BAYESIAN MIXTURES OF TRIANGULAR DISTRIBUTIONS WITH APPLICATION TO GOODNESS-OF-FIT TESTING

R.MCVINISH, J. ROUSSEAU, AND K. MENGERSEN

Abstract. Two forms of mixtures of triangular distributions are considered as an alternative to the Bernstein polynomials for Bayesian nonparametric density estimation on $[0,1]$. Conditions for weak and strong consistency of the posterior distribution and a rate of convergence are established. This class of priors is applied to the problem of testing a parametric family against a nonparametric alternative.

1. Introduction

The Bernstein polynomial prior for nonparametric density estimation on $[0,1]$ has been studied by Petrone (1999a,b), Ghoshal (2001) and Petrone and Wasserman (2002). Perron and Mengersen (2001) have shown that the Bernstein polynomials are a poor approximation to the space of distribution functions on $[0,1]$. A better approximation can be achieved by mixtures of Triangular distributions.

For a given $k \geq 1$, let the sequence $0 = x_{-1} = x_0 < x_1 < \ldots < x_{k-1} < x_k = x_{k+1} = 1$ be a partition of $[0,1]$. The function $h_i(x)$ is the triangular density function with support on the interval $[x_{i-1}, x_{i+1}]$ and mode at x_i. The mixture of triangular distributions has the density function

$$p(x) = \sum_{i=0}^{k} w_i h_i(x),$$

where $w_i \geq 0$, $\sum_{i=0}^{k} w_i = 1$. Note that $p(x)$ is a piecewise linear function on $[0,1]$ interpolating the points $(x_i, w_i h_i(x_i))$. The two cases considered by Perron and Mengersen (2001) were:

Date: June 14, 2005.

1991 *Mathematics Subject Classification.* Primary: 62C10 Secondary: 62G07, 62F15.

Key words and phrases. Bayesian nonparametrics, consistency, goodness-of-fit.

This work was supported by the ARC Center of Excellence - Center for Complex Dynamical Systems and Control CEO348165.

I For each k, the partition of $[0,1]$ is assumed fixed and the weights w_i are to be estimated.

II For each k, the weights w_i are fixed at $w_0 = w_k = \frac{1}{2k}$, $w_i = \frac{1}{k}$, $i = 1, \ldots, k-1$ and the partition is to be estimated.

The advantage of these Triangular distributions as priors is that they are both simple and flexible. In particular their implementation is quite straightforward, see Perron and Mengersen (2001). On the other hand they provide a good approximation to smooth densities, so that the posteriors associated with these priors have good frequentist asymptotic properties. We prove in particular that they lead to minimax estimators of the densities for some class of densities, see Section 2.

In this paper we study the asymptotic properties of the posterior distributions associated with these priors, which leads to the asymptotic properties of the related Bayes estimators. We also construct Goodness of fit tests using these priors. We first study the asymptotic behaviour of the Bayes factor since it is widely used as a test statistic in the Bayesian community. To begin with we consider the simple test against a fixed known distribution, which is equivalent to testing against the uniform. In this case consistency of the Bayes factor is proved under very general conditions and rates of convergence are obtained which are as anticipated: $n^{-1/2}$ under the null hypothesis and exponential rates for fixed alternatives. These results are stated in Section 3. We then extend our results to testing against a parametric model $\mathcal{Q} = \{p_\theta, \theta \in \Theta\}$. This is a widely studied problem, see for example Zhang (2002), Fortiana and Grane (2003), Munk and Czado (1998) and references therein. Here instead of embedding the parametric family into a nonparametric family as was done by Florens, Richard and Rolin (1996), Carota and Parmigiani (1996), Berger and Guglielmi (2001) and Verdinelli and Wasserman (1998) we consider the test of the parametric family (which is considered as a distribution on $[0,1]$ without loss of generality) against the nonparametric model defined by the mixture of triangular densities. This has the advantage of substantial simplification of the models but we show that it has limitations with respect to the Bayes factor. Indeed, we prove that if both models are *separated enough*, in other words if no density of the parametric family can be represented exactly by a finite mixture of triangular density, then the Bayes factor is consistent; otherwise it is not consistent. Bayes factors are to some extent the Bayesian answers to 0-1 types of losses. Goodness of fit tests are seldom phrased as such; the question is rather whether the true density can be reasonably

well represented by a parametric family of densities. Therefore, we also propose another test based on a distance approach, similar to that considered by Robert and Rousseau (2004). This is the Bayesian solution to the loss function function defined by, for any density p on $[0,1]$

$$L(\delta,p) \;=\; \begin{cases} a_0 d(p,\mathcal{Q}) & \text{if} \quad \delta=0 \\ a_1(1-d(p,\mathcal{Q})) & \text{if} \quad \delta=1 \end{cases} \tag{1.1}$$

where $d(p,\mathcal{Q}) = \inf_{\theta\in\Theta} d(p,p_\theta)$ and $d(p,p_\theta) = \int_0^1 |p(x)-p_\theta(x)|dx$. The Bayes estimator is then given by $\delta(Y^n)=0$ (we accept the parametric family) if and only if

$$T(Y^n) = \mathbb{E}^\pi[d(p,\mathcal{F})|Y^n] \leq a_1/(a_0+a_1),$$

where $Y^n=(Y_1,...,Y_n)$ is the vector of observations. Since the choice of (a_0,a_1) is often arbitrary we propose, as in Robert and Rousseau (2004), to calibrate the test procedure using a Bayesian p-value, namely the conditional predictive p-value p_{cpred} based on the maximum likelihood estimator under the parametric family: let $\hat\theta$ be the maximum likelihood estimator based on the observations Y^n

$$p_{\text{cpred}}(\hat\theta,Y^n) = \int_\Theta P_\theta\left[T(X^n)>T(Y^n)|\,T(Y^n),\hat\theta\right]\pi_0(\theta|\hat\theta)d\theta.$$

This p-value has been proposed by Bayarri and Berger (2000) and discussed by Robbins *et al.* (2000), Robert and Rousseau (2004) and Fraser and Rousseau (2005). In Section 4.3 we study the asymptotic behaviour of this test procedure and we prove that it is also optimal in the frequentist sense.

The following section states the consistency results for these mixtures. Since the methods of proof for the first case of mixtures are very similar to that of the Bernstein polynomials, the proofs will be omitted or sketched. For both cases of mixtures the resulting density estimates can be shown to converge at the minimax rate for Hölder continuous density functions, up to a $(\log n)$ term. This rate of convergence is significantly faster than for the Bernstein polynomials.

2. Asymptotic properties of the posterior distributions

We first give some definitions that will be used throughout this paper. 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. The Kullback-Leibler divergence is defined by $K(p,q) = \int \log(p/q)p\, d\mu$ and let $V(p,q) = \int \log(p/q)^2 p\, d\mu$. The space of densities

will be denoted by $\mathcal{F}$. The prior on $\mathcal{F}$ is denoted by π and the posterior distribution given data $Y^n = (Y_1, \ldots, Y_n)$ is denoted by $\pi(\cdot \mid Y^n)$. The true distribution function is denoted by P_0 and its density by p_0.

2.1. **Type I mixture.** For the Type I mixture of triangular distributions the partition on $[0,1]$ is defined as $x_i = i/k,\ i = 0, 1, \ldots, k$. This choice of partition is made for simplicity. Alternative partitions could be more suitable to different density functions. The prior for these mixtures is assumed to satisfy the following basic assumptions:

- For all $k = 1, 2, \ldots\ \pi(k) > 0$
- Given k, the support of the prior on $\mathbf{w} = (w_0, \ldots, w_k)$ is the entire k-dimensional simplex. Hence, for any set U of positive Lebesgue measure $\pi(U \mid k) > 0$.

The aim of this subsection is to give sufficient conditions on the prior leading to a given rate of convergence for the posterior. This result will play a significant role in the later discussions on goodness-of-fit testing. In the process we shall recall some theorems used in proving convergence of the posterior. As with the Bernstein polynomial priors, for a given k, this first class of mixtures is a simple convex combination of density functions which are bounded by a multiple of k. Thus the following theorems can be proved using very similar techniques to those used in the corresponding theorems for Bernstein polynomials(Petrone and Wasserman (2002), Ghosal (2001)). Hence the proofs of this section will either be omitted or sketched.

Consider first weak consistency of the posterior distribution, that is the posterior probability of all weak neighbourhoods of the true density converges to one, almost surely. Since this is sufficient for the posterior predictive distribution function to converge weakly to the true distribution function. Hjort (2003) argues that this is sufficient for most inferential questions. Schwartz (1965) established the following sufficient condition for weak consistency of the posterior. See Walker (2003) for a different proof.

Theorem 2.1 (Schwartz (1965))**.** *If the prior places positive probability on every Kullback-Leibler neighbourhood of p_0 then the posterior is weakly consistent at p_0, almost surely.*

The proof of the following theorem requires only very minor modifications to the proof of Theorem 2 in Petrone and Wasserman (2002).

Theorem 2.2. *Suppose P_0 has a continuous density p_0 on $[0,1]$. If the prior satisfies the above assumption then the resulting posterior is weakly consistent at P_0.*

Often we are interested not only in functionals of the posterior predictive density but also in an estimate of the density function. It is well known that weak convergence of a distribution does not imply convergence of the respective density function without additional conditions on the prior. For the posterior predictive density to converge to the true density, with respect to the L_1 or Hellinger distances, it is sufficient to show that the posterior probability of all strong (i.e. Hellinger) neighbourhoods of the true density converge to one, almost surely (Ghosh and Ramamoorthi (2003), proposition 4.2.1). Sufficient conditions for strong consistency of the posterior distribution have been established by Baron *et al.* (1999), Ghosh and Ramamoorthi (2003) and Walker (2004). For a subset $\mathcal{A}$ of $\mathcal{F}$ and $\delta > 0$ let $D(\delta, \mathcal{A}, d)$ denote the minimum number of points in $\mathcal{A}$ such that the distance between each pair is no greater than δ, sometimes called the δ-covering number.

Theorem 2.3 (Ghosh and Ramamoorthi (2003), Theorem 4.4.4)**.** *Let π be a prior on $\mathcal{F}$. Suppose the conditions of Theorem 2.1 are satisfied. If for each $\epsilon > 0$, there is a $\delta < \epsilon, c_1, c_2 > 0, \beta < \epsilon^2/2$, and $\mathcal{F}_n \subset \mathcal{F}$ such that, for all n large,*

- $\pi(\mathcal{F}_n^c) < c_1 e^{-nc_2}$,
- $\log D(\delta, \mathcal{F}_n, |\cdot|) < n\beta$,

then the posterior is strongly consistent at p_0.

The following theorem may be proved by a minor modification of the proof of theorem 3 in Petrone and Wasserman (2002).

Theorem 2.4. *Suppose there exists $k_n \longrightarrow \infty$ such that $k_n = O(n)$ and $\sum_{k \geq k_n} \pi(k) \leq \exp(-nr)$ for some $r > 0$. Further suppose that the assumptions of Theorem 2.2 hold. Then the resulting posterior is Hellinger consistent at P_0.*

Ghosal (2001) showed that under certain conditions the rate of convergence of the posterior distribution from a Bernstein polynomial prior is of the order $n^{-1/3}(\log n)^{5/6}$ provided the true density is bounded away from zero and has a bounded second derivative. This rate is much slower than can be achieved by other methods, though is believed to be sharp except for the $(\log n)$ term. Theorem 2.3 in Ghosal (2001) can be adapted to determine the rate of convergence of the posterior

for mixtures of triangulars. The conditions that need to be checked are given by Ghosal's Theorem 2.1 which is stated below.

Theorem 2.5 (Ghosal (2001) Theorem 2.1)**.** *Let π_n be a sequence of priors on $\mathcal{F}$. Suppose that for positive sequences $\bar{\epsilon}_n, \tilde{\epsilon}_n \longrightarrow 0$ with $n \min\left(\bar{\epsilon}_n^2, \tilde{\epsilon}_n^2\right) \longrightarrow \infty$, constants $c_1, \ldots, c_4 > 0$ and sets $\mathcal{F}_n \subset \mathcal{F}$, we have*

$$\log D\left(\tilde{\epsilon}_n, \mathcal{F}_n, d\right) \leq c_1 n \bar{\epsilon}_n^2, \tag{2.1}$$

$$\pi_n\left(\mathcal{F}_n^c\right) \leq c_3 e^{-(c_2+4)n\tilde{\epsilon}_n^2}, \tag{2.2}$$

$$\pi_n\left(N\left(\tilde{\epsilon}_n, p_0\right)\right) \geq c_4 e^{-c_2 n\tilde{\epsilon}_n^2}, \tag{2.3}$$

where $N\left(\epsilon, p_0\right) = \left\{p : K\left(p_0, p\right) < \epsilon^2, V\left(p_0, p\right) < \epsilon^2\right\}$. Then for $\epsilon_n = \max\left(\bar{\epsilon}_n, \tilde{\epsilon}_n\right)$ and a sufficiently large $M > 0$, the posterior probability

$$\pi_n\left(p : \ d\left(p, p_0\right) > M\epsilon_n \mid Y^n\right) \longrightarrow 0, \tag{2.4}$$

in P_0^n-probability.

Theorem 2.6. *Assume that p_0 belongs to the Hölder class $\mathcal{H}(L, \beta)$, with $\beta, L > 0$ and satisfies $p_0(x) \geq ax(1-x)$ for some constant $a > 0$. Assume the Type I mixture of triangular distributions prior for p satisfies $b_1 e^{-\beta_1 k} \leq \pi\left(k\right) \leq b_2 e^{-\beta_2 k}$ for all k for some constants $b_1,\ b_2,\ \beta_1,\ \beta_2 > 0$ and for each k the prior on $\mathbf{w}$ is a Dirichlet distribution with parameters bounded for all k. Then for a sufficiently large constant M,*

- *If $\beta \leq 2$ then*

$$E_0[\Pi\left(p : d\left(p, p_0\right) > Mn^{-\beta/(2\beta+1)}(\log n) \mid Y^n\right)] \leq n^{-H}, \quad \forall H > 0,$$

  when n is large enough.

- *If $\beta > 2$ then*

$$E_0\left[\Pi\left(p : d\left(p, p_0\right) > Mn^{-2/5}(\log n) \mid Y^n\right)\right] \leq n^{-H}, \quad \forall H > 0,$$

  when n is large enough.

