Bayesian Goodness-of-Fit Testing with Mixtures of Triangular Distributions

Ross Mcvinish, Judith Rousseau, Kerrie Mengersen

BAYESIAN GOODNESS OF FIT TESTING WITH MIXTURES OF TRIANGULAR DISTRIBUTIONS

ROSS McVINISH
School of Mathematical Sciences, Queensland University of Technology
JUDITH ROUSSEAU
Ceremade, Université Paris Dauphine
KERRIE MENGERSEN
School of Mathematical Sciences, Queensland University of Technology

ABSTRACT. We consider the consistency of the Bayes factor in goodness of fit testing for a parametric family of densities against a nonparametric alternative. Sufficient conditions for consistency of the Bayes factor are determined and demonstrated with priors using certain mixtures of triangular densities.

Running headline: Bayesian Goodness of Fit Testing

Key words: Bayesian nonparametrics, consistency, goodness of fit

1. Introduction

A problem common to many statistical analyses is to determine if a sample of n independent and identically distributed observations have been generated from a distribution described by a finite dimensional parametric model. This problem may be stated formally as a test of hypotheses on a density p_*;

$$H_0 : p_* \in \mathcal{F}_0 \quad \text{against} \quad H_1 : p_* \in \mathcal{F}_1 \backslash \mathcal{F}_0,$$

where $\mathcal{F}_0$ denotes a set of density functions with a particular finite dimensional parametric representation and $\mathcal{F}_1$ is some encompassing set of density functions such as the set of bounded and continuous densities. All densities are assumed to be with respect to the same dominating measure μ.

Central to the Bayesian approach to hypothesis testing is the Bayes factor which requires the specification of a prior for p_*. Let the prior probabilities on H_0 and H_1 be denoted by α and $(1-\alpha)$, respectively, and let the prior distributions on the sets $\mathcal{F}_0$ and $\mathcal{F}_1$ be denoted by π_0 and π_1. It is assumed that $\pi_1(\mathcal{F}_0) = 0$ and we will consider π_1 as a prior on $\mathcal{F}_1 \backslash \mathcal{F}_0$. The overall prior for the density p_* may be written as $\pi(\mathcal{A}) = \alpha \cdot \pi_0(\mathcal{A}) + (1-\alpha) \cdot \pi_1(\mathcal{A})$, for any $\mathcal{A} \subset \mathcal{F}_1$. The Bayes factor is then defined by

$$B_n = \frac{\pi(\mathcal{F}_0 \mid Y^n)}{\pi(\mathcal{F}_1 \mid Y^n)} \times \frac{\pi(\mathcal{F}_1)}{\pi(\mathcal{F}_0)},$$

where $\pi(\cdot \mid Y^n) = \pi(\cdot \mid Y_1, \ldots, Y_n)$ is the posterior distribution formed with the prior π.

There have been some examinations of the asymptotic properties of the Bayes factor for various $\mathcal{F}_0, \mathcal{F}_1$. Gelfand & Dey (1994) studied the case where both $\mathcal{F}_0$ and $\mathcal{F}_1$ have a parametric representation, in the setup of model choice. In a goodness of fit setup Verdinelli & Wasserman (1998) considered the case where $\mathcal{F}_0$ is a singleton and $\mathcal{F}_1$ is nonparametric. Using an infinite dimensional exponential family prior on $\mathcal{F}_1$ they were able to determine sufficient conditions for consistency. This problem was also studied by Dass & Lee (2004) who, based on an application of Doob's theorem, were able to give sufficient conditions for consistency with a general prior π_1. Walker *et al.* (2004) studied the related problem where $\mathcal{F}_0$, $\mathcal{F}_1$ may be arbitrarily defined but the true density p_* is not contained in either set.

The objective of this paper is to study the case where $\mathcal{F}_0$ is parametric and $\mathcal{F}_1$ is non-parametric, which is of considerable practical importance. The consistency of the Bayes factor in this general framework is a more delicate problem than the consistency of the Bayes factor in a goodness of fit problem where the null hypothesis is a singleton. Very recently Ghosal *et al.* (2008) have provided sufficient conditions for consistency of the Bayes factor when $\mathcal{F}_0$ is parametric. Their main condition given in equation (4.1) is tighter than necessary as can be seen when applying it to nested parametric models. In section 2 we provide weaker sufficient conditions for the Bayes factor to be consistent in testing a parametric family against a nonparametric alternative and demonstrate that they are close to being necessary.

In practice, the choice of the nonparametric prior is not only determined by theoretical properties but also by the ease of implementation. Various priors have been used for

goodness of fit problems in the literature including Dirichlet processes (Carota & Parmigiani, 1996), Polya trees (Berger & Guglielmi, 2001) and infinite dimensional exponential families (Verdinelli & Wasserman, 1998). With priors such as these, verifying that the general conditions given in section 2 hold is far from trivial. In order to overcome this difficulty we consider in section 3 a class of priors based on certain mixtures of triangular distributions which were proposed in Perron & Mengersen (2001). Using these priors, we discuss in section 4 the consistency of the Bayes factor for testing a parametric family, giving simple conditions on the parametric family for the Bayes factor to be consistent. An example is provided where one of the conditions given in section 2 does not hold and the Bayes factor is inconsistent. The rate of convergence for the Bayes factor in the case where $\mathcal{F}_0$ is a singleton is also studied. We conclude with a discussion in section 5. The technical details are presented in the Appendix.

We now give some notations that will be used throughout this paper. For a distribution P with density p, let P^n and P^∞ denote its n-fold product and infinite product distribution, respectively. Expectations with respect to these measures will be denoted E^n and E^∞, respectively. The most common measures of distance between two densities p, q are the L_1-distance, denoted by $\| p - q \|_1$, and the Hellinger distance $h(p,q) = \| p^{1/2} - q^{1/2} \|_2$. We allow d to stand for either of these distances. Finally, we let $\mathcal{H}(L,\beta)$ denote the set of Hölder continuous functions, that is the set of functions f such that $\left|f^{(m)}(x) - f^{(m)}(y)\right| < L\,|x-y|^\beta$ for $\beta \in (m, m+1]$.

2. Consistency of Bayes Factors

Recall that the Bayes factor is said to be consistent if

$$\lim_{n\to\infty} B_n = \begin{cases} \infty, \text{ in } P_*^\infty \text{ probability} & \text{if } p_* \in \mathcal{F}_0 \\ 0, \text{ in } P_*^\infty \text{ probability} & \text{if } p_* \in \mathcal{F}_1 \backslash \mathcal{F}_0 \end{cases}.$$

In other words any decision in the form : H_0 is accepted if $B_n > t$ for some fixed level t gives asymptotically the right answer. In goodness of fit tests (or other tests of nested hypotheses) the null distribution can be regarded as being part of the alternative, therefore inconsistency of the Bayes factor in this case means inconsistency of its answer to the question: is the null hypothesis true?

Here we shall consider the case where $\mathcal{F}_0 = \{p : p = p_\theta, \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}^d$. As the prior probabilities on H_0 and H_1 have no effect on the consistency we will take $\alpha = 1/2$.

The Bayes factor for our hypothesis test is given by

$$B_n = \left\{ \int_\Theta \prod_{i=1}^{n} p_\theta\left(Y_i\right) \pi_0\left(d\theta\right) \right\} \left\{ \int_{\mathcal{F}_1} \prod_{i=1}^{n} p\left(Y_i\right) \pi_1\left(dp\right) \right\}^{-1}.$$

To study the asymptotic behaviour of B_n we need to introduce some assumptions. Our first two assumptions were also used by Ghosal *et al.* (2008) in their examination of the consistency of Bayes factors.

Assumption A1: The nonparametric posterior from π_1 is strongly consistent at p_* with rate ϵ_n, that is

$$\pi_1\left(p : d(p, p_*) > \epsilon_n \mid Y^n\right) \to 0,$$

in P_*^∞ probability.

Assumption A2: For any $\theta \in \Theta \subset \mathbb{R}^d$

$$\pi_0\left(\theta' : K(p_\theta, p_{\theta'}) < cn^{-1}, V(p_\theta, p_{\theta'}) < cn^{-1}\right) > Cn^{-d/2},$$

where $K\left(p, q\right)$ is the Kullback-Leibler divergence defined by $K\left(p, q\right) = \int \log(p/q)p \; d\mu$ and $V\left(p, q\right) = \int \log(p/q)^2 p \; d\mu$.

