a proof. The proof is similar to the one given above for $f = f_0$. □

Lemma 2 *Fix* $f \in H_1$*. Let*

$$\Theta^* = \{(x_1, x_2, \ldots) : B(\underline{x}_n) \longrightarrow 0\}$$

and

$$\Theta_f = \{(x_1, x_2, \ldots) : \pi^*(\cdot|\underline{x}_n) \longrightarrow \delta_f(\cdot) \textit{ weakly.}\}.$$

Then, $\Theta^* \supseteq \Theta_f$*.*

Proof. Choose a $(x_1, x_2, \ldots)$ in Θ_f, and a sufficiently small weak neighborhood of f, N, not intersecting H_0. Since

$$B(\underline{x}_n) = \frac{\pi^*(H_0|\underline{x}_n)}{\pi^*(H_1|\underline{x}_n)}$$

and $N \subseteq H_1$, we have $\pi^*(H_1|\underline{x}_n) \longrightarrow 1$ and $\pi^*(H_0|\underline{x}_n) \longrightarrow 0$. It follows that $B(\underline{x}_n) \longrightarrow 0$. □

Proof of Theorem 2. Define

$$\Theta_0 = \{(f, x_1, x_2, \ldots) : \pi^*(\cdot|\underline{x}_n) \longrightarrow \delta_f(\cdot) \text{ weakly } \}$$

and Θ_f as in Lemma 2. By Lemma 1, we have $Q_{\pi^*}(\Theta_0) = 1$. Since

$$Q_{\pi^*}(\Theta_0) = \frac{1}{2} \cdot P_{f_0}^{\infty}(\Theta_{f_0}) + \frac{1}{2} \cdot \int_{\mathcal{F}} P_f^{\infty}(\Theta_f)\, \pi(df),$$

we have that $P_{f_0}^{\infty}(\Theta_{f_0}) = 1$ and $P_f^{\infty}(\Theta_f) = 1$, $\pi - a.s.$. By Lemma 2, $P_f^{\infty}(\Theta^*) = 1, \pi - a.s.$. □

To prove Theorem 3, we need the following lemma.

Lemma 3 *Suppose $X_1, X_2, \cdots$ are iid from f and f is in the Kullback-Leibler support of π. Then, for all $\epsilon > 0$,*

$$\liminf_{n\to\infty} e^{n\epsilon} \int \frac{\prod_{i=1}^n g(x_i)}{\prod_{i=1}^n f(x_i)} \pi(dg) = \infty.$$

Proof. Let $\epsilon > 0$ be given.

$$\begin{aligned}
& \liminf_{n\to\infty} e^{n\epsilon} \int \exp(-n\frac{1}{n}\sum_{i=1}^n \log \frac{f(X_i)}{g(X_i)})\pi(dg) \\
\geq\ & \liminf_{n\to\infty} e^{n\epsilon} \int_{g\in K_{\epsilon/2}(f)} \exp(-n\frac{1}{n}\sum_{i=1}^n \log \frac{f(X_i)}{g(X_i)})\pi(dg) \\
=\ & \liminf_{n\to\infty} \int_{g\in K_{\epsilon/2}(f)} \exp\{n(\epsilon - \frac{1}{n}\sum_{i=1}^n \log \frac{f(X_i)}{g(X_i)})\}\pi(dg) \\
\geq\ & \int_{g\in K_{\epsilon/2}(f)} \liminf_{n\to\infty} \exp\{n(\epsilon - \frac{1}{n}\sum_{i=1}^n \log \frac{f(X_i)}{g(X_i)})\}\pi(dg) \\
=\ & \infty.
\end{aligned}$$

The second to last inequality is due to Fatou's lemma and the last equality is by that fact that, for all $g \in K_{\epsilon/2}(f)$,

$$\lim_{n\to\infty} \epsilon - \frac{1}{n}\sum_{i=1}^n \log \frac{f(X_i)}{g(X_i)} = \epsilon - K(f,g) > \epsilon/2, a.s.,$$

by the choice of g and the strong law of large numbers.

