A Mixture Approach to Bayesian Goodness of Fit

Christian P. Robert

CREST, INSEE, and CEREMADE, Université Paris Dauphine, 75775 Paris cedex 16

Judith Rousseau

CREST, INSEE, and Université Paris 5, 75232 Paris cedex 05

Summary. We consider a Bayesian approach to goodness of fit, that is, to the problem of testing whether or not a given parametric model is compatible with the data at hand. We thus consider a parametric family $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$, where F_θ denotes a cumulative distribution function with parameter θ. The null hypothesis is $H_0 : X \sim F_\theta$ for an unknown θ, that is, there exists θ such that $F_\theta(X) \sim \mathcal{U}(0,1)$. If H_0 does not hold, $F_\theta(X)$ is a random variable on $(0,1)$ which is not distributed as $\mathcal{U}(0,1)$. The alternative nonparametric hypothesis can thus be interpreted as $F_\theta(X)$ being distributed from a general cdf G_Ψ on $(0,1)$, where Ψ is infinite dimensional. Instead of using a functional basis as in Verdinelli and Wasserman (1998), we represent G_Ψ as the (infinite) mixture of Beta distributions,

$$p_0\mathcal{U}(0,1) + (1-p_0)\sum_{k\geq 1} p_k \mathcal{B}(\alpha_k, \beta_k)\,.$$

Estimation within both parametric and nonparametric structures are implemented using MCMC algorithms that estimate the number of components in the mixture. Since we are concerned with a goodness of fit problem, it is more of interest to consider a functional distance to the tested model $d(F, \mathcal{F})$ as the basis of our test, rather than the corresponding Bayes factor, since the later puts more emphasis on the parameters. We therefore propose a new test procedure based on the posterior conditional predictive p-value associated with $E^\pi[d(f,\mathcal{F})|X^n]$, with both an asymptotic justification and a finite sampler implementation.

AMS 1991 classification. Primary 62C05. Secondary 60J05, 62F15, 65D30, 65C60.

Keywords: Bayesian inference, Beta mixture distribution, birth-and-death process, consistency, nonparametric estimation, posterior conditional predictive p-value, variable dimension model

1. Introduction

It is both of high interest and of strong difficulty to come up with a satisfactory notion of a Bayesian test for goodness of fit to a distribution or to a family of distributions

$$\mathcal{F} = \{F_\theta,\ \theta \in \Theta\}\,,$$

where F_θ denotes a cumulative distribution function with parameter θ, for a given sample $X^n = (x_1, \ldots, x_n)$. The interest of the problematic being self-explanatory, let us rather insist on the difficulty.

In regular testing problems, the usual Bayesian solution, as described in most textbooks (see, e.g., Robert, 2001), is to build a prior distribution on each model and to derive the *Bayes factor*, ratio of the marginal distributions for both models: the magnitude of this factor is then interpreted as a degree of plausibility (or implausibility) of the hypothesis being tested. In a goodness of fit setting, there is no such clearcut separation between two possibilities: outside the case when $X \sim F_\theta$, the set of alternatives simply is the whole set of probability distributions, with no obvious structure on which to base the derivation of a reference prior. Since we do not want to engage in the difficult and disputed construction of nonparametric priors, we will use the device of Verdinelli and Wasserman (1998), which reduces the problem to finding a prior distribution on $[0,1]$, rather than on $\mathbb{R}$ or $\mathbb{R}^p$, through the use of the probability transform, that is, considering $F_\theta(X)$. If H_0 does not hold, $F_\theta(X)$ is a random variable on $(0,1)$ which is *not* distributed as $\mathcal{U}(0,1)$ for any value of θ. If H_0 is true then there exists $\theta \in \Theta$ for which $F_\theta(X) \sim \mathcal{U}(0,1)$. The alternative nonparametric hypothesis can thus be interpreted as $F_\theta(X)$ being distributed from a general cdf G_ψ on $[0,1]$, where ψ is infinite dimensional. The parameter in the alternative hypothesis is then (θ, ψ). In this setup, an acceptable resolution of the nonparametric problem on $[0,1]$, is to use mixtures of Beta distributions, of the form

$$p_0\mathcal{U}(0,1) + (1-p_0)\sum_{k\geq 1} p_k \mathcal{B}e(a_k, b_k)\,, \tag{1}$$

where $\mathcal{B}e(a_k, b_k)$ denotes a Beta random variable with parameters (a_k, b_k). Their shapes are variable enough to allow for an approximation of almost any arbitrary distribution on $[0,1]$. We believe that this approach is relevant since it models naturally the distortions from the uniform distribution. In this respect, Petrone and Wasserman (2002) have studied Bernstein priors based on Bernstein polynomials. The advantage of Bernstein polynomials over general mixtures of Beta distributions is that this modeling is easier to implement since the parameters of the Beta distributions which appear in the modeling of the nonparametric density are fixed integers; the weights are the only quantities to estimate. However, since the parameters of the Beta distributions are also allowed to vary in $]0,1]$ and are not restricted to be greater than 1, the mixtures of Beta distributions such as (1) should need less components to approximate a given density on $[0,1]$.

As ψ is infinite dimensional, subjective priors cannot be entirely justified, contrarywise to parametric settings. In other words, non parametric priors are highly arbitrary. It is therefore necessary to assess the consistency of the posterior distribution, as a validation for our prior. Diaconis and Freedman (1986) advocate this approach and maintain that this property is important even for a subjectivist. In this paper, this assessment is paramount given that we are concerned with a goodness of fit perspective.

The quantity of interest is then the distance between the true density and the proposed model, $d(f, \mathcal{F})$. As is often the case in nonparametric inference, we consider the Hellinger distance between two distributions F and G, defined as

$$d(F,G) = \left\{ \int \left(\sqrt{dF} - \sqrt{dG}\right)^2 \right\}^{1/2} .$$

Since we are only concerned with distributions absolutely continuous with respect to Lebesgue measure, we also denote $d(f,g)$ the Hellinger distance between F and G, where f, g are the densities with respect to the Lebesgue measure of F and G respectively. Then we define

$$d(f, \mathcal{F}) = \inf_{\theta\in\Theta} d(f, f_\theta)\,.$$

We approximate this quantity using its posterior expectation $E^\pi[d(f,\mathcal{F})|X^n]$, for some prior π on (θ,ψ). To test the parametric model, we must therefore compare the above posterior expectation with some reference quantity. Actually, the Bayes estimate under the loss function :

$$L(\delta, f) \quad = \quad \left\{ \begin{array}{ccc} a_0 d(f,\mathcal{F}) & \text{if} & \delta = 0 \\ a_1(2 - d(f,\mathcal{F})) & \text{if} & \delta = 1 \end{array} \right. \tag{2}$$