Proof. Assume initially that $\beta \leq 2$; the case of $\beta > 2$ follows exactly the same lines. Let C denote a generic finite, positive constant which may be different in each instance. First, define the function $p\left(x; \hat{\mathbf{w}}_0, k\right)$ by

$$p\left(x; \hat{\mathbf{w}}_0, k\right) = p_i^0(1 - k(x - i/k)) + p_{i+1}^0 k(x - i/k), \quad \forall x \in [i/k, (i+1)/k],$$

where $p_i^0 = p_0(i/k)$ for $k-1 \geq i \geq 1$ and if $p_0(0) = 0$, (resp. $p_0(1) = 0$) $p_0^0 = k^{-\beta} \wedge p_0(1/k)$ else $p_0^0 = P_0(0)$ (resp. $p_0^k = k^{-\beta} \wedge p_0(1-1/k)$ else $p_0^k = p_0(1)$). It is easily seen that the density defined by $p(x; \mathbf{w}_0, k) = p(x; \hat{\mathbf{w}}_0, k)/S$, with $S = (p_0 + p_k)/(2k) + \sum_{i=1}^{k-1} p_i/k$ is a mixture of triangular densities. The function $p(x; \hat{\mathbf{w}}_0, k)$ is, with minor modification at $x = 0$ and $x = 1$, simply the linear interpolation of p_0 so $\sup_{0\leq x\leq 1} |p_0(x) - p(x; \hat{\mathbf{w}}_0, k)| \leq Ck^{-\beta}$. It follows that $S = 1 + O\left(k^{-\beta}\right)$ and $\sup_{0\leq x\leq 1} |p_0(x) - p(x; \mathbf{w}_0, k)| \leq Ck^{-\beta}$. Then the Kullback-Leibler distance between p_0 and p_w is bounded by

$$K(p_0, p(.; \mathbf{w}_0, k)) \leq \int \frac{(p(x; \mathbf{w}_0, k) - p_0(x))^2}{p(x; \mathbf{w}_0, k)} dx.$$

This implies that

$$\begin{aligned} K(p_0, p(.; \mathbf{w}_0, k)) &\leq (1-S)^2 + Ck^{-2\beta-1} \sum_{i=1}^{k-2} p(x; \hat{\mathbf{w}}_0, k)^{-1} \\ &\quad + Ck^{-2\beta-1} \left(\int_0^1 \frac{u^2}{(p_0^1 + up_0^0)} du + \int_0^1 \frac{u^2}{(p_0^{k-1} + up_0^k)} du \right) \\ &\leq (1-S)^2 + Ck^{-2\beta} \log k. \end{aligned}$$

Let $p(x; \mathbf{w}, k)$ be another mixture of triangular densities, then

$$\begin{aligned} |p(x; \mathbf{w}_0, k) - p(x; \mathbf{w}, k)| &\leq 2k \max_{1\leq j\leq k} |w_{0,j} - w_j| \\ &\leq 2k \sum_{j=1}^{k} |w_{0,j} - w_j| \end{aligned}$$

Therefore, if $\| \mathbf{w}_0 - \mathbf{w} \|_1 \leq \epsilon^{1+1/\beta} |\log \epsilon|^{-1/2(1+1/\beta)}$ and $d_1 \epsilon^{-1} |\log \epsilon|^{1/2} \leq k^\beta \leq d_2 \epsilon^{-1} |\log \epsilon|^{1/2}$ for some constants d_1, d_2 then

$$\sup_{0\leq x\leq 1} |p_0(x) - p(x; \mathbf{w}, k)| \leq C\epsilon |\log \epsilon|^{-1/2}.$$

For ϵ small enough $p(x; \mathbf{w}, k)$ will be bounded away from zero on $[1/k, 1-1/k]$ and near the boundary $p(x; \mathbf{w}, k) \geq ck^{-\beta}$ for some constant $c > 0$. It follows that $\sup_x |p_0(x) - p(x; \mathbf{w}, k)| \leq Ck^{-\beta}$ and we can apply the same calculations as with $p(x; \mathbf{w}_0, k)$ so that

$$p(x; \mathbf{w}, k) \in N(C\epsilon, p_0)$$

and hence

$$\left\{ p(x; \mathbf{w}_0, k) : \|\mathbf{w} - \mathbf{w}_0\|_1 \leq \epsilon^{1+1/\beta} |\log \epsilon|^{-1/2(1+1/\beta)} \right\} \subset N(C\epsilon, p_0).$$

Now let ϵ depend on n by taking $\tilde{\epsilon}_n = n^{-\beta/(2\beta+1)}(\log n)^{1/2}$ and let

$$d_1 n^{1/(2\beta+1)} \leq k_n \leq d_2 n^{1/(2\beta+1)},$$

for constants d_1, d_2. Applying lemma A.1 of Ghosal (2001) gives

$$\pi\left(N\left(C\tilde{\epsilon}_n, p_0\right)\right) \geq Ce^{-ck_n \log(1/\tilde{\epsilon}_n)},$$

and thus condition (2.3) is satisfied.

Define the sets $\mathcal{F}_n = \{p\left(\cdot; \mathbf{w}, k\right) : k \leq s_n\}$, i.e. those mixtures with s_n or less components. The condition (2.2) of Theorem 2.5 is satisfied by taking s_n to be an integer satisfying

$$d_1 n^{1/(2\beta+1)}\left(\log n\right) \leq s_n \leq d_2 n^{1/(2\beta+1)}\left(\log n\right),$$

as $\pi\left(\mathcal{F}_n^c\right) \leq Ce^{-cn\tilde{\epsilon}_n^2}$ where c can be made arbitrarily large by taking d_1 sufficiently large. Finally, to check the condition (2.1) we follow the arguments in Ghosal's Theorem 2.3 and note that $\log D\left(\epsilon, \mathcal{F}_n, d\right)$ is bounded by

$$s_n \log\left(\frac{C}{\epsilon}\right) + \log s_n.$$

Upon taking $\bar{\epsilon}_n = n^{-\beta/(2\beta+1)}\left(\log n\right)$ condition (2.1) is satisfied. Note that Theorem 2.5 only gives convergence in probability. However the bound on the expectation of the posterior probability as given in Theorem 2.6 comes from their proof and from the fact that when $p\left(\cdot; \mathbf{w}, k\right)$ is considered as above,

$$\begin{aligned} P_0^n\left[\log p(Y^n) - \log p_0(Y^n) \leq -nC\epsilon^2\right] &\leq e^{-snC\epsilon^2}\left(E_0\left[e^{s\log\left(p_0/p\right)(Y)}\right]\right)^n \\ &\leq Cn^{-H}, \end{aligned}$$

when n is large enough and $\epsilon^2 \leq \epsilon_0 n^{-2\beta/(2\beta+1)}$, by choosing correctly s. This implies using Shen and Wasserman (2001)'s type of proof that

$$P_0^n\left[\int \frac{p(Y^n; \mathbf{w}, k)}{p_0(Y^n)} d\pi(\mathbf{w}, k) \leq e^{-2nC\epsilon_n^2}\right] \leq n^{-H}$$

for all $H > 0$, when n is large enough. This completes the proof. □

This Theorem implies in particular that if $\hat{p}(x) = E^{\pi}\left[p(x)|\, Y^n\right]$, then when n is large enough,

$$E_0^n\left[d(\hat{p}, p_0)\right] \leq 2Cn^{-\beta/(2\beta+1)}(\log n),$$

which is the minimax rate of convergence for the class of Hölder continuous functions $\mathcal{H}\left(\beta, L\right)$ (up to a $\log n$ term).

The assumption of $p_0(x) \geq ax(1-x)$ in Theorem 2.6 can be removed at the expense of having a slower rate of convergence. Without the lower bound on p_0 a small modification of the above proof will yield a rate of convergence, up to a $(\log n)$, of $n^{-\beta/(2\beta+2)}$. While this is significantly slower than the minimax rate, it is the same as that achieved by the Bernstein polynomials (Ghosal (2001)) for twice differentiable densities and will be sufficient for our examination of the Bayes factor in section 4.1.

Simulating from the posterior distribution with a fixed number of components poses no difficulty. Using a Dirichlet prior on the weights associated with each component in the mixture, a two-stage Gibbs sampling strategy can be implemented by taking advantage of the usual latent variable structure;

$$\mathbf{z}_i \sim \mathcal{M}(w_0, \ldots, w_k), \qquad Y_i \mid \mathbf{z}_i, k \sim h_{\mathbf{z}_i}(y),$$

where $\mathcal{M}$ is the multinomial distribution and if $\mathbf{z}_i$ has a 1 in the j-th position then $h_{\mathbf{z}_i}$ is the triangular density h_j as described in the introduction.

We need to deal with the case of an unknown number of components in the mixture. Two prominent approaches to the problem are Richardson and Greens (1997) reversible jump MCMC algorithm and the birth-and-death process described in Stephens (2000). The lack of a nested structure in these mixtures makes the construction of suitable proposal distributions difficult. The approach adopted in this paper will be to simulate from the posterior conditional on the number of components and then determine posterior probability for the number of components using the marginal likelihood calculations described in Chib (1995). Given k, the marginal likelihood $m(\mathbf{y} \mid k)$ is computed from the Gibbs output $\{\mathbf{z}_1, ..., \mathbf{z}_M\}$ as

$$\log \hat{m}(\mathbf{y} \mid k) = \log f(\mathbf{y} \mid \mathbf{w}^*, k) + \log \pi(\mathbf{w}^* \mid k) - \log \left\{ M^{-1} \sum_{i=1}^{M} \pi(\mathbf{w}^* \mid \mathbf{y}, \mathbf{z_j}, k) \right\}$$

for a suitable $\mathbf{w}^*$ such as the posterior mode.

These computations are demonstrated on a small collection of simulated data sets. The data comprise 1000 independent realizations from each one of the following six densities: (A) Beta(3,3), (B) Beta(3,5), (C) Uniform, (D) Triangular, (E) Truncated Normal(1/2,1/36), and (F) 0.3×Beta(2,5) + 0.7×Beta(6,1). In each case a geometric distribution with parameter 0.6 was taken for the prior on the number of components k in the mixture (truncated at $k = 10$). Given k, a Dirichlet prior was placed on

the weights with parameters $(4/(k+1), ..., 4/(k+1))$. The posterior mean density estimates are plotted in Figure 1.

FIGURE 1 NEAR HERE.

2.2. **Type II Mixtures.** 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. In this section we study the asymptotic properties of the posterior distributions associated with type II mixtures and give sufficient conditions for consistency.

The following assumptions on the prior for the mixture of triangular distributions are made:

- For all $k = 1, 2, \ldots$ $\pi(k) > 0$.
- Given k, the support of the prior on the partition increments

$$\{x_1 - x_0, x_2 - x_1, \ldots, x_k - x_{k-1}\}$$

  is the entire k-dimensional simplex. Hence, for any set U of positive Lebesgue measure, $\pi(U \mid k) > 0$.

These conditions on the prior are sufficient to establish weak consistency when p_0 is continuous.

Theorem 2.7. *Suppose p_0 is a continuous density on $[0,1]$. The Type II mixture of triangular distributions posterior is weakly consistent at p_0.*

Proof. From Schwartz's theorem we only need to check that every Kullback-Leibler neighbourhood of any continuous density function has positive prior probability. It is initially assumed that

$$\inf_{x \in [0,1]} p_0(x) = a > 0.$$

We now show that any continuous density on $[0,1]$ that is bounded away from zero can be uniformly approximated by a sequence of mixtures of triangular densities $p_{\psi(k)}$, where $\psi(k)$ denotes a partition of size k. For a given k the partition

$$\psi(k) = \{x_0, x_1, \ldots, x_k\}$$

of $[0,1]$ is chosen to satisfy the equations $P_0(x_i) = i/k,\ i = 0, \ldots, k$. From the mean value theorem

$$P_0(x_{i+1}) - P_0(x_{i-1}) = 2/k = p_0(x_i^*)(x_{i+1} - x_{i-1}), \quad x_i^* \in (x_{i-1}, x_{i+1}),$$

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

$$\begin{aligned} P_0(x_1) - P_0(x_0) &= 1/k = p_0(x_0^*)(x_1 - x_0), \quad x_0^* \in (x_1, x_0), \\ P_0(x_k) - P_0(x_{k-1}) &= 1/k = p_0(x_k^*)(x_k - x_{k-1}), \quad x_k^* \in (x_{k-1}, x_k). \end{aligned}$$

It follows that $p_{\psi(k)}$ is the linear interpolation of the points $(x_i, p_0(x_i^*))$, $i = 0, \ldots, k$. Take any $x \in (x_i, x_{i+1})$, the density $p_{\psi(k)}$ in this interval is given by

$$p_{\psi(k)}(x) = \frac{p_0(x_{i+1}^*) - p_0(x_i^*)}{x_{i+1} - x_i}(x - x_i) + p_0(x_i^*).$$

and hence

$$\begin{aligned} \left|p_0(x) - p_{\psi(k)}(x)\right| &\leq |p_0(x) - p_0(x_i^*)| + \left|p_0(x_i^*) - p_{\psi(k)}(x)\right| \\ &\leq |p_0(x) - p_0(x_i^*)| + \left|p_0(x_{i+1}^*) - p_0(x_i^*)\right| \frac{x - x_i}{x_{i+1} - x_i}. \end{aligned}$$

We note that

$$\frac{1}{Mk} \leq |x_i - x_{i-1}| \leq \frac{1}{ak},$$

where $M = \sup_{x \in [0,1]} p_0(x)$. As a continuous function on $[0,1]$ is uniformly continuous it follows that any continuous density bounded away from zero can be uniformly approximately by a mixture of triangular densities. Similar arguments to those used by Petrone and Wasserman (2002) in the proof of their Theorem 2 can be applied to conclude that

$$\lim_{k \longrightarrow \infty} \int \log\left[\frac{p_0(x)}{p_{\psi(k)}(x)}\right] p_0(x)\, dx = \int \lim_{k \longrightarrow \infty} \left(\log\left[\frac{p_0(x)}{p_{\psi(k)}(x)}\right] p_0(x)\right) dx = 0.$$

Thus, for any δ there exists an k_0 such that for all $k > k_0$, $K\left(p_0, p_{\psi(k)}\right) < \delta$.

Now consider k to be fixed and so we denote $\psi(k)$ by ψ. Let $\eta = (x_0, \tilde{x}_1, \ldots, \tilde{x}_{k-1}, x_k)$ and for $\epsilon > 0$ define the neighbourhood around ψ by

$$N_\epsilon = \left\{\eta : \max_{j=1,\ldots,k} |\tilde{x}_j - x_j| < \frac{\epsilon}{kM}\right\}.$$

We aim to show that for $\eta \in N_\epsilon$,

$$\lim_{\eta \longrightarrow \psi} K(p_0, p_\eta) \longrightarrow K(p_0, p_\psi).$$

From lemma 6.1

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

for any $\epsilon < 1/4$ and a finite constant C which is independent of the partition and k.

Again, following similar arguments to Theorem 2 of Petrone and Wasserman (2002) it follows that $K(p_0, p_\eta)$ is continuous in η at ψ. Hence, for any δ we can choose an ϵ such that $K(p_0, p_\eta) < \delta$ for any $\eta \in N_\epsilon$. Therefore, the prior probability on any Kullback-Leibler neighbourhood of a continuous density function bounded away from zero is strictly positive. The condition that $\inf_{x\in[0,1]} p_0(x) = a > 0$ can be removed using lemma 6.3.2 of Ghosh and Ramamoorthi (2003). □

It was noted in Perron and Mengersen (2001) that any continuous distribution function can be well approximated by a finite Type II mixture of triangular distributions. This was in contrast to Type I mixtures of triangular distributions and the Bernstein polynomials which could perform poorly, particular for non-differentiable distribution functions. The ability of Type II mixtures to always approximate a continuous distribution function well arise as these mixtures can place a large amount of probability on rather small intervals. Such behaviour, while being advantageous in the context studied in Perron and Mengersen (2001), poses a difficulty for consistent density estimation as for any given $k > 3$ the set of Type II mixtures has an infinite δ-covering number. For this let $p_1, \ldots, p_m$ be any finite collection of densities on $[0, 1]$ and let $p_{\eta(k)}$ be a mixture of triangular densities. For $|x_3 - x_1|$ in the partition $\eta(k)$ sufficiently small the distance between $p_{\eta(k)}$ and p_i will be greater than $1/k$ for any $i = 1, \ldots, m$. Hence, the δ-covering number is infinite for any $\delta < 1/k$. To overcome this the prior needs to be formed so that the prior probability of any points in the partition being close to each other is very small. The precise statement of this is given in the following theorem.