□

Proof of Theorem 3. The Bayes factor for testing (1) can be written as

$$\begin{aligned}
B &= \frac{\prod_{i=1}^n f_0(X_i)/f(X_i)}{\int \prod_{i=1}^n g(X_i)/f(X_i)\pi(dg)} \\
&= \frac{\exp(-n\frac{1}{n}\sum_{i=1}^n \log f(X_i)/f_0(X_i))}{\int \prod_{i=1}^n g(X_i)/f(X_i)\pi(dg)}. \qquad (11)
\end{aligned}$$

Let $\epsilon = K(f, f_0)/2 > 0$. We will show the numerator in (11) multiplied by $e^{n\epsilon}$ goes to 0 and the denominator multiplied by $e^{n\epsilon}$ goes to ∞. First, since $\lim_{n\to\infty}(\epsilon - \frac{1}{n}\sum_{i=1}^n \log f(X_i)/f_0(X_i)) \leq$

$-\epsilon/6$,

$$\limsup_{n\to\infty} \exp\{n(\epsilon - \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i)}{g(X_i)})\} = 0 \ a.s.$$

Second, by Lemma 3,

$$\liminf_{n\to\infty} e^{n\epsilon} \int \prod_{i=1}^{n} g(X_i)/f(X_i)\pi(dg) = 0 \ a.s.$$

Combining these two, we get the conclusion

$$B \to 0, P_g^\infty - a.s.$$

□

5 Discussion

In this paper, we only considered the problem of testing a point null versus non-parametric alternative and showed that under very weak conditions, the resulting Bayes factor was consistent. Of course, what is more interesting is to see if the consistency results hold for the more general composite testing of

$$\begin{aligned} H_0: &\quad f \text{ belongs to the } N(\mu, \sigma^2) \text{ family} \\ &\quad \text{versus} \\ H_1: &\quad f \text{ does not belong to the } N(\mu, \sigma^2) \text{ family,} \end{aligned}$$

for example. The consistency of the Bayes factors for composite hypotheses testing situations such as the above is still an open question.

References

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

Berger, J.O., Guglielmi, A., 2001. Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. Jour. Amer. Statist. Assoc. 96, 174-184.

Carota, C., Parmigiani, G., 1996. On Bayes factors for nonparametric alternatives. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 5, London, Oxford University Press, pp. 507-511.

Doob, J.L., 1949. Application of the theory of martingales. Colloque Internal du CNRS, 22-28.

Doss, H., Sellke, T., 1982. The tails of probabilities chosen from a Dirichlet prior. Ann. Statist. 10, 1302-1305.

Diaconis, P., Freedman, D., 1986. On the consistency of Bayes estimates. Ann. Statist. 14, 1-67.

Ferguson, T.S., 1973. A Bayesian analysis of some non-parametric problems. Ann. Statist. 1, 209-230.

Florens, J.P., Richard, J.F., Rolin, J.M., 1996. Bayesian encompassing specifications tests of a parametric model against a nonparametric alternative. Technical Report 96.08, Institut de Statistique, Université Catholique de Louvain, Belgium.

Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V., 1999a. Consistency issues in Bayesian nonparametrics. In: Ghosh, S. (Eds.), Asymptotic Nonparametrics and Time Series: A Tribute to Madan Lal Puri, Marcel Dekker, Inc., pp. 639-667.

Ghosal, S., Ghosh, J., amamoorthi, R., 1999b. Posterior consistency of Dirichlet mixtures in density estimation. Ann. Statist., 27, 143-158.

MacEachern, S.N., and Müller, P., 1998. Estimating mixture of Dirichlet process models. J. Comp. Graph. Statist., 7, 223-239.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica, 4, 639-650.

Verdinelli, I., Wasserman, L., 1998. Bayesian goodness of fit testing using infinite dimensional exponential families. Ann. Statist. 20, 1203-1221.