The Annals of Statistics
1998, Vol. 26, No. 4, 1215–1241

BAYESIAN GOODNESS-OF-FIT TESTING USING INFINITE-DIMENSIONAL EXPONENTIAL FAMILIES

BY ISABELLA VERDINELLI[1] AND LARRY WASSERMAN[1,2]

University of Rome and Carnegie Mellon University and Carnegie Mellon University

We develop a nonparametric Bayes factor for testing the fit of a parametric model. We begin with a nominal parametric family which we then embed into an infinite-dimensional exponential family. The new model then has a parametric and nonparametric component. We give the log density of the nonparametric component a Gaussian process prior. An asymptotic consistency requirement puts a restriction on the form of the prior, leaving us with a single hyperparameter for which we suggest a default value based on simulation experience. Then we construct a Bayes factor to test the nominal model versus the semiparametric alternative. Finally, we show that the Bayes factor is consistent. The proof of the consistency is based on approximating the model by a sequence of exponential families.

1. Introduction. Many statistical analyses begin with the assumption that the data are generated from a distribution that belongs to a finite-dimensional parametric model. Usually, the model is only an approximation and is used mainly for convenience. It is thus important to check the fit of the assumed model and, when the fit is poor, replace the nominal model with a more flexible one. There exist many frequentist methods for checking fit. Recent examples include Dümbgen (1998), Eubank and LaRiccia (1992), Härdle and Mammen (1993), Hart (1997) and Inglot and Ledwina (1996), among others. These tests generally posit a parametric null against an infinite-dimensional alternative. However, Bayesian methods for testing a parametric model against an infinite-dimensional alternative are lacking and it seems reasonable to see how such a test can be constructed. This paper proposes such a test. Of course, this test can also be used as a frequentist test. Although we examine some of the frequentist properties of the test, we leave more detailed questions about comparisons with other methods to future work.

Received January 1996; revised April 1998.
[1]Supported by NIH Grant R01-CA54852-01
[2]Supported by NSF Grants DMS-9303557 and DMS-9357646.
AMS 1991 *subject classifications*. Primary 62F15, 62G10.
Key words and phrases. Bayes factor, consistency, Gaussian process prior, Markov chain Monte Carlo, nonparametric Bayesian inference, sieve.

Our aim is to construct a nonparametric alternative model $\mathscr{M}_1$ to a given parametric model $\mathscr{M}_0$ and to find the Bayes factor

$$B = \frac{\Pr(\mathscr{M}_0|\text{Data})}{\Pr(\mathscr{M}_1|\text{Data})} \div \frac{\Pr(\mathscr{M}_0)}{\Pr(\mathscr{M}_1)}.$$

The Bayes factor can be used directly as a test of fit. Alternatively, one can make inferences using a mixture of the two models. For example, predictions of a future observation can be based on

$$\begin{aligned}\Pr(Y \le y|\text{Data}) = {} & \Pr(Y \le y|\text{Data}, \mathscr{M}_0)\Pr(\mathscr{M}_0|\text{Data}) \\ & + \Pr(Y \le y|\text{Data}, \mathscr{M}_1)\Pr(\mathscr{M}_1|\text{Data}).\end{aligned}$$

However, the latter require one to find $\Pr(\mathscr{M}_0|\text{Data})$, which is a function of B. In either case, one must find B. Thus, our focus will be on B without regard to whether the ultimate goal is testing or model averaging.

Most Bayesian methods for assessing the fit of parametric models generally fall into three categories. The first category consists of informal methods, such as predicting deleted observations [Gelfand, Dey and Chang (1992), Gelfand and Dey (1994)] which, though simple, are difficult to justify [Raftery (1992)] and are not guaranteed to lead to an asymptotically consistent test. The second consists of p-values and related techniques, perhaps cast in some partially Bayesian way. Examples include Box (1980), Good (1967, 1992), Rubin (1984), Gelman, Meng and Stern (1996) and Meng (1994). Again, the simplicity of these procedures is appealing, but they are difficult to interpret in a Bayesian framework. A third approach, which we use in this paper, is to embed the parametric family in a larger family, which we call the extended model. For example, Box and Tiao (1973), Chapter 3, embedded the normal family within the power exponential family. This family has an extra parameter β and the normal corresponds to $\beta = 0$. Neyman (1937) proposed an exponential family extension of the nominal model; see also Ledwina (1994) and Rayner and Best (1990). This is the approach we follow. A criticism of this third approach is that the larger model could itself be wrong. We address this problem by using an infinite-dimensional (nonparametric) model, thus making the extended model nonparametric. We test the nominal model using a Bayes factor approach. Delampady and Berger (1990) consider Bayes factors for testing fit based on partitioning the real line. Kass and Raftery (1995) discuss Bayes factors in greater detail; see also Berger and Pericchi (1994) and O'Hagan (1995).

To implement the method, we need to construct a specific extended model. There are many infinite-dimensional models used in Bayesian inference such as Dirichlet processes [Ferguson (1973)], mixtures of Dirichlet processes [Antoniak (1974)], Pólya trees [Lavine (1994)] and Gaussian and log Gaussian processes [Barron (1998), Lenk (1988, 1991), Leonard (1978), Thorburn (1986)]. Alternatively, one may use high-dimensional parametric families such as mixtures of normals [Diebolt and Robert (1994), Nobile (1994), Roeder and Wasserman (1995), West (1992)]. We use a log Gaussian process

built from orthogonal polynomials. This model was used in a different context in Barron (1988), Barron and Sheu (1991), Barron and Cover (1991) and Crain (1974, 1976). Briefly, our approach is as follows.

We begin with a random variable U on the unit interval. We model the log density of this random variable as an infinite series of orthogonal polynomials. We then place Gaussian priors on the coefficients of the polynomials. The prior variances of the coefficients die off rapidly to ensure consistency of both the density estimate and the Bayes factor. The log density is thus a Gaussian process as in Barron [(1988), Section 8], Lenk (1988, 1991) and Leonard (1978). Next, we perform an inverse integral transform using the original parametric family; that is, we set $Y = F_\theta^{-1}(U)$ where θ is the parameter of the nominal model. The result is a semiparametric family focused around the parametric family F_θ. Although previous authors have used inverse integral transforms to transfer the distribution to the unit interval, it seems that this transformation is usually assumed to be known. Instead, we allow parameters in the transformation. In our formulation, gross features of the distribution such as location and scale are picked up parametrically, and the nonparametric component accommodates deviations from the parametric model. If a fixed inverse integral transform is used, then the nonparametric part must also estimate location and scale. In a sense, this places a much greater burden on the nonparametric component of the analysis. We also discuss a simple approximation based on a method of Brunk (1978).

Obviously, it is not possible to carry out computations with the infinite-dimensional model, so in practice, we truncate the infinite series at a fixed number of terms m. However, the theory still works even when m is infinite. We have found it is usually not necessary to let m be very large; often taking $m = 5$ to $m = 10$ suffices.

Other authors have built nonparametric models around parametric models. Some recent approaches are discussed in Hjort (1994), Hjort and Glad (1994) and Hjort and Jones (1994). Efron and Tibshirani (1995) consider a method which might be seen as the dual of our method. They start with a nonparametric estimate, such as a kernel density estimate, and multiply this by a correction factor consisting of a parametric exponential family. In contrast, we multiply a parametric component by an infinite dimensional (nonparametric) exponential family. Evans and Schwartz (1994) construct a family consisting of a normal times a polynomial. A related idea is discussed in Geweke (1989). All these papers emphasize density estimation. We know of none that deal with Bayes factors for testing fit.

Our approach may also be viewed as a Bayesian version of Neyman's (1937) smooth goodness-of-fit test against uniformity. Neyman used a finite-dimensional exponential family as an alternative against the uniform. This test has further been developed in a frequentist setting by many people; recent results along with a review of the literature are given in Ledwina (1994) and Rayner and Best (1990). However, it appears that little has been done in the way of formal Bayesian goodness-of-fit testing using Bayes factors. Delampady and Berger (1990) discuss a Bayesian goodness-of-fit test

obtained by dividing the real line into a finite partition and then using a multinomial model. This is an important step forward but still leaves open many questions such as the choice of partition. Bayarri (1985) used a one-parameter extension model together with a decision theoretic framework in place of a Bayes factor approach.

We present the model in Section 2 and we discuss the construction of the priors in Section 3. The prior has a hyperparameter that acts as a smoothing parameter. A further prior is placed on this parameter so that the data can adaptively choose the amount of smoothing. We choose the prior variances so that nonparametric consistency is guaranteed for a large class of densities. We completely specify the prior apart from a single hyperparameter. We then suggest a default value for the hyperparameter based on simulation experience. In Section 4 we consider methods for computing the posterior. In Section 5 we discuss the Bayesian goodness-of-fit test. Like all Bayesian tests, ours has the virtue that it quantifies evidence for and against the nominal model in contrast to frequentist tests, which can only reject a nominal model. In Section 6 we discuss density estimation. We study some examples and simulations in Section 7. In Section 8 we show that the Bayes factor and the density estimate are consistent. In Section 9 we give some closing remarks.

Before proceeding, we need to say a word about the purpose of simulation studies and consistency results in Bayesian inference. We believe that nonparametric Bayesian methods have not become popular partly because the user needs to specify numerous hyperparameters on a case-by-case basis. This renders the methods impractical. Indeed, some papers on Bayesian nonparametric inference present examples where the hyperparameters are tuned specifically for each example. We think it is important to provide suggested default values for the hyperparameters so that the prior is completely specified. Moreover, it is also important to provide at least some simulation evidence that the method works reasonably well under a variety of conditions as well as theoretical results to show good large sample behavior.

From a strict Bayesian point of view, the sampling properties of a Bayesian procedure are not relevant and the hyperparameters should be chosen subjectively. From a more pragmatic perspective, poor sampling properties suggest that the model or prior is not well chosen [Diaconis and Friedman (1986)]. We acknowledge that some statisticians may object to choosing the hyperparameters of a prior this way, but we feel it is simply impractical to choose them subjectively. Berger and Bernardo (1989) argue that frequentist behavior of Bayes procedures is useful in choosing priors.

2. Extending a parametric model. Consider a family of distributions $\mathscr{F} = \{F(\cdot|\theta)\colon \theta \in \Omega\}$ and let Y be a random variable such that $Y \sim F(\cdot|\theta)$. We refer to $\mathscr{F}$ as the *nominal model*. We begin by reexpressing the distribution for Y in the following way:

$$Y = F^{-1}(U|\theta) \quad \text{where } U \sim \mathscr{U}(0,1) \text{ and } \theta \in \Omega. \tag{1}$$

Let $\mathscr{G} = \{G(\cdot|\psi);\ \psi \in S\}$ be a family of distributions on $[0, 1]$ and let $G(\cdot|\psi_0)$ be the uniform $(0, 1)$ distribution which we assume is a member of $\mathscr{G}$. Here, S is the parameter space for ψ. We use $\mathscr{G}$ to model departures from the family $\mathscr{F}$ and we refer to $\mathscr{G}$ as the *extended model*. Specifically, we replace (1) with

$$Y = F^{-1}(U|\theta) \quad \text{where } U \sim G(\cdot|\psi),\ \theta \in \Omega \text{ and } \psi \in S. \tag{2}$$

From (2), the distribution function for Y is $H(y|\theta, \psi) = G(F(y|\theta)|\psi)$ and, assuming $F(y|\theta)$ has density $f(y|\theta)$ and $G(u|\psi)$ has density $g(u|\psi)$, the density for Y is

$$h(y|\theta, \psi) = f(y|\theta)\, g(F(y|\theta)|\psi). \tag{3}$$

Thus, the new density is simply the product of the original density and a perturbation factor. The new family of distribution functions is

$$\mathscr{H} = \{H(\cdot|\theta, \psi) = G(F(\cdot|\theta)|\psi);\ \theta \in \Omega,\ \psi \in S\}, \tag{4}$$

which we call the *hybrid of $\mathscr{F}$ and $\mathscr{G}$*. Of course, the nominal model $\mathscr{F}$ is contained as a special case when $\psi = \psi_0$.