is given by $\delta(X^n) = 0$, i.e. we accept the null hypothesis, if and only if $E^\pi[d(f,\mathcal{F})|X^n] \leq 2a_1/(a_0 + a_1)$. In the general case, the choice of (a_0, a_1) is quite arbitrary. We therefore propose in this paper a way to calibrate $2a_1/(a_0 + a_1)$. In particular, a_0 should increase as the number of observations becomes larger. The informal perspective on this point is that if the parametric model is not far from the *true* model, it is better to use such a model, especially when the number of observations is not large. In other words, the smaller the sample size is, the more relevant the parametric model might get. The idea is then to compare $E^\pi[d(f,\mathcal{F})|X^n]$ with a quantity that would characterize its behaviour under the null hypothesis. A usual way to do it is to use posterior predictive p-values, see for instance Meng (1994). These p-values have interesting features in average, however in practice they can behave quite poorly due to a strong double use of the data. In this paper, we investigate the use of the conditional predictive p-value, as defined by Bayarri and Berger (2000), and associated with the test statistic $E^\pi[d(f,\mathcal{F})|X^n]$. We compute the distribution of $E^\pi[d(f,\mathcal{F})|Y^n]$, when Y^n is distributed according to

$$m_0(y^n|\hat{\theta}_x) = \int_\Theta f(y^n|\hat{\theta}_x;\theta)\pi_0(\theta|\hat{\theta}_x)d\theta\,, \quad \text{with} \quad \pi_0(\theta|\hat{\theta}_x) \propto g(\hat{\theta}_x|\theta)\pi_0(\theta), \tag{3}$$

where π_0 is the prior distribution of θ under the parametric model $\mathcal{F}$,

$$\hat{\theta}_x = \arg\min_\theta(l_n(\theta, x^n))$$

is the maximum likelihood estimator in the parametric model $\mathcal{F}$ associated with the observations $x^n = (x_1, ..., x_n)$ and $f(y^n|\hat{\theta}_x;\theta)$ is the conditional distribution of the sample y^n given the parameter θ under the fixed mle constraint

$$\hat{\theta}_y = \hat{\theta}_x\,.$$

The test consists in evaluating

$$p_{cpred} = \Pr\left[E^{m_0(.|\hat{\theta}_x)}[d(f,\mathcal{F})|y^n] \geq E^\pi[d(f,\mathcal{F})|X^n]\right],$$

where the probability is calculated under $m_0(y^n|X^n;\hat{\theta}_x)$. This quantity is the conditional predictive p-value, where the statistics on which we condition is the maximum likelihood estimator.

We prove, in Section 3.2 that such a test is equivalent to using a conditional p-value (conditional on the mle) and is consistent, in the sense that the above probability goes to zero as n goes to infinity under the alternative, see Theorem 4. This test procedure is therefore also satisfying from a frequentist point of view.

The paper is organised as follows: in Section 2, we study the problem of nonparametric estimation of a density in $[0, 1]$ using mixtures of Beta densities. In Section 3 we consider the goodness of fit test of a parametric model $\mathcal{F}$. We prove first that the posterior distribution of the full model is consistent almost surely and we deduce from that the consistency of the test procedure. In both sections simulations are given to illustrate the behaviour of the estimates of the density and of the test statistic.

2. Nonparametric estimation via mixtures of Beta distributions

As in Verdinelli and Wasserman (1998), to test the appropriateness of a parametric family, we need to consider the estimation of a general density on $[0, 1]$, hence, in this section we only consider the problem of estimating a given density in $[0, 1]$, which amounts to deriving a goodness of fit test for a specific density. Let thus $U_1, \cdots, U_n$ be a sample of random variables in $[0, 1]$ distributed from a density g_0. Our aim is mainly to test if g_0 is a uniform random variable, but at this stage we first consider the estimation of the density g_0. (The test can be derived from the general description in Section 3.)

2.1. Representation of the alternative hypothesis

Given that almost any distribution on $[0, 1]$ can be expressed as an infinite mixture of Beta distributions,

$$\sum_{k\geq 1} p_k \mathcal{B}(\alpha_k, \beta_k)\,,$$

as shown in Theorem 1 below, we define the general alternative to $U \sim \mathcal{U}([0, 1])$ to be

$$U \sim \sum_{k\geq 1} p_k \mathcal{B}(\alpha_k, \beta_k) \qquad \sum_{k\geq 1} p_k = 1\,.$$

We are thus facing a rather standard mixture estimation problem where the number of components is unknown, as in Richardson and Green (1997) or Stephens (2000). (The approach we follow is Stephen's (2000), as detailed below.) Due to the specificity of the testing problem, we reparameterise the mixture as follows:

$$p_0 \mathcal{U}(0, 1) + (1 - p_0) \sum_{k=1}^{K} p_k \mathcal{B}(\alpha_k \epsilon_k, \alpha_k(1 - \epsilon_k)) \qquad \sum_{k\geq 1} p_k = 1\,, \tag{4}$$

to signify that the null hypothesis corresponds to $p_0 = 1$ and that the alternative corresponds to $p_0 \neq 1$, under the identifiability constraint that none of the other components $\mathcal{B}(a_k, b_k)$ is equal to $\mathcal{U}(0, 1)$.

Given the difficult identifiability issues connected with mixtures (see Celeux *et al.*, 2000) and this representation of H_0, we circumvent this difficulty by (a) resorting to the estimation of the distance between (4) and $\mathcal{U}(0, 1)$, bypassing parameterisation problems, and by (b) selecting an appropriate prior distribution.

For simulation reasons discussed in Cappé *et al.* (2003), we also choose to replace the weights p_k with their unscaled version, ω_k, namely $(k = 1, \ldots, K)$

$$p_k = \frac{\omega_k}{\sum_{\ell=1}^{K} \omega_\ell}\,, \qquad 0 \leq \omega_k \leq 1\,.$$

Note at last that the representation of a Beta distribution as $\mathcal{B}e(\alpha_k \epsilon_k, \alpha_k(1 - \epsilon_k))$ is chosen to distinguish between the scale $\alpha_k > 0$ and the position $0 < \epsilon_k < 1$. The set of parameters of the mixture is then denoted by $\psi = (K, \omega_1, \cdots \omega_K, \alpha_1, \epsilon_1, \cdots, \alpha_K, \epsilon_K)$ which varies in the space $\mathcal{S}$.

We prove in Section 2.2 that the posterior mean of (4) is a consistent estimate of the density g_0, in terms of Hellinger distance.

2.2. Consistency of the posterior distribution

In this Section we give general conditions on the prior π_1 on ψ to achieve the convergence of the posterior distribution. This will then imply the consistency of Bayes estimates of the density such as the posterior mean as well as the consistency of the test procedure.