Assumption *A1* is satisfied by many nonparametric priors. Consistency of the posterior is actually a minimal condition to require since there cannot be a full subjective (informative) construction of a prior on an infinite dimensional parameter set. Diaconis & Freedman (1986) discuss the importance of consistency of the posterior distribution. Sufficient conditions for establishing rates of convergence of the posterior distribution in the setting of density estimation have been determined by Ghosal *et al.* (2000). This assumption is very weak, considering that there is no constraint on ϵ_n other than $\lim_n \epsilon_n = 0$. Assumption *A2* is used to provide a lower bound on the marginal likelihood under the parametric model and it is satisfied by any regular parametric model with positive continuous prior density. It is possible to replace this assumption with a Laplace expansion of the marginal likelihood, however any model with a valid Laplace expansion will most likely satisfy *A2*.

The next assumption is weaker than the condition given in equation (4.1) of Ghosal *et al.* (2008).

Assumption A3: If $A_{\epsilon_n}(\theta) = \{p : \ d(p, p_\theta) < C\epsilon_n\}$, where ϵ_n is the rate of convergence given in Assumption *A1*, then

$$\sup_\theta \pi_1(A_{\epsilon_n}(\theta)) = o(n^{-d/2}).$$

Assumption *A3* compares the amount of probability the nonparametric prior π_1 places near the parametric family with the prior mass of the parametric prior near each parameter value $\theta \in \Theta$. This is the key assumption as will be demonstrated in section 4.1.

Our final assumption is a technical assumption which will be satisfied by most regular models.

Assumption A4: Θ is a compact subset of $\mathbb{R}^d$ and p_θ is continuous in L_1 as a function of θ.

The compactness of Θ is used here to make the proof clearer. In the non compact case, we can go back to the compact case by assuming for instance some regularity conditions on the model so that the maximum likelihood estimator under the parametric model converges to the projection of p_* on $\{p_\theta, \theta \in \Theta\}$ (Arcones, 2002).

The following theorem holds for general classes of prior distributions both on the alternative and on the parametric models.

Theorem 1. *Assume that given P_* the data $Y_1, \ldots, Y_n$ are independent and identically distributed on $[0, 1]$. Assume also that assumptions A1-A4 hold.*

- *If $p_* \in \mathcal{F}_0$ then $B_n \to \infty$ in P_*^∞ probability.*
- *If $p_* \in \mathcal{F}_1 \backslash \mathcal{F}_0$ and p_* is in the Kullback-Leibler support of π_1 then $B_n \to 0$, exponentially fast, almost surely with respect to P_*^∞.*

Proof. Assume that $p_* \in \mathcal{F}_0$, i.e. there exists θ_* such that $p_* = p_{\theta_*}$. Let $A_{\epsilon_n} = A_{\epsilon_n}(\theta_*)$, then the Bayes factor can be written as

$$\begin{aligned} B_n^{-1} &= \left\{ \int_{\mathcal{F}_1} \prod_{i=1}^n p(Y_i) \, \pi_1(dp) \right\} \left\{ \int_\Theta \prod_{i=1}^n p_\theta(Y_i) \, \pi_0(d\theta) \right\}^{-1} \\ &= \left\{ \int_{A_{\epsilon_n}} \prod_{i=1}^n p(Y_i) \, \pi_1(dp) \right\} \left\{ \int_\Theta \prod_{i=1}^n p_\theta(Y_i) \, \pi_0(d\theta) \right\}^{-1} \times \pi_1^{-1}(A_{\epsilon_n} \mid Y_1, \ldots, Y_n). \end{aligned}$$

Under assumption *A1*, $\pi_1\left(A_{\epsilon_n} \mid Y_1, \ldots, Y_n\right)$ converges to one in P_*^∞ probability. From *A3* we may apply lemma 1 of Shen & Wasserman (2001) to give that when δ is small enough

$$P_*^n\left\{\int_\Theta \prod_{i=1}^n \frac{p_\theta(y_i)}{p_*(y_i)}\pi_0(d\theta) < \delta n^{-d/2}\right\} \leq \frac{C}{(-\log\delta)}.$$

Applying the Markov inequality, for any $\epsilon > 0$,

$$P_*^n\left\{\int_{A_{\epsilon_n}} \prod_{i=1}^n \frac{p\left(y_i\right)}{p_*\left(y_i\right)}\pi_1\left(dp\right) > \epsilon\delta n^{-d/2}\right\} < \left(\epsilon\delta\right)^{-1} n^{d/2}\pi_1\left(A_{\epsilon_n}\right).$$

Combining the above two inequalities and letting $\epsilon, \delta \to 0$ at an appropriate rate, we see that under *A4* this probability will converge to zero and hence $B_n^{-1} \to 0$ in P_*^∞ probability.

Now assume that $p_* \in \mathcal{F}_1 \backslash \mathcal{F}_0$. Define for $i = 1, 2, \ldots$ $\Phi_i = \int_0^1 y^i p_*\left(y\right) dy$ and $\Phi_i(\theta) = \int_0^1 y^i p_\theta\left(y\right) dy$, for all $\theta \in \Theta$. Define $\Theta_0 = \Theta$ and $\Theta_n = \Theta_{n-1} \cap \{\theta : \Phi_n(\theta) = \Phi_n\}$. If there exists a $\theta' \in \Theta_n$ for all $n = 1, 2, \ldots$ then $p_{\theta'} = p_*$, $\mu-$a.e. by the uniqueness of the Hausdorff moment problem and as $p_* \in \mathcal{F}_1 \backslash \mathcal{F}_0$ we have a contradiction. Therefore, there exists an m such that $\Theta_m = \emptyset$, i.e. the first m moments of p_* and p_θ can not be equal for any $\theta \in \Theta$. By *A2* $\Phi_i(\theta)$ is a continuous function of θ for all i and hence $g(\theta) = \sup_{i=1,\ldots,m} \left|\Phi_i - \Phi_i(\theta)\right|$ is also a continuous function of θ. As there is no $\theta \in \Theta$ such that $g(\theta) = 0$ and Θ is a closed set then there exists an $\epsilon > 0$ such that $g(\theta) > \epsilon$ for all $\theta \in \Theta$.

The set Θ can be partitioned as

$$\Theta = \bigcup_{i=1}^m \left[\left\{\theta : \Phi_i - \Phi_i\left(\theta\right) > \epsilon\right\} \bigcup \left\{\theta : \Phi_i\left(\theta\right) - \Phi_i > \epsilon\right\}\right].$$

The functions y^i can be used to form a strictly unbiased test of $p = p_*$ against $p \in \{p_\theta : \Phi_i\left(\theta\right) - \Phi_i > \epsilon\}$ and of $p = p_*$ against $p \in \{p_\theta : \Phi_i\left(\theta\right) - \Phi_i < \epsilon\}$. Applying proposition 4.4.1 of Ghosh & Ramamoorthi (2003) there exists an exponentially consistent test. By applying their lemmas 4.4.1 and 4.4.2 it follows that

$$\frac{\int_{\{\theta : h_i - h_i(\theta) > \epsilon\}} \prod_{i=1}^n \frac{p_\theta(Y_i)}{p_*(Y_i)}\pi\left(d\theta\right)}{\int_{\mathcal{F}_1} \prod_{i=1}^n \frac{p(Y_i)}{p_*(Y_i)}\pi_1\left(dp\right)} \longrightarrow 0, \quad P_*^\infty - \text{almost surely},$$

exponentially fast. The Bayes factor is bounded by a finite sum of these terms and hence it must converge to zero $P_*^\infty-$ almost surely, exponentially fast.

Remark 1. The restriction of Y_i to $[0,1]$ was only introduced to simplify the construction of a weak neighbourhood around p_* which did not intersect $\mathcal{F}_0$. It is possible to alter this assumption for random variables on other ranges. For example, if the support of the Y_i is $[0,\infty)$ then using functions $e^{-my},\ m=1,2,\ldots$ can be used to form weak neighbourhood of p_* and the same arguments hold. We need only consider $m=1,2,\ldots$ since a distribution on $[0,\infty)$ is uniquely identified by the sequence of values of its Laplace transform at integer values (Feller, 1939).