For $\mathscr{G}$ we use an infinite-dimensional exponential family considered by Barron (1988) and Lenk (1988, 1991). The family is based on orthogonal series which has been used in many different settings; see Brunk (1978), Wahba (1981), Whittle (1958) and the book by Tarter and Lock (1993).

Let $\{\phi_0, \phi_1, \phi_2, \dots,\}$ be a sequence of bounded, orthonormal functions on $[0, 1]$ with respect to Lebesgue measure where $\phi_0 \equiv 1$. In what follows, any basis could be used. In our implementation we use Legendre polynomials. The polynomials have been rescaled to live on the unit interval and to have mean 0 and variance 1 with respect to the uniform probability. In other words, $\phi_j(u) = \sqrt{2j + 1}\,\tilde{\phi}_j(2u - 1)$ where $\tilde{\phi}_j$ are defined on $[-1, 1]$ by

$$\tilde{\phi}_j(x) = \frac{1}{2^j j!}\frac{d^j}{dx^j}(x^2 - 1)^j.$$

Let $\psi = (\psi_1, \psi_2, \dots)$ and define

$$g(u|\psi) = \exp\left\{\sum_{j=1}^{\infty} \psi_j \phi_j(u) - c(\psi)\right\}, \tag{5}$$

where $c(\psi) = \log \int_0^1 \exp\{\sum_j \psi_j \phi_j(u)\}\, du$. The choice of prior for ψ is important; we discuss this in Section 3.

Barron and Sheu (1991) consider the model in (5) but with the summation extending only to a finite number of terms m. Then they let m increase with sample size n thus creating a sieve in the sense of Grenander (1981), and they obtain the best rates of convergence for the sieve maximum likelihood estimator. The infinite-dimensional version is considered in Barron (1988).

Although we do not pursue them here, we now briefly mention some alternatives. First, one can put a prior on m and include m as a parameter in the estimation procedure. Asymptotically, this is the same as using the

Schwartz criterion [Schwartz (1978)] to choose m; Barron and Cover (1991) discuss this possibility. Still another alternative is to let $0 \leq m \leq M$ where M grows at an appropriate rate and a prior is placed on m with support $\{1, \ldots, M\}$. This combines the sieve idea with the Bayesian approach and has the advantage of allowing the choice of dimension to be data dependent.

The methods in this paper are very general and can be applied to any parametric model. For concreteness, we shall mostly concern ourselves with the normal family. Thus we take $f(x|\theta) = \sigma^{-1}\{2\pi\}^{-1/2} \exp\{-(x - \mu)^2/2\sigma^2)\}$ where $\theta = (\mu, \gamma)$ and $\gamma = \log \sigma$.

3. The prior. In this section we discuss the choice of prior. We shall take θ and ψ to be independent, that is, $p(\theta|\psi) = p(\theta)$. This is a nontrivial assumption and there are good reasons for thinking it may be inappropriate. However, the crucial part of the prior is $p(\psi)$ so we shall content ourselves with the independence assumption. The prior for θ seems not to be too important and we shall use standard reference priors. In the normal, writing $\theta = (\mu, \gamma)$ where μ is the mean and $\sigma = e^{\gamma}$ is the standard deviation, we use $p(\mu, \gamma) \propto 1$.

The prior for ψ is more important. Since nonzero values of ψ_j represent deviations from the nominal model, it seems reasonable to use priors that are symmetric, unimodal and centered at 0. Note that had we not allowed the inverse integral transform to have free parameters, we could not interpret the coefficients this way, and the construction of a reasonable prior would be much more difficult. For simplicity, we take $\psi_1, \psi_2, \ldots,$ to be independent. The orthogonality of the polynomials $\phi_1, \phi_2, \ldots$ suggests that independence is reasonable. A natural choice is $\psi_j \sim N(0, \tau_j^2)$. Thus, $\Sigma_j \psi_j \phi_j(u)$ is a Gaussian process. One guiding principle in choosing τ_j is to require some sort of consistency. One possibility is to require that the predictive density (which is the Bayes estimate of the unknown density under a variety of loss functions) be consistent, in the sense that it converges in total variation distance, with probability 1, to the true density p, for a large class of p's. Based on work of Barron (1988), a sufficient condition for consistency of the predictive distribution is $E(\exp(\alpha b_k|\theta_k|) \leq \exp(a_k)$ for some $\alpha > 0$ where b_k is the supremum of the derivative of the kth Legendre polynomial and $\{a_k;\ k = 1, 2, \ldots\}$ is an absolutely summable sequence; see Section 8. We achieve this consistency by fixing $\varepsilon > 0$ and choosing $\tau_j = \tau/c_j$ where $c_j = j^{3+\varepsilon}$ or $c_j = (1 + \varepsilon)^j$. Alternatively, we might require consistency of the Bayes factor. (The Bayes factor is defined in Section 5.) We show in Section 8 that consistency of the Bayes factor is implied by $c_j = j^{8+\varepsilon}$ or $c_j = (1 + \varepsilon)^j$. For the numerical results of this paper we used $c_j = 2^j$.

The parameter τ controls the amount of smoothing. Rather than fixing this value, we have found it better to add a further stage to the model by placing a prior on τ and letting the data choose the amount of smoothing. Doing so has another important benefit: treating τ as an unknown parameter simplifies the goodness-of-fit test. For, instead of testing $(\psi_1, \psi_2, \ldots) = (0, 0, \ldots)$, we

now only need to test $\tau = 0$. In Section 5, we show that there is a simple method for doing this one-dimensional test.

Finally, we need a prior $p(\tau)$ for τ. We would like a prior that decreases monotonically from 0. Further, it is important that $p(0)$ be finite and strictly greater than 0 at $\tau = 0$; otherwise it may not make good sense to test $\tau = 0$. We choose $p(\tau)$ to be a normal distribution with mean 0, variance w^2, truncated to the positive part of the real line. This leaves only the choice of w. Our numerical experience, documented in Section 7, suggests that $w = 1$ works well. With w specified, the prior is now completely determined and requires no subjective input.

4. The posterior. We now need to obtain the posterior distribution of ψ, θ, τ given $Y = y$ where $Y = (Y_1, \ldots, Y_n)$ and $y = (y_1, \ldots, y_n)$. For computation, we truncate the infinite sums to m. By a direct application of Bayes' theorem, the posterior has a density $p(\theta, \tau, \psi | y_1, \ldots, y_n)$ on the space $\Omega \times \mathbb{R}^+ \times \mathbb{R}^m$ given by

$$
\begin{aligned}
p(\theta, \tau, \psi | y_1, \ldots, y_n) \propto & \left[\prod_i f(y_i|\theta)\right]\left[\prod_i g(u_i|\psi)\right] \\
& \times \left[\tau^{-m} \exp\left\{-\sum_{j=1}^{m} c_j^2 \psi_j^2/(2\tau^2)\right\}\right] \\
& \times \exp(-\tau^2/(2w^2))
\end{aligned} \tag{6}
$$

where $u_i = F(y_i|\theta)$. When the nominal family is normal, and the prior $p(\mu, \gamma) \propto 1$ is used, (6) becomes

$$
\begin{aligned}
& p(\mu, \gamma, \tau, \psi | y_1, \ldots, y_n) \\
& \quad \propto \exp(-n\gamma) \exp\left\{-\frac{1}{2\exp(2\gamma)}\left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]\right\} \\
& \qquad \times \exp\left(n\left[\overline{\phi}^T \psi - c(\psi)\right]\right) \tau^{-m} \exp\left\{-\sum_{j=1}^{m} c_j^2 \psi_j^2/(2\tau^2)\right\} \\
& \qquad \times \exp(-\tau^2/(2w^2)),
\end{aligned}
$$

where $\bar{y}$ is the sample mean, s^2 is the sample variance, $\overline{\phi}^T = (\overline{\phi}_1, \ldots, \overline{\phi}_m)$ and $\overline{\phi}_j = n^{-1} \sum_{i=1}^{n} \phi_j(u_i)$.

Computing $p(\theta, \tau, \psi | y)$ is intractable so we shall draw a random sample from the posterior using Markov chain Monte Carlo. This technique has now become quite standard and has been discussed in so many contexts that we shall not go into detail; some key references are Tanner and Wong (1987), Gelfand and Smith (1990) and Tierney (1994).

To draw from the posterior we use a Metropolis algorithm embedded in a Gibbs sampling scheme. Let $\delta = \log(\tau)$. At the ith step in the algorithm we draw

$$
\begin{aligned}
\mu^{(i)} &\sim p\big(\mu|\gamma^{(i-1)}, \delta^{(i-1)}, \psi_1^{(i-1)}, \ldots, \psi_m^{(i-1)}\big),\\
\gamma^{(i)} &\sim p\big(\gamma|\mu^{(i)}, \delta^{(i-1)}, \psi_1^{(i-1)}, \ldots, \psi_m^{(i-1)}\big),\\
\delta^{(i)} &\sim p\big(\delta|\mu^{(i)}, \gamma^{(i)}, \psi_1^{(i-1)}, \ldots, \psi_m^{(i-1)}\big),\\
\psi_1^{(i)} &\sim p\big(\psi_1|\mu^{(i)}, \gamma^{(i)}, \delta^{(i)}, \psi_2^{(i-1)}, \ldots, \psi_m^{(i-1)}, y\big),\\
&\vdots\\
\psi_m^{(i)} &\sim p\big(\psi_m|\mu^{(i)}, \gamma^{(i)}, \delta^{(i)}, \psi_2^{(i)}, \ldots, \psi_{m-1}^{(i)}, y\big).
\end{aligned}
\tag{7}
$$

The vector $(\mu^{(i)}, \gamma^{(i)}, \delta^{(i)}, \psi_1^{(i)}, \ldots, \psi_m^{(i)})$ has a distribution which, under weak conditions converges to the posterior as the number of iterations N goes to infinity. To draw from each of the conditional distributions in (7) we use a Metropolis algorithm driven by a Gaussian random walk. Suppose that α is one of the parameters and its conditional density is $p(\alpha)$. We draw $\alpha' \sim N(\alpha_i, t)$ for some fixed t. Let $r = \max\{p(\alpha')/p(\alpha_i), 1\}$. We then set $\alpha_i = \alpha'$ with probability r and $\alpha_i = \alpha_{i-1}$ with probability $1 - r$. The efficiency of this method depends on the choice of t. We have found that the following choices of t work well for this model: $\mu(t = s/\sqrt{n})$, $\gamma(t = 1/\sqrt{n})$, $\delta(t = 10/\sqrt{n})$ and $\psi_j(t = 2/\sqrt{n})$.