Let π_1 be a prior on $\psi \in S$; π_1 is assumed to satisfy the following conditions:

(a) $K \sim P(K)$. We assume that $\forall t > 0$, $\exists\, r > 0$ such that

$$P(K \geq tn/\log n) \leq e^{-rn}, \tag{5}$$

(b) $p_0 \sim \pi(p_0)$ a.c. wrt Lebesgue measure and with support $[0,1]$.

(c) Conditional on K, we denote $h(\omega_1, ..., \omega_K)$ the prior density on $(\omega_1, ..., \omega_K)$ wrt Lebesgue measure on $[0,1]^K$. We assume that h is continuous.

(d) Conditional on K, we assume that the (a_j, b_j)'s are independent with identical priors with density derived from (7) under the reparameterisation $a_j = \alpha_j \epsilon_j$ and $b = \alpha_j(1 - \epsilon_j)$.

Obviously we need not assume that K and p_0 are independent; however we consider such a prior as the basis of the following results. Note that the condition (5) is satisfied in particular by the Poisson distribution. In addition, the type of priors on ψ described in Section 2.3 and considered in the simulations will satisfy these conditions.

Let thus $U_1, ..., U_n$ be n iid observations from a distribution with density g_0 on $[0,1]$, wrt Lebesgue measure. Let

$$A_\varepsilon(g_0) = \{g : d(g_0, g) \leq \varepsilon\} \qquad \text{and} \qquad N_\varepsilon = \{g : \mathcal{I}(g_0, g) \leq \varepsilon\}\,,$$

where d is the Hellinger distance and $\mathcal{I}$ is the Kullback divergence,

$$\mathcal{I}(g_0, g) = \int_0^1 g_0 \log\left[\frac{g_0}{g(u)}\right] du\,.$$

First, we prove that the set of densities that can be approximated, in the sense of the Kullback-Leibler divergence, by a mixture of Beta distributions, contains the set Ω of densities g for which

$$\int_0^1 g(x)\log g(x) dx < \infty\,,$$

and which satisfy : $\forall M > 0$, the function is piecewise continuous on $\{x; g(x) \leq M\}$. It is in fact well-known (Petrone and Wasserman, 2002) that any continuous density on the closed set $[0,1]$ can be approximated by Bernstein polynomials, which constitute a subset of Ω. The following Theorem proves that more general densities can in fact be approximated.

Theorem 1. *Let $g \in \Omega$, then, for every $\varepsilon > 0$, there exists g_ψ, with $\psi \in \mathcal{S}$, such that*

$$\mathcal{I}(g, g_\psi) \leq \varepsilon.$$

The idea of the proof is the following: We approximate g by a continuous function on $\{g(x) \leq M\}$, when M is large enough, and we use the uniform approximation of a continuous function by Bernstein densities to obtain the Kullback-Leibler approximation.

Proof. Let $g \in \Omega$ and define $g_1 = (g \vee \varepsilon/12)/(\int(g \vee \varepsilon/3)(x)dx)$, then since

$$\int (g \vee \varepsilon/12)(x)dx \le 1 + \varepsilon/3,$$

$g \le g_1(1+\varepsilon/3)$ and we can work with g_1 instead of g, using Ghosal, Ghosh and Ramamoorthi's result (1999, Lemma 5.2). Let M be such that

$$\int_{g_1(x) \ge M} g_1(x) \log g_1(x) dx < \varepsilon/3,$$

and define $\tilde{g} = g_1 \mathbb{I}_{g_1 \le M} + M\mathbb{I}_{g_1 > M}$, then, since $\tilde{g}$ has a finite number of first order discontinuities there exists a continuous function $\tilde{g}_\varepsilon$ such that

$$\sup_{x\in[0,1]} |\tilde{g}(x) - \tilde{g}_\varepsilon(x)| \le \varepsilon/12.$$

Ghosal, Ghosh and Ramamoorthi's result (1999, Lemma 5.2) implies that there exists a Bernstein density $b(x)$ such that

$$\sup_{x\in[0,1]} |\tilde{g}(x) - b(x)| \le \sup_{x\in[0,1]} |\tilde{g}_\varepsilon(x) - b(x)| + \varepsilon/12 \le \varepsilon/6.$$

Therefore,

$$\begin{aligned} \int_{g_1(x) \le M} g_1(x) \log(g_1(x)/b(x)dx &= \int_{g_1(x) \le M} g_1(x) \log(\tilde{g}(x)/b(x)dx \\ &\le \int_{g_1(x) \le M} g_1(x) \log \frac{\tilde{g}(x)}{\tilde{g}(x) - \varepsilon/6} dx \le \frac{\varepsilon}{3}, \end{aligned}$$

since $\tilde{g}(x) \ge \varepsilon/3$. We also have

$$\begin{aligned} \int_{g_1(x) > M} g_1(x) \log(g_1(x)/b(x)dx &\le \int_{g_1(x) > M} g_1(x) \log \left(\frac{g_1(x)}{M - \varepsilon/6} \right) dx \\ &\le \int_{g_1(x) > M} g_1(x) \log \left(\frac{g_1(x)}{M} \right) + \frac{\varepsilon}{3M} \le \frac{\varepsilon}{3}, \end{aligned}$$

and Theorem (1) is proved. □

Note that Bernstein polynomials can approximate any function in Ω, however it is our belief that general mixtures of Betas would require less components in practice to approximate densities in Ω, in particular when the density is discontinuous.

We then have the following result on the posterior distribution:

Theorem 2. *Let $U_1, ..., U_n$ be independent and identically distributed r.v.'s from $g_0 \in \Omega$. Consider the prior π_1 satisfying the above conditions (a)–(d), then the posterior distribution of π converges in the following strong sense: $\forall \varepsilon > 0$,*

$$\pi[A_\varepsilon(g_0)|U_1, \ldots, U_n] \to 1, \quad g_0 \ a.s. \tag{6}$$

The consistency of the posterior mean of the density (which is a standard Bayesian estimate for the density) follows from (6) in terms of the Hellinger distance d. The proof of this theorem is obtained using Theorem 1 of Barron, Schervish and Wasserman (1998) *[hereafter BSW]* and is given in Appendix B.

Note also that

$$E^{\pi}[d(g_{\psi},1)|U_1,...,U_n] \rightarrow d(g_0,1), \quad g_0 \text{ a.s.}$$

since $E^{\pi}[d(g_0,g_{\psi})|U_1,...,U_n] \rightarrow 0$ g_0 a.s. as n goes to infinity.

We describe in the following section the type of priors on ψ we have considered in our simulations.

2.3. Priors for Beta mixtures

Although a regular conjugate prior could be used in this setting just as in Diebolt and Robert (1994) or Richardson and Green (1997), we now build a specific prior distribution in order to oppose the uniform component of the mixture (4) with the other components.