Remark 2. Note that we obtain also an upper bound on the rate of convergence of the Bayes factor under H_0 since we have proved that

$$\lim_{C\to\infty}\limsup_n P_*^n\left\{B_n^{-1} > Cn^{-d/2}\pi(A_{\epsilon_n})\right\} = 0.$$

We therefore obtain the significant result that B_n^{-1} is controlled, when H_0 is true, by the ratio of the prior probabilities of effective neighbourhoods of the true density under π_1 and π_0 respectively.

The conditions of theorem 1 can be difficult to verify as they require finding upper bounds on the prior probabilities in non-regular cases. Despite this difficulty, Rousseau (2008) has provided a general framework where assumptions *A1* and *A3* hold. This is the case of the embedded prior, that is where the nonparametric prior is constructed on the embedded model

$$\{p_{\theta,g}(x) = p_\theta(x)g(P_\theta(x)), \theta\in\Theta, g\in\mathcal{G}\},$$

where $\mathcal{G}$ is the set of density functions on [0,1]. In such cases the prior mass of neighbourhoods under the nonparametric model will be typically much smaller than those under the parametric model, as described in section 4.1.1 of Rousseau (2008).

The following two sections consider the mixture of triangular priors as a prior for which these conditions can be verified. They also provide an illustration of difficulties that can be encountered.

3. Mixtures of triangular priors

3.1. **Definitions.** A mixture of triangular distributions is defined as follows. Let the sequence $0 = t_0 < t_1 < \ldots < t_{k-1} < t_k = 1$ be a partition of $[0,1]$. The function $\Delta_i(x)$ is the triangular density function with support on the interval $[t_{i-1}, t_{i+1}]$ and mode at

t_i for $i = 1, \ldots, k-1$. The density $\Delta_0(x)$ is the triangular distribution on $[t_0, t_1]$ with mode at t_0, similarly $\Delta_k(x)$ has support $[t_{k-1}, t_k]$ and mode t_k. A mixture of triangular distributions then has the density function $p(x) = \sum_{i=0}^{k} w_i \Delta_i(x)$, where $w_i \geq 0$ and $\sum_{i=0}^{k} w_i = 1$. As in Perron & Mengersen (2001) we consider the two cases:

I For each k, the partition of $[0,1]$ is assumed fixed so that $t_i = i/k$ and the weights w_i are varied. The density is denoted $p(x; w(k))$ where $w(k) = (w_0, \ldots, w_k)$.

II For each k, the weights w_i are fixed at $w_0 = w_k = 1/(2k)$, $w_i = 1/k$, $i = 1, \ldots, k-1$ and the partition is varied. The density is denoted by $p(x; \psi(k))$ where $\psi(k) = (t_0, \ldots, t_k)$ denotes the partition of $[0,1]$.

When needed, we generically denote by $\xi(k)$ the vector of parameters of a mixture of triangulars with k components and by S_k the set of these parameters. The log-likelihood shall be denoted by $l_n(\xi(k)) = \sum_{i=1}^{n} \log p(Y_i; \xi(k))$ and the prior on the parameter $(\xi(k), k)$ is written $\pi(\xi(k), k) = \pi(\xi(k) \mid k)\pi(k)$, in the mixture of triangular distributions prior.

There are a number of reasons for choosing to work with mixtures of triangular distributions. Firstly, as the resulting density functions are piecewise linear functions on $[0,1]$ interpolating the points $(t_i, w_i \Delta_i(t_i))$, they are easy to manipulate and simplify some necessary calculations. The flexibility of the densities allows them to approximate smooth density functions well which leads to good asymptotic properties for the posterior distribution. Finally, they can be relatively easy to implement in practice.

Since the rate of convergence of the posterior is relevant to obtaining consistency of the Bayes factor, we first give a few results on posterior rates of convergence under such priors. When there is no possibility of confusion we shall denote the mixture of triangular prior by π instead of π_1.

3.2. **Type I mixture - rates of convergence.** As with the Bernstein polynomial priors, for a given k, this class of mixtures is a simple convex combination of density functions which are bounded by a multiple of k. Thus consistency (strong and weak) and rates of convergence can be proved using very similar techniques to those used for Bernstein polynomials (Ghosal, 2001, Petrone & Wasserman, 2002).

Theorem 2. *Assume that p_* belongs to the Hölder class $\mathcal{H}(L, \beta)$ with $\beta \leq 2$ and satisfies $p_*(x) \geq ax(1-x)$ for some constant $a > 0$. Assume also that the Type I mixture of*

triangular distributions prior satisfies for all k, $c_1 e^{-C_1 k \log k} \leq \pi(k) \leq c_2 e^{-C_2 k}$ *for some constants* $c_1,\ c_2,\ C_1,\ C_2 > 0$ *and for each* k *the prior on* $w(k)$ *is a Dirichlet distribution with parameters uniformly bounded in* k. *Then there exists an* $R > 0$ *such that*

$$E_*^n[\pi\left(p : d(p, p_*) > Rn^{-\beta/(2\beta+1)}(\log n)^{(4\beta+1)/4\beta} \mid Y^n\right)] \leq n^{-H},$$

for all $H > 0$ *and all* n *sufficiently large.*

The proof of theorem 2 is given in Appendix (i).

3.3. **Type II Mixtures - rates of convergence.** The type II mixture is slightly more complicated to study than the type I, in the same way as the free knot splines are more complicated to study than fixed splines estimators. However here, the problem is made easier since the weights are fixed. As in the case of fixed partition and free weights we obtain the minimax rate of convergence up to a $\log n$ term.

Theorem 3. *Assume that* p_* *belongs to the Hölder class* $\mathcal{H}(\beta, L)$ *with* $\beta \leq 2$ *Assume also that* $p_* \geq a > 0$ *on* $[0, 1]$. *In addition, assume that the prior satisfies the following conditions:*

- *The prior on the number of components* k *is such that there exists* $c_1, c_2 > 0$ *satisfying* $e^{-c_1 n \log n} \leq \pi(k > n) \leq e^{-c_2 n \log n}$ *for all* n *sufficiently large.*
- *For any* $k \leq k_0 n^{1/(2\beta+1)}$, *where* k_0 *is some positive constant, the prior places very small probability on two points in the partition being close. Specifically, there exist* $\alpha, \gamma > 0$ *satisfying, for all* $c > 0$

  $$\pi\left(\max_i |t_i - t_{i-1}| < e^{-\alpha n^\gamma} \mid k\right) < \exp(-cn^{1/(2\beta+1)} \log n),$$

  for all n *sufficiently large. Moreover, for any* k *the prior has a positive density with respect to the Lebesgue measure* μ_k *on* $\{\psi(k) = (t_0, ..., t_k); 0 < t_1 < ... < t_{k-1} < 1\}$*: there exists* $r > 0$ *such that for all* $k > 1$

  $$\pi(\psi(k) \mid k) \geq c_k \mu_k(\psi(k)), \quad c_k > c/\Gamma(k)^r.$$

Then there exists an $R > 0$ *such that*

$$E_*^n\left[\pi\left(p : d(p, p_*) > Rn^{-\beta/(2\beta+1)} \log n \middle| Y^n\right)\right] \leq n^{-H},$$

for all $H > 0$ *and all* n *sufficiently large.*

The proof of theorem 3 is given in point (ii) of the Appendix.

Remark 3. As a simple example of a prior satisfying the conditions of theorem 3 consider a Poisson process with continuous intensity on $[0, 1]$. This prior can also be viewed as a Poisson prior on k and conditional on k, any density absolutely continuous with respect to the distribution of the order statistic of a k sample of uniforms on $[0, 1]$.

Remark 4. Note that the above classes of priors, both for the Types I and II mixtures of triangular densities lead to adaptive estimators with respect to the smoothness parameter β, on $\beta \in (0, 2]$, up to a $\log n$ term.

4. Bayes factors with mixtures of triangular prior

4.1. **Bayes factor with a parametric null hypothesis.** Now consider the hypothesis test of the introduction where the prior on $\mathcal{F}_1$ is one of the mixture of triangular priors that we have described. We note that it is also possible to specify the prior on $\mathcal{F}_1$ in a goodness of fit setting by embedding the parametric model p_θ in the nonparametric model through $p_\theta(y)g(P_\theta(y))$ where g is a nonparametric density on $[0, 1]$ and P_θ is the cumulative distribution function. This is the approach taken in Verdinelli & Wasserman (1998) and Robert & Rousseau (2004). Although the former can have some desirable properties (Rousseau, 2008), we think that, due to the popularity of using mixtures to directly model density functions, it is of interest to investigate if such direct modeling can be applied to goodness of fit problems. To apply theorem 1 with a mixture of triangular distributions prior to theorem we need to verify assumptions *A1* - *A4* hold. *A1* holds by either theorem 2 or theorem 3 depending on the choice of prior. Assumptions *A2* and *A4* depend on the parametric family. To verify that *A3* holds we introduce a new assumption on the parametric family to be tested.