Our ultimate goal is to compute a Bayes factor. Drawing a random sample from the posterior does not lead immediately to an estimate of the Bayes factor needed in the goodness-of-fit test. Discussion on this matter is postponed until Section 5.

4.1. *Brunk's method.* The procedures described above are computationally intensive. Brunk (1978) proposed a simpler method that may be regarded as a crude approximation of the current method. Assume that ψ is not too far from 0. A first order approximation of g is then $g(u|\psi) \approx 1 + \sum_{j=1}^{\infty} \psi_j \phi_j(u)$. Let $\hat{\theta}$ be a point estimate of θ under the nominal model and fix θ at $\hat{\theta}$. Let $U_i = F_{\hat{\theta}}(X_i)$ and treat the U_i as a sample from g. Then we employ a normal approximation from Brunk (1978) to the resulting posterior and fix τ at its prior expectation $(\sqrt{2/\pi})w$. This leads to the approximation $\psi|y_1, \ldots, y_n \approx N(\hat{\psi}, \Sigma)$ where $\hat{\psi}_j = n\overline{\phi}_j/(n + c_j^2/\tau^2)$ and Σ is diagonal with the jth diagonal element being b_j^2 where $b_j^{-2} = n + c_j^2/\tau^2$. The behavior of this method is examined in Section 7.

5. Goodness of fit. We are now ready to test the null hypothesis that the nominal model is correct. We begin by giving a brief, general description of Bayesian testing. Then we discuss a method for computing the Bayes factor.

5.1. *Bayesian testing*. Jeffreys' [(1961), Chapter 5], method for testing the model $\mathscr{F}_1 = \{F(\cdot|\theta_1);\ \theta_1 \in \Omega_1\}$ versus the model $\mathscr{F}_2 = \{F(\cdot|\theta_2);\ \theta_2 \in \Omega_2\}$ is to compute the Bayes factor

$$B = \frac{\int f(x|\theta_1)p_1(\theta_1)\,d\theta_1}{\int f(x|\theta_2)p_2(\theta_2)\,d\theta_2}.$$

This may be interpreted as the ratio of the posterior odds in favor of model 1 to the prior odds in favor of model 1 under a prior that puts mass π on model 1 ($0 < \pi < 1$) and mass $1 - \pi$ on model 2 and spreads the mass in each model according to the densities p_1 and p_2, respectively. Large values are interpreted as evidence for model 1. In this paper, we take $B < 1$ as a criterion for rejecting the nominal model, which corresponds to putting equal prior probability on each model. Under weak conditions, it can be shown that B is consistent in the sense that the probability of choosing the wrong model tends to 0 as n tends to infinity, almost surely. If the parameter in $\mathscr{F}_2$ can be written as $\theta_2 = (\theta, \alpha)$ and if $\mathscr{F}_1$ corresponds to $\mathscr{F}_2$ with $\alpha = \alpha_0$, then the models are nested and the Bayes factor is testing the hypothesis that $\alpha = \alpha_0$.

Bayes factors are sometimes criticized on the grounds that it may be unrealistic to assume that one of the two models holds. In particular, in the nested model case, it may seem unrealistic to place a lump of positive mass on a lower-dimensional submodel. Although it is not our goal to defend Jeffreys' approach to testing, we do want to remark that one need not believe that a given model has a positive probability of exactly being true to use a Bayes factor. Rather, the Bayes factor may be interpreted as testing whether a hypothesis is a reasonable working hypothesis. For example, when we test whether the data are normal, say, we interpret this as a test of whether normality is a reasonable working hypothesis. Thus, regardless of the fact that we would not seriously expect the data to be exactly normally distributed, we do think it is reasonable to test for normality. This view is expressed in Jeffreys and is discussed by Kass and Raftery (1995) and Raftery (1992).

5.2. *Computing Bayes factors*. Note that the Bayes factor is the ratio of the normalizing constants. Most methods for simulating from the posterior avoid computing the normalizing constant of the posterior so standard Monte Carlo methods do not give a direct estimate of B. The problem of estimating these constants by simulation is an active area of research; see Gelfand and Dey (1994), DiCiccio, Kass, Raftery and Wasserman (1995) and Meng and Wong (1993). Most of these methods are variations on importance sampling and require making delicate choices of importance samplers. A method that avoids these difficulties was given in Verdinelli and Wasserman (1995) and is applicable to this case. Here, that method reduces to the following observation: after some simple algebra we may write $B = p_\tau(0|y_1, \ldots, y_n)/p_\tau(0)$. The subscripts indicate that these are marginal densities for τ. The numerator is unknown but can be estimated by kernel density estimation from the sampled values $\tau_1, \ldots, \tau_N$. Since we are interested in the posterior at the bound-

ary we use a reflected normal kernel as in Silverman (1986), page 30. We also use Silverman's (1986), page 45, rule of thumb for choosing the bandwidth, namely, $h = 1.06s/N^{1/5}$ where s is the standard deviation of $(\tau_1, -\tau_1, \tau_2, -\tau_2, \ldots, \tau_N, -\tau_N)$.

5.3. *Approximate Bayes factor*. Arguing as in Section 4.1 and noting that when τ is taken as fixed the Bayes factor is simply $p_\psi(0|y_1, \ldots, y_n)/p_\psi(0)$, we get the following crude approximation to the log Bayes factor:

$$\log B \approx \sum_{j=1}^{m} \log \frac{\tau_j}{b_j} - \frac{1}{2} \sum_{j=1}^{m} \frac{\hat{\psi}_j^2}{b_j^2}, \tag{8}$$

where b_j and $\hat{\psi}_j$ are defined as in Section 4.1.

6. Density estimation. Although density estimation is not our prime goal, it is a virtue of the current method that, in addition to a goodness-of-fit test, we also get a semiparametric density estimator. The usual Bayesian density estimate is the predictive distribution given by

$$\hat{h}(s) = \int h(s|\theta, \psi)\, p(\theta, \psi|y_1, \ldots, y_n)\, d\theta\, d\psi, \tag{9}$$

where $h(s|\theta, \psi)$ was defined in (3). As explained in Section 8, for a large class of distributions P, $d(\hat{h}, p)$ tends to 0 almost surely under i.i.d. sampling from P where $d(f, g) = \int |f(s) - g(s)|\, ds$. Now (9) is easily estimated from the Monte Carlo by

$$\hat{h}(s) \approx \frac{1}{N} \sum_{j=1}^{N} h\big(s|\theta_j, \psi_j\big).$$

However, the evaluation of this estimate involves computing $h(s|\theta_j, \psi_j)$ over a grid of values of s at each iteration. This can be very time consuming. A cruder estimate is simply $\hat{h}(s) \approx h(s|\hat{\theta}, \hat{\psi})$ for some point estimates $\hat{\theta}$ and $\hat{\psi}$ such as the posterior mean. In practice, this is usually accurate enough and the examples reported in this paper were computed this way.

7. Examples and simulation study. We now consider some examples and simulations. In particular, the simulations in Section 7.3 provide guidance for choosing the hyperparameter w.

7.1. *Kevlar pressure vessels*. For comparison with Evans and Swartz (1994), we consider the logarithms of 100 stress-rupture lifetimes of Kevlar pressure vessels [Andrews and Herzberg (1985), page 183]. Using $w = 1$ and $m = 10$, we obtained a Bayes factor of 0.10, that is, 10 to 1 odds against the normal. The first plot in Figure 1 shows a kernel density estimate (solid line) along the normal estimate (dotted line) of the density. The bandwidth for the kernel was chosen using Silverman's rule of thumb [Silverman (1986), page 45]. The second plot shows the density estimate using our method (solid line) together with the normal estimate (dotted line) and the Brunk estimate

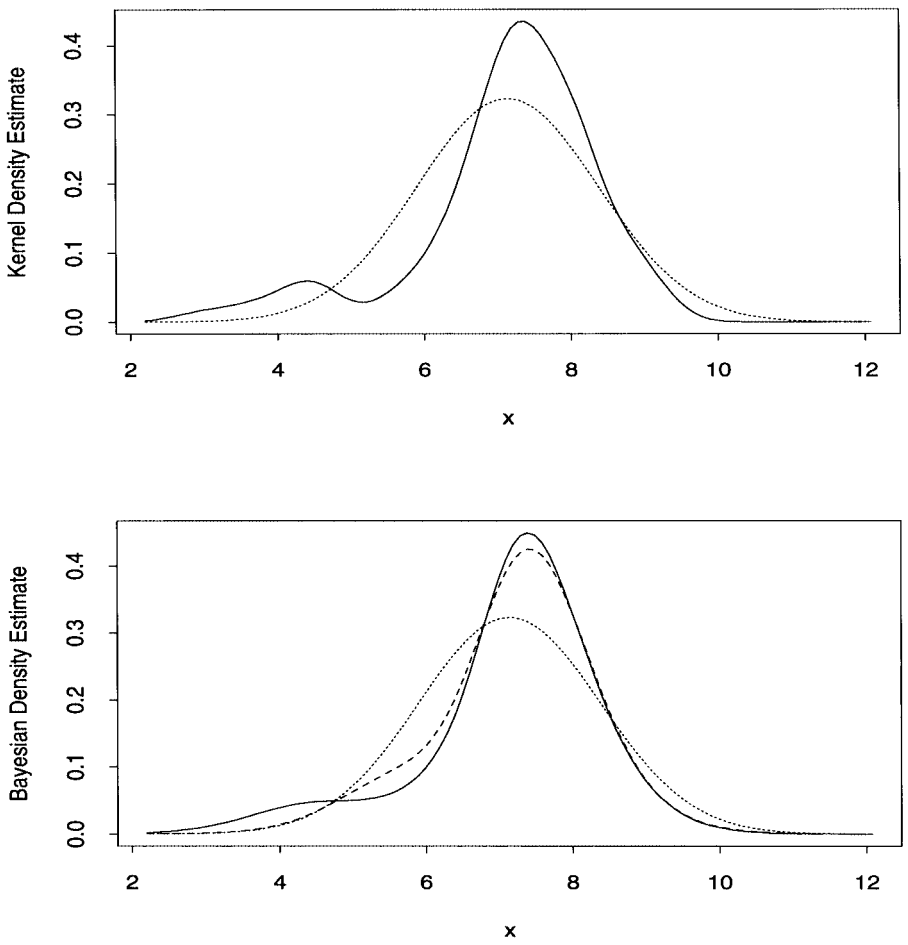

FIG. 1. *Density estimates for Kevlar pressure vessels data* (*Section* 7.1). *First plot is kernel density estimate* (*solid line*) *and normal estimate* (*dotted line*). *Second plot is exponential family estimate* (*solid line*), *Brunk's method* (*dashed line*) *and normal* (*dotted line*).

(dashed line). Our method seems to smooth a little more in the left tail than the kernel. The Brunk estimate smooths the tail even more. The estimate of Evans and Swartz is similar but the bump in the left tail is more pronounced than in the kernel estimate. All the methods suggest that the normal is a poor fit.