Although we have considered a fully nonparametric approach, for the alternative, we have in practice chosen a uniform $\{1,\ldots,K_{\max}\}$ distribution on the number of components, K, the prior

$$p_0 \sim Be(0.8,1.2),$$

on p_0 *[in order to favour small values of* p_0*, since the distribution* $Be(0.8,1.2)$ *has an infinite mode at* 0*]*, the prior

$$\omega_k \sim Be(1,k), \qquad k=1,\ldots,K,$$

on the ω_k's for parsimony reasons *[so that higher order components are less likely]*, and a prior of the form

$$\begin{aligned}(\alpha_k,\epsilon_k) \quad \sim \quad & \{1-\exp\left[-\left\{\beta_1(\alpha_k-2)^{c_3}+\beta_2(\epsilon_k-.5)^{c_4}\right\}\right]\} \\ & \exp\left[-\tau_0\alpha_k^{c_0}/2-\tau_1/\{\alpha_k^{2c_1}\epsilon_k^{c_1}(1-\epsilon_k)^{c_1}\}\right] ,\end{aligned} \tag{7}$$

on the (α_k,ϵ_k)'s, where $c_0,\ldots,c_4,\tau_0,\tau_1,\beta_1,\beta_2$ are hyperparameters. This choice is purposely designed to avoid the $(\alpha,\epsilon)=(2,1/2)$ region for the parameters of the other components. There obviously is a fair amount of arbitrariness there, but it fits our purpose that (a) the extra-components should avoid the uniform distribution as much as they can and (b) that small values of $\alpha\epsilon$ and $\alpha(1-\epsilon)$ should also be excluded.

In the following simulations, we took the specific form

$$(\alpha_k,\epsilon_k) \sim \left\{1-\exp\left[-\xi\left\{(\alpha_k-2)^2+(\epsilon_k-.5)^2\right\}\right]\right\}\exp\left[-\zeta/\{\alpha_k^2\epsilon_k(1-\epsilon_k)\}-\kappa\alpha_k^2/2\right] \tag{8}$$

illustrated by Figure 1 for a series of values of (ξ,ζ,κ). Our specific choice in the following, unless otherwise specified, is $(\xi,\zeta,\kappa)=(5,.01,.01)$, which corresponds to Figure 2.

2.4. Estimating the number of components

Although we are not aware of mixtures of Beta distributions being estimated in the past, there is nothing inherently complicated in the estimation of a mixture model

$$\sum_{k=1}^{K} p_k\mathcal{B}(\alpha_k,\beta_k)\,,$$

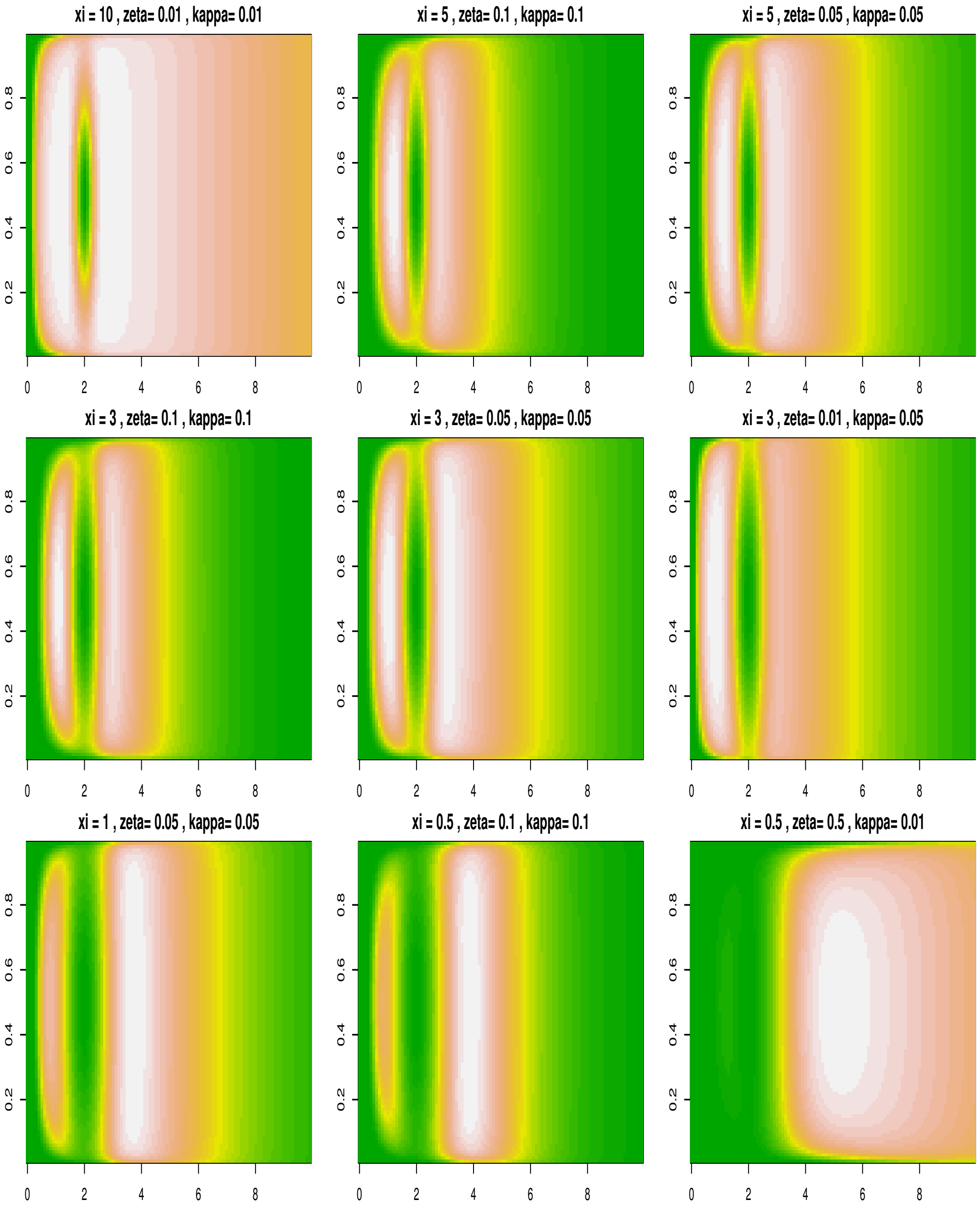

Fig. 1. R's `filled.contour` representation of the prior distribution (7) for various values of (ξ, ζ, κ)