Assumption A3′: For each $\theta \in \Theta$, p_θ has a bounded third derivative and its second derivative is non-zero on some interval $[a, b]$, $0 < a, b < 1$.

The set $A_{\epsilon_n}(\theta)$ comprises of all densities which are within ϵ_n of p_θ as measured by Hellinger or L_1 distance. In the proof of lemma 4.2 of McVinish *et al.* (2005) it was shown that for any density $p(x)$ satisfying *A3′* and any mixture of triangular densities $p(x; \xi(k))$ (type I or II), then there exists constants α, c such that

$$\int_0^1 |p(x) - p(x; \xi(k))| \, dx \geq ck^{-\alpha}. \tag{1}$$

As we have assumed that for all $\theta \in \Theta$, p_θ satisfies *A3′* then

$$\inf_{\theta\in\Theta} \inf_{\xi(k)\in S^k} d\left(p_\theta, p(\cdot\,;\xi(k))\right) \geq ck^{-\alpha},$$

where $d(.,.)$ can be either L_1 or Hellinger distance. Therefore, for a mixture of k triangular densities to be an element of $A_{\epsilon_n}(\theta)$ we need $k > \epsilon_n^{-1/\alpha} c = k_n$. Since theorem 2 and theorem 3 together with the regularity condition of $A3'$ imply that $\epsilon_n < Cn^{-3/7}(\log n)^2$, it follows that a mixture of triangular densities will need at least $Cn^{3/(7\alpha)}(\log n)^{-2/\alpha}$ components. Thus

$$\pi\left(A_{\epsilon_n}\right) \leq \pi(k \geq k_n) \leq e^{-C'k_n},$$

from the conditions on the priors on k and we finally obtain

$$\pi\left(A_{\epsilon_n}\right) \leq Be^{-bn^\gamma},$$

for some $B, b > 0$ and any $\gamma < 3/(7\alpha)$. Therefore, $n^{d/2}\pi\left(A_{\epsilon_n}\right) \to 0$ for any finite d and assumption *A3* is satisfied.

As an example of a parametric family to be tested, consider the two-dimensional exponential family

$$p_\theta(y) = f(y)\exp\left(\theta_1 b_1(y) + \theta_2 b_2(y) + \alpha(\theta_1, \theta_2)\right), \tag{2}$$

where Θ is a compact subset of $\mathbb{R}^2$, b_1, b_2, α are smooth functions and $y \in [0,1]$. It is easily seen that *A2* is satisfied. Also, since for all $\theta, \theta' \in \Theta$, $h(p_\theta, p_{\theta'}) < C|\theta - \theta'|$ it follows that *A4* is satisfied. If on some interval $f(y)$ has a bounded third derivative and its second derivative is non-zero then assumption *A3′* is satisfied. Also, if $f(y)$ is a finite mixture of triangular densities then assumption *A3′* will be satisfied provided $(0,0) \notin \Theta$. The Bayes factor therefore can provide a consistent test of this parametric family.

We now give a situation where failure of assumption *A3* to hold leads to an inconsistency in the Bayes factor, that is under H_0, $B_n \to 0$ in P_*^∞ probability. Consider again the parametric density (2) where $(0,0) \in \Theta$ and $f \equiv 1$; in other words there exists $\theta \in \Theta$ such that p_θ is the uniform density. In this case assumption *A3* is no longer satisfied since from theorem 2 or theorem 3

$$\pi_1\left(A_{\epsilon_n}(\theta)\right) > cn^{-2/5}(\log n)^{9/8},$$

for $\theta = (0, 0)$ and some $c > 0$ sufficiently small. When P_* is the uniform distribution we can bound the marginal likelihood for the parametric model using a Laplace approximation so that

$$\int_\Theta \prod_{i=1}^{n} p_\theta(y_i)\pi_0(d\theta) < Cn^{-1}(\log n)$$

with P_*^∞ probability tending to one. Sufficient conditions for the Laplace approximation to be valid are given in theorem 8 of Kass *et al.* (1990) and can easily be verified for (2). Inconsistency of the Bayes factor will have been demonstrated if we can show that

$$\int_{\mathcal{F}_1} \prod_{i=1}^{n} p\left(Y_i\right) \pi_1\left(dp\right) > c_n n^{-1}(\log n), \tag{3}$$

with P_*^∞ probability tending to one, for some sequence $c_n \to \infty$. We answer this question in the following subsection where we obtain a lower bound on the marginal likelihood under the prior π_1 by studying the rate of convergence of the Bayes Factor under a point null hypothesis.

4.2. **Bayes Factor with a point null hypothesis.** We now give a few results on the rate of convergence of the Bayes factor when the null hypothesis is the singleton $\{p_0\}$, where p_0 is the uniform density. The test of hypothesis becomes $H_0 : p_* = p_0$ against $H_1 : p_* \neq p_0$ and the Bayes factor is given by

$$B_n^{-1} = \int_{\mathcal{F}_1} \prod_{i=1}^{n} p\left(Y_i\right) \pi_1\left(dp\right).$$

The following theorem demonstrates that the Bayes factor goes to infinity under the null at a rate smaller than $n^{1/2}$ and hence verifies that inequality (3) holds. This completes the example of inconsistency of the Bayes factor at the end of section 4.1 where assumption *A3* fails to hold.

Theorem 4. *Assume that for Type I mixtures the prior on $\pi(w \mid k = 1)$ has a strictly positive and continuously differentiable density in a neighbourhood of $w = 1/2$, then under H_0 ($p_0 \equiv 1$)*

$$P_0^n \left[B_n^{-1} \leq C_0/\sqrt{n}\right] \leq n^{-H}, \tag{4}$$

for all $H > 0$ and all n sufficiently large. Assume that for Type II mixtures the prior on $\pi(w \mid k = 2)$ has a strictly positive and continuously differentiable density in a neighbourhood of $w = 1/2$, then under H_0 (4) holds.

Remark 5. The case $k = 1$ in the Type II mixtures corresponds to a point mass at the uniform density, so that it does not make sense to put positive mass on it under the alternative.

Proof. The proof is only given for type I mixture of triangular distributions prior since the proof for type II mixtures follows essentially the same argument. For a type I mixture of triangular distributions prior

$$\begin{aligned} B_n^{-1} = \sum_{k=1}^{\infty} B_{n,k} &= \sum_{k=1}^{\infty} \pi(k) \int_{S_k} \prod_{i=1}^{n} p(Y_i; w(k))\, d\pi(w(k) \mid k) \\ &\geq B_{n,1} = \pi(k=1) \int_0^1 \prod_{i=1}^{n} p(Y_i; w(1))\, d\pi(w(1) \mid 1). \end{aligned} \tag{5}$$

The true distribution corresponds to $w(1) = w_1 = 1/2$. This integral is bounded from below by considering only the integral on $w_1 \in (1/2 - \delta_n, 1/2 + \delta_n)$, $\delta_n = K \log n/\sqrt{n}$. We now take a Taylor expansion of the log-likelihood $l_n(w_1)$ and the log of the prior around the maximum likelihood estimator $\hat{w}_1$ and we bound the remaining term. This leads to

$$l_n(w_1) - l_n(\hat{w}_1) + \log \pi(w_1|1) - \log \pi(\hat{w}_1|1) = -\frac{n(w_1 - \hat{w}_1)^2 \hat{j}_1}{2} + R_n,$$

where

$$\begin{aligned} |R_n| &\leq Cn|w_1 - \hat{w}_1|^3 \sup_{|w-\hat{w}_1| \leq 4\delta_n} \left| \frac{\partial^3 (l_n/n)(w)}{\partial w^3} \right| + C|w_1 - \hat{w}_1| \sup_{|w-\hat{w}_1| \leq 4\delta_n} \left| \frac{\partial \log \pi(w)}{\partial w} \right| \\ &= R_{n,1} + R_{n,2}, \end{aligned}$$

where $\hat{j}_1$ is the empirical Fisher information. In a neighbourhood of $w_1 = 1/2$, $\log p(y; w_1)$ is 3 times continuously differentiable with finite moments of all order (uniformly bounded for w_1 near $1/2$. Therefore, for all $\delta > 0, H > 0$ and all n sufficiently large

$$P_0^n \left[\sup_{|w_1 - \hat{w}_1| \leq \delta_n} \frac{|R_{n,1}|}{n(w_1 - \hat{w}_1)^2} > \delta \hat{j}_1 \right] \leq Cn^{-H}.$$