7.2. *Marron–Wand examples*. Marron and Wand (1992) provided a set of 15 test cases for density estimation. Each is a finite mixture of normal densities. The first eight are reasonably smooth. The ninth is a trimodal mixture with two large modes and one small mode in the middle. The remaining six are rather bizarre densities which are of some theoretical

interest but are not the typical deviations from normality that we are hoping to detect. We thus focus on the nine fairly nice cases and we include one bizarre case just to see how our estimate behaves. Marron and Wand call this last case the "claw density," though Michael Escobar (personal communication) has suggested that the name "Bart Simpson density" might be more appropriate. We do not expect the method to work for the Bart Simpson density though we include this for completeness.

Figure 2 shows these 10 densities. The first column in Figure 2 is the true density, the second is our estimate based on 100 observations from the

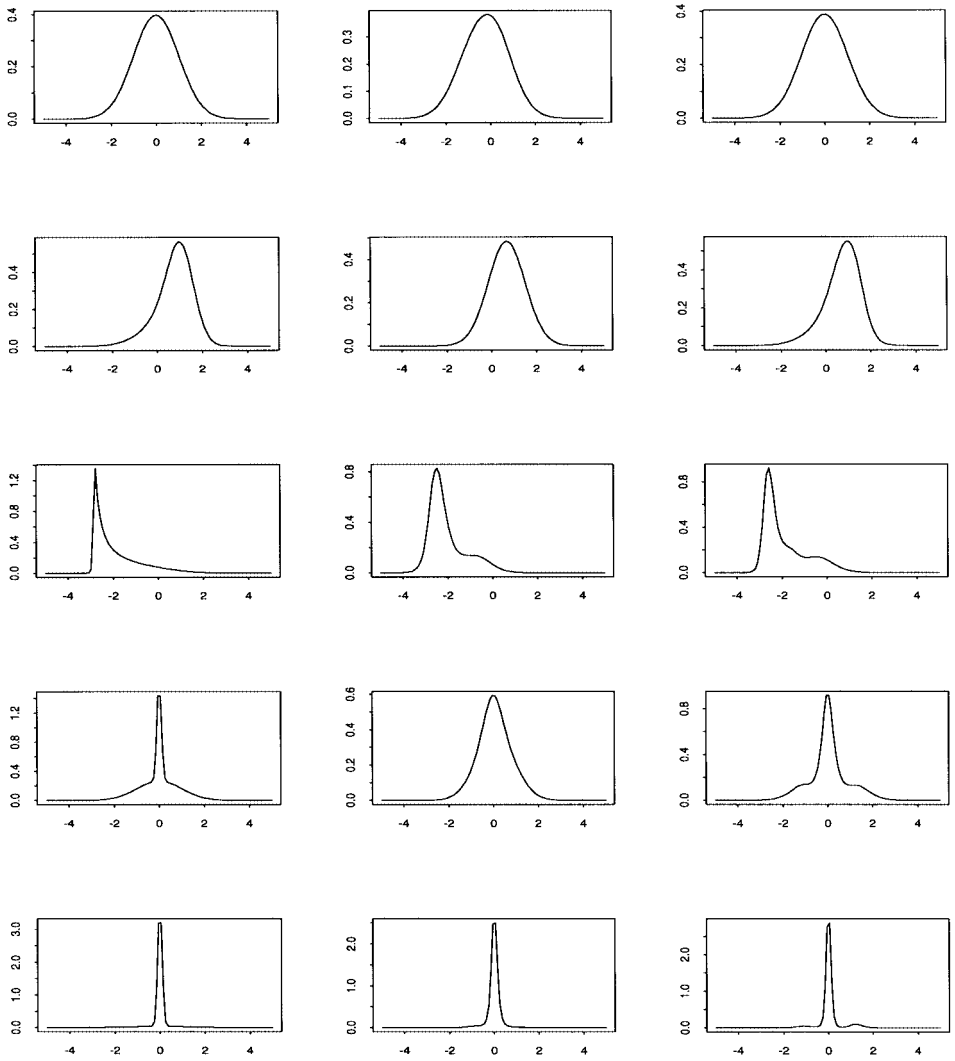

FIG. 2a. *Marron–Wand examples, densities* 1–5. *Column* 1 *is true density. Column* 2 *is estimate based on* $n = 100$. *Column* 3 *is estimate based on* $n = 1000$.

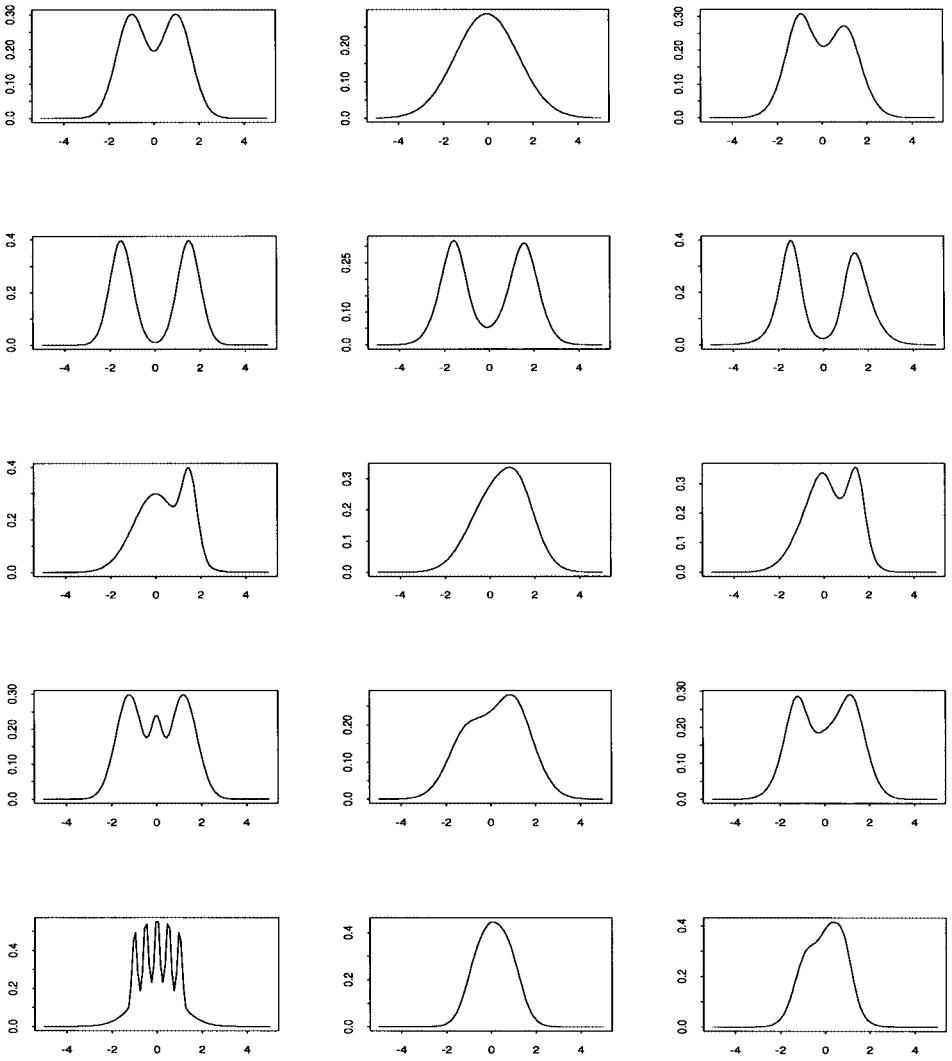

FIG. 2b. *Marron–Wand examples, densities* 6–10. *Column* 1 *is true density. Column* 2 *is estimate based on* $n = 100$. *Column* 3 *is estimate based on* $n = 1000$.

density and the third is our estimate based on 1000 observations. In all cases we took $m = 10$ and $w = 1$. The examples in Figure 2 are typical. The well-behaved densities are reasonably well estimated, especially for large sample sizes. The small node in the trimodal density is not picked up without a large sample size. The method fails to estimate the Bart Simpson density; to successfully handle this case it may be necessary to tweak the hyperparameter w and increase m. Figure 3 shows the behavior of the Brunk estimates. These do not do as well, but are much easier to compute.

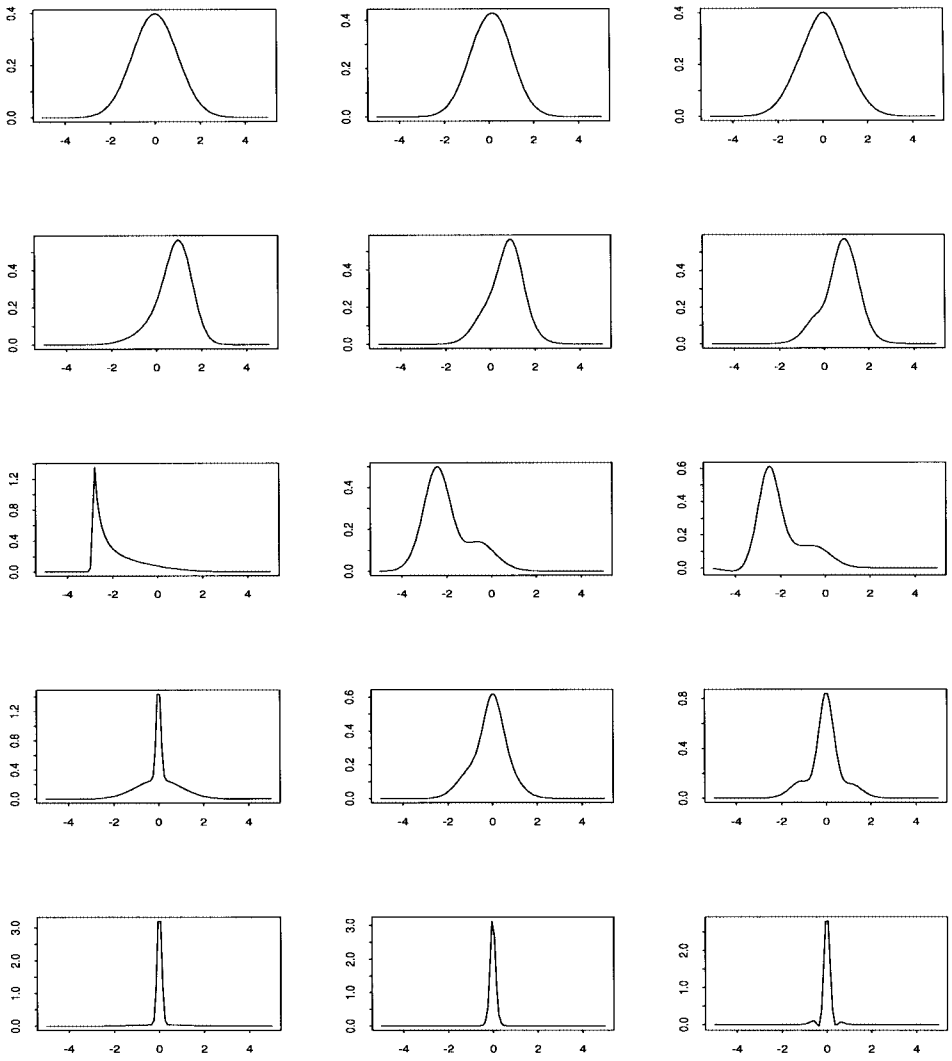

FIG. 3a. *Marron–Wand examples using Brunk method. Densities* 1–5. *Column* 1 *is true density. Column* 2 *is estimate based on* $n = 100$. *Column* 3 *is estimate based on* $n = 1000$.