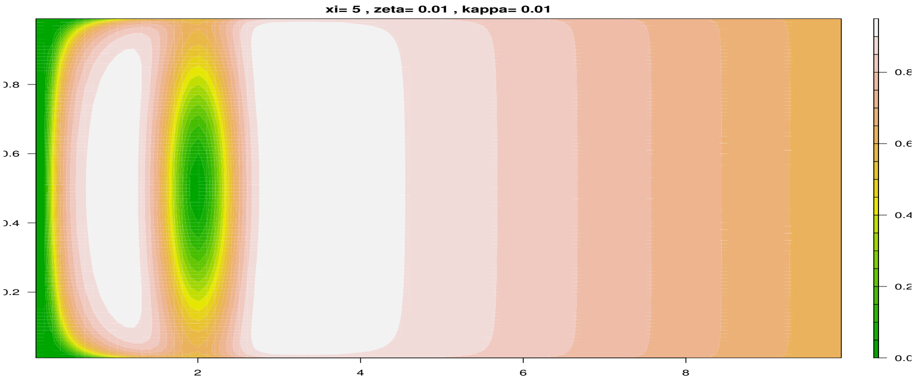

Fig. 2. R's `filled.contour` representation of the prior distribution (7) for $(\xi, \zeta, \kappa) = (5, 0.01, 0.01)$

with a fixed number of components K. For instance, a Gibbs sampling strategy as in Diebolt and Robert (1994) can be implemented, based on a completion of the sample $x_1, \ldots, x_n$ into $(x_1, z_1), \ldots, (x_n, z_n)$ where the z_i's are the component indicators,

$$z_i \sim \mathcal{M}(p_1, \ldots, p_K), \qquad x_i | z_i = k \sim \mathcal{B}(\alpha_k \epsilon_k, \alpha_k (1 - \epsilon_k))\,.$$

The simulation of the parameters (α, ϵ) is then based on either an accept-reject algorithm adapted to the distribution

$$\left\{1 - \exp\left[-\xi\left\{(\alpha - 2)^2 - (\epsilon - .5)^2\right\}\right]\right\} \exp\left[-\zeta / \{\alpha^2 \epsilon (1 - \epsilon)\} - \kappa \alpha^2 / 2\right]$$
$$\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha\epsilon)\Gamma(\alpha(1-\epsilon))}\right)^{n_k} \left\{\prod_{z_i = k} x_i\right\}^{\alpha\epsilon} \left\{\prod_{z_i = k} (1 - x_i)\right\}^{\alpha(1-\epsilon)}\,,$$

based on a $\mathcal{N}(0, 10) \times \mathcal{U}([0, 1])$ proposal, or more simply on a random walk Metropolis–Hastings proposal on $(\log\alpha, \log\epsilon/(1 - \epsilon))$. As noted in Celeux *et al.* (2000), the posterior distribution of a mixture problem is available in close form, except for the normalizing constant, and, therefore, direct [meaning, *without completion*] Metropolis–Hastings algorithms can be implemented.

The difficulty with this model arises when the number of components K is unknown. The setting is, however, familiar, in that several solutions for this problem have been proposed in the past, the two most prominent being Richardson and Green's (1997) reversible jump MCMC algorithm and Stephens' (2000) birth-and-death process algorithm, who both dealt with normal mixtures. Although both solutions are intrinsically equivalent, as discussed in Cappé *et al.* (2003), we chose to implement the birth-and-death process solution here, because the birth-and-death process approach is somehow simpler when no additional "split" and "combine" moves are required, borrowing Richardson and Green's (1997) terminology. In the case of normal mixtures, Stephens (2000) showed that the mixing properties of the algorithm were fairly good and we confirmed through simulations that this is equally the case here. Note that, in the case of hidden Markov models, Cappé *et al.* (2003) found that the "birth" and "death" steps were not sufficient to ensure proper moves for the MCMC chain of the K's and the θ's, thus requiring additional "split" and "combine" moves with a complexity then equivalent to Richardson and Green's (1997) algorithm.

We will not describe in detail Stephens' (2000) birth-and-death algorithm, nor will we give the corresponding description for Richardson and Green's (1997), enough details being available either in the original papers, or in Cappé *et al.* (2003). It is sufficient to mention here that the algorithm is based on a continuous time jump process that changes K at each jump by $+1$ *[birth]* or -1 *[death]*, with a fixed birth intensity λ_0 and a death intensity proportional to the sum of the likelihood ratios corresponding to the removal of one of the K components. The durations between jumps are exponential variates with inverse expectation the sum of the birth and the death intensities, that is, with $\theta_{(-k)} = (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_K)$,

$$T_i - T_{i-1} \sim \mathcal{E}xp\left(\lambda_0 + \frac{\lambda_0}{K}\sum_{k=1}^{K}\frac{L(K-1, \theta_{(-k)}|X^n)}{L(K, \theta|X^n)}\right),$$

except at the endpoints $K = 1$ and $K = K_{\max}$. Observation of the jump process chain at fixed time (or at every jump weighted by the duration time $T_i - T_{i-1}$) then leads to a stationary evaluation of the posterior distribution on (K, θ) (see Cappé *et al.*, 2003).

2.5. Simulations

The purpose of this paper being far from studying the performances of a birth and death jump process to evaluate the number of components in a mixture of Beta distributions, we simply report here some basic facts that ensure that the MCMC sampler is working well for our purpose. The illustration is thus based on 3 simulated data sets, the first one being artificially made of 1000 equidistant values on $[0, 1]$ which correspond to a flat histogram, the second one being made of random iid observations from a Beta distribution, and the third one being made of random iid observations from a mixture of two Beta distributions.

In the first case (Figures 3 and 4), the algorithm does capture the uniform structure of the sample and it produces an estimate of K equal to 0, as shown by the upper left and upper central graphs in the monitoring plots. The other graphs are not particularly relevant when $K = 0$, since they were designed for the non-uniform case $K > 0$. One can still notice that the posterior distribution on (α_k, ϵ_k) (lower left and lower center graph) is quite similar to the prior distribution (see Figure 2) and also that the posterior distribution on the p_0's when $K > 0$ (central left) is quite concentrated at 1.

In the second case (Figures 5 and 6), the unimodality of the distribution is again well-captured by the algorithm since the estimate of K is 1 (upper left and upper central graphs of Figure 5). In addition, the uniform part of the mixture is estimated as negligible (center left graph of Figure 5) and the position parameter ϵ_1 is well concentrated around 0.4, while α_1 has a wider variation due to the heavy tails of the histogram. (Note on Figure 5 (central left and center) the artifact induced by the fact that, when $K = 1$, ω_1 is taken equal to 1.) The fit by the "plugg-in" estimate

$$E^\pi[p_0|X^n]\mathcal{U}([0,1]) + E^\pi[(1-p_0)|X^n]\mathcal{B}\left(E^\pi[\alpha_1\epsilon_1|X^n], E^\pi[\alpha_1(1-\epsilon_1)|X^n]\right)$$

is quite satisfactory, as shown in Figure 6. It does not exhibit the poor tail fit of the average of the densities, which is due to the fact that the values of $\alpha_k\epsilon_k$ and of $\alpha_k(1-\epsilon_k)$ that are less than 1 pull the tails up.