Similarly, we have assumed that π is positive at $1/2$ and is continuously differentiable around $1/2$. Therefore, for all $\delta' > 0, H' > 0$ and all n sufficiently large

$$P_0^n \left[\sup_{|w_1 - \hat{w}_1| \leq \delta_n} |R_{n,2}| > \delta' \right] \leq Cn^{-H'}.$$

It follows that

$$\int_0^1 e^{l_n(w_1)} \pi(w_1 \mid k=1) dw_1 \geq e^{-\delta}(1+\delta)^{-1/2} \exp\{l_n(\hat{w}_1)\} \pi(\hat{w}_1 \mid k=1) \hat{j}_1^{-1/2} n^{-1/2},$$

with probability greater than $1 - n^{-H}$ for all $H > 0$ and all n sufficiently large. Since $e^{l_n(\hat{w}_1)} \geq 1$ and $P_0^n\left[\hat{j}_1^{-1/2} > c\right] \leq n^{-H}$ for all $H > 0$ and n sufficiently large, we may now take C_0 sufficiently small so that

$$P_0^n\left(B_n^{-1} \leq C_0 n^{-1/2}\right) \leq P_0^n\left(B_{n,1} \leq C_0 n^{-1/2}\right) \leq n^{-H}, \tag{6}$$

for all n sufficiently large.

By imposing additional mild conditions on the priors it is possible to strengthen the result of the above theorem to show that the Bayes factor is actually of order $n^{1/2}$ under the null up to a $\log n$ term.

Theorem 5. *Assume the prior on k satisfies $\pi\left(k > n/\log n\right) < e^{-nc}$ for some $c > 0$. For a Type I mixtures prior assume $\pi(w_0, \ldots, w_k \mid k)$ is absolutely continuous with respect to the Lebesgue measure on the simplex with a density bounded by $M\Gamma(k+1)$ for all k and some $M > 0$. Then there exists $C, C' > 0$ such that for all $\delta > 0$ and all $n \geq 1$,*

$$P_0^n\left(B_n^{-1} \geq C(\log n)^2 n^{-1/2}\right) \leq C'(\log n)^{-\frac{(1-\delta)}{2}}. \tag{7}$$

For a type II mixtures prior assume that $\pi(t_0, \ldots, t_k \mid k)$ is absolutely continuous with respect to the Lebesgue measure on the simplex with a density bounded by $M\Gamma(k+1)$ for all $k \geq 2$ and some $M > 0$. Assume also that there exist $\alpha, \gamma > 0$ satisfying,

$$\pi\left(\max_i |t_i - t_{i-1}| < e^{-\alpha n^\gamma} \mid k\right) < \exp(-cn^{1/(2\beta+1)} \log n),$$

for all $c > 0$ and all n sufficiently large. Then there exists $C, C' > 0$ such that (7) holds.

Proof. Using the decomposition (5) of B_n then for any $v_n = v_0 n^{-1/2}(\log n)$

$$P_0^n\left[B_n^{-1} \geq v_n\right] \leq P_0^n\left[B_{n,1} \geq v_n/2\right] + P_0^n\left[\sum_{k=2}^{\infty} B_{n,k} \geq v_n/2\right].$$

The proof is first given for type I mixtures. Define the sets $\mathcal{G}_n = \{w(k) : 2 \geq k \geq k_n\}$, $V_{n,k} = \{w(k) : \|p_0 - p\| < r_n\}$ and $V_n = \cup_k V_{n,k}$ where $r_n = r_0 n^{-1/2} \log n$ and $k_n =$

$k_0(\log n)^2$. Denote

$$\begin{aligned} I_1 &= \int_{V_n \cap \mathcal{G}_n} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right) = \sum_{k=2}^{k_n} \pi(k) \int_{V_{n,k}} \left[\prod_{i=1}^{n} p(Y_i; w(k))\right] \pi(w(k)|k), \\ I_2 &= \int_{V_n^c \cap \mathcal{G}_n} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right) = \sum_{k=2}^{k_n} \pi(k) \int_{V_{n,k}^c} \left[\prod_{i=1}^{n} p(Y_i; w(k))\right] \pi(w(k)|k), \\ I_3 &= \int_{\mathcal{G}_n^c} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right) = \sum_{k=k_n+1}^{\infty} B_{n,k}. \end{aligned}$$

Applying the Markov inequality

$$P_0^n\left[I_1 \geq v_n/6\right] \leq 6 v_n^{-1} \sum_{k \geq 2}^{k_n} \pi(k) \pi(V_{n,k} \mid k).$$

Applying lemma 2 in Appendix (iii) and the fact that the prior density on the weights is bounded, it follows that

$$\pi(V_{n,k} \mid k) \leq M r_n^k \Gamma(k+1) \pi^{k+1/2} / \Gamma(k/2 + 3/2).$$

From Stirling's approximation of the Gamma function

$$\sum_{k=2}^{k_n} M r_n^k \Gamma(k+1) \pi^{k+1/2} / \Gamma(k/2 + 3/2) \leq C r_n^2 \sum_{k=2}^{k_n} \exp\left((k-2)\log(k^{1/2} r_n) + \log k\right),$$

and hence

$$P_0^n\left[I_1 \geq v_n/6\right] \leq C v_n^{-1} r_n^2.$$

An application of lemma 4 in Appendix (iii) yields $P_0^n\left[I_2 \geq v_n/6\right] \leq n^{-H}$ for any $H > 0$ and n sufficiently large. Also

$$P_0^n\left[I_3 \geq v_n/6\right] \leq v_n^{-1} \sum_{k \geq k_n} \pi(k) \leq v_n^{-1} \exp(-r k_n) < n^{-H},$$

for any $H > 0$ and n sufficiently large.

Finally, we treat $B_{n,1}$ using a Laplace expansion in a similar manner to (6), in other words we can bound the integral on $|w_1 - 1/2| < \delta_n$ similarly to the proof of theorem 3.5. However we must bound from above the integral outside the δ_n neighbourhood of $\hat{w}_1$. To do so note that the model is regular and that for all $\epsilon > 0$ there exists $\delta > 0$ such that if $|w_1 - 1/2| > \epsilon$, $\|1 - p_{w_1}\|_1 > \delta$. Moreover $w_1 \in (0,1)$ a bounded interval in $\mathbb{R}$ so

that we can construct tests (Ghosal *et al.*, 2000) such that $E_0^n[\phi_n] \leq e^{-cn\delta_n^2}$ and for all $|w_1 - 1/2| > \delta_n$, $E_{w_1}^n[1-\phi_n] \leq e^{-cn\delta_n^2}$. This implies that for all $H > 0$,

$$P_0^n\left[\int_{|w_1-1/2|>\delta_n} e^{l_n(w_1)}\pi(w_1)dw_1 > e^{-nc\delta_n^2/2}\right] \leq e^{-nc\delta_n^2/2} = O(n^{-H}).$$

We need also to control $l_n(\hat{w}_1)$, which under the uniform follows asymptotically a Chi-square random variable and satisfies for all $\delta > 0$,

$$\begin{aligned} P_0^n[l_n(\hat{w}_1) > \log\log n] &= P_0^n\left[\chi_1^2 > \log\log n - \delta\right] + n^{-H} \\ &\leq \frac{C}{\sqrt{(\log n)^{1-\delta}}}, \end{aligned}$$

leading finally to (7).