7.3. *Simulation study*. To investigate the goodness-of-fit test we conducted a small simulation. Further, as discussed at the end of the Introduction, the simulation will be used to provide guidance in selecting a reasonable value of the hyperparameter w.

The simulation involves 30 conditions: three choices for w $(1/5, 1, 5)$, two choices for n $(25, 100)$, and five densities: densities $1, 2, 3, 5, 7$ from the Marron–Wand examples. In each of the 30 conditions, we carried out 100 replications using $m = 10$. Because the computing is so intensive, we calcu-

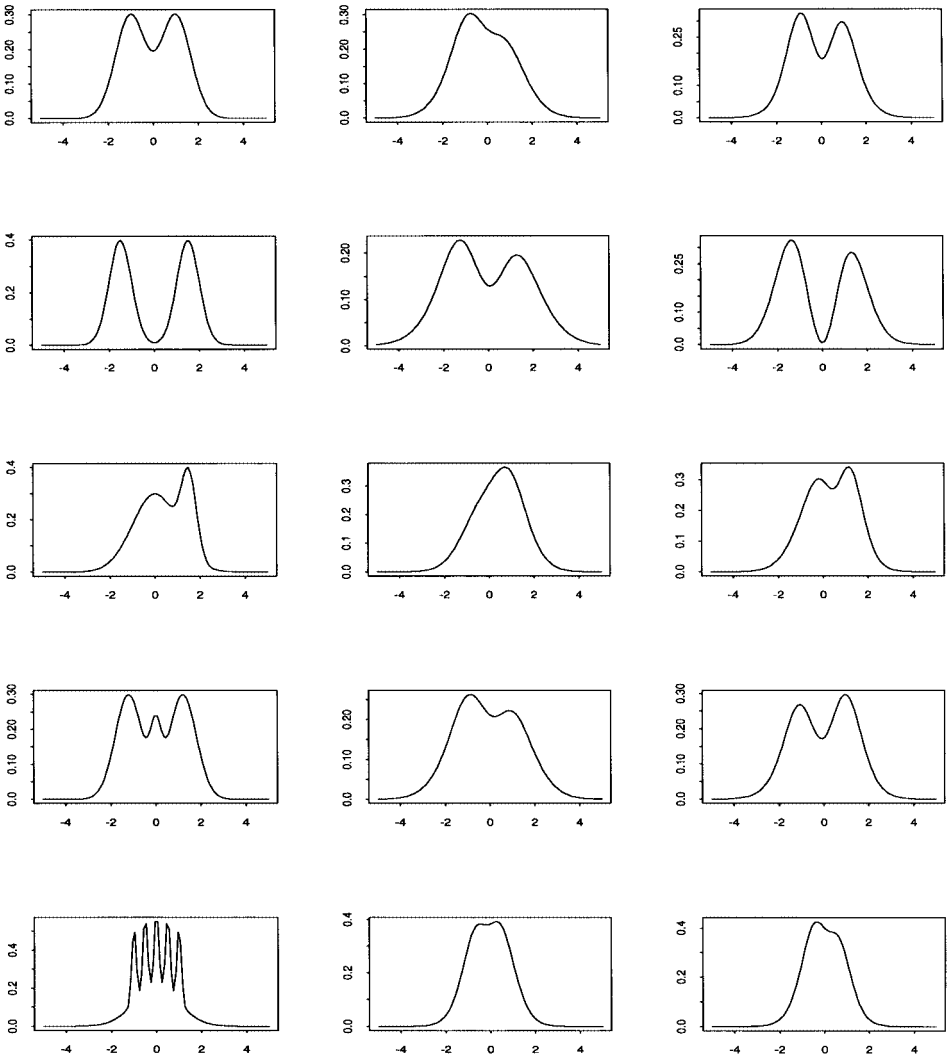

FIG. 3b. *Marron–Wand examples using Brunk method. Densities* 6–10. *Column* 1 *is true density. Column* 2 *is estimate based on* $n = 100$. *Column* 3 *is estimate based on* $n = 1000$.

lated the Legendre polynomials on a grid of size 10. This is quite coarse but limited experimentation suggests that increasing the grid size does not have a dramatic effect. The Markov chain was run for 1000 iterations in each case. Table 1 shows $\Pr(B < 1)$. For density 1, this corresponds to type 1 error; for the others it represents power.

Density 2 is very similar to a normal so we expect poor power; this is verified in the simulation. In all other cases, the power is quite good for $n = 100$; much lower powers are obtained for $n = 25$. This is not too surpris-

TABLE 1
Simulation study for the Marron–Wand examples: values of Prob{*Bayes factor* < 1}, *for* 100 *replications*

Density Type	**1**		**2**		**3**		**5**		**7**	
Sample Size	**25**	**100**	**25**	**100**	**25**	**00**	**25**	**100**	**25**	**100**
Values of *w*										
1/5	0/29	0/38	0/29	0/36	0/27	0/88	0/40	0/97	0/36	0/91
1	0.16	0.16	0.23	0.33	0.17	0.95	0.50	0.96	0.36	0.95
5	0.01	0.00	0.03	0.03	0.07	0.78	0.32	0.96	0.25	0.93

ing given that we are using test densities that are mixtures of normals for which the method was not tailored. The values of the type I error suggest choosing w between 1 and 5. A conservative test which rejects too often might be preferable since the method automatically provides a suitable alternative model. From this point of view, a value of $w = 1$ seems reasonable.

We carried out a similar simulation for the Bayes factor from Brunk's method (Section 5.3). The computations are much less demanding so we used 1000 replications. The results are in Table 2.

This test has excellent type I error. Its power is worse for density 2 but otherwise seems to have better power than we saw in Table 1. Again, $w = 1$ appears to be a reasonable value. We found the good performance of this approximation rather surprising. It is orders of magnitude simpler to compute than the previous Bayes factor. The main advantage of the previous method is that it produces a full posterior distribution for all the parameters. In cases where a simple goodness-of-fit test is all that is needed, the Brunk approximation to the Bayes factor may suffice.

8. Consistency. In this section we discuss the consistency of the method. We are interested in two types of consistency: consistency of the Bayes factor

TABLE 2
Simulation study for Brunk's method: values of Prob{*Bayes factor* < 1}, *for* 1000 *replications*

Density Type	**1**		**2**		**3**		**5**		**7**	
Sample Size	**25**	**100**	**25**	**100**	**25**	**100**	**25**	**100**	**25**	**100**
Values of *w*										
1/5	0.001	0.001	0.013	0.186	0.404	1.000	0.724	0.997	0.077	1.000
1	0.003	0.006	0.027	0.240	0.531	1.000	0.717	0.995	0.381	1.000
5	0.003	0.000	0.008	0.067	0.507	1.000	0.630	0.987	0.801	1.000

and consistency of the density estimate. For the former, we shall approximate the infinite-dimensional problem with a sequence of finite-dimensional exponential families as in Barron and Sheu (1991) and Portnoy (1988).

8.1. *Consistency of the Bayes factor*. Recall that $Y_1, \ldots, Y_n$ are independent observations from a distribution with density $h(y|\theta, \psi) = f(y|\theta)g(F(y|\theta)|\psi)$ where $\log g(u|\psi) = \sum_{j=1}^{\infty} \psi_j \phi_j(u) - c(\psi)$. Let π be the prior for the ψ_j's. Recall from Section 3 that π makes the ψ_j's independent with $\psi_j \sim N(0, \tau^2/c_j^2)$. In Section 3 we also placed a prior on τ. Here, we prove consistency for a simplified version of the model in which τ and θ are fixed. In this case, there is no loss of generality in transforming the problem to the unit interval. Thus we assume $Y_1, Y_2, \ldots$ are observations from a distribution on the unit interval.

Let $D(p\|q) \equiv D(P\|Q) = \int p \log p/q$, let $N_\varepsilon(p) = \{q; D(p\|q) \le \varepsilon\}$ and let $D_n(p, q) = (1/n)\sum_{i=1}^{n} \log p(Y_i)/q(Y_i)$. Let p_0 be the uniform density and let P_0 be the corresponding measure. We are interested in testing:

H_0: $Y_1, \ldots, Y_n$ are independent observations from a $U(0, 1)$ distribution versus

H_1: $Y_1, \ldots, Y_n$ are independent observations from a distribution with density $g(u|\psi) = \exp\{\sum_{j=1}^{\infty} \psi_j \phi_j(u) - c(\psi)\}$, for some ψ such that $\sum_{j=1}^{\infty} |\psi_j| < \infty$, with $0 < D(p_0\|g) < \infty$.

The parameter space is $\Omega = \{\psi = (\psi_1, \psi_1, \ldots); \sum_j |\psi_j| < \infty\}$. Equivalently, we can regard the parameter space as $\Omega = \{q(\cdot) = g(\cdot|\psi); \sum_j |\psi_j| < \infty\}$. We will use both q and ψ to denote the parameter.

Assuming $0 < \Pr(H_0) < 1$, the Bayes factor B_n is defined by

$$B_n = \frac{\Pr(H_0|Y_1, \ldots, Y_n)}{\Pr(H_1|Y_1, \ldots, Y_n)} \div \frac{\Pr(H_0)}{\Pr(H_1)} = \left\{ \int_\Omega \prod_{i=1}^{n} q(Y_i)\, \pi(dq) \right\}^{-1}. \tag{10}$$

Let P be the true sampling distribution and let P^∞ be the corresponding infinite product measure.

THEOREM 8.1. *Let $c_j = j^k$ where $k > 8$.*

(i) *If H_1 is true then $B_n \to 0$ exponentially quickly, almost surely with respect to P^∞.*

(ii) *If H_0 is true, then $B_n^{-1} \to 0$ in probability.*

Case (i) is the easier one to prove. For case (ii), we will approximate B_n^{-1} with a sequence of finite-dimensional integrals which tend to 0 in probability; related calculations can be found in Barron and Sheu (1991), Portnoy (1988) and Shun and McCullagh (1994).

Before proving Theorem 8.1, we first need the following two lemmas from Barron (1988). Since that paper is unpublished, we provide a proof for one of the lemmas.

LEMMA 8.1 Barron (1988). *If* $D(p_0\|f) < \infty$ *and* π *is as described above with* $\tau > 0$, *then, for every* $\varepsilon > 0$, *we have* $\pi(N_\varepsilon(f)) > 0$.

LEMMA 8.2 Barron (1988). *Suppose that* $\pi(N_\varepsilon(p)) > 0$ *for every* $\varepsilon > 0$. *Then, for every* $\delta > 0$, *we have* $\int_\Omega \exp(-nD_n(p,q))\pi(dq) > \exp(-n\delta)$ *for large* n *with probability* 1.