Theorem 2.8. *Assume that the prior satisfies,*

- *The prior on the number of components k is such that $\pi(k > n/(\log n)) \leq e^{-nr}$.*
- *Given k, the prior places very small probability on two points in the partition being close. Specifically, $\pi(|x_i - x_{i-1}| < n^{-a} \mid k) < e^{-nr}$,*

for some $a, r > 0$ and all n sufficiently large. Suppose p_0 is a continuous density on $[0, 1]$. The Type II mixture of triangular distributions posterior is strongly consistent at p_0.

Proof. Strong consistency of the posterior is proved by checking the conditions stated in Theorem 2.3 hold. From Theorem 2.7 we have that under these conditions any continuous probability density function is in the Kullback-Leibler support of the

prior. For a given k consider the set formed by the collection of all partitions $\psi(k)$ of $[0,1]$ such that $\sup_{x\in[0,1]} p_{\psi(k)}(x) \leq n^a$ and define $\mathcal{F}_n$ to be the union of these sets for $k = 1, \ldots, k_n$, i.e.

$$\mathcal{F}_n = \bigcup_{k=1}^{k_n} \left\{ \psi(k) : \sup_{x\in[0,1]} p_{\psi(k)}(x) \leq n^a \right\}.$$

The prior probability on $\mathcal{F}_n^c$ needs to decrease exponentially quickly.

$$\begin{aligned} \pi(\mathcal{F}_n^c) &= \pi(\{k > k_n\}) + \sum_{k=1}^{k_n} \pi\left(\left\{ \psi(k) : \sup_{x\in[0,1]} p_{\psi(k)}(x) > n^a \right\} \right) \\ &\leq \pi(\{k > k_n\}) + \sum_{k=1}^{k_n} \pi\left(\left\{ \psi(k) : \min_i |x_i - x_{i-1}| < n^{-a} \right\} \right) \end{aligned}$$

The assumptions made on the prior will ensure that with the appropriate choice of $k_n = k_0 n/(\log n)$, $\pi(\mathcal{F}_n^c) < Ce^{-nc}$ for some c and all sufficiently large n. Now to show that the L_1 δ-covering number of $\mathcal{F}_n$ grows at the correct rate. Consider for a fixed k, two mixtures of triangular distributions p_ψ and p_η which are bounded by M. From Lemma 6.1 if $\| \psi - \eta \|_{(\infty)} \leq \frac{\delta}{CkM^2}$ then $\| p_\psi - p_\eta \|_{(1)} \leq \delta$. The number of density functions required to cover $\left\{ \psi(k) : \sup_{x\in[0,1]} p_{\psi(k)}(x) \leq n^a \right\}$ is then

$$\left(\frac{Ckn^{2a}}{\delta} \right)^k$$

and hence to cover $\mathcal{F}_n$ we need

$$k_n \left(\frac{Ck_n n^{2a}}{\delta} \right)^{k_n}$$

density functions. The L_1 entropy is then

$$\log D(\delta, \mathcal{F}_n, \|\cdot\|_1) = k_n \left(2a \log n + \log k_n + \log \delta^{-1} + \log C \right).$$

Letting $k_n = k_0 n/(\log n)$ we have $\log D(\delta, \mathcal{F}_n, \|\cdot\|_1) < nk_0(2a+1)$, for sufficiently large n and hence, if $k_0 < \epsilon^2/(2a+1)$ the conditions of Theorem 2.3 are satisfied. It follows now that the posterior is strongly consistent at p_0. □

As in the case of fixed partition and free weights we obtain the minimax rate of convergence up to a $\log n$ term. The result is presented in the following theorem.

Theorem 2.9. *Assume that the true density p_0 belongs to the Hőlder class $\mathcal{H}(\beta, L)$ with regularity β. Assume also that $p_0 \geq a > 0$ on $[0,1]$. In addition, assume that the prior satisfies,*

- *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.*
- *Given k, the prior places very small probability on two points in the partition being close. Specifically, $\pi(|x_i - x_{i-1}| < n^{-a} \mid k) < e^{-rn}$, for some $a, r > 0$ and all n sufficiently large.*

Then there exists $C > 0$ such that

$$E_0^n \left[\Pi \left(p_\psi : d(p_\psi, p_0) > C n^{-\beta/(2\beta+1)} \log n \middle| Y^n \right) \right] \leq n^{-H}, \quad \forall H > 0,$$

when n is large enough.

Proof. The proof follows the line of the proof on the strong consistency. If $\beta \leq 1$, using the construction of $p_{\psi(k)}$, $\psi(k) = (x_1, ..., x_{k-1})$ described in Theorem 2.6, we obtain that

$$|p_0 - p_{\psi(k)}|_\infty \leq C k^{-\beta}, \quad if \quad p_0 \geq a > 0.$$

If $\beta \in [1, 2]$, we note that for each $i \leq k - 1$,

$$p_0(x) = p_0(x_i) + \frac{(x - x_i)}{x_{i+1} - x_i}(p_0(x_{i+1}) - p_0(x_i)) + O(k^{-\beta}).$$

Therefore, for $x \in (x_i, x_{i+1})$

$$|p_0 - p_{\psi(k)}| = \left| (p_0(x_i) - p_0(x_i^\star)) \frac{(x_{i+1} - x)}{x_{i+1} - x_i} + (p_0(x_{i+1}) - p_0(x_{i+1}^\star)) \frac{(x - x_i)}{x_{i+1} - x_i} \right|.$$

We also have, using a Taylor expansion of $\int_{x_i}^{x_{i+1}} p_0(x)dx$ and of $\int_{x_{i-1}}^{x_i} p_0(x)dx$, both equal to $1/k$:

$$\begin{aligned} & p_0(x_i)[(x_{i+1} - x_i) - (x_i - x_{i-1})] \\ &= -p_0'(x_i)[(x_{i+1} - x_i)^2 + (x_i - x_{i-1})^2]/2 + O(k^{-\beta-1}) \\ &= O(k^{-2}), \end{aligned}$$

and using a Taylor expansion of $\int_{x_{i-1}}^{x_{i+1}} p_0(x)dx$,

$$\begin{aligned} & p_0(x_i^\star)(x_{i+1} - x_{i-1}) \\ &= p_0(x_i)(x_{i+1} - x_{i-1}) + \frac{p_0(x_i)'}{2}[(x_{i+1} - x_i)^2 - (x_i - x_{i-1})^2] + O(k^{-\beta-1}), \end{aligned}$$

which implies that

$$p_0(x_i) - p_0(x_i^\star) = -p_0'(x_i)/2[(x_{i+1} - x_i) - (x_i - x_{i-1})] + O(k^{-\beta}) = O(k^{-\beta}).$$

The same argument can be applied to show $p_0(x_i) - p_0(x_i^\star) = O(k^{-\beta})$. Therefore, the absolute difference of p_0 and $p_{\psi(k)}$ is bounded by $Ck^{-\beta}$ for some $C > 0$. Let $\eta(k)$ be another partition and $p_{\eta(k)}$ the resulting density. From Lemma 6.1 we have

$$|p_0(x) - p_{\eta(k)}(x)| \leq C(k^{-\beta} + \epsilon),$$

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

$$N(C\epsilon, p_0) \supset \left\{\eta(k) : |\eta(k) - \psi(k)| < c\epsilon k^{-1}\right\}$$

Allow k to depend on n by taking

$$d_1\left(\frac{n}{\log n}\right)^{1/(2\beta+1)} \leq k_n \leq d_2\left(\frac{n}{\log n}\right)^{1/(2\beta+1)}$$

and $\tilde{\epsilon}_n = k_n^{-\beta}$. The prior probability on $N(C\tilde{\epsilon}_n, p_0)$ can be bounded below

$$\begin{aligned}
\pi(N_{\epsilon_n}) &\geq \pi(\kappa_n)\pi_{\kappa_n}[\eta : |\tilde{x}_i - x_i| \leq \epsilon_n/(M\kappa_n)] \\
&\geq \frac{\pi(\kappa_n)\Gamma(\kappa_n)(c\epsilon_n/\kappa_n)^{\kappa_n - 1}}{\Gamma(\kappa_n/2)} \\
&\geq Ce^{c_1 n^{1/(2\beta+1)}\log n},
\end{aligned}$$

for some constants $c, c_1, C > 0$. Hence, condition (2.3) of Theorem 2.4 is satisfied. Similar to Theorem 2.7 define

$$\mathcal{F}_n = \bigcup_{k=1}^{k_n}\left\{\eta(k) : \sup_{x\in[0,1]} p_{\eta(k)}(x) \leq n^{a/(2\beta+1)}\right\}.$$

Under the assumptions of this theorem $\pi(\mathcal{F}_n^c) \leq Ce^{-cn^{1/(2\beta+1)}\log n}$, and so condition (2.2) is satisfied. Finally, the entropy is obtained using the same calculations as Theorem 2.7, replacing δ by ϵ_n. This leads to the L_1 entropy being bounded from above by $Cn^{1/(2\beta+1)}\log n$. Taking $\overline{\epsilon}_n = n^{-\beta/(2\beta+1)}(\log n)^{\beta/(2\beta+1)}$ condition (2.1) is satisfied and the proof is achieved. □

Calculations can be performed by Metropolis-Hastings sampling for fixed k and marginal likelihoods calculated from the output as described in Chib and Jeliazkov (2001). We see from this section that mixtures of triangulars are an easy and efficient tool for density estimation. It is in particular adaptive for β at least in the range $\beta \leq 2$. In the next section we use these priors in a different context than estimation, namely testing.

3. Application to Goodness-of-Fit Testing - Point Null Hypothesis

Now consider the problem of testing the point null hypothesis that the data Y^n belong to some completely specified distribution. Without loss of generality, this distribution is assumed to be uniform on $[0,1]$. The alternative hypothesis is that the data belong to some distribution with a continuous density function, other than the uniform.

H_0: $Y_1, \ldots, Y_n$ are independent observations from a $U(0,1)$ distribution.
H_1: $Y_1, \ldots, Y_n$ are independent observations from a distribution with a continuous density p_0 (not uniform).

Verdinelli and Wasserman (1998) modeled the density in H_1 with an infinite dimensional exponential family in testing the above hypothesis. In their Theorem 8.1, Verdinelli and Wasserman (1998) show that under rather strong conditions the Bayes factor defined by

$$B_n = \frac{\Pr(H_0 \mid Y_1, \ldots, Y_n)}{\Pr(H_1 \mid Y_1, \ldots, Y_n)} \div \frac{\Pr(H_0)}{\Pr(H_1)} = \left\{ \int_\Omega \prod_{i=1}^n p(Y_i)\, \pi(dp) \right\}^{-1}$$

is consistent. The following is an improvement and generalization of their result.

Theorem 3.1. *Assume that the prior on the density in H_1 places positive probability on all Kullback-Leibler neighbourhoods of all continuous density functions. In addition, assume that this prior places zero probability on the uniform density.*

(i) *If H_1 is true then $B_n \longrightarrow 0$ exponentially quickly, almost surely with respect to P^∞.*
(ii) *If H_0 is true then $B_n^{-1} \longrightarrow 0$, almost surely with respect to P^∞.*

Proof. The proof of (i) is the same as the proof of Theorem 8.1 (i) in Verdinelli and Wasserman (1998). For part (ii) we first show that B_n^{-1} forms a martingale. Applying Fubini's theorem

$$\begin{aligned}
\mathbb{E}\left[B_{n+1}^{-1} \mid Y_n, \ldots, Y_1\right] &= \int_\Omega \mathbb{E}\left[\prod_{i=1}^{n+1} p(Y_i) \mid Y_n, \ldots, Y_1\right] \pi(dp) \\
&= \int_\Omega \mathbb{E}\left[p(Y_{n+1})\right] \prod_{i=1}^{n} p(Y_i)\, \pi(dp) \\
&= B_n^{-1}.
\end{aligned}$$

Similarly, $\mathbb{E}\left(B_1^{-1}\right) = 1$ and hence B_n^{-1} is a martingale with respect to the filtration $\sigma\left(Y_n, \ldots, Y_1\right)$. Using Doob's Theorem on non-negative (sub)martingales, since $E(B_n^{-1}) = 1 < \infty$, there exists an integrable random variable X_∞ such that as $n \longrightarrow \infty,\ B_n^{-1} \longrightarrow B_\infty^{-1},\ P_\infty$ almost surely.

Now let V be a weak neighbourhood of the uniform distribution which we shall denote by P_0. Specifically, let m be some arbitrary but finite integer, $\epsilon_j > 0,\ f_j, j = 1, \ldots, m$ be bounded continuous functions and define V as

$$\begin{aligned} V &= \bigcap_{j=1}^{m} \left\{ P : \left| \int f_j dP - \int f_j dP_0 \right| < \epsilon_j \right\} \\ &= \bigcap_{j=1}^{m} \left\{ \left\{ P : \int f_j dP - \int f_j dP_0 < \epsilon_j \right\} \bigcap \left\{ P : \int f_j dP_0 - \int f_j dP < \epsilon_j \right\} \right\} \end{aligned}$$

We can then write

$$B_n^{-1} = \int_V + \int_{V^c} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right) = I_1 + I_2$$

Fix $a > 0$. We can apply Markov's inequality and then the Fubini theorem to I_1 to get

$$\Pr\left(\int_V \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right) > a \right) \leq a^{-1} \pi\left(V\right)$$

For I_2 let U be one of the sets in the finite intersection forming the set V. There exists a strictly unbiased function for testing $P = P_0$ versus $P \in U^c$ and so by proposition 4.4.1 in Ghosh and Ramamoorthi (2003) there also exists a uniformly exponentially consistent sequence of test functions. Lemma 4.4.2 of Ghosh and Ramamoorthi (2003) can be applied to show that for some $\beta > 0$,

$$\lim_{n \longrightarrow \infty} e^{n\beta} \int_{U^c} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right) = 0,$$

almost surely with respect to P^∞. It follows that for any V

$$\lim_{n \longrightarrow \infty} I_2 \leq \lim_{n \longrightarrow \infty} \sum_{j=1}^{m} \int_{U_{1,j}^c} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right) + \lim_{n \longrightarrow \infty} \sum_{j=1}^{m} \int_{U_{2,j}^c} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right) = 0,$$

almost surely with respect to P^∞. From the Portmanteau theorem

$$\Pr\left(B_\infty^{-1} > a\right) \leq \liminf_{n \longrightarrow \infty} \Pr\left(B_n^{-1} > a\right) \leq a^{-1} \pi\left(V\right)$$

and hence as $\pi\left(V\right)$ can be made arbitrarily small $B_\infty^{-1} = 0$, almost surely with respect to P^∞. □

In the case of mixtures of triangulars as described in the previous section, the uniform corresponds to the parameterisation: $x_i = i/k$ and $w_i = 1/k$ (apart from $w_0 = w_k = 1/(2k)$) for each value of k. The prior puts a null mass on the uniform distribution but a positive mass on any Kullback-Leibler neighbourhood of the uniform so these priors satisfy the assumptions of Theorem 3.1 and the Bayes factor is consistent. Under the alternative it decreases exponentionally quickly. We now give the rate of convergence of the Bayes Factor to infinity under the null hypothesis.