For Type II mixtures it is necessary to make some small changes to the proof. Define $\mathcal{G}_n = \{\psi(k); 3 \leq k \leq k_n, |t_i - t_{i+1}| > n^{-a(\log n)^2}, i \leq k-1\}$ and let $k_n = k_0(\log n)^2$ $r_n = r_0 n^{-1/2}(\log n)^{-3}$ and $v_n = v_0 n^{-1/2}(\log n)$. From remark 5 $B_{n,2}$ plays the role of $B_{n,1}$ in the proof of type I mixtures. The main difference in the proof being in the treatment of $P_0^n[I_1 \geq v_n/6]$. To control this term, we use lemma 3 in Appendix (iii) so that

$$\begin{aligned} &\sum_{j=1}^{k-2} \frac{(t_{j+1}-t_j)}{4}\left(\frac{|(t_{j+2}-t_j)-2/k|}{(t_{j+2}-t_j)} + \frac{|(t_{j+1}-t_{j-1})-2/k|}{(t_{j+1}-t_{j-1})}\right) \\ &+\frac{1}{2kt_2}|t_1 - t_2/2| + \frac{1}{2k(1-t_{k-2})}|(1-t_{k-1}) - (1-t_{k-2})/2| \leq r_n. \end{aligned}$$

Since $k_n r_n = o(1)$, $|1 - t_{k-1} - (t_{k-1} - t_{k-2})| \leq 2r_n(1+o(1))$ so that $|t_{k-1} - (1-1/k)| \leq 8r_n(1+o(1))$ and $|t_{k-2} - (1-2/k)| \leq 8r_n(1+o(1))$. We then iterate the formula using $|t_{k-j-1} - (1-(j+1)/k)| \leq 8r_n(1+o(1)) + |t_{k-j} - (1-j/k)|$ so that when $k \leq k_n$,

$$\pi\left(\| p_0 - p_{\psi(k)} \|_1 \leq r_n \mid k\right) \leq \pi\left(|t_j - j/k| \leq 8kr_n, j \leq k-1 \mid k\right).$$

The bound on $P_0^n[I_1 \geq v_n/6]$ becomes

$$P_0^n[I_1 \geq v_n/6] \leq v_n^{-1}\sum_{k=3}^{k_n}\pi(V_{n,k} \mid k)\pi(k) \leq Cv_n^{-1}r_n^2.$$

The proof now follows the same arguments as for the type I mixture.

Remark 6. The above theorems state that when testing the point null hypothesis, p_0 is uniform, against the nonparametric alternative, the inverse Bayes factor B_n^{-1} converges to zero in probability under the null at a rate of $n^{-1/2}$. In order to increase the rate at

which the Bayes factor converges to zero it is necessary to restrict the amount of prior probability placed near the uniform distribution.

Remark 7. Another application of the results of this section is when the null hypothesis is the singleton $\{p_0\}$, with p_0 any density, not necessarily uniform. In this case one can apply the cumulative distribution function transform of p_0 to transform the data. This brings us back to the case where the null hypothesis is the uniform on $[0, 1]$. The effect is similar to the embedding of a parametric family in the nonparametric model. Note that the rate of convergence of the marginal likelihood only applies to the given p_0.

5. Discussion

This paper has provided sufficient conditions for the consistency of the Bayes factor in testing goodness of fit of a parametric density function and the conditions are verified for the mixture of triangular distributions prior. We have also shown that if these conditions are not satisfied then the Bayes factor may be inconsistent. These results complement the result of Rousseau (2008) who has given a necessary condition for consistency of the Bayes factor. We believe the study of consistency is an important issue in goodness of fit and other testing problems. Not only does it provide a frequentist validation of the Bayesian procedure, but it also aids our understanding of how intergrating over the parameter space accounts for parameter uncertainty.

Mixture of triangular distributions can be useful priors since they have good theoretical properties and are simple to implement. An interesting problem for further study is to establish conditions for consistency of the Bayes factor which are both necessary and sufficient. We believe that the conditions presented here are close to being so. Another problem of interest is to consider goodness of fit testing in other contexts such as regression. Using the results in Ghosal & van der Vaart (2006) it may be possible to establish a result similar to our theorem 1. As before, the main challenge in applying such a result would be to determine appropriate upper bounds on the prior probability in a neighbourhood of the parametric model. Both of these problems deserve further consideration.

Acknowledgements

This work was supported by the ARC Centre for Complex Dynamic Systems and Control. We thank the Editor, Associate Editor and two referees for their careful reading and helpful suggestions.

References

Arcones, M.A. (2002) Moderate deviations for M-estimators. *TEST* **11**, 465-500.

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

Carota, C. & Paramigiani, G. (1996) On Bayes factors for nonparametric alternatives. In *Bayesian statistics 5* (eds. Bernardo, J.M., Berger, J.O., Dawid, A.P. & Smith, A.F.M.), 507-511. Oxford University Press, USA.

Dass, S.C. & Lee, J. (2004) A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives. *J. Statist. Plann. Inference* **119**, 143-152.

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

Gelfand, A.E. & Dey, D.K. (1994) Bayesian model choice: Asymptotics and exact calculations. *J. R. Stat. Soc. ser. B Stat. Methodol.* **56**, 501-514.

Feller, W. (1939) Interpolation of completely monotone functions. *Duke Math. J.* **5**, 661-674.

Ghosal, S. (2001) Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.* **29**, 1264-1280.

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

Ghosal, S., Lember, J. & van der Vaart, A.W. (2008) Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.* **2**, 63-89.

Ghosal, S. & van der Vaart, A.W. (2006) Convergence rates of posterior distributions for noniid observations. *Ann. Statist.* **35**, 192-223.

Ghosh, J.K. & Ramamoorthi, R.V. (2003) *Bayesian nonparametrics*, Springer, New York.

Kass, R.E., Tierney, L. & Kadane, J.B. (1990) The validity of posterior expansions based on Laplace's method. In *Bayesian and likelihood methods in statistics and econometrics*

(eds. Geisser, S., Hodges, J.S., Press, S.J., & Zellner, A.) Elsevier Science Publishers, The Netherlands.

McVinish, R., Rousseau, J. & Mengersen, K. (2005) Bayesian mixtures of triangular distributions with application to goodness of fit testing. *Les cahiers du CEREMADE* (2005-31). http://www.ceremade.dauphine.fr/preprints/CMD/2005-31.pdf

Perron, F. & Mengersen, K. (2001) Bayesian nonparametric modeling using mixtures of triangular distributions. *Biometrics* **57**, 518-528.

Petrone, S. & Wasserman, L. (2002) Consistency of Bernstein polynomial posteriors. *J. R. Stat. Soc. ser. B Stat. Methodol.* **64**, 79-100.

Robert, C.P. & Rousseau, J. (2004) A mixture approach to Bayesian goodness of fit. *Les cahiers du CEREMADE* (2005-31). http://www.ceremade.dauphine.fr/preprints/CMD/2002-9.ps.gz

Rousseau, J. (2008) Approximating Interval hypothesis : p-values and Bayes factors. In *Bayesian statistics 8* (eds. Bernardo, J.M., Berger, J.O., Dawid, A.P. & Smith, A.F.M.), 417-452. Oxford University Press, USA.

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

Verdinelli, I. & Wasserman, L. (1998) Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.* **26**, 1215-1241.

Walker, S., Damien, P. & Lenk, P. (2004) On priors with a Kullback-Leibler property. *J. Amer. Statist. Assoc.* **99**, 404-408.

Corresponding author's address (R. McVinish): School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane Q4001, Australia.
E-mail: r.mcvinish@qut.edu.au

Appendix

(i) *Proof of theorem 2.* The proof follows similar lines to theorem 2.3 of Ghosal (2001) and so we shall only provide the lower bound for the prior probability on the set

$$\left\{p : K(p_*, p) < \tilde{\epsilon}_n^2, V(p_*, p) < \tilde{\epsilon}_n^2\right\}.$$

If p_* is bounded away from zero then the proof is essentially the same as for theorem 2.3 of Ghosal (2001). Therefore, we shall assume $p_*(0) = p_*(1) = 0$. From

lemma 8.3 of Ghosal *et al.* (2000) we need only bound the prior probability on the set $\{p : h^2(p_*, p) \parallel p_*/p \parallel_\infty < \tilde{\epsilon}_n^2\}$. Define $p_*^i = p_*(i/k) \vee k^{-\beta}$ for $i = 0, \ldots, k$. Consider the set of densities $N(k, \epsilon; p_*)$ defined by

$$p\left(x\ ; w(k)\right) = S^{-1}(p^i(1 - k(x - i/k)) + p^{i+1}k(x - i/k)), \quad x \in [i/k, (i+1)/k],$$

where $S = (p^0 + p^k)/(2k) + \sum_{i=1}^{k-1} p^i/k$ and $|p^i - p_*^i| \leq C p_*^i \epsilon$. It is seen that these densities are type I mixtures of triangular densities. Let $p(\cdot; w_0(k))$ denote the density where $p^i = p_*^i$. It is seen that $S \cdot p(x; w_0(k))$ is, with minor modification near $x = 0$ and $x = 1$, the linear interpolation of p_* so $\sup_{0 \leq x \leq 1} |p_*\left(x\right) - S \cdot p\left(x; w_0(k)\right)| \leq C k^{-\beta}$. In this case $S = 1 + O(k^{-\beta})$ and $\sup_{0 \leq x \leq 1} |p_*\left(x\right) - p\left(x; w_0(k)\right)| \leq C k^{-\beta}$. For densities on $N(k, \epsilon; p_*)$ we have $S = 1 + O(k^{-\beta} + \epsilon)$ and $\sup_{0 \leq x \leq 1} |p_*\left(x\right) - p\left(x; w(k)\right)| \leq C\left(k^{-\beta} + \epsilon\right)$.