PROOF. Let $A = N_{\delta/2}(p)$ and note that

$$\begin{aligned} &\exp(n\delta)\int_\Omega \exp(-nD_n(p,q))\pi(dq) \\ &\quad \geq \exp(n\delta)\int_A \exp(-nD_n(p,q))\pi(dq). \end{aligned} \tag{11}$$

For each q, $D_n(p,q) \to D(p\|q)$ almost surely $[P^\infty]$ by the strong law of large numbers. By Fubini's theorem, there exists a set of sequences of probability 1 such that, for each $x = (x_1, x_2, \ldots,)$ in the set we have $D_n(p,q) \to D(p\|q)$ for π-almost all q. By Fatou's lemma,

$$\begin{aligned} &\liminf \exp(n\delta)\int_A \exp(-nD_n(p,q))\pi(dq) \\ &\quad = \liminf \int_A \exp\bigl(n[\delta - D_n(p,q)]\bigr)\pi(dq) \\ &\quad \geq \int_A \liminf \exp\bigl(n[\delta - D_n(p,q)]\bigr)\pi(dq). \end{aligned} \tag{12}$$

Since $D_n(p,q) \to D(p\|q) < \delta/2$, for π-almost all q in A the integrand tends to ∞. Furthermore, $\pi(A) > 0$ by assumption. Thus,

$$\liminf \exp(n\delta)\int_\Omega \exp(-nD_n(p,q))\pi(dq) = \infty$$

almost surely. In particular, $\exp(n\delta)\int_\Omega \exp(-nD_n(p,q))\pi(dq) \geq 1$ for large n with probability 1. Hence $\int_\Omega \exp(-nD_n(p,q))\pi(dq) \geq \exp(-n\delta)$ for large n with probability 1. □

PROOF OF CASE (i) OF THEOREM 8.1. We claim that there exists $r > 0$ such that, for all large n, $B_n < e^{-nr}$ with P^∞ probability 1. Write B_n as

$$B_n = \frac{\exp(-nD_n(p,p_0))}{\int_\Omega \exp(-nD_n(p,q))\pi(dq)}.$$

By the law of large numbers, $D_n(p,p_0)$ tends to $D(p\|p_0)$ almost surely. By assumption, $D(p\|p_0) = c > 0$. Thus, for large n, with probability 1, $D_n(p,p_0) > c/2$. Let $\delta = c/4$. By Lemma 8.2, for large n with probability 1, the denominator of B_n is larger than $e^{-nc/4}$. Thus, for large n with probability 1, $B_n < e^{-nc/4}$. □

REMARK. In the more general model used throughout the paper, with θ not fixed, a similar proof holds. In that case, B_n has an integral in the numerator which is shown to be exponentially small by appropriately covering the parameter space with finitely many sets as in the usual Wald approach.

We now turn to case (ii) of Theorem 8.1. The idea is to approximate the infinite-dimensional exponential family with a sequence of finite-dimensional families. Let $m \equiv m(n)$ and let $A_n = \{\psi = (\psi_1, \psi_2, \ldots);\ \sum_{j=m}^{\infty} \sqrt{2j+1}|\psi_j| < 1/n\}$. We will need the following two lemmas.

LEMMA 8.3. *Let $c_j = j^k$ where $k = 8 + \varepsilon$ for some $\varepsilon > 0$. Let $m = n^{1/7-a}$ where $0 < a < \varepsilon/[7(7+\varepsilon)]$. Then $\pi(A_n^c) = o(1)$ as $n \to \infty$.*

PROOF. Let $R_n = \sum_{j=m}^{\infty} \sqrt{2j+1}|\psi_j|$. Let $z_n = \pi(R_n > 1/n)$. By Chernoff's inequality,

$$z_n \le \inf_{t \ge 0} \exp(-t) E \exp(nt) \sum_{j=m}^{\infty} \sqrt{2j+1}|\psi^j|$$

$$= b \inf_{t \ge 0} \exp(-t) \exp\bigl(n^2 t^2 \tau^2 U_n/2\bigr)$$

where $U_n = \sum_{j=m}^{\infty} (2j+1)/c_j^2$ and b is a positive constant. The inf occurs at $t = 1/(n^2\tau^2 U_n)$. Thus, $z_n \le \exp(-1/(2n^2\tau^2 U_n))$. Now,

$$U_n = 2 \sum_{j=m}^{\infty} \frac{1}{j^{2k-1}} + \sum_{j=m}^{\infty} \frac{1}{j^{2k}}$$

$$= -\frac{2}{(2k-2)!}\Psi(2k-2, m) + \frac{1}{(2k-1)!}\Psi(2k-1, m),$$

where $\Psi(r, y)$ is the rth derivative of the psi function evaluated at y. In other words, $\Psi(r, y) = d^r \Psi(x)/dx^r|_{x=y}$ where $\Psi(x) = d \log \Gamma(x)/dx$. Since $\Psi(r, y) = (-1)^{r-1}[(r-1)!/y^r + o(y^{-r})]$ for $y \to \infty$, then $U_n = O(m^{-(2k-2)})$ and $n^2 U_n = O(1/n^{(1/7-a)(2k-2)-2})$. Thus, since $k > 8$, $n^2 U_n = o(1)$ and $z_n = o(1)$. □

LEMMA 8.4. *Let $\psi = (\psi_1, \ldots, \psi_m)$ and let $\mathrm{Cov}_\psi(\phi_i, \phi_j)$ denote the covariance between functions ϕ_i and ϕ_j under the distribution with density $g(u|\psi) = \exp\{\sum_{j=1}^m \psi_j \phi_j(u)\}/\int_0^1 \exp\{\sum_{j=1}^m \psi_j \phi_j(t)$. Also, let $\|\cdot\|$ denote the Euclidean norm. Then,*

$$\mathrm{Cov}_\psi(\phi_i, \phi_j) = O\bigl(m^2 \|\psi\|\bigr).$$

For the proof, see the Appendix.

Lemma 8.4 implies that the covariance matrix $\mathrm{Cov}_{\hat\psi}(\phi_i, \phi_j)$ can be written as $I + A$, where A is a matrix whose elements a_{ij} are $O(\sqrt{m^5/n})$ where $\hat\psi$ is the maximum likelihood estimate of ψ in the family defined in Lemma 8.4.

PROOF OF CASE (ii) OF THEOREM 8.1. Let m and A_n be as in Lemma 8.3. Denote the n-fold product of the sample space by $\mathcal{Y}^n$, the n-fold product measure of P by P^n and let $Y^n = (Y_1, \ldots, Y_n)$. We write $p(y^n)$ to mean $\prod_{i=1}^n p(y_i)$. Then

$$B_n^{-1} = \int_{A_n} \exp(-nD_n(p_0, q))\pi(dq) + \int_{A_n^c} \exp(-nD_n(p_0, q))\pi(dq)$$
$$= I_1 + I_2, \quad \text{say.}$$

First we show that $I_2 = o_P(1)$. Fix $c > 0$, and apply Markov's inequality and then Fubini's theorem to get

$$\Pr\left(\int_{A_n^c} \exp(-nD_N(p_0, q))\pi(dq) > c\right)$$
$$\leq c^{-1} \int_{\mathcal{Y}^n} \int_{A_n^c} \exp(-nD_n(p_0, q))\pi(dq)\, dP_0^n$$
$$= c^{-1} \int_{A_n^c} \int_{\mathcal{Y}^n} q(y^n)/p_0(y^n)\, dP_0^n(y^n)\pi(dq)$$
$$= c^{-1}\pi(A_n^c).$$

The latter quantity goes to zero by Lemma 8.3. Thus $I_2 = o_P(1)$.

Next we bound $q(Y^n)$, where $q(Y^n) = \prod_{i=1}^n g(Y_i|\psi)$, $\psi \in A_n$. From the definition of A_n and the fact that $\sup_u |\phi_j(u)| \leq \sqrt{2j+1}$, we see that $-1/n < \sum_{j=m}^\infty \psi_j \phi_j(u) < 1/n$. let $\psi^m = (\psi_1, \ldots, \psi_m)$ and define $g_m(u|\psi^m) = \exp\{\sum_{j=1}^m \psi_j \phi_j(u)\}/\int_0^1 \exp\{\sum_{j=1}^m \psi_j \phi_j(t)\}\, dt$. Then,

$$q(Y^n) = \frac{\exp\left(n\sum_{j=1}^\infty \psi_j \overline{\phi}_j\right)}{\left[\int_0^1 \exp\left(\sum_{j=1}^\infty \psi_j \phi_j(u)\right) du\right]^n}$$
$$\leq \exp(2n(1/n)) \frac{\exp\left(n\sum_{j=1}^m \psi_j \overline{\phi}_j\right)}{\left[\int_0^1 \exp\left(\sum_{j=1}^m \psi_j \phi_j(u)\right) du\right]^n}$$
$$= \exp(2) g_m(Y^n|\psi^m),$$

where $\overline{\phi}_j = n^{-1}\sum_{i=1}^n \phi_j(Y_i)$.

Let π_m be the measure on the Borel sets of $\mathbb{R}^m$ that makes ψ_j independent normals with mean 0 and variance τ_j^2, $j = 1, \ldots, m$. Let $f_m = d\pi_m/d\mu$ where μ is Lebesgue measure on $\mathbb{R}^m$. Then

$$\int_{A_n} q(Y^n)\pi(dq) \leq e^2 \int_{A_n} g_m(Y^n|\psi^m)\pi(dq)$$
$$= e^2 \int_{\mathbb{R}^m} g_m(Y^n|\psi^m) f_m(\psi^m)\mu(d\psi^m). \tag{13}$$

Our strategy is to bound this last integral using Laplace's method. A complication is that the dimension m is increasing along with n.

For notational convenience, we drop the superscript and subscript m. Also, let $\hat{\psi}$ denote the maximum likelihood estimator for the m-dimensional exponential family. Since $m = o(n^{1/7})$, from Lemma 5 and Theorem 2 of Barron and Sheu (1991) we conclude that $\|\hat{\psi}\| = O_P(\sqrt{m/n})$. Let $l_n = n[\sum_{j=1}^m \psi_j \overline{\phi}_j(Y_i) - c(\psi)]$ be the log likelihood corresponding to the $m = m(n)$-dimensional exponential family. Let $l_n''(\psi) = -nc''(\psi)$ be the matrix of second derivatives with respect to ψ. Note that $c''(\psi) = \mathrm{Cov}(\psi)$, the covariance matrix of the ϕ_j's for fixed ψ. Let $\Sigma_n = [-l_n''(\hat{\psi})]^{-1} = n^{-1}\widehat{\mathrm{Cov}}^{-1}$. From Lemma 8.4, $\widehat{\mathrm{Cov}} = I_n + A$ where I_n is the identity matrix and A has entries which are $O_P(\sqrt{m^5/n})$.