3.1. **Rate of convergence.** Let p be a mixture of triangulars, either of type I or of type II. In the first case, the parameters defining p are the weights $\mathbf{w} = (w_0, ..., w_k)$ and k the number of components and in the second case the parameters are the partition $\psi(k) = (x_1, ..., x_{k-1})$ and k. 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. Finally, denote by $l_n(\xi(k)) = \log p_{\xi(k)}(Y^n) = \sum_{i=1}^n \log p_{\xi(k)}(Y_i)$ the log-likelihood.

We prove in this section that the Bayes Factor is convergent with rate $n^{-1/2}$ up to a $\log n$ term: in the following sense

$$P_0 \left[B_n^{-1} \geq C \log n^q / \sqrt{n} \right] \leq \epsilon_n, \quad \epsilon_n \to 0$$

We also find a lower bound.

Theorem 3.2. *Assume P_0 is the uniform distribution on $[0,1]$ and the prior on k satisfies $\pi(k > n/\log n) < e^{-nr}$ for some $r > 0$.*

- *Fixed partition, free weights: If $\pi_k(w_1, ..., w_k)$ is absolutely continuous with respect to the Lebesgue measure on the simplex and has a density bounded by a constant times the Dirichlet$(1, \cdots, 1)$, then there exists $C, C' > 0$ such that*

$$P_0 \left[B_n^{-1} \geq C \log n^2 / \sqrt{n} \right] \leq \frac{C' \log n^q}{\sqrt{n}}, \quad \forall n \geq 1 \tag{3.1}$$

- *Free partition, fixed weights: If $\pi_k(\psi(k))$ satisfies the assumptions of Theorem 3.5, if the prior for $k = 1$ satisfies $\pi(|x_1 - 1/2| \leq \delta) \leq c\delta^2$, for some constant $c > 0$ and any $\delta < \delta_0$, then there exists $C, C' > 0$ such that (3.1) holds.*

REMARK: If, in the case of the type I mixtures, the prior probability

$$\pi_k\left(\mathbf{w}=(w_1,...,w_k), w_i\geq 0, \sum_{i=1}^{k-1}|w_i-1/k|+|w_0-1/(2k)|+|w_k-1/(2k)|\leq\delta\right)$$

is less than $C_k\delta^{rk}$ with $\sum_{k\geq 1}p(k)C_k<\infty$ and $r>1$ for δ small enough, then the above probability is bounded by

$$P_0\left[B_n^{-1}\geq C\log n^2/\sqrt{n}\right]\leq\frac{C'\log n^q}{n^{r/2}},\quad\forall n \text{ large enough}$$

Similarly, in the case of the type II mixtures, if the prior on the partition, at fixed k satisfies

$$\pi_k\left(\psi(k)=(x_1,...,x_{k-1});\forall j,|x_j-j/k|<\delta\right)\leq C\delta^{r(k-1)},$$

for δ small enough and $r>1$, then

$$P_0\left[B_n^{-1}\geq C\log n^2/\sqrt{n}\right]\leq\frac{C'\log n^q}{n^{r/2}},\quad\forall n \text{ large enough}$$

In the case of strictly positive priors on either the simplex (type I prior) or the sets of partitions $\psi(k)$, k fixed, we have the following lower bound on the Bayes factor, which proves that the upper bound obtained in Theorem 3.2 is sharp.

Theorem 3.3. *Assume that for $k=1$ the prior on w_0 has a strictly positive and continuously differentiable density on $(0,1)$, then*

$$P_0\left[B_n^{-1}\leq C_0/\sqrt{n}\right]\leq C/n$$

Both Theorems imply that essentially the Bayes factor for testing against the uniform is of order $1/\sqrt{n}$ unless we prevent the prior to put mass around the uniform distribution.

We first prove Theorem 3.3.

Proof. Recall that we can write the Bayes Factor as

$$B_n^{-1} \quad = \quad \left\{\sum_{k=1}^{\infty}\pi(k)\int_{S_k}p_{\xi(k)}(Y^n)d\pi_k(\xi(k))\right\}^{-1}.$$

Denote by $B_{n,k}=\pi(k)\int_{S_k}p_{\xi(k)}(Y^n)d\pi_k(\xi(k))$, then

$$P_0\left[B_n^{-1}\leq C_0/\sqrt{n}\right] \quad\leq\quad P_0\left[B_{n,1}\leq C_0/\sqrt{n}\right]$$

In the case of type I mixtures, at $k = 1$, the model is regular so that we can apply a Laplace expansion:

$$\int_0^1 e^{l_n(w_1)}\pi_1(w_1)dw_1 = \frac{e^{l_n(\hat{w}_1)}\pi_1(\hat{w}_1)}{\hat{j}_1^{1/2}\sqrt{n}}\left(1 + O_P(n^{-1})\right)$$

where $\hat{w}_1$ is the maximum likelihood estimator and $\hat{j}_1$ is the empirical Fisher information. Therefore by considering C_0 large enough and since $e^{l_n(\hat{w}_1)} \geq 1$

$$P_0\left[B_{n,1} \leq C_0/\sqrt{n}\right] \leq C/n,$$

for some C large enough. In the case of type II mixtures, the model is not regular, but it is regular in quadratic mean, so that we also obtain a Laplace expansion to the first order and the same argument holds. □

We now present the proof of Theorem 3.2.

Proof. Split the integral defining B_n^{-1} into two parts: a shrinking L_1 neighbourhood of the uniform density and its complementary. The first integral will be small since the neighbourhood has small prior probability and the second integrand quickly decreases with n. Special care needs to be taken for the smaller subspace corresponding to $k = 1$. Using the decomposition, $B_n^{-1} = B_{n,1} + \sum_{k\geq 2} B_{n,k}$, we have, if $v_n = C\log n^2/\sqrt{n}$

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

Denote by $V_{n,k} = \{\xi(k) \in S_k : |1 - p_{\xi(k)}|_1 \leq \rho_0 \log n^p/\sqrt{n}\}$, $k \geq 1$. We first consider the integrals over $V_{n,k}$:

$$I_1 = \sum_{k\geq 2}\pi(k)\int_{\xi(k)\in V_{n,k}} p_{\xi(k)}(Y_1, ..., Y_n)d\pi_k(\xi(k))$$

then

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

We therefore need to bound the terms $\pi_k(V_{n,k})$. To do so, we consider separately the two types of mixtures.

The fixed partition case: We have the following Lemma

Lemma 3.4. *For each $k \geq 2$,*

$$|1 - p_{w^k}|_1 \leq \rho_0 \log n^p/\sqrt{n}$$
$$\Rightarrow \sum_{j=1}^{k-1} |w_j - 1/k| + |w_0 - 1/(2k)| + |w_k - 1/(2k)| \leq 4\rho_0 \log n^p/\sqrt{n}$$

The proof of Lemma 3.4 is postponed to the Appendix.

Lemma 3.4, together with the fact that the prior on the weights is bounded by the Dirichlet density times a constant M, implies that

$$\pi_k(V_{n,k}) \leq M \left(4\rho_0 \log n^p/\sqrt{n}\right)^k \Gamma(k+1)\pi^{k+1/2}/\Gamma(k/2+3/2).$$

Since $\sum_k \pi(k)\Gamma(k+1)/\Gamma(k/2+3/2) < \infty$ if $\pi(k)$ satisfies the assumption of Theorem 3.2,

$$P_0\left[I_1 \geq v_n^{-1}/6\right] \leq C\frac{\log n^{2(p-1)}}{\sqrt{n}}.$$

The free partition case: We then have the following Lemma

Lemma 3.5. *For each $k \geq 2$,*

$$|1 - p_{\psi(k)}|_1 \leq \rho_0 \log n^p/\sqrt{n}$$
$$\Rightarrow \sum_{j=1}^{k-1} \frac{(x_{j+1} - x_j)}{4}\left(\frac{|(x_{j+2} - x_j) - 2/k|}{(x_{j+2} - x_j)} + \frac{|(x_{j+1} - x_{j-1}) - 2/k|}{(x_{j+1} - x_{j-1})}\right)$$
$$+\frac{1}{2kx_2}|x_1 - x_2/2| + \frac{1}{2k(1-x_{k-2})}|(1 - x_{k-1}) - (1 - x_{k-2})/2| \leq \rho_0 \log n^p/\sqrt{n}$$

The proof of this Lemma is given in Appendix 3. When $k = 2, 3, 4$, Lemma 3.5 implies that

$$\pi_1\left(|1 - p_{\psi(k)}|_1 \leq \rho_0 \log n^q/\sqrt{n}\right) \leq C(\log n^q/\sqrt{n})^k$$

and by a recursive argument, if $k \leq (\rho_0 \log n^p/\sqrt{n})^{-1/2}$,

$$\pi_k\left(|1 - p_{\psi(k)}|_1 \leq \rho_0 \log n^q/\sqrt{n}\right) \leq \pi_k\left(|x_j - j/k| \leq (\rho_0 \log n^q/\sqrt{n})^{1/2}, \forall j \leq k-1\right)$$

so that

$$\sum_{n^{1/4}\log n^{-q/2}\sqrt{\rho_0} \geq k \geq 3} \pi_k(V_{n,k})\pi(k) \leq C\frac{\log n^{2q}}{n}.$$

Since $P(k \geq \rho_0 n^{1/4}/\log n^{p/2}) = o(1/n)$,

$$P_0\left[I_1 \geq v_n^{-1}/6\right] \leq C\frac{\log n^{2(q-1)}}{\sqrt{n}}.$$

The following Lemma together with the above results, completes the proof of Theorem 3.2.

Lemma 3.6. *Let* $I_2 = \int_{V_n^c} p_\eta(Y^n) d\pi(\eta)$*, where* $V_n = \cup_k V_{n,k}$*, then*

$$P_0^n \left[I_2 \geq v_n^{-1} \right] \quad \leq \quad n^{-H}$$

for any positive H*, when* n *is large enough.*

The proof is given in Appendix 4. □

We have thus obtained good properties of the Bayes factor for testing against the uniform, using a nonparametric alternative represented as a mixture of triangular densities. In this case the uniform is a member of the nonparametric model of mixtures of triangular densities.

We now consider the more general problem of testing against a parametric family.

4. Goodness-of-Fit Testing - Parametric families

The case of the null hypothesis being 'fixed' without nuisance parameters is unlikely to occur in practice. Now consider the case of H_0 including some finite dimensional nuisance parameter. This situation is typical for testing if the data come from some family of distributions. A recent paper by Walker *et al.* (2004) has given a rate of convergence for the Bayes factor assuming the variance of $\log(p_0(X)/q(X))$ is bounded uniformly in q. However, their result is not helpful when testing a parametric family against a nonparametric alternative. In this case, if the true density were a member of the parametric family then their result states that $\frac{1}{n} \log B_n$ would converge to zero. From this we are unable to determine which model to select.

Here we give sufficient conditions for the Bayes factor to be consistent and give a rate of convergence. Contrary to the approaches of Verdinelli and Wasserman (1998) or Robert and Rousseau (2004) we do not embed the parametric family in a nonparametric one. Instead, if the density of the parametric family can not be represented as a finite mixture of triangular densities, the Bayes factor works well since we are essentially testing two separate families and a $0-1$ type of loss is therefore relevant. More generally, we consider a distance approach resembling that taken by Robert and Rousseau (2004) with no embedding.

We state formally the hypothesis to be tested:

H_0: $Y_1, \ldots, Y_n$ are independent observations from a finitely parameterised distribution $\mathcal{Q} = \{p_\theta(y), \theta \in \Theta\}$ on $[0,1]$.

H_1: $Y_1, \ldots, Y_n$ are independent observations from a distribution with a continuous density and not a member of that family.

4.1. **The Bayes factor.** The Bayes factor is given by,

$$B_n = \left\{ \int_\Theta \prod_{i=1}^n p_\theta(Y_i)\, \pi(d\theta) \right\} \left\{ \int_\Omega \prod_{i=1}^n p(Y_i)\, \pi(dp) \right\}^{-1}.$$

Theorem 4.1. *Assume that the following conditions hold:*

(i) *The nonparametric posterior is strongly consistent at p_0 with rate ϵ_n.*

(ii) *For all $\theta \in \Theta$, $p_\theta \in L_r$ for some $r > 1$.*

(iii) *The probability placed on the set*

$$A_{\epsilon_n} = \left\{ p : \| p - p_0 \|_{(1)} < C\epsilon_n \right\}$$

by the nonparametric prior converges to zero faster than $n^{-d/2}$ where $d = \dim(\theta)$.

(iv) *$p_\theta : \theta \in \Theta$ is a regular model.*

Then

(1) *If H_1 is true then $B_n \longrightarrow 0$, $P_0^\infty - a.s.$*

(2) *If H_0 is true then $B_n^{-1} \longrightarrow 0$, in P_0^∞ probability.*

Proof. Under H_0: The Bayes factor can be written as

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

Under assumption (i) the probability $\Pi(A_{\epsilon_n} \mid Y_1, \ldots, Y_n)$ converges to one. From (iv) $\left\{ \int_\Theta \frac{p_\theta(Y^n)}{p_0(Y^n)} \pi(d\theta) \right\}$ goes to zero with rate $n^{-d/2}$. Applying the Markov inequality to $\left\{ \int_{A_{\epsilon_n}} \frac{p(Y^n)}{p_0(Y^n)} \pi(dq) \right\}$ we have

$$\Pr\left(\int_{A_{\epsilon_n}} \frac{p(Y^n)}{p_0(Y^n)} \pi(dp) > n^{-d/2} \right) < n^{d/2} \pi(A_{\epsilon_n})$$

Under the assumptions of the theorem this probability will converge to zero and hence $B_n^{-1} \longrightarrow 0$ in probability.

Under H_1: Suppose the true density function is $p_0(y)$ and let $\phi_i, i = 1, 2, \ldots$ be a sequence of test functions which form a basis for $L_s[0,1], \frac{1}{s} + \frac{1}{r} = 1$. Let

$$\begin{aligned} \int_0^1 \phi_i(y) p_\theta(y) dy &= h_i(\theta) \\ \int_0^1 \phi_i(y) p_0(y) dy &= h_i \end{aligned}$$

If there exists θ_0 such that $h_i(\theta_0) = h_i$ for all $i = 1, 2, \ldots$ then by the Riesz theorem $p_0 \equiv p_{\theta_0}$. Therefore, there exists a finite collection of the ϕ_i such that for all $\theta \in \Theta$, $\sup_i |h_i - h_i(\theta)| > \epsilon$, for some $\epsilon > 0$. We write the Bayes factor as

$$B_n = \frac{\int_\Theta \prod_{i=1}^n \frac{p_\theta(Y_i)}{p_0(Y_i)} \pi(d\theta)}{\int_\Omega \prod_{i=1}^n \frac{p(Y_i)}{p_0(Y_i)} \pi(dp)}$$

The parameter space Θ can be partitioned into

$$\Theta = \bigcup_{i=1}^m \Big\{ \{\theta : h_i - h_i(\theta) > \epsilon\} \bigcup \{\theta : h_i(\theta) - h_i > \epsilon\} \Big\}$$

The ϕ_i form a strictly unbiased test of $p = p_0$ against $p = p_\theta$, $\theta \in \{\theta : h_i - h_i(\theta) > \epsilon\}$ and so by proposition 4.4.1 of Ghosh and Ramamoorthi (2003) there exists an exponentially consistent test. Hence we may apply their lemmas 4.4.1 and 4.4.2 to show

$$\frac{\int_{\{\theta : h_i - h_i(\theta) > \epsilon\}} \prod_{i=1}^n \frac{p_\theta(Y_i)}{p_0(Y_i)} \pi(d\theta)}{\int_\Omega \prod_{i=1}^n \frac{p(Y_i)}{p_0(Y_i)} \pi(dp)} \longrightarrow 0, \quad P - a.s.$$

The Bayes factor is bounded by a sum of these terms and hence it must converge to zero $P - a.s.$

□

The main condition that needs to be checked in Theorem 4.1 for it to be applicable is that the nonparametric prior places a sufficiently small amount of probability near the true density. A basic condition for this to hold for Type I mixtures of triangular distributions is given in the following lemma.