In the third case (Figures 7 and 8), the two components are again well identified, the algorithm allocating approximately the same posterior weight to the cases $K = 2$ and $K = 3$ (upper left graph of Figure 7) but clearly exhibiting the bimodality of the distribution (lower

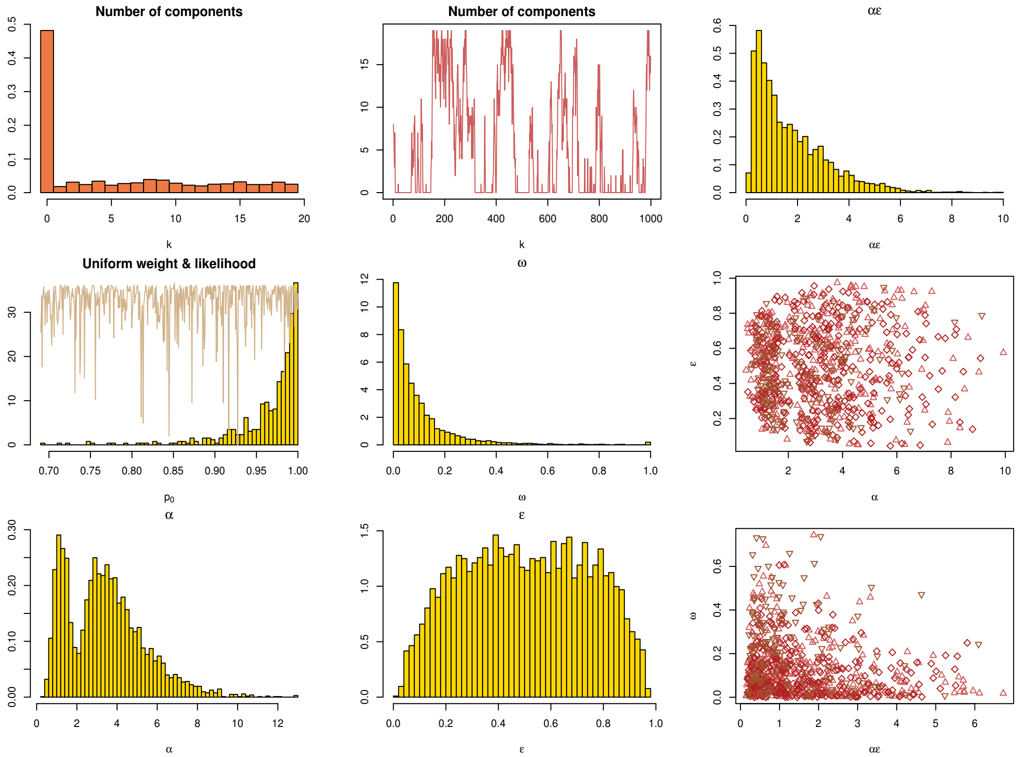

Fig. 3. Monitorings of the convergence of the birth and death sampler for an equidistributed sequence of 1000 points: *(upper left)* histogram of K; *(upper center)* sequence of the simulated K's; *(upper right)* histogram of the $\alpha_k\epsilon_k$'s; *(central left)* histogram of p_0 for $K > 0$ and sequence of the log-likelihoods; *(central center)* histogram of the ω_k's; *(central right)* plot of the (α_k, ϵ_k)'s for the three most likely values of $K > 0$; *(lower left)* histogram of the α_k's; *(lower center)* histogram of the ϵ_k's; *(lower right)* plot of the $(\alpha_k\epsilon_k, \omega_k)$'s for the three most likely values of $K > 0$.

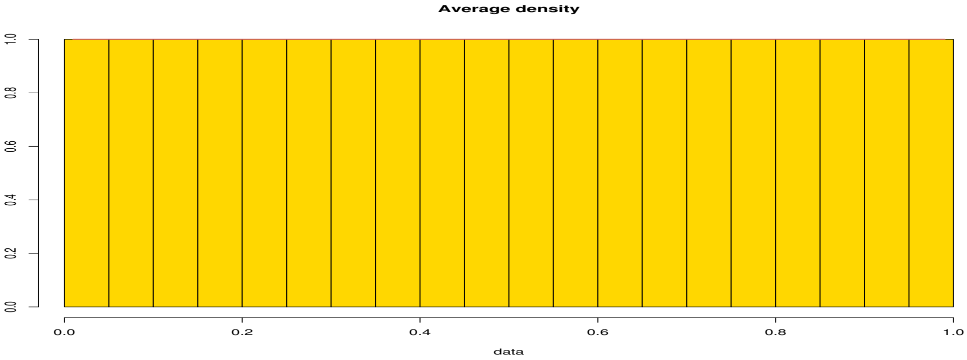

Fig. 4. Histogram of the equidistributed sequence of 1000 points and averaged density estimator.

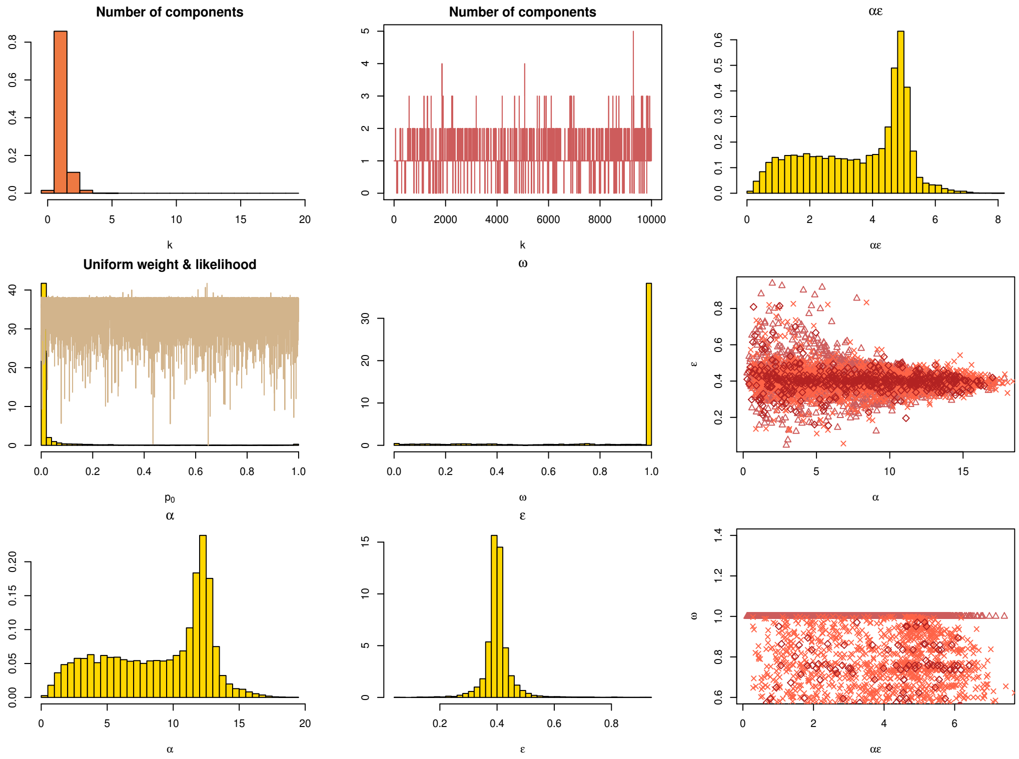

Fig. 5. Monitorings of the convergence of the birth and death sampler for a random sample of 1500 points (same legend as Figure 3).