Now we approximate (13) with a normal integral. Let δ_n be a sequence of numbers such that (a) $\delta_n = o(1/m^2)$ and (b) $\sqrt{m/n}\,/\delta_n \to 0$. Such sequences exist since $m = o(n^{1/7})$. Let $N_n = \{\psi; \|\psi\| \le \delta_n\}$. Then

$$\begin{aligned}\int_{\mathbb{R}^m} & g_m(Y^n|\psi) f(\psi)\,\mu(d\psi) \\ &= \int_{N_n} g_m(Y^n|\psi) f(\psi)\,\mu(d\psi) + \int_{N_n^c} g_m(Y^n|\psi) f(\psi)\,\mu(d\psi) \\ &= J_1 + J_2.\end{aligned}$$

First we show that $J_2 = o_P(1)$.

Since $J_2 \le \exp[\sup_{\psi \in N_n^c} l_n(\psi)]$ it suffices to show that $\sup_{\psi \in N_n^c} l_n(\psi)$ tends to $-\infty$ in probability. Recall that $l_n(\psi) = n[\psi^T \overline{\phi} - c(\psi)]$ is strictly concave and is maximized at $\hat{\psi}$. Since $\|\hat{\psi}\| = O_P(\sqrt{m/n})$ it follows that $\|\hat{\psi}\| = O_P(\delta_n)$ and hence $\hat{\psi} \in N_n$ with probability tending to 1. And when $\hat{\psi} \in N_n$, $l_n(\psi)$ is maximized over N_n^c at some point ψ_n satisfying $\|\psi_n\| = \delta_n$. Expanding the log likelihood around 0 we have

$$l_n(\psi_n) = n\left[\psi_n^T \overline{\phi} - (1/2)\psi_n^T\, \mathrm{Cov}(\psi^+)\psi_n\right], \tag{14}$$

where ψ^+ lies on the line segment joining 0 and ψ_n. Since $\psi^+ \in N_n$, it follows from Lemma 8.4 that $\mathrm{Cov}(\psi^+) = I + B$ where the elements of B are of order $O_P(\delta_n m^2)$. Given a matrix C, Let $\underline{\lambda}(C)$ and $\overline{\lambda}(C)$ denote the smallest and largest eigenvalues of C. Note that $\underline{\lambda}(B) = O_P(m^2\delta_n)$. From (14), the Cauchy–Schwarz inequality and the fact that $\|\overline{\phi}\| = O_P(\sqrt{m/n})$,

$$\begin{aligned} l_n(\psi_n) &\le n\left\{\|\psi_n\|\,\|\overline{\phi}\| - (1/2)\|\psi_n\|^2 \underline{\lambda}[\mathrm{Cov}(\psi^+)]\right\} \\ &= n\left\{\delta_n a_n - (1/2)\|\psi_n\|^2\left[1 + O\left(m^2\delta_n\right)\right]\right\},\end{aligned}$$

where $a_n \ge 0$ and $a_n = O_P(\sqrt{m/n})$. Since $m = o(n^{1/7})$ and $\delta_n = o(1/m^2)$, the last expression goes to $-\infty$ in probability. Thus, $J_2 = o_P(1)$ as claimed.

Next we show that, with probability tending to 1, for each $\varepsilon > 0$,

$$\sup_{\psi \in N_n,\, \|\gamma\| = 1} \gamma^T \widehat{\mathrm{Cov}}^{-1/2}\, \mathrm{Cov}(\psi) \widehat{\mathrm{Cov}}^{-1/2} \gamma \ge 1 - \varepsilon. \tag{15}$$

Recall that $\widehat{\mathrm{Cov}} = I + A$ where the elements of A are $O_P(\sqrt{m^5/n})$ and that for any $\psi \in N_n$, $\mathrm{Cov}(\psi) = I + B$ where the elements of B are $O(\delta_n m^2)$. Now,

for any $\psi \in N_n$,

$$\begin{aligned}
\gamma^T \widehat{\mathrm{Cov}}^{-1/2} \mathrm{Cov}(\psi) \widehat{\mathrm{Cov}}^{-1/2} \gamma &\geq \underline{\lambda}\Big(\widehat{\mathrm{Cov}}^{1/2} \mathrm{Cov}(\psi) \widehat{\mathrm{Cov}}^{-1/2}\Big) \\
&= \underline{\lambda}\Big[(I+A)^{-1/2}(I+B)(I+A)^{-1/2}\Big] \\
&\geq \underline{\lambda}(I+A)^{-1/2} \underline{\lambda}(I+B)\, \underline{\lambda}(I+A)^{-1/2} \\
&= \overline{\lambda}(I+A)\, \underline{\lambda}(I+B) \\
&= \big[1 + \overline{\lambda}(A)\big]\big[1 + \underline{\lambda}(B)\big] = 1 + o_P(1),
\end{aligned}$$

since $\overline{\lambda}(A) = O_P(\sqrt{m^7/n})$ and $\underline{\lambda}(B) = O(m^2\delta_n)$. The second inequality follows from Anderson and DasGupta (1963), Theorem 2.2. Hence, (15) holds except on a set of probability tending to 0. It follows from the definition of Σ_n that $-\gamma^T Q(\psi)\gamma \leq -(1-\varepsilon)$ for all $\psi \in N_n$, with probability tending to 1, where $Q(\psi) = -\Sigma_n^{1/2} l_n''(\psi) \Sigma_n^{1/2}$.

Now $l(\psi) = l(\hat{\psi}) + \frac{1}{2}(\psi - \hat{\psi})^T l''(\tilde{\psi})(\psi - \hat{\psi})$ where $\tilde{\psi}$ is between ψ and $\hat{\psi}$. Hence $L(\psi) = L(\hat{\psi})\exp((1/2)(\psi - \hat{\psi})^T l''(\tilde{\psi})(\psi - \hat{\psi}))$. Let

$$Q = -\Sigma_n^{1/2} l''\big(\tilde{\psi}\big)\Sigma_n^{1/2}, \quad b^T = \big(\psi - \hat{\psi}\big)^T \Sigma_n^{-1/2} \quad \text{and} \quad \gamma = b/\|b\|.$$

Then

$$\begin{aligned}
J_1 &= \int_{N_n} L(\psi) f(\psi)\, d\psi = L\big(\hat{\psi}\big) \int_{N_n} \exp\Big\{-\tfrac{1}{2}\|b\|^2 \gamma^T Q \gamma\Big\} f(\psi)\, d\psi \\
&\leq L\big(\hat{\psi}\big) \int_{N_n} \exp\Big\{-\tfrac{1}{2}(1-\varepsilon)\|b\|^2\Big\} f(\psi)\, d\psi \\
&\leq L\big(\hat{\psi}\big) \int_{\mathbb{R}^m} \exp\Big\{-\tfrac{1}{2}(1-\varepsilon)\|b\|^2\Big\} f(\psi)\, d\psi \qquad (16) \\
&= L\big(\hat{\psi}\big)|T|^{-1/2} \big|(1-\varepsilon)\Sigma_n^{-1} + T^{-1}\big|^{-1/2} \\
&\quad \times \exp\Big\{-\tfrac{1}{2}\hat{\psi}^T \Big(T + (1-\varepsilon)^{-1}\Sigma_n\Big)^{-1} \hat{\psi}\Big\} \\
&\leq L\big(\hat{\psi}\big)|T|^{-1/2} \big|(1-\varepsilon)\Sigma_n^{-1} + T^{-1}\big|^{-1/2},
\end{aligned}$$

where T is a diagonal matrix with jth diagonal element equal to $\tau_j^2 = \tau^2/j^{2k}$.

Now we examine $l(\hat{\psi}) = n[\hat{\psi}^T \overline{\phi} - c(\hat{\psi})]$. First, expand $c(\psi)$ around 0, noting that $c(0) = 0$ and that $c'(\hat{\psi}) = E_{\hat{\psi}}(\phi) = \overline{\phi}$. Thus, for some ψ^+ between 0 and $\hat{\psi}$,

$$\begin{aligned}
c\big(\hat{\psi}\big) &= \hat{\psi}^T \overline{\phi} + \tfrac{1}{2}\hat{\psi}^T \mathrm{Cov}(\psi^+)\hat{\psi} \\
&= \hat{\psi}^T \overline{\phi} + \tfrac{1}{2}\hat{\psi}^T (I+A)\hat{\psi} \\
&\geq \hat{\psi}^T \overline{\phi} + \tfrac{1}{2}\|\hat{\psi}\|^2 \underline{\lambda}(I+A) \\
&= \hat{\psi}^T \overline{\phi} + \tfrac{1}{2}\|\hat{\psi}\|^2 \Big(1 + O_P\Big(\sqrt{m^5/n}\Big)\Big).
\end{aligned}$$

We conclude that $l(\hat{\psi}) \leq -(n/2)a_n[1 + b_n]$ where $a_n \geq 0$, $a_n = O_P(m/n)$ and $b_n = O_P(\sqrt{m^5/n})$. Hence, $L(\hat{\psi})$ is bounded in probability. Finally, we use the next lemma, which is proved in the Appendix.

LEMMA 8.5. *As* $n \to \infty$,

$$|(1 - \varepsilon)T\Sigma_n^{-1} + I|^{-1/2} = o_P(1).$$

Thus, $J_1 = o_P(1)$ and the proof is complete. □

COROLLARY 8.1. *The conclusions of Theorem* 8.1 *continue to hold if* $c_j = k^j$ *with* $k > 1$.

The proof is the same as Theorem 8.1 except that one takes $m = c \log n$ with $c = (2/\log k)$.

8.2. *Consistency of the density estimate*. Although density estimation is not the main goal of this paper, we take this opportunity to point out that the Bayes estimate from our model is consistent under mild conditions. To show this, we appeal to the results in Barron (1988). [An alternative proof can be constructed along the lines of Shen (1995).] Barron gives two conditions which are sufficient to guarantee that $\int |\hat{f}_n(x) - f(x)|\, \mu(dx) \to 0$ almost surely, where $\hat{f}_n$ is the predictive density and f is the true density. Our model is essentially his Example 2. The consistency follows from Barron's calculations; we state relevant results without proof.

Let T_n be a sequence of partitions defined by $T_n = \{[0, 1/n), [1/n, 2]n), \ldots, [(n-1)/n, 1]\}$. For every density q define

$$\tilde{q}(x) = \frac{\int_A q(s)\, d\mu(s)}{\mu(A)} \quad \text{for } x \in A \in T_n \tag{17}$$

and let $\tilde{Q}(dx) = \tilde{q}(x)\mu(dx)$. Let $d(p, q) = \int |p - q|$ and let $C_\varepsilon = \{q; d(p, q) \leq \varepsilon\}$, where p represents the true density. For every $\delta > 0$ define $B_n(\delta) = \{q; d(q, \tilde{q}) \leq \delta\}$. The following is a specialization of a more general theorem in Barron (1988).

THEOREM 8.2 Barron (1998). *Let* $X_1, X_2, \ldots$ *be i.i.d.* P *and suppose that*:

(a) *For every* $\varepsilon > 0$, $\pi(N_\varepsilon(p)) > 0$;
(b) *For every* $\delta > 0$, $\pi(B_n^c(\delta)) \leq e^{-nr}$ *for large* n, *for some* $r > 0$.

Then, $\pi(C_\varepsilon | X^n) \to 1$, *a.s.* $[P^\infty]$ *and* $\int |\hat{f}_n(x) - f(x)|\, \mu(dx) \to 0$ *a.s.* $[P^\infty]$.