Lemma 4.2. *Assume the true density p_0 has a bounded third derivative and the second derivative is non-zero on some interval $\mathcal{I}$. A Type I mixture of triangular distributions prior satisfying the conditions of Theorem 2.6 satisfies $\pi(A_{\epsilon_n}) < e^{-cn^\alpha}$ for some $c, \alpha > 0$.*

The proof of this lemma is given in Appendix 5.

The rate of convergence for the Bayes factor under the alternative hypothesis is easily seen to be exponential as in the case of testing a point null hypothesis. To determine the rate of convergence under the null hypothesis we first write,

$$B_n^{-1} = \left\{ \int_{A_{\epsilon_n}} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right) \right\} \left\{ \int_{\Theta} \prod_{i=1}^{n} p_\theta\left(Y_i\right) \pi\left(d\theta\right) \right\}^{-1} \times \Pi^{-1}\left(A_{\epsilon_n} \mid Y_1, \ldots, Y_n\right),$$

From the assumptions of Theorem 4.1 it is known that $\pi^{-1}\left(A_{\epsilon_n} \mid Y_1, \ldots, Y_n\right)$ converges to one in probability and so, from Slutsky's lemma, we only need determine the rate at which $\left\{ \int_{A_{\epsilon_n}} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dq\right) \right\} \left\{ \int_{\Theta} \prod_{i=1}^{n} p_\theta\left(Y_i\right) \pi\left(d\theta\right) \right\}^{-1}$ goes to zero.

$$\begin{aligned} & P_0^n \left[\left\{ \int_{A_{\epsilon_n}} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dq\right) \right\} \left\{ \int_{\Theta} \prod_{i=1}^{n} p_\theta\left(Y_i\right) \pi\left(d\theta\right) \right\}^{-1} < \nu_n \right] \\ \leq \; & P_0^n \left[\int_{A_{\epsilon_n}} \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right) \geq C n^{-d/2} \nu_n \right] + O(n^{-1}) \\ \leq \; & \frac{\pi\left(A_{\epsilon_n}\right) n^{d/2}}{C \nu_n} + O(n^{-1}). \end{aligned}$$

From Lemma 4.2 we have $\pi\left(A_{\epsilon_n}\right) < e^{-cn^\alpha}$ for some $c, \alpha > 0$ and hence taking $\nu_n = e^{-n^\kappa}$ for any $\kappa < \alpha$ we have established

$$P_0^n \left[B_n^{-1} \geq e^{-n^\kappa} \right] \longrightarrow 0.$$

4.2. **Simulation Study.** We note that if p_0 is the uniform distribution then conditions of Theorem 4.1 will not be satisfied. Furthermore, it can be seen that in this case $B_n \longrightarrow 0, \; a.s.$ To see this note that the Beta family forms a regular model and so the numerator of B_n is $O\left(n^{-1}\right)$. From Theorem 3.2 it can be seen that $\int_\Omega \prod_{i=1}^{n} p\left(Y_i\right) \pi\left(dp\right)$ is $o\left(n^{-1/2}\right)$. Combination of the two results yields the inconsistency. When p_0 is a Beta distribution with both parameters greater than one, Lemma 4.2 can be applied to show that the conditions of Theorem 4.1 hold and so the Bayes factor will behave appropriately.

In this small simulation study we apply the Bayes factor to test the hypothesis:

H_0: $Y_1, \ldots, Y_n$ are independent observations from a Beta distribution.

H_1: $Y_1, \ldots, Y_n$ are independent observations from a distribution with a continuous density but not a Beta distribution.

In all cases the parameters of the Beta distribution are assumed independent a priori with a Gamma distribution with parameters (3,0.5). A sample was generated from the posterior distribution using the Metropolis-Hastings algorithm and the marginal likelihood was computed from the output using the method described in Chib and Jeliazkov (2001). Sample sizes of n = 25 and n = 500 were used in the simulation study. The same densities (A)–(F) from section 2 are used. The results are summarized in the table below. Histograms of the Bayes factors are given in Figures 2 and 3.

Proportion $BF < 1$ replications: 100

sample size	A	B	C	D	E	F
25	0.12	0.06	1.00	0.79	0.14	0.96
500	0.0	0.0	1.00	0.92	0.45	1.00

The behaviour of the Bayes factor for these sample size is as predicted by the asymptotics with the possible exception of the truncated normal distribution. This density was included in the simulation study as it appears very close to a Beta density and so should be a challenging case for the Bayes factor.

FIGURE 2 NEAR HERE.
FIGURE 3 NEAR HERE.

4.3. **A distance approach.** Now consider as in Robert and Rousseau (2004) the distance approach, based on the loss function:

$$L(\delta, p) = \begin{cases} a_0 d(p, \mathcal{Q}) & \text{if} \quad \delta = 0 \\ a_1(1 - d(p, \mathcal{Q})) & \text{if} \quad \delta = 1 \end{cases} \tag{4.1}$$

where $\delta = 0$ corresponds to choosing the null hypothesis and $\delta = 1$ corresponds to the alternative. The Bayes estimator is $\delta(Y^n) = 0$ if and only if

$$T(Y^n) = \mathbb{E}^\pi[d(p, \mathcal{F})|Y^n] \leq a_1/(a_0 + a_1)\,,$$

and $\delta(Y^n) = 1$ otherwise. In a non informative set-up, where there is no prior information on how to calibrate (a_0, a_1), we use the Bayesian p-value p_{cpred} defined by

$$p_{cpred}(\hat{\theta}, T) = \int_\Theta P_\theta \left[T(X^n) \geq T|\hat{\theta}, T\right] d\pi(\theta|\hat{\theta}),$$

where $\hat{\theta}$ is the maximum likelihood estimator under the parametric family and associated with the data (Y^n) and

$$\pi_0(\theta|\hat{\theta}) \propto d\pi(\theta)p_\theta(Y^n).$$

In this Section we prove that under the assumptions of Theorems 2.6 or 2.9, and if the parametric family is smooth enough, we have:

$$\sup_{p_0\in\mathcal{G}(\beta),d(p_0,\mathcal{Q})\geq Cn^{-\beta/(2\beta+1)}} E_0^n\left[p_{\mathrm{cpred}}(\hat{\theta},T)\right] = O(n^{-1/2}), \tag{4.2}$$

where $\mathcal{G}(\beta)$ is the subclass of Hölder functions with regularity β, satisfying the assumptions of Theorems 2.6 or 2.9, and if $p_0 \in \mathcal{Q}$, $p_{\mathrm{cpred}}(\hat{\theta},T) = \mathcal{U}(0,1) + O_P(n^{-1/2})$.

This implies that the test procedure is asymptotically optimal.

The result under the null hypothesis comes from Robert and Rousseau (2004). We therefore only have to study the behaviour of this test procedure under the alternative hypothesis.

Let $\theta^\perp = \inf_{\theta\in\Theta} \int p_0(x)\log p_0/p_\theta(x)dx$, then the result of Robert and Rousseau (2004) implies that

$$p_{cpred}(\hat{\theta},T) \quad = \quad P_{\theta^\perp}\left[T(X^n) \geq T|\hat{\theta},T\right] + R_n,$$

where R_n goes to zero in probability. More precisely,

$$P_0^n\left[|R_n| > M\log n^t/\sqrt{n}\right] \leq P_0^n\left[P_{\theta^\perp}[A_n^c|\hat{\theta}] > M\log n^t/\sqrt{n}\right] + O(n^{-1/2}),$$

where A_n is a set on which the Laplace approximation is valid. Under moment conditions on the derivatives of the log-likelihood, it is well known that

$$P_{\theta^\perp}\left[A_n^c\right] = O(n^{-H}),$$

with H large enough (depending on the number of moments of the second and third derivatives of the log-likelihood).

Assume first that $h(p_0, p_{\theta^\perp}) \leq h_0 \log n^{-2}$, where $h(p,p')$ denotes the Hellinger distance between p and p'. Then using Lemma 6.2 in Appendix 6, we obtain that

$$P_0\left[P_{\theta^\perp}[A_n^c|\hat{\theta}]\right] \leq P_{\theta^\perp}\left[A_n^c\right] + O(n^{-H}) = O(n^{-H}).$$

We therefore need only consider the quantity

$$p_{\theta^\perp,n} = P_{\theta^\perp}\left[T(X^n) \geq T|\hat{\theta},T\right].$$

From the inequality

$$T(Y^n) \geq d(p_0, \mathcal{F}) - E^\pi[d(p_0, p)|Y^n],$$

we obtain that

$$P_{\theta^\perp}\left[T(X^n) \geq T|\hat{\theta}, T\right] \leq P_{\theta^\perp}\left[E^\pi\left[d(p_0,p)|Y^n\right] + E^\pi\left[d(p_{\theta^\perp},p)|X^n\right] \geq d(p_0,\mathcal{F})|\hat{\theta}, Y^n\right].$$

Therefore, as soon as $d(p_0, \mathcal{Q}) > O(n^{-\beta/(2\beta+1)} \log n^2)$ this quantity is small. Indeed, define $v_n = v_0 n^{-\beta/(2\beta+1)} \log n^2$, for some constant v_0 large enough, which corresponds to the rate of convergence of $E^\pi[d(p,q)|Y^n]$ to zero (as obtained in Theorems 2.6 and 2.9), then

$$\begin{aligned} E_0[p_{\theta^\perp,n}] &\leq E_0\left[P_{\theta^\perp}\left(E^\pi\left[d(p_{\theta^\perp},p)|X^n\right] \geq d(p_0,\mathcal{Q}) - v_n|\hat{\theta}\right)\right] \\ &\quad + P_0^n\left[E^\pi\left[d(p_0,p)|Y^n\right] \geq v_n\right] \\ &\leq E_0\left[E^\pi\left[d(p_{\theta^\perp},p)|X^n\right] \geq d(p_0,\mathcal{Q}) - v_n|\hat{\theta}\right] + O(n^{-H}). \end{aligned}$$

Theorems 2.6 and 2.9 imply that

$$E_{\theta^\perp}\left[P_{\theta^\perp}\left(E^\pi\left[d(p_{\theta^\perp},p)|X^n\right] \geq d(p_0,\mathcal{Q}) - v_n|\hat{\theta}\right)\right] \leq n^{-H}$$

for any H, as soon as $d(p_0, \mathcal{Q}) \geq 2v_n$. Using Lemma 6.2 in Appendix 6, we have that

$$\begin{aligned} &E_0\left[P_{\theta^\perp}\left(E^\pi\left[d(p_\theta^\perp,p)|X^n\right] \geq d(p_0,\mathcal{F}) - v_n(p_0)|\hat{\theta}\right)\right] \\ &\leq E_{\theta^\perp}\left[P_{\theta^\perp}\left(E^\pi\left[d(p_\theta^\perp,p)|X^n\right]\right)\right] + O(n^{-1/2}) \\ &\leq O(n^{-1/2}). \end{aligned}$$

Finally, as soon as $2v_n \leq d(p_0, \mathcal{Q})$ and $h(p_0, p_{\theta^\perp}) \leq c_0 \log n^{-2}$, equation (4.2) is proved under the above constraint.

If $h(p_0, p_{\theta^\perp}) \geq c_0 \log n^{-2}$, then $d(p_0, \mathcal{Q}) \geq c_0 \log n^{-2}$. Using

$$\begin{aligned} E_0^n\left[p_{\text{cpred}}\right] &= \int_\Theta \pi(\theta) P_\theta\left[T(X^n) \geq T|\hat{\theta}, T\right] d\theta \\ &\leq \int_{|\theta-\theta^\perp|<n^{-1/4}} \pi(\theta) P_\theta\left[T(X^n) \geq T|\hat{\theta}, T\right] d\theta + O(n^{-1}), \end{aligned}$$

together with the fact that uniformly in θ, such that $|\theta - \theta^\perp| < n^{-1/4}$, we have

$$P_\theta\left[E^\pi\left[d(p_\theta,p)|X^n\right] \geq c_0 \log n^{-2}/2\right] \leq n^{-H}, \quad \text{and } g_\theta(\hat{\theta}) = 1 + O_{P_\theta}(n^{-1}),$$

we obtain, up to a term of order $O(n^{-1})$:

$$\begin{aligned} & E_0^n \left[p_{\text{cpred}} \right] \\ & \leq \int_{|\theta - \theta^\perp| < n^{-1/4}} \pi(\theta) \left(\int g_0(\hat{\theta}) P_\theta \left[E^\pi \left[d(p_\theta, p) | X^n \right] \geq \frac{c_0 \log n^{-2}}{2} \middle| \hat{\theta} \right] g(\hat{\theta}|\theta) d\hat{\theta} \right) d\theta \\ & \leq \int_{|\theta - \theta^\perp| \leq n^{-1/4}} \pi(\theta) P_\theta \left[E^\pi \left[d(p_\theta, p) | X^n \right] \geq c_0 \log n^{-2}/2 \right] d\theta \\ & = O(n^{-1}), \end{aligned}$$

which achieves the proof of equation (4.2).

This proves that the test procedure using p_{cpred} and the statistic $T(Y^n)$ has an optimal rate of convergence and good behaviour under the null hypothesis asymptotically. The interest in such a procedure is that if a priori one is able to determine $a_1/(a_0 + a_1)$, which appears in the loss function, there is no need to use the conditional predictive p-value and the test procedure has non-asymptotic properties such as admissibility. However if this a priori knowledge is not available, one can always the p-value p_{cpred} as described above and the test procedure has good asymptotic properties. Note that if the parametric family that is being tested belongs to the class of exponential families then under the null hypothesis the p-value is exactly uniform and the test procedure has good non-asymptotic properties.

5. Discussion

This paper has examined the asymptotic properties of Bayesian density estimates using a mixture of triangular prior. It was seen that the density estimates, under certain conditions, possess the desirable properties of weak and strong consistency. Both Type I and Type II mixture priors were seen to attain the minmax rate of convergence (up to a $\log n$ term) for Hölder continuous density functions when the true density is bounded below. From the results in Perron and Mengersen (2001) we believe that the Type II mixtures are better suited to situations where weak convergence of the estimate is of greater interest. This may be the case when the true distribution is believed to possess some multifractal property; see Falconer (1997). To the best of our knowledge, general conditions for determining weak rates of convergence have not been previously determined.

While general conditions for good priors have been established in Section 2 of this paper we have not discussed specific prior choice. Viewing the mixture of triangulars as a parametrization of a piecewise linear density function enables us to clearly see

the role of the prior distribution. In a Type I mixture the density can be seen to be the linear interpolation of the point $(0, 2kw_0), \{(x_i, w_i k)\}_{i=1}^{k-1}$ $(1, 2kw_k)$ and hence the role of $\mathbf{w}$ is clear. For computational simplicity a Dirichlet prior could be used with the parameters chosen to reflect an *a priori* belief on the shape of the density. Alternatively, the prior on $\mathbf{w}$ could be chosen to reflect a strong belief about other features of the density such as monotonicity or unimodality. A similar statement is possible for the Type II mixture of triangular distributions.