Now we determine a bound on the Hellinger distance between p_* and $p(\cdot; w(k))$. The squared Hellinger distance is bounded by

$$h^2(p_*, p(\cdot; w(k))) \leq \int_0^1 \frac{\left(p\left(x; w(k)\right) - p_*(x)\right)^2}{p\left(x; w(k)\right)} dx.$$

The range of integration is divided into small intervals to obtain the bound. Assuming $\epsilon < k^{-\beta}(\log k)^{-1}$,

$$\begin{aligned}
&\int_0^{k^{-1}} \frac{\left(p\left(x; w(k)\right) - p_*(x)\right)^2}{p\left(x; w(k)\right)} dx \\
&= \int_0^{k^{-1}} \frac{\left(p_*(x) - p^0(1 - kx) - p^1 kx\right)^2}{p^0(1 - kx) + p^1 kx} dx, \\
&\leq \int_0^{k^{-1}} \frac{\left(p_*(x) - p_*^1 kx + O(\epsilon + k^{-\beta})\right)^2}{p^0(1 - kx) + p^1 kx} dx, \\
&\leq C \int_0^{k^{-1}} \frac{\left(\epsilon + k^{-\beta}\right)^2}{p^0(1 - kx) + p^1 kx} dx, \\
&= C(k^{-\beta} + \epsilon)^2 \frac{\log p^1 - \log p^0}{k(p^1 - p^0)},
\end{aligned}$$

with the convention that $(\log x - \log y)/(x - y) = x^{-1}$ when $x = y$. The second last inequality is obtained by taking Taylor expansions for $p_*(x)$ and $p_*(1/k)$ around zero. Noting that if $\beta < 1$, then $p^0 = p^1 = k^{-\beta}$ and if $\beta \geq 1$, then $p^0 = k^{-\beta}$ and $p^1 = p(1/k)$ we obtain that the above term is bounded by a constant times $k^{-2\beta} \log k$. The integral

over $[1-1/k, 1]$ can be bounded in a similar manner. Finally

$$\begin{aligned}\int_{k^{-1}}^{1-k^{-1}} \frac{\left(p\left(x; w(k)\right) - p_*(x)\right)^2}{p\left(x; w(k)\right)} dx &\leq C(k^{-\beta}+\epsilon)^2 \int_{k^{-1}}^{1-k^{-1}} p\left(x; w(k)\right)^{-1} dx \\ &\leq C(k^{-\beta}+\epsilon)^2 k^{-1} \sum_{i=1}^{k-1} \frac{1}{p^i} \leq C(k^{-\beta}+\epsilon)^2 \log k.\end{aligned}$$

Hence, for densities in $N(k, \epsilon; p_*)$,

$$h^2(p_*, p(\cdot; w(k))) \leq C(k^{-\beta}+\epsilon)^2 \log k.$$

For $x \in (i/k, (i+1)/k)$ we can bound $p_*(x)/p(x)$ for all $p \in N(k, \epsilon; p_*)$ as

$$\frac{p_*(x)}{p(x)} \leq \frac{p_*(x)}{p^i \wedge p^{i+1}} \leq \frac{2p_*(x)}{(k^{-\beta} \vee p_*(i/k)) \wedge (k^{-\beta} \vee p_*((i+1)/k))}.$$

Taking a Taylor expansion of $p_*(x)$ around $x = i/k$ and $x = (i+1)/k$ it is seen that $\| p_*/p \|_\infty < C$ for k sufficiently large. The prior probability on $N(k, \epsilon; p_*)$ can be bounded below using lemma A.1 of Ghosal (2001) and the fact that $\pi(k = k_n) \geq e^{-n\epsilon^2}$ for $c_1 n^{1/(2\beta+1)} < k_n < c_2 n^{1/(2\beta+1)}$ and n sufficiently large. The remainder of the proof follows theorem 2.3 of Ghosal (2001) to give convergence in probability. To bound the expectation we follow the proof of theorem 2.2 in Ghosal *et al.* (2000) (also lemma 8.4) so that

$$P_*^\infty \left\{ \Pi\left(p : d\left(p, p_*\right) > Rn^{-\beta/(2\beta+1)} (\log n)^{(4\beta+1)/4\beta} \mid Y^n\right) > \exp(-B_1 n\epsilon_n^2) \right\} < \exp(-B_2 n\epsilon_n^2).$$

(ii) *Proof of theorem 3.*

This result is proved by verifying that the conditions of theorem 2.1 in Ghosal (2001) hold. The first step is to provide a lower bound on the prior probability for the set of densities

$$N(\epsilon, p_*) = \left\{ p : K(p_*, p) < \epsilon^2, V(p_*, p) < \epsilon^2 \right\}.$$

For a given k we take $\psi(k) = (t_0, t_1, ..., t_k)$ where $P_*(t_i) = i/k,\ i = 0, \ldots, k$. From the mean value theorem

$$P_*\left(t_{i+1}\right) - P_*\left(t_{i-1}\right) = 2/k = p_*\left(t_i^*\right)\left(t_{i+1} - t_{i-1}\right), \quad t_i^* \in \left(t_{i-1}, t_{i+1}\right),$$

for $i = 1, \ldots, k-1$. For $i = 0, k$

$$\begin{aligned}P_*\left(t_1\right) - P_*\left(t_0\right) &= 1/k = p_*\left(t_0^*\right)\left(t_1 - t_0\right), \quad t_0^* \in \left(t_1, t_0\right), \\ P_*\left(t_k\right) - P_*\left(t_{k-1}\right) &= 1/k = p_*\left(t_k^*\right)\left(t_k - t_{k-1}\right), \quad t_k^* \in \left(t_{k-1}, t_k\right).\end{aligned}$$

It follows that $p(\cdot;\psi(k))$ is the linear interpolation of the points $(t_i, p_*(t_i^*))$, $i = 0, \ldots, k$. It is noted that $(Mk)^{-1} \leq |t_{i+1} - t_i| \leq (ak)^{-1}$, where $M = \sup p_*(x)$. If $\beta \leq 1$, we obtain that

$$\sup_{x\in[0,1]} |p_*(x) - p(x;\psi(k))| \leq Ck^{-\beta}, \quad \text{if } p_* \geq a > 0.$$

If $\beta \in (1, 2]$, we note that for $i = 0, \ldots, k-1$,

$$p_*(x) = p_*(t_i) + \frac{(x - t_i)}{t_{i+1} - t_i}(p_*(t_{i+1}) - p_*(t_i)) + O(k^{-\beta}).$$

Therefore, for $x \in (t_i, t_{i+1})$

$$\begin{aligned}
&\sup_{x\in[0,1]} |p_*(x) - p(x;\psi(k))| \\
&= \left|(p_*(t_i) - p_*(t_i^*))\frac{(t_{i+1} - x)}{t_{i+1} - t_i} + (p_*(t_{i+1}) - p_*(t_{i+1}^*))\frac{(x - t_i)}{t_{i+1} - t_i}\right| + O(k^{-\beta}).
\end{aligned}$$