THEOREM 8.3. *Consider the prior in Section* 3 *and let* $\varepsilon > 0$. *If* $c_j = j^{3+\varepsilon}$ *or* $c_j = (1 + \varepsilon)^j$ *then* (a) *and* (b) *hold for any density* q *such that* $D(p_0 \| q) < \infty$.

We thus see that the conditions needed to make the Bayes factor consistent are strong enough to make the density estimate consistent too.

9. Discussion. We have taken the nominal family to be normal. One would hope that, if the underlying density is heavy tailed and produces many outliers, the nonparametric component of the model will suitably accommodate the extreme observations. This is indeed the case as was born out by density five of the example in Section 7. Nonetheless, there are advantages to handling extreme observations in another way, namely, by using a heavy-tailed density such as a t-density, for the nominal model. This may improve the efficiency of the density estimator. It also allows us to test, by way of Bayes factors whether the fully nonparametric model is necessary or whether a long-tailed parametric nominal model suffices. We do not pursue such an extension here, although it appears that it would be straightforward. Also, we have not attempted to get rates of convergence of the density estimate. The results of Shen (1995) would be useful for this. Finally, the Brunk method works well, more in terms of the Bayes factor than its density estimate. This model deserves further investigation.

APPENDIX

PROOF OF LEMMA 8.4. Consider $\mathrm{Cov}_\psi(\phi_i, \phi_j) = E_\psi(\phi_i\phi_j) - E_\psi(\phi_i)E_\psi(\phi_j)$. Let $f(\psi) = E_\psi(\phi_i\phi_j)$ and let $\alpha_r(\psi) = E_\psi(\phi_r)$. Also, note that for $1 \le r \le m$, $\sup_{0 \le u \le 1}|\phi_r(u)| = \sqrt{2r+1} \le \sqrt{2m+1}$. We show that $f(\psi)$ has the appropriate order; the proof for $\alpha_r(\psi)$ is similar. A first-order expansion of $f(\psi)$ around $\psi = 0$ yields, for some ψ^+ between 0 and ψ,

$$\begin{aligned} f(\psi) &= \sum_{r=1}^{m} \psi_r \int_0^1 \phi_i(u)\phi_j(u)\big[\phi_r(u) - \alpha_r(\psi)\big] g(u|\psi^+)\, du \\ &\le \sum_{r=1}^{m} |\psi_r| \int_0^1 |\phi_i(u)||\phi_j(u)||\phi_r(u) - \alpha_r(\psi)| g(u|\psi^+)\, du \\ &\le 2\{2m+1\}^{3/2} \sum_{r=1}^{m} |\psi_r| \le \{2m+1\}^{3/2}\sqrt{m}\,\|\psi\| = O\big(m^2\|\psi\|\big). \qquad \square \end{aligned}$$

PROOF OF LEMMA 8.5. Note that

$$H \equiv (1-\varepsilon)T\Sigma_n^{-1} + I = (1-\varepsilon)Tn(I+A) + I$$

where A has entries $O_P(\sqrt{m^5/n})$. Thus, $|H| = h_1 h_2$ where

$$h_1 = \frac{|nT_0(I+A) + I|}{|nT_0 + I|},$$

$h_2 = |nT_0 + I|$ and $T_0 = (1-\varepsilon)T$. A simple calculation shows that $h_2 \to \infty$ as $n \to \infty$. Now we show that h_1 is bounded away from 0 in probability. Given a matrix C, let $\lambda_1(C) \ge \cdots \lambda_m(C)$ be the ordered eigenvalues. Using Theorem

2.2 from Anderson and DasGupta (1963), we have

$$\begin{aligned} h_1 &= \frac{\prod_{i=1}^m \lambda_i\big[nT_0(I+A)+I\big]}{\prod_{i=1}^m \lambda_i\big[nT_0+I\big]} \\ &\geq \frac{\prod_{i=1}^m \big[\lambda_m(I+A)\lambda_i(nT_0)+1\big]}{\prod_{i=1}^m \big[\lambda_i(nT_0)+1\big]} \\ &= \prod_{i=1}^m \left[1+\frac{\lambda_m(A)}{1+i^{2k}/\big(n(1-\varepsilon)\tau^2\big)}\right] \\ &\geq \left[1-\frac{O_P\Big(\sqrt{m^5/n}\Big)}{1+1/\big(n(1-\varepsilon)\tau^2\big)}\right]^m. \end{aligned}$$

The last term is bounded away from 0 in probability both when $m = o(n^{1/7})$ and when $m = c \log n$. □

Acknowledgments. The authors thank the Editor, Associate Editor and referee for helpful comments.

REFERENCES

ANDERSON, T. W. and DAS GUPTA, S. (1963). Some inequalities on characteristic roots of matrices. *Biometrika* **50** 522–524.

ANDREWS, D. F. and HERZERG, A. (1985). *Data: A Collection of Problems From Many Fields for the Student and Research Worker*. Springer, New York.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to non-parametric problems. *Ann. Statist.* **2** 1152–1174.

BARRON, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, Dept. Statistics, Univ. Illinois.

BARRON, A. R. and COVER, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37** 1034–1054.

BARRON, A. R. and SHEU, C. H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19** 1347–1369. [Correction (1991) **19** 2284.]

BAYARRI, M. J. (1985). A Bayesian test for goodness-of-fit. Unpublished manuscript.

BERGER, J. O. and BERNARDO, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84** 200–207.

BERGER, J. O. and PERICCHI, L. (1996). The intrinsic Bayes factor for model selection and prediction. J. Amer. Statist. Assoc. **91** 109–122.

BOX, G. E. P. (1980). Sampling and Bayes inference in scientific modeling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430.

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Menlo Park, NY.

BRUNK, H. D. (1978). Univariate density estimation by orthogonal series. *Biometrika* **65** 521–528.

CRAIN, B. (1974). Estimation of distributions using orthogonal expansions. *Ann. Statist.* **2** 454–463.

CRAIN, B. (1976). More on the estimation of distributions using orthogonal expansions. *J. Amer. Statist. Assoc.* **71** 741–745.

DELAMPADY, M. and BERGER, J. O. (1990). Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit. *Ann. Statist.* **18** 1295–1316.

DIACONIS, P. and FRIEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–67.

DICICCIO, T., KASS, R. E., RAFTERY, A. and WASSERMAN, L. (1995). Simulation methods for computing Bayes factors. Technical report, Dept. Statistics, Carnegie Mellon Univ.

DIEBOLT, J. and ROBERT, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56** 363–375.

DÜMBGEN, L. (1998). New goodness of fit tests and their application to nonparametric confidence sets. *Ann. Statist.* **26** 288–314.

EFRON, B. and TIBSHIRANI, R. (1996). Specially designed exponential families for density estimation. *Ann. Statist.* **24** 2431–2461.

EUBANK, R. and LARICCIA, V. (1992). Asymptotic comparison Cramér–Von Mises and nonparametric function estimation techniques for testing goodness of fit. *Ann. Statist.* **20** 2071–2086.

EVANS, M. and SWARTZ, T. (1994). Distribution theory and inference for polynomial-normal densities. *Comm. Statist. Theory Methods* **23** 1123–1148.

FERGUSON, T. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

GELFAND, A. and DEY, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–515.

GELFAND, A., DEY, D. and CHANG, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 147–167. Oxford Univ. Press.

GELFAND, A. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GELMAN, A., MENG, X. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807.

GEWEKE, J. (1989). Modelling with normal polynomial expansions. In *Economic Complexity. Chaos, Sunspots and Nonlinearity* (W. A. Bennett, J. Geweke and K. Shell, eds.) Cambridge Univ. Press.

GOOD, I. J. (1967). A Bayesian significance test for multinomial distributions (with discussion), *J. Roy. Statist. Soc. Ser. B* **29** 399–431.

GOOD, I. J. (1992). The Bayes/non-Bayes compromise: a brief review. *J. Amer. Statist. Assoc.* **87** 597–606.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** 1926–1947.

HART, J. (1997). *Nonparametric Smoothing and Lack of Fit Tests*. Springer, New York.

HJORT, N. L. (1996). Bayesian approaches to non- and semiparametric density estimation. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 223–253. Oxford Univ. Press.

HJORT, N. L. and GLAD, I. (1995). Nonparametric estimation with a parametric start. *Ann. Statist.* **23** 882–904.

HJORT, N. L. and JONES, M. C. (1996). Locally parametric/nonparametric density estimation. *Ann. Statist.* **24** 1619–1647.

INGLOT, T. and LEDWINA, T. (1996). Asymptotic optimality of data-driven Neyman tests for uniformity. *Ann. Statist.* **24** 1982–2019.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford.

KASS, R. E. and RAFTERY, A. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–793.

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

LEDWINA, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.* **89** 1000–1005.

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

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

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

MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736.

MENG, X. L. (1994). Posterior predictive *p*-values. *Ann. Statist.* **22** 1142–1160.

MENG, X. and WONG, W. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6** 831–860.

NEYMAN, J. (1937). "Smooth" test for goodness of fit. *Skand. Aktvarietidskr.* **20** 150–199.

NOBILE, A. (1994). Bayesian analysis of finite mixture distributions. Ph.D. dissertation, Dept. Statistics, Carnegie Mellon Univ.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 99–138.

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameter tends to infinity. *Ann. Statist.* **16** 356–366.

RAFTERY, A. (1992). Discussion of, "Model determination using predictive distributions with implementation via sampling-based methods" by A. Gelfand, D. Dey and H. Chang. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 160–163. Oxford Univ. Press.

RAYNER, J. C. W. and BEST, D. J. (1990). Smooth tests of goodness of fit: an overview. *Internat. Statist. Rev.* **58** 9–17.

ROEDER, K. and WASSERMAN, L. (1995). Practical Bayesian density estimation using mixtures of normals. Technical Report 633, Dept. Statistics, Carnegie Mellon Univ.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SHEN, X. (1995). On the properties of Bayes procedures in general parameter spaces. Technical report, Dept. Statistics, Ohio State Univ.

SHUN, Z. and MCCULLAGH, P. (1995). Laplace approximation of high dimensional integrals. *J. Roy. Statist. Soc. B* **57** 749–760.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

TANNER, M. and WONG, W. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–540.

TARTER, M. and LOCK, M. (1993). *Model-Free Curve Estimation*. Chapman and Hall, New York.

THORBURN, D. (1986). A Bayesian approach to density estimation. *Biometrika* **73** 65–76.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762.

VERDINELLI, I. and WASSERMAN, L. (1995). Computing Bayes factors by using a generalization of the Savage–Dickey density ratio. *J. Amer. Statist. Assoc.* **90** 614–618.

WAHBA, G. (1981). Data-based optimal smoothing of orthogonal series density estimates. *Ann. Statist.* **9** 146–156.

WEST, M. (1992). Modelling with mixtures. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 503–524. Oxford Univ. Press.

WHITTLE, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. Ser. B* **20** 334–343.

DIPARTIMENTO DI STATISTICA
UNIVERSITA DI ROMA "LA SAPIENZA"
PIAZZALE A. MORO 5
00185 ROMA,
ITALY

DEPARTMENT OF STATISTICS
232 BAKER HALL
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213-3890
E-MAIL: larry@stat.cmu.edu