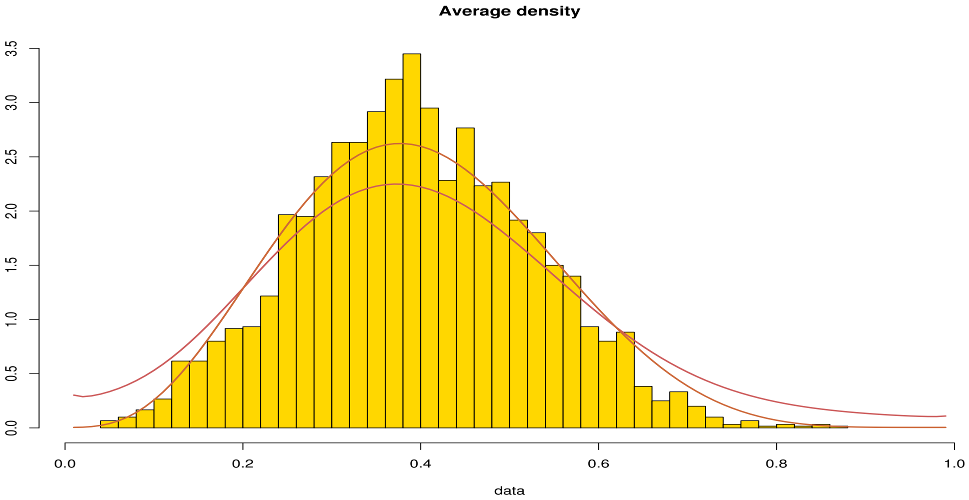

Fig. 6. Histogram of a random sample of 1500 points and averaged density estimators. The curve with the fatter tails corresponds to the average of the densities over the MCMC simulations and the curve with the thinner tails corresponds to the plugg-in estimate of the density where the parameters $(p_k, \alpha_k, \epsilon_k)$ are replaced by their estimates. These two curves are estimated conditional on $K = 1$.

central graph of Figure 7 and Figure 8). The same poor fit in the tail of the average of the densities can be observed in Figure 8, as well as the very good performances of the plugg-in estimate

$$E[p_0|X^n]\mathcal{U}([0,1]) + (1 - E[p_0|X^n])\left\{E[p_1|X^n]\mathcal{B}(E[\alpha_1\epsilon_1|X^n], E[\alpha_1(1-\epsilon_1)|X^n]) \right.$$
$$\left. + E[(1-p_1)|X^n]\mathcal{B}(E[\alpha_2\epsilon_2|X^n], E[\alpha_2(1-\epsilon_2)|X^n])\right\} .$$

We now go back to our original problem, namely the goodness of fit test for a parametric model.

3. General goodness of fit model

As in Verdinelli and Wasserman (1998) *[hereafter VW]*, we rewrite the problem of testing the appropriateness of a family of distributions

$$\mathcal{F} = \{F_\theta, \theta \in \Theta\},$$

a parameter $\theta \in \Theta$, for a given sample $x_1, , \ldots, x_n$, as a test of *uniformity* for the transforms $u_1 = F_\theta(x_1), \ldots, u_n = F_\theta(x_n)$. The alternative to this null hypothesis, also called *full model*, is that the sample $u_1, \ldots, u_n$ is distributed from an arbitrary distribution on $[0,1]$, represented as a possibly infinite mixture of Beta distributions, given by (1).

We first give a result on the consistency of the posterior distribution in terms of Hellinger neighbourhoods under some general conditions on the model $\mathcal{F}$. This implies in particular that the posterior mean is a consistent estimate of the density. Using this consistency result, we will then prove that the test procedure defined in Section 3.2 is consistent and that it is asymptotically equivalent to using a conditional p-value.

3.1. Consistency of the posterior distribution

The full model on the observations $x_1, , \ldots, x_n$ is thus given, in terms of densities, as

$$\mathcal{H} = \{f(x) = f_\theta(x) g_\psi(F_\theta(x)), \theta \in \Theta, \psi \in S\}, \tag{9}$$

where g_ψ is defined as in the previous section. We now establish the strong consistency of the posterior distribution, under some regularity conditions on f_θ and under the same conditions on $\pi_1(\psi)$ as in Section 2.2.

We denote by f_0 the true density of the x_i's and assume that $\Theta \subset \mathbb{R}^p$ is compact.; $\pi(\theta)$ is the marginal prior density on Θ. We consider the following assumptions:

H1 For all $\theta \in \Theta$, $\text{supp}(f_\theta) = \mathcal{X}$, independent of θ and $\text{supp}(f_0) \subset \mathcal{X}$.

H2 For all $\theta \in \Theta$, $\epsilon > 0$, $\exists\ \psi \in S$ such that

$$\mathcal{I}(f_0, f_\theta g_\psi(F_\theta)) \leq \epsilon.$$

H3 Almost all densities within the parametric family are at a finite Kullback divergence from f_0, that is,