Sections 3 and 4 were devoted to the application of these prior distributions to goodness-of-fit testing. We have shown that Bayes factor will be consistent under very general conditions when testing a point null hypothesis. The case of a simple null hypothesis has received a detailed theoretical treatment in the literature, for example Verdinelli and Wasserman (1998) and Fortiana and Grané (2003). However, in practice a test against a parametric family with unknown parameters is required. We have been able to give sufficient conditions for consistency of the Bayes factor in this case and determined a rate of convergence under the null. Importantly situations in which the Bayes factor is inconsistent were identified. The inconsistency in the Bayes factor can arise if the nonparametric prior places too much probability near the true density relative to the parametric prior. For this reason, among others, a distance based approach was considered for the goodness-of-fit test. The distance approach produces a $p-$value which is asymptotically uniformly distributed under the null hypothesis and converges to zero under the alternative. Rates of convergence were also established. As the test statistic is based on the distance between densities the procedure is less sensitive to the prior than the Bayes factor. In particular, the distance approach is able to avoid the inconsistency that occurs with the Bayes factor.

6. Appendix

6.1. **Appendix 1: Lemma 6.1.**

Lemma 6.1. *Let $p_{\psi(k)}$ and $p_{\eta(k)}$ be two type II mixtures of triangular densities where $\psi(k) = (x_0, x_1, \ldots, x_{k-1}, x_k)$ and $\eta(k) = (x_0, \tilde{x}_1, \ldots, \tilde{x}_{k-1}, x_k)$. For $0 < \epsilon < 1/4$, if*

$$\max_i |x_i - \tilde{x}_i| \leq \frac{\epsilon}{Mk}, \quad M = \sup_{x \in [0,1]} p_{\psi(k)}(x),$$

then

$$\sup_{x\in[0,1]} \left|p_{\psi(k)}(x) - p_{\eta(k)}(x)\right| \leq CM\epsilon,$$

where C is a constant independent of the partitions and k.

Proof. This is a tedious calculation, hence its relegation to an appendix. Suppose that $\tilde{x}_i = x_i$ for all $i \neq j$ and $\tilde{x}_j < x_j$. As we have seen in the proof of Theorem 2.7

$$\frac{1}{Mk} < |x_i - x_{i-1}| < \frac{1}{ak},$$

so under the assumptions of this lemma $\tilde{x}_j \in [x_{j-1}, x_{j+1}]$. If the partitions forming $p_{\psi(k)}$ and $p_{\eta(k)}$ differ only at x_j then the two densities will differ only on the interval $[x_{j-2}, x_{j+2}]$. We now calculate the difference between the component triangular densities which form $p_{\psi(k)}, p_{\eta(k)}$ on $[x_{j-2}, x_{j+2}]$. The triangular density on $[x_{j-1}, x_{j+1}]$ with mode at x_j is

$$\Delta_j(x) = \begin{cases} \frac{2}{(x_{j+1}-x_{j-1})}\frac{(x-x_{j-1})}{(x_j-x_{j-1})} & x \in [x_{j-1}, x_j] \\ \frac{2}{(x_{j+1}-x_{j-1})}\frac{(x_{j+1}-x)}{(x_{j+1}-x_j)} & x \in [x_j, x_{j+1}] \end{cases}$$

and the triangular density on $[x_{j-1}, x_{j+1}]$ with mode $\tilde{x}_j$ is

$$\widetilde{\Delta}_j(x) = \begin{cases} \frac{2}{(x_{j+1}-x_{j-1})}\frac{(x-x_{j-1})}{(\tilde{x}_j-x_{j-1})} & x \in [x_{j-1}, \tilde{x}_j] \\ \frac{2}{(x_{j+1}-x_{j-1})}\frac{(x_{j+1}-x)}{(x_{j+1}-\tilde{x}_j)} & x \in [\tilde{x}_j, x_{j+1}] \end{cases}.$$

The greatest difference between these two densities occur at x_j and $\tilde{x}_j$.

$$\begin{aligned} \left|\Delta_j(x_j) - \widetilde{\Delta}_j(x_j)\right| &= \frac{2}{(x_{j+1}-x_{j-1})}\left|1 - \frac{(x_j - x_{j-1})}{(\tilde{x}_j - x_{j-1})}\right| \\ &= \frac{2|\tilde{x}_j - x_j|}{(x_{j+1}-x_{j-1})(\tilde{x}_j - x_{j-1})} \\ \left|\Delta_j(\tilde{x}_j) - \widetilde{\Delta}_j(\tilde{x}_j)\right| &= \frac{2}{(x_{j+1}-x_{j-1})}\left|\frac{(x_{j+1} - \tilde{x}_j)}{(x_{j+1} - x_j)} - 1\right| \\ &= \frac{2|\tilde{x}_j - x_j|}{(x_{j+1}-x_{j-1})(x_{j+1} - x_j)} \end{aligned}$$

Now consider the triangular densities on $[x_j, x_{j+2}]$ and $[\tilde{x}_j, x_{j+2}]$ with mode at x_{j+1}.

$$\Delta_{j+1}(x) = \begin{cases} \frac{2}{(x_{j+2}-x_j)}\frac{(x-x_j)}{(x_{j+1}-x_j)} & x \in [x_j, x_{j+1}] \\ \frac{2}{(x_{j+2}-x_j)}\frac{(x_{j+2}-x)}{(x_{j+2}-x_{j+1})} & x \in [x_{j+1}, x_{j+2}] \end{cases}$$

and

$$\widetilde{\Delta}_{j+1}(x) = \begin{cases} \frac{2}{(x_{j+2}-\tilde{x}_j)}\frac{(x-\tilde{x}_j)}{(x_{j+1}-\tilde{x}_j)} & x \in [\tilde{x}_j, x_{j+1}] \\ \frac{2}{(x_{j+2}-\tilde{x}_j)}\frac{(x_{j+2}-x)}{(x_{j+2}-\tilde{x}_j)} & x \in [x_{j+1}, x_{j+2}] \end{cases}$$

The greatest difference between these two densities occur at x_{j+1} and $\tilde{x}_j$.

$$\begin{aligned}\left|\Delta_{j+1}(x_{j+1}) - \widetilde{\Delta}_{j+1}(x_{j+1})\right| &= 2\left|\frac{1}{(x_{j+2}-x_j)} - \frac{1}{(x_{j+2}-\tilde{x}_j)}\right| \\ &= \frac{2|\tilde{x}_j - x_j|}{(x_{j+2}-x_j)(x_{j+2}-\tilde{x}_j)} \\ \left|\Delta_{j+1}(\tilde{x}_j) - \widetilde{\Delta}_{j+1}(\tilde{x}_j)\right| &= \frac{2|\tilde{x}_j - x_j|}{(x_{j+2}-x_j)(x_{j+1}-x_j)}\end{aligned}$$

Similar calculations show that the difference between the two triangular densities on $[x_{j-2}, x_j]$ and $[x_{j-2}, \tilde{x}_j]$ with modes at x_{j-1} is greatest at x_j and x_{j-1} giving

$$\begin{aligned}\left|\Delta_{j-1}(x_j) - \widetilde{\Delta}_{j-1}(x_j)\right| &= \frac{2|\tilde{x}_j - x_j|}{(\tilde{x}_j - x_{j-2})(\tilde{x}_j - x_{j-1})} \\ \left|\Delta_{j-1}(x_{j-1}) - \widetilde{\Delta}_{j-1}(x_{j-1})\right| &= \frac{2|\tilde{x}_j - x_j|}{(x_j - x_{j-2})(\tilde{x}_j - x_{j-2})}\end{aligned}$$

Combining these bounds give

$$\sup_{x\in[0,1]} \left|p_{\psi(k)}(x) - p_{\eta(k)}(x)\right| < \frac{19}{3} M^2 k \left|\tilde{x}_j - x_j\right|.$$

Now define $\eta_j(k) = (x_0, \tilde{x}_1, \ldots, \tilde{x}_j, x_j, \ldots, x_{k-1}, x_k)$, $j = 1, \ldots, k-1$. As each triangular density shares support with at most two others, the two densities $p_{\eta_{j-1}}, p_{\eta_j}$ will differ only on the interval $[\tilde{x}_{j-2}, x_{j+2}]$. Under the assumptions of the lemma is easily seen that $\sup_x p_{\eta_j(k)}(x) < \frac{4}{3}M$ for all $j = 1, \ldots, k-1$. Hence,

$$\sup_{x\in[0,1]} \left|p_{\eta_j(k)}(x) - p_{\eta_{j-1}(k)}(x)\right| < \frac{19}{3} \times \left(\frac{4}{3}\right)^2 M^2 k \left|\tilde{x}_j - x_j\right|.$$

As the intervals $[\tilde{x}_{j-2}, x_{j+2}]$ overlap only four times we have

$$\sup_{x\in[0,1]} |p_\psi(x) - p_\eta(x)| \leq C M^2 k \max_i |\tilde{x}_i - x_i| \leq C M \epsilon.$$

This completes the proof.

□

6.2. **Appendix 2 : Proof of Lemma 3.4.**

Proof. Let $p(.;\mathbf{w},k)$ be such that $|1-p(.;\mathbf{w},k)|_1 \leq \sqrt{\rho_n}$ and denote $w'_j = w_j$ for $1 \leq j \leq k-1$ and $w'_0 = 2w_0$, $w'_k = 2w_k$, then

$$\begin{aligned}
|1-p(.;\mathbf{w},k)|_1 &= \sum_{j=1}^{k-2} \int_{j/k}^{(j+1)/k} \left| w'_{j+1}k^2(x-j/k) + w'_j k^2((j+1)/k - x) - 1 \right| dx \\
&= \sum_{j=0}^{k-1} \int_{j/k}^{(j+1)/k} |a_j + b_j x| \, dx
\end{aligned}$$

where $a_j = k[w'_j(j+1) - w'_{j+1}j] - 1$ and $b_j = k^2(w'_{j+1} - w'_j)$. Let $\bar{x}_j = -a_j/b_j = j/k - w'_j/k(w'_{j+1} - w'_j) + 1/k^2(w'_{j+1} - w'_j)$,

- If $j/k < \bar{x}_j < (j+1)/k$ then

$$\begin{aligned}
\Delta_j &= \int_{j/k}^{(j+1)/k} |a_j + b_j x| \, dx \\
&= \left| a_j(2\bar{x}_j - 2j/k - 1/k) + \frac{b_j}{2}\left(2\bar{x}_j^2 - 2j^2/k^2 - 1/k^2 - 2j/k^2\right) \right| \\
&= |b_j| \left| -\bar{x}_j^2 + \frac{(2j+1)\bar{x}_j}{k} - \frac{j^2}{k^2} - \frac{1}{2k^2} - \frac{j}{k^2} \right| \\
&= |b_j| \left| -\left(\bar{x}_j - \frac{2j+1}{2k}\right)^2 - \frac{1}{4k^2} \right| \\
&= |b_j| \left(\left(\bar{x}_j - \frac{2j+1}{2k}\right)^2 + \frac{1}{4k^2} \right) \\
&\geq \frac{|w'_{j+1} - w'_j|}{4} \\
&\geq \frac{|w'_{j+1} - 1/k| + |w'_j - 1/k|}{8},
\end{aligned}$$

  since in this case we have either $w'_j < 1/k < w'_{j+1}$ or $w'_j > 1/k > w'_{j+1}$.
- If $\bar{x}_j \notin (j/k, (j+1)/k)$,

$$\begin{aligned}
\Delta_j &= \left| \int_{j/k}^{(j+1)/k} a_j + b_j x dx \right| \\
&= \left| [w'_j(j+1) - w'_{j+1}j] - \frac{1}{k} + \frac{(w_{j+1} - w_j)}{2}[1+2j] \right| \\
&= \left| w'_j - \frac{1}{k} + \frac{(w'_{j+1} - w'_j)}{2} \right| \\
&\geq \frac{|w'_j - 1/k|}{2}
\end{aligned}$$

The latter inequality comes from the fact that when $\bar{x}_j \notin (j/k, (j+1)/k)$, then we have either $w'_{j+1} > w'_j > 1/k$ or $w'_{j+1} < w'_j < 1/k$.

Finally we obtain the result of Lemma 3.4 □

6.3. **Appendix 3: Proof of Lemma 3.5.**

Proof. We consider the same type of calculations as in the case of type I mixtures.

$$\Delta_j = \int_{x_j}^{x_{j+1}} |a_j + b_j x| \, dx$$

but this time

$$a_j = -2\left[\frac{p_{j+1}x_j}{D_{j+1}} - \frac{p_j x_{j+1}}{D_j}\right]\frac{1}{x_{j+1}-x_j} - 1, \quad b_j = 2\left[\frac{p_{j+1}}{D_{j+1}} - \frac{p_j}{D_j}\right]\frac{1}{x_{j+1}-x_j}$$

where $D_j = x_{j+1} - x_{j-1}$ for $1 \le j \le k-1$, $D_0 = x_1$ and $D_k = (1 - x_{k-1})$. We still have $\bar{x}_j = -a_j/b_j$ and the following two conditions $\bar{x}_j \in (x_j, x_{j+1})$ or $\bar{x}_j \notin (x_j, x_{j+1})$

- If $\bar{x}_j \in (x_j, x_{j+1})$.

$$\begin{aligned}\Delta_j &= \left|a_j(2\bar{x}_j - x_j - x_{j+1}) + \frac{b_j}{2}\left(2\bar{x}_j^2 - x_j^2 - x_{j+1}^2\right)\right| \\ &= |b_j|\left((\bar{x}_j - (x_j + x_{j+1})/2)^2 + (x_j - x_{j+1})^2\right).\end{aligned}$$

  If $k > j \ge 1$

$$\begin{aligned}\Delta_j &= \left((\bar{x}_j - (x_j + x_{j+1})/2)^2 + (x_j - x_{j+1})^2\right)\frac{2}{k}\left|D_{j+1}^{-1} - D_j^{-1}\right|(x_{j+1} - x_j)^{-1} \\ &\ge \frac{2(x_{j+1} - x_j)}{k}\left|D_{j+1}^{-1} - D_j^{-1}\right|.\end{aligned}$$

  and Moreover $\bar{x}_j \in (x_j, x_{j+1})$ if and only if $D_j \wedge D_{j+1} < 2/k < D_j \vee D_{j+1}$, so that

$$\Delta_j \ge \frac{2(x_{j+1} - x_j)}{k}\left|D_{j+1}^{-1} - k/2\right| + \left|D_j^{-1} - k/2\right|.$$

- If $\bar{x}_j \notin (x_j, x_{j+1})$,

$$\begin{aligned}\Delta_j &= \left|a_j(x_{j+1} - x_j) + \frac{b_j}{2}\left(x_{j+1}^2 - x_j^2\right)\right| \\ &= \frac{(x_{j+1} - x_j)}{k}\left|\frac{1}{D_{j+1}} + \frac{1}{D_j} - k\right| \\ &\ge \frac{(x_{j+1} - x_j)}{k}\left(\left|\frac{1}{D_{j+1}} - \frac{k}{2}\right| + \left|\frac{1}{D_j} - \frac{k}{2}\right|\right).\end{aligned}$$

We now consider Δ_0 and Δ_k. Similar calculations imply that

$$\Delta_0 \geq \frac{1}{2kx_2}\left|x_1 - x_2/2\right|, \Delta_1 \geq \frac{1}{2k(1-x_{k-2})}\left|(1-x_{k-1}) - (1-x_{k-2})/2\right|$$