Using a Taylor expansion of $\int_{t_i}^{t_{i+1}} p_*(x)dx$ and of $\int_{t_{i-1}}^{t_i} p_*(x)dx$, both equal to $1/k$:

$$\begin{aligned}
&p_*(t_i)[(t_{i+1} - t_i) - (t_i - t_{i-1})] \\
&= -p_*'(t_i)[(t_{i+1} - t_i)^2 + (t_i - t_{i-1})^2]/2 + O(k^{-\beta-1}) \\
&= O(k^{-2}).
\end{aligned}$$

Similarly, using a Taylor expansion of $\int_{t_{i-1}}^{t_{i+1}} p_*(x)dx$,

$$\begin{aligned}
&p_*(t_i^*)(t_{i+1} - t_{i-1}) \\
&= p_*(t_i)(t_{i+1} - t_{i-1}) + \frac{p_*(t_i)'}{2}[(t_{i+1} - t_i)^2 - (t_i - t_{i-1})^2] + O(k^{-\beta-1}).
\end{aligned}$$

Together these equations imply that

$$p_*(t_i) - p_*(t_i^*) = -p_*'(t_i)/2[(t_{i+1} - t_i) - (t_i - t_{i-1})] + O(k^{-\beta}) = O(k^{-\beta}).$$

The same argument can be applied to show $p_*(t_i) - p_*(t_i^*) = O(k^{-\beta})$. Therefore, the absolute difference of p_* and $p(\cdot;\psi(k))$ is bounded by $Ck^{-\beta}$ for some $C > 0$. Let $\eta(k)$ be another partition and $p(\cdot;\eta(k))$ the resulting density. From lemma 1 we have

$$|p_*(x) - p(x;\eta(k))| \leq C(k^{-\beta} + \epsilon),$$

where $|\psi(k) - \eta(k)| < c\epsilon k^{-1}$. Taking k to satisfy $d_1\epsilon^{-1} < k^\beta < d_2\epsilon^{-1}$ for constants d_1, d_2 and applying lemma 8.2 of Ghosal *et al.* (2000) it is seen that

$$N(C\epsilon, p_*) \supset \left\{\eta(k) : |\eta(k) - \psi(k)| < c\epsilon k^{-1}\right\},$$

which leads to the correct lower bound for the probability of the set $N(C\epsilon, p_*)$.

Now define the sets $\mathcal{G}_n = \{\psi(k) : k \leq k_0 n^{1/(2\beta+1)}, |t_j - t_{j+1}| \geq e^{-\alpha n^\gamma}, j \leq k-1\}$. An upper bound on the entropy of $\mathcal{G}_n$ now needs to be determined. We follow Ghosal (2001) with the remark that

$$\|p(\cdot;\psi(k)) - p(\cdot;\eta(k))\|_1 \leq C\epsilon_0\epsilon_n^* \log n,$$

as soon as the partitions $\psi(k) = (t_j, j = 0, ..., k)$ and $\eta(k) = (\tilde{t}_j, j = 0, ..., k)$ with $\tilde{t}_0 = t_0 = 0$ and $\tilde{t}_k = t_k = 1$ satisfy

$$|t_j - \tilde{t}_j| \leq \epsilon_0\epsilon_n^* \log n(|t_j - t_{j-1}| \wedge |t_j - t_{j+1}|).$$

Therefore, the number of points in the net defined by the above constraint is bounded by

$$N_n \;=\; k_n \left(\frac{C\alpha n^\gamma}{\epsilon_0\epsilon_n^* \log n}\right)^{k_n} \leq e^{k_0 C n^{1/(2\beta+1)} \log n},$$

with $k_n = k_0 n^{1/(2\beta+1)}$ and C some generic constant. Finally, the bound on the prior probability for the set $\mathcal{G}_n^c$ is easily determined. These bounds can be used to verify that the conditions of theorem 2.1 of Ghosal (2001) hold and so the rate of convergence in probability is found. The bound in expectation can be obtained by similar arguments as used in the proof of Equation (2).

(iii) A few lemmas that are considered in the study of the mixtures of triangular distributions. The proofs of the following lemmas can be found in McVinish *et al.* (2005) or are small modifications of these as indicated.

Lemma 1. *Let $p(\cdot;\psi(k))$ and $p(\cdot;\eta(k))$ be two type II mixtures of triangular densities where $\psi(k) = (t_0, t_1, \ldots, t_{k-1}, t_k)$ and $\eta(k) = \left(t_0, \tilde{t}_1, \ldots, \tilde{t}_{k-1}, t_k\right)$. Let C denote a constant that is independent of the partitions and k. For $0 < \epsilon < 1/4$, if*

$$\max_i \left|t_i - \tilde{t}_i\right| \leq \frac{\epsilon}{Mk}, \quad M = \sup_{x\in[0,1]} p(x;\psi(k)),$$

then

$$\sup_{x\in[0,1]} |p(x;\psi(k)) - p(x;\eta(k))| \leq CM\epsilon.$$

Furthermore, if

$$\max_i \left|t_i - \tilde{t}_i\right| \leq \epsilon(|t_i - t_{i-1}| \wedge |t_i - t_{i+1}|),$$

then

$$\|p(\cdot;\psi(k)) - p(\cdot;\eta(k))\|_1 \leq C\epsilon.$$

Proof. Let $\psi_j = \left(t_0, \tilde{t}_1, \ldots, \tilde{t}_j, t_{j+1}, \ldots, t_k\right)$, for $j = 0, ..., k-1$ then

$$\|p(\cdot;\psi(k)) - p(\cdot;\eta(k))\|_1 \quad \leq \quad \sum_{j=0}^{k-1} \|p(\cdot;\psi_j) - p(\cdot;\psi_{j+1})\|_1,$$

and the difference between the two consecutive functions occurs only on the interval $[\tilde{t}_{j-2}, t_{j+2}]$. Now if $|t_j - \tilde{t}_j| \leq \epsilon(|t_j - t_{j-1}| \wedge |t_j - t_{j+1}|)$ for all j, $|t_j - \tilde{t}_j| \leq (1/3|t_j - \tilde{t}_{j-1}| \wedge |t_j - t_{j+1}|)$, so that $\tilde{t}_j \in (\tilde{t}_{j-1}, t_{j+1})$. Using tedious but straightforward calculations, we obtain that

$$\|p(\cdot;\psi_j) - p(\cdot;\psi_{j+1})\|_1 \leq \frac{C|t_j - \tilde{t}_j|}{k}\left[\frac{1}{t_{j+1} - \tilde{t}_{j-1}} + \frac{1}{t_j - \tilde{t}_{j-2}} + \frac{1}{t_{j+2} - t_j}\right].$$

Applying the triangle inequality and summing over j we have

$$\|p(\cdot;\psi(k)) - p(\cdot;\eta(k))\|_1 \quad \leq \quad C\frac{2\epsilon}{1-\epsilon}.$$

Lemma 2 (type I mixtures)**.** *Let p_0 be the uniform density. For each $k \geq 2$, if*

$$\| \, p_0 - p(\cdot; w(k)) \, \|_1 \leq \epsilon,$$

then

$$\sum_{j=1}^{k-1} |w_j - 1/k| + |w_0 - 1/(2k)| + |w_k - 1/(2k)| \leq 4\epsilon.$$

Lemma 3 (type II mixtures)**.** *Let p_0 be the uniform density. For each $k \geq 2$, if*

$$\| \, p_0 - p(\cdot; \psi(k)) \, \|_1 \leq \epsilon,$$

then

$$\sum_{j=1}^{k-2} \frac{(t_{j+1} - t_j)}{4}\left(\frac{|(t_{j+2} - t_j) - 2/k|}{(t_{j+2} - t_j)} + \frac{|(t_{j+1} - t_{j-1}) - 2/k|}{(t_{j+1} - t_{j-1})}\right)$$
$$+\frac{1}{2kt_2}\,|t_1 - t_2/2| + \frac{1}{2k(1 - t_{k-2})}\,|(1 - t_{k-1}) - (1 - t_{k-2})/2| \leq \epsilon.$$

Lemma 4. *Let $I_2 = \int_{V_n^c \cap \mathcal{G}_n} p_\eta(Y^n) d\pi(\eta)$, where $V_n = \cup_k V_{n,k}$, then*

$$P_0^n \left[I_2 \geq v_n\right] \quad \leq \quad n^{-H}$$

for all $H > 0$ and n sufficiently large.

Proof. The proof is based on the construction of tests based on the L_1 distance as in Ghosal *et al.* (2000).