$$\pi(\{\theta, \mathcal{I}(f_0, f_\theta) < \infty\}) = 1.$$

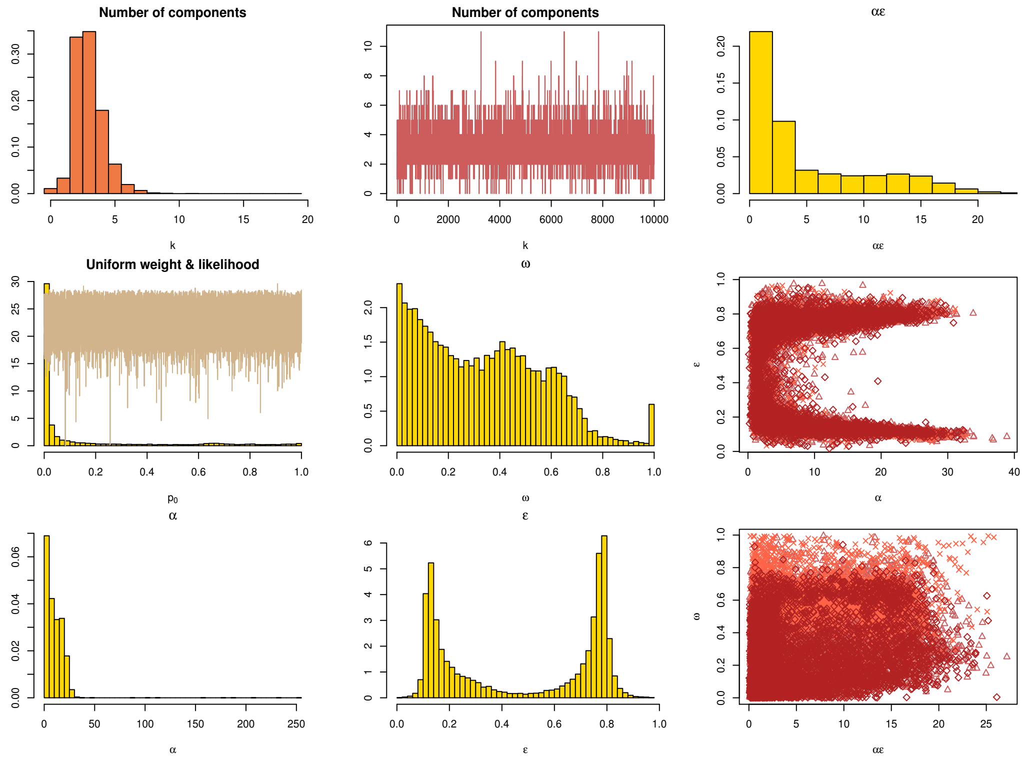

Fig. 7. Monitorings of the convergence of the birth and death sampler for a random sample of 1250 points (same legend as Figure 3).

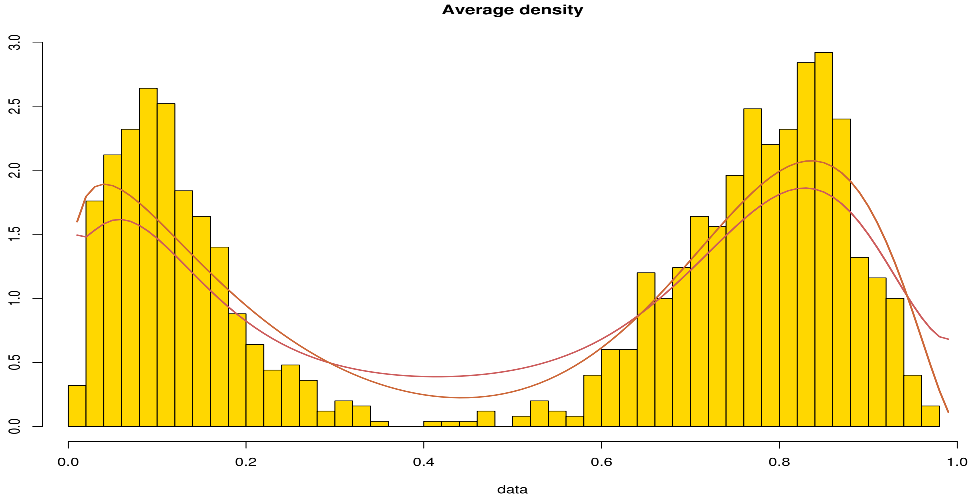

Fig. 8. Histogram of a random sample of 1250 points and averaged density estimators conditional on $K = 2$ (same legend as Figure 6).

H4 Assume that $\forall\theta \in \Theta$, f_θ is bounded and that $\exists\, \tau_0 > 0$ such that

$$\int_{\mathcal{X}} \sqrt{f_\theta^{1-\tau_0}}(x)dx < \infty.$$

H5 Assume that $\forall\theta \in \Theta$, $\exists d_0, \tau_1, C, \beta > 0$, such that $\forall d \leq d_0$, $\exists\, 0 < \tau \leq \tau_1 d^\beta$, $\exists\, m_1, m_2 > 0$ such that

$$m_1 f_\theta(x)^{1+\tau} \leq f_{\theta'}(x) \leq m_2 f_\theta(x)^{1-\tau}, \quad \forall|\theta' - \theta| < d,$$

and

$$m_2 \int_{-\infty}^{\infty} f_\theta(x)^{1-\tau} dx \leq 1 + Cd^\beta.$$

Hypothesis **H2** is the equivalent of Theorem 1 modulo the transformations F_θ of the data for all values of θ. While **H3** does not always hold (take for instance the case when f_0 is a Cauchy density and $\mathcal{F}$ is the family of Gaussian distributions), it is a general prerequisite for any hope of consistency of a test based on the likelihood. When **H2** holds, **H3** is very mild, given the compacity of Θ. Although the expressions of **H4** and **H5** are rather unusual, they are in essence fairly general and are satisfied for most known models, when the parameter space is compact. For instance, if $f_\theta(x) = \theta\, e^{-\theta x}$, with $\theta \in [\epsilon, E]$, $0 < \epsilon < E$, then if $|\theta' - \theta| \leq d\theta$,

$$(1-d)\theta^{-d} f_\theta(x)^{1+d} \leq f_{\theta'}(x) \leq (1+d)\theta^d f_\theta(x)^{1-d}.$$

More generally, **H4** holds for all bounded exponential families such that $f_\theta^{(1-\tau_0)/2}$ is still integrable and **H5** is a consequence of the boundedness assumption. Heavy tailed distributions also often satisfy **H4** and **H5**, at least when they have moments of order greater than 2 (for **H4** to be satisfied). We have chosen such an expression for the above hypotheses because it is more appropriate to the mixture of Beta distributions.

We then obtain the following consistency theorem :

THEOREM 3. *Under the conditions* **H1**–**H5**, $\forall\epsilon > 0$,

$$\pi[A_\epsilon(f_0)|X^n] \to 1, \quad as \quad n \to \infty, \quad f_0\, a.s. \tag{10}$$

This result implies that

$$E^\pi\left[d(f_0, f_{\theta,\psi})|X^n\right] \to 0, \quad f_0\, a.s.$$

as n goes to ∞. Moreover, since $|d(f_{\theta,\psi}, \mathcal{F}) - d(f_0, \mathcal{F})| \leq d(f_0, f_{\theta,\psi})$, Theorem 3 also implies that

$$E^\pi\left[d(f, \mathcal{F}))|X^n\right] \to d(f_0, \mathcal{F}), \quad f_0\, a.s. \tag{11}$$

as n goes to ∞. (The proof of Theorem 3 is given in Appendix C.)

The condition that Θ is compact could be relaxed, but that would imply conditions on the regularity of the f_θ's stronger than those considered here as well as conditions on the prior π.

3.2. Test of Goodness of Fit

One would like to obtain a test such that, if $d(f, \mathcal{F})$ is *small*, the parametric model is chosen. The difficulty is obviously to decide what *small* means. It could be