We finally obtain that

$$\begin{aligned}|1-f_k|_1 &\geq \sum_{j=1}^{k-1}\frac{(x_{j+1}-x_j)}{2k}\left(\left|\frac{1}{D_{j+1}}-\frac{k}{2}\right|+\left|\frac{1}{D_j}-\frac{k}{2}\right|\right)\\ &\quad+\frac{1}{2kx_2}\left|x_1-x_2/2\right|+\frac{1}{2k(1-x_{k-2})}\left|(1-x_{k-1})-(1-x_{k-2})/2\right|\end{aligned}$$

and Lemma 3.5 is proved. □

6.4. **Appendix 4: Proof of Lemma 3.6.** Let $\rho_n = \rho_0 \log n^{2p}/n$. We consider tests as defined in Ghosal *et al.* (2000) based on the L_1 distance: We know that there exists tests ϕ_i associated with densities $p_i \in V_n^c$ such that

$$\begin{aligned}E_0^n\left[\phi_i\right] &\leq e^{-nK|1-p_i|^2}\\ \sup_{p:|p-p_i|\leq|1-p_i|/5} E_f^n\left[(1-\phi_i)\right] &\leq e^{-nK|1-p_i|^2}.\end{aligned}$$

Consider $\mathcal{F}_n = \{p_{\xi(k)}, k \leq k_n, \xi(k) : \sup_{x\in[0,1]} p_{\xi(k)}(x) \leq n^a\}$ where $k_n = k_0 \log n^{p-1}$, with $p \geq 2$. Following Lemma 6.1 we have that the number of balls required to recover the set $\mathcal{F}_n \cap V_n^c$ is bounded by

$$D_n = k_n\left(\frac{ck_n n^{2a}}{\rho_n}\right)^{k_n} \leq e^{k_0(2a+1)\log n^p}$$

so that, by considering k_0 small enough we have that if $\phi_n = \max_i \phi_i$

$$\begin{aligned}E_0^n\left(\phi_n\right) &\leq e^{k_0 \log n^p (2a+1)} e^{-K\rho_0^2 \log n^p}\\ &\leq e^{-c\log n^p}\end{aligned}$$

Therefore, if

$$I_{21} = \int_{V_n^c\cap\mathcal{F}_n} p_\eta(Y^n) d\pi(\eta),$$

we obtain that

$$\begin{aligned}P_0^n\left[I_{21} \geq v_n^{-1}\right] &\leq E_0^n\left[\phi_n\right] + E_0^n\left[(1-\phi_n)\mathbb{1}_{(I_{21}\geq v_n^{-1})}\right]\\ &\leq E_0^n\left[\phi_n\right] + v_n\int_{V_n^c\cap\mathcal{F}_n} E_{p(\xi)}(1-\phi_n) d\pi(\xi)\\ &\leq E_0^n\left[\phi_n\right] + v_n\pi\left(V_n^c\cap\mathcal{F}_n\right)e^{-K\log n^p}\\ &\leq n^{-H}\end{aligned}$$

for any positive H, when n is large enough.

We finally consider

$$I_{22} = \int_{\mathcal{F}_n^c} p_\eta(Y^n) d\pi(\xi);$$

and

$$\begin{aligned} P_0^n \left[I_{22} \geq v_n^{-1} \right] &\leq v_n \pi(\mathcal{F}_n^c) \\ &\leq v_n \pi[k \geq k_0 \log n^{p-1}] + O(n^{-H}) \\ &\leq O(n^{-H}) \end{aligned}$$

for any $H > 0$.

6.5. **Appendix 5: Proof of lemma 4.2.** We are trying to give a lower bound on the L_1 norm between the mixture of triangular densities and p_0. If we can show that

$$\int_0^1 |p_0(x) - p_k(x)| \, dx \geq ck^{-\alpha} \tag{6.1}$$

where p_k is the mixture of triangular densities with k components that best approximates p_0 in the L_1 distance and c, α are constants, then from the discussion following Theorem 2.6 the prior probability on A_{ϵ_n} will be bounded by

$$\begin{aligned} \pi(A_{\epsilon_n}) &\leq \pi\left\{k : ck^{-\alpha} < \epsilon_n\right\} \\ &\leq \pi\left\{k : k > (\epsilon_n/c)^{-1/\alpha}\right\} \\ &\leq \left\{k : k > cn^{\beta/(2\beta+2)\alpha}(\log n)\right\} \\ &\leq Be^{-\beta n^\kappa}, \end{aligned}$$

for any $\kappa < \beta/((2\beta+2)\alpha)$. Hence, $n^{d/2}\pi(A_{\epsilon_n}) \longrightarrow 0$ for any finite d provided we can show that $\alpha < \infty$.

From the assumptions of lemma, for all k sufficiently large there exists a j such that $[j/k, (j+1)/k] \subset \mathcal{I}$. It will be assumed that the second derivative is negative on $\mathcal{I}$, the proof will follow the same arguments if it is positive. We proceed now to find a lower bound on the L_1 norm.

$$\begin{aligned} \int_0^1 |p_0(x) - p_k(x)| \, dx &> \int_{j/k}^{(j+1)/k} |p_0(x) - p_k(x)| \, dx \\ &> \int_{j/k}^{(j+1)/k} |p_0(x) - a - bx| \, dx, \end{aligned}$$

where a and b are chosen to minimize the integral. As p_0 is strictly concave on the interval $[j/k, (j+1)/k]$ the line $a+bx$ will intersect p_0 at two points x_1, x_2. The line joining $(x_1, p_0(x_1))$ and $\left(\frac{x_1+x_2}{2}, p_0\left(\frac{x_1+x_2}{2}\right)\right)$ is

$$l_1(x) = \left(\frac{p_0\left(\frac{x_1+x_2}{2}\right) - p_0(x_1)}{x_2 - x_1}\right) 2(x - x_1) + p_0(x_1),$$

and satisfies

$$l_1(x) < p_0(x), \quad x \in \left(x_1, \frac{x_1+x_2}{2}\right).$$

Similarly, the line joining $\left(\frac{x_1+x_2}{2}, p_0\left(\frac{x_1+x_2}{2}\right)\right)$ and $(x_2, p_0(x_2))$ is

$$l_2(x) = \left(\frac{p_0(x_2) - p_0\left(\frac{x_1+x_2}{2}\right)}{x_2 - x_1}\right) 2(x - x_2) + p_0(x_2),$$

and satisfies

$$l_2(x) < p_0(x), \quad x \in \left(\frac{x_1+x_2}{2}, x_2\right).$$

The L_1 norm between p_0 and p_k is bounded below by the area of the triangle formed by $(x_1, p_0(x_1))$, $\left(\frac{x_1+x_2}{2}, p_0\left(\frac{x_1+x_2}{2}\right)\right)$ and $(x_2, p_0(x_2))$. From basic geometry this area is

$$\left[p_0\left(\frac{x_1+x_2}{2}\right) - \frac{p_0(x_1)+p_0(x_2)}{2}\right]\left(\frac{x_2-x_1}{2}\right)$$

Taking a Taylor expansions of p_0;

$$\begin{aligned}
&p_0\left(\frac{x_1+x_2}{2}\right) - \frac{p_0(x_1)+p_0(x_2)}{2} \\
&= \frac{1}{2}\left(\frac{x_2-x_1}{2}\right)\left[p_0'(x_1) - p_0'(x_2) + \left(\frac{x_2-x_1}{2}\right)\left(\frac{p_0''(x_1)}{2} + \frac{p_0''(x_2)}{2}\right)\right. \\
&\quad \left. + O(x_2-x_1)^2\right] \\
&= \frac{1}{2}\left(\frac{x_2-x_1}{2}\right)^2\left[-2p_0''(x_2) + \frac{p_0''(x_1)}{2} + \frac{p_0''(x_2)}{2} + O(x_2-x_1)^2\right] \\
&= \frac{1}{2}\left(\frac{x_2-x_1}{2}\right)^2\left[-p_0''(x_2) + O(x_2-x_1)\right]
\end{aligned}$$

Hence,

$$\int_0^1 |p_0(x) - p_k(x)|\, dx > \frac{1}{2}\left(\frac{x_2-x_1}{2}\right)^3\left[-p_0''(x_2) + O(x_2-x_1)\right]$$

and assuming $(x_2 - x_1) > k^{-2}$ then

$$\int_0^1 |p_0(x) - p_k(x)|\, dx > Ck^{-6}$$

for some constant C, as required.

Now suppose that $|x_2 - x_1| \leq k^{-2}$. For ease of notation assume that $j = 0$. Consider the tangent line at $x = 0$ given by

$$t(x) = p_0(0) + p_0'(0)\, x.$$

As p_0 is strictly concave on the interval $[j/k, (j+1)/k]$ it satisfies

$$t(x) > p_0(x), \quad x \in (0, 1/k).$$

The line joining the points $(x_1, p_0(x_1))$ and $(x_2, p_0(x_2))$ is given by

$$l_3(x) = \frac{p_0(x_2) - p_0(x_1)}{x_2 - x_1}(x - x_1) + p_0(x_1),$$

which is the line which minimizes the L_1 norm on the interval $[j/k, (j+1)/k]$. The point of intersection between $t(x)$ and $l_3(x)$ is

$$\begin{aligned} x^* &= \frac{p_0(x_1) - x_1 \frac{p_0(x_2) - p_0(x_1)}{x_2 - x_1} - p_0(0)}{p_0'(0) - \frac{p_0(x_2) - p_0(x_2)}{x_2 - x_1}} \\ &= \frac{p_0(x_1) - p_0(0) - x_1 (p_0'(x_1) + O(x_2 - x_1))}{p_0'(0) - p_0'(x_1) + O(x_2 - x_1)} \\ &= x_1 \frac{p_0''(x_1)}{p_0''(0)} + O\left(x_1^2\right) + O(x_2 - x_1), \end{aligned}$$

provided $(x_2 - x_1) \leq k^{-2}$ and $x_1 > k^{-3/2}$. The area of the triangle formed by $(0, t(0)), (x^*, t(x^*))$ and $(0, l_3(0))$ is

$$\begin{aligned} \frac{x^*}{2} |l_3(0) - t(0)| &= \frac{x^*}{2} \left| p_0(x_1) - x_1 \frac{p_0(x_2) - p_0(x_1)}{x_2 - x_1} - p_0(0) \right| \\ &= \frac{x^*}{2} \left| p_0(x_1) - p_0'(x_1) x_1 - p_0(0) + O(x_1(x_2 - x_1)) \right| \\ &= \frac{x^*}{2} \left| -\frac{x_1^2}{2} p_0''(x_1) + O(x_1(x_2 - x_1)) + O\left(x_1^3\right) \right| \\ &= \frac{-(p_0''(x_1))^2}{2 p_0''(0)} x_1^3 + O\left(k^{-5}\right) \end{aligned}$$

Thus,

$$\int_0^1 |p_0(x) - p_k(x)|\, dx > C k^{-4.5}$$

if $x_1 > k^{-3/2}$ and $x_2 - x_1 \leq k^{-2}$. For the final case of $x_1 \leq k^{-3/2}$ and $x_2 - x_1 \leq k^{-2}$ then $1/k - x_2 > k^{-3/2}$ and so a similar argument can be applied starting with a

tangent line at $x = 1/k$. Thus, it has been established that for all k there are constants c, α such that (6.1) holds.

6.6. **Appendix 6: Lemma 6.2.**

Lemma 6.2. *Let B by any measurable set, and let p_0 be such that $h(p_0, p_{\theta^\perp}) \leq h_0(\log n)^{-2}$, for some constant h_0, then under the usual regularity conditions on the parametric model $\mathcal{Q} = \{p_\theta, \theta \in \Theta\}$,*

$$E_0^n\left[P_{\theta^\perp}\left[B|\hat{\theta}\right]\right] \leq 2P_{\theta^\perp}[B] + O(n^{-1/2})$$

Proof. Let $g_{\theta^\perp}(\hat{\theta})$ and $g_0(\hat{\theta})$ be the density of the mle, under $P_{\theta^\perp}$ and P_0 respectively. Under the usual regularity conditions

$$\begin{aligned} g_{\theta^\perp}(\hat{\theta}) &\propto e^{-\sqrt{n}(\hat{\theta}-\theta^\perp)i_{\theta^\perp}\sqrt{n}(\hat{\theta}-\theta^\perp)/2} + O_P(n^{-1/2}), \\ g_0(\hat{\theta}) &\propto e^{-\sqrt{n}(\hat{\theta}-\theta^\perp)i_0\sqrt{n}(\hat{\theta}-\theta^\perp)/2} + O_P(n^{-1/2}), \end{aligned}$$

where $i_{\theta^\perp} = E_{\theta^\perp}\left[-D^2 \log f_{\theta^\perp}(X)\right]$ and $i_0 = E_0\left[-D^2 \log f_{\theta^\perp}(X)\right]$.

$$E_0\left[P_{\theta^\perp}\left(B|\hat{\theta}\right)\right] = E_{\theta^\perp}\left[P_{\theta^\perp}\left(B|\hat{\theta}\right)\right] + \Delta_n,$$

If $i_{\theta^\perp} \geq i_0$ (in the sense that the difference between both matrices is semi-definite positive), then $\Delta_n \leq 0 + O_P(n^{-1/2})$. Else, let $u = \sqrt{n}(\hat{\theta} - \theta^\perp)$

$$\begin{aligned} e^{-u'i_0u/2} \leq\ & e^{-u'i_{\theta^\perp}u/2}\left(\mathbb{1}_{u'(i_0-i_{\theta^\perp})u\geq 0} + A\mathbb{1}_{2\geq u'(-i_0+i_{\theta^\perp})u\geq 0}u'(i_{\theta^\perp} - i_0)u\right) \\ & + \mathbb{1}_{2\leq u'(-i_0+i_{\theta^\perp})u}e^{-u'i_0u/2}. \end{aligned}$$

Moreover,

$$\begin{aligned} |u'(i_{\theta^\perp} - i_0)u| &= \left|\int (f_{\theta^\perp} - f_0)[u'D^2 f_{\theta^\perp}(x)u]dx\right| \\ &\leq h(p_{\theta^\perp}, p_0)u'Ju \end{aligned}$$

where $J = 2\left(\int (p_{\theta^\perp} + p_0)|D^2 \log f_{\theta^\perp}(x)|dx\right)^{1/2}$. Since $h(p_{\theta^\perp}, p_0) \leq h_0 \log n^{-2}$,

$$\begin{aligned} \Delta_{n,2} &= \int_u \mathbb{1}_{2\leq u'(-i_0+i_{\theta^\perp})u}e^{-u'i_0u/2}P_{\theta^\perp}(u)du \\ &\leq e^{-2c^2/h(p_{\theta^\perp},p_0)^2} \\ &= O(n^{-H}), \quad \forall H > 0. \end{aligned}$$

Moreover, if $2 \leq u'(-i_0 + i_{\theta^\perp})u$ then $u'Ju \geq C \log n^2$ and $e^{-u'i_0u} \leq n^{-c \log n}$, for some constant $c > 0$, so that

$$\begin{aligned}
\Delta_{n,1} &= h(p_{\theta^\perp}, p_0) \int_u \mathbb{1}_{0 \leq u'(-i_0+i_{\theta^\perp})u \leq 2} u'Jue^{-u'i_0u/2} P_{\theta^\perp}[B \mid \theta^\perp + u/\sqrt{n}]du \\
&\leq CP_{\theta^\perp}[B]^{1-1/L} h(p_{\theta^\perp}, p_0), \quad \forall L > 0 \\
&\leq CP_{\theta^\perp}[B] h(p_{\theta^\perp}, p_0).
\end{aligned}$$

Therefore, when $h(p_{\theta^\perp}, p_0) \leq h_0 \log n^{-2}$

$$E_0\left[P_{\theta^\perp}\left(B \mid \hat{\theta}\right)\right] \leq 2P_{\theta^\perp}(B) + O(n^{-1/2})$$

when n is large enough, which achieves the proof of Lemma 6.2. □

REFERENCES

Baron, A., Schervish, M.J., and Wasserman, L. (1999) The consistency of distributions in nonparametric problems. *Ann. Statist.* **27** 536-561.

Bayarri, M.J., and Berger, J.O. (2000) P-values for composite null models (with discussion). *J. Amer. Statist. Assoc.* **95** 1127-1142.

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

Carota, C., and Paramigiani, G. (1996) On Bayes factors for nonparametric alternatives. *Bayesian Statistics* vol. 5 (Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M. eds.) Oxford University Press.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313-1321.

Chib, S., and Jeliazkov, I. (2001) Marginal likelihood from the Meteropolis-Hastings output. *J. Amer. Statist. Assoc.* **96** 270-281.

Falconer, K. (1997) *Fractal geometry* John Wiley and sons, Great Britain.

Florens, J.P., Richard, J.F., and Rolin, J.M. (1996) Bayesian encompassing specification tests of a parametric model against a nonparametric alternative. *Technical Report 96.08* Université Catholique de Louvian, Institut de Statistique.

Fortiana, J. and Grané, A. (2003) Goodness-of-fit tests based on maximum correlations and their orthogonal decompositions. *J. R. Stat. Soc. B Stat. Methodol.* **65** 115-126.

Fraser, D., and Rousseau, J. (2005) Developing p-values: a Bayesian-frequentist

convergence. *preprint*

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

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

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

Hjort, N.L. (2003) Topics in non-parametric Bayesian statistics. *Highly Structured Stochastic Systems* (Green, P.J., Hjort, N.L., and Richardson, S. eds.) Oxford Statistical Science Series, vol. 27, Oxford University Press.

Munk, A. and Czado, C. (1998) Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc. ser. B Stat. Methodol.* **60** 223-241.

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

Petrone, S. (1999a) Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.* **27** 105-126.

Petrone, S. (1999b) Random Bernstein polynomials. *Scand. J. Statist.* **26** 373-393.

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

Richardson, S. and Green, P. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Stat. Soc. ser. B Stat. Methodol.* **59** 731-792.

Robert, C.P., and Rousseau, J. (2004) A mixture approach to Bayesian goodness of fit. *preprint*

Robins, J.M., van der Vaart, A., and Ventura, V. (2000). Asymptotic distribution of P values in composite null models (with discussion). *J. Amer. Statist. Assoc.* **95** 1143-1167

Schwartz, L. (1965) On Bayes procedures. *Z. Wahrscheinlichkeitstheor. Verw. Geb.* **4** 10-26.

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

Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump method. *Ann. Statist.* **28** 40-74.

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

Walker, S. (2003) On sufficient conditions for Bayesian consistency. *Biometrika* **90** 482-488.

Walker, S. (2004) New approaches to Bayesian consistency. *Ann. Statist.* **32** 2028-2043.

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

Zhang, J. (2002) Powerful goodness-of-fit tests based on the likelihood ratio. *J. R. Stat. Soc. ser. B Stat. Methodol.* **64** 281-294.

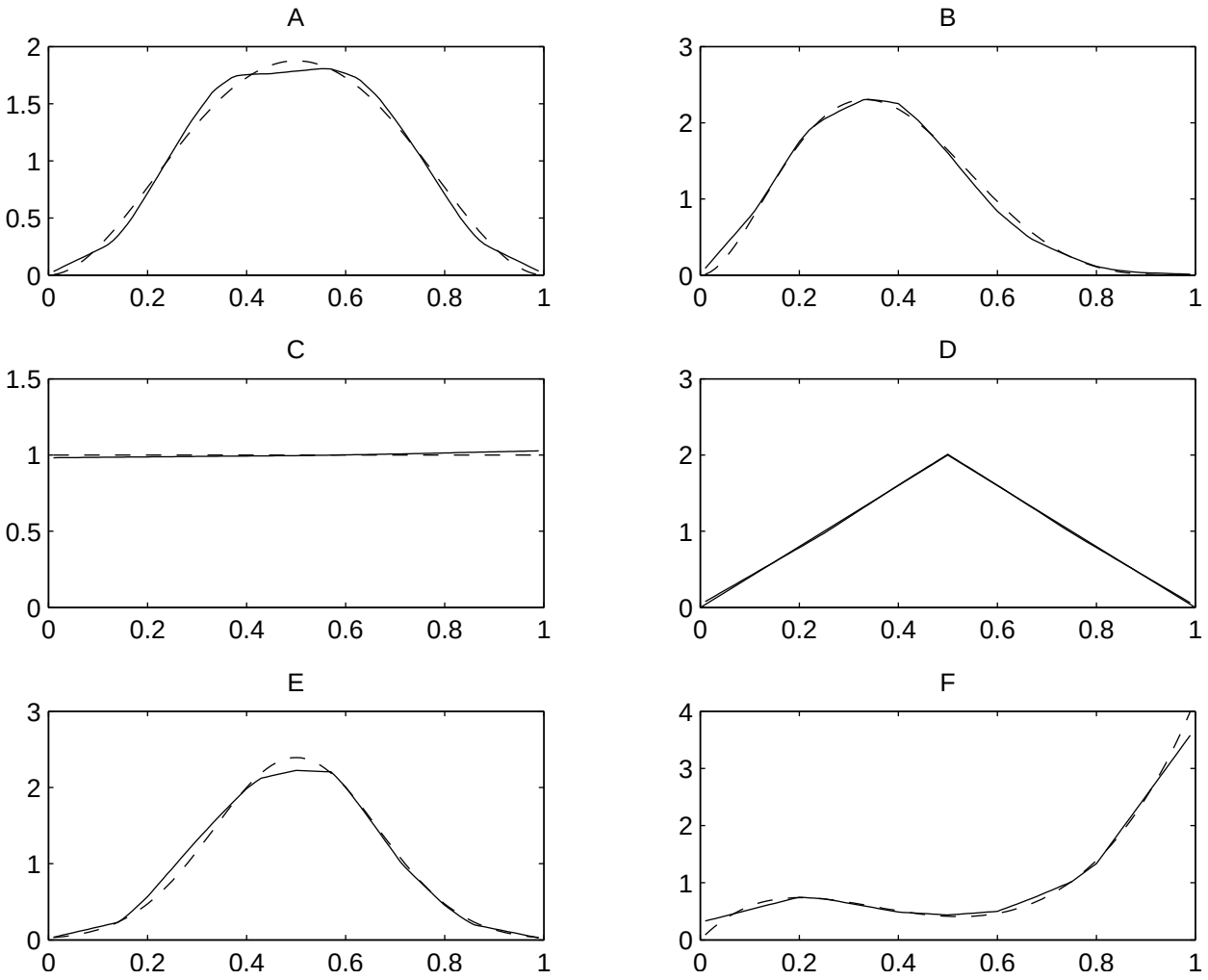

FIGURE 1. Application of the mixture of triangulars prior to density estimation. The solid line is the posterior mean density estimate. The dashed line is the true density from which the data was generated.

(R. McVinish and K. Mengersen) SCHOOL OF MATHEMATICAL SCIENCES, QUEENSLAND UNIVERSITY OF TECHNOLOGY, GPO BOX 2434, BRISBANE, Q4001, AUSTRALIA

(J. Rousseau) UNIVERSITÉ PARIS DAUPHINE, 75016 PARIS AND CREST

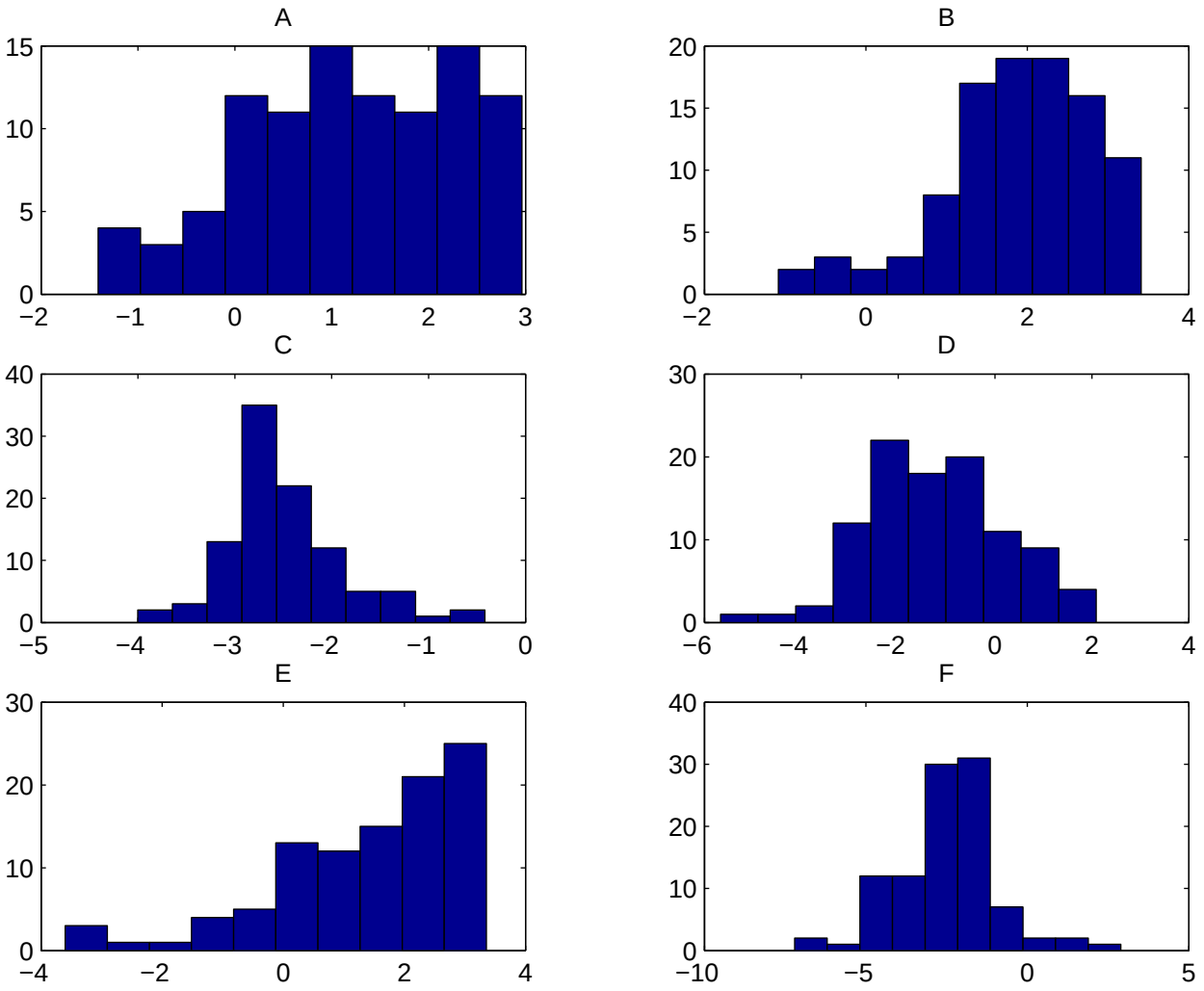

FIGURE 2. Histograms of the log Bayes Factor in each of the six cases study with sample size of 25.

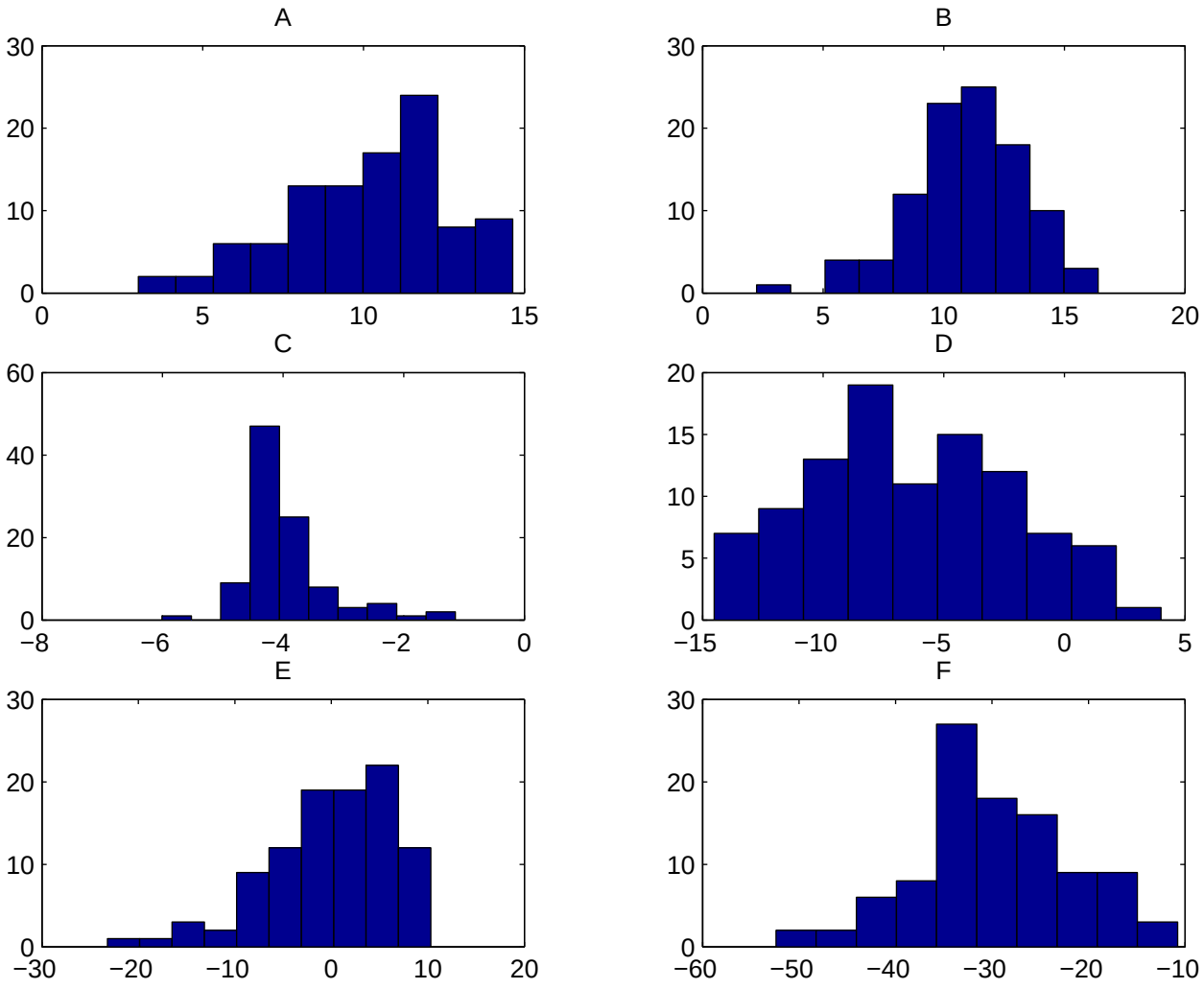

FIGURE 3. Histograms of the log Bayes Factor in each of the six cases study with sample size of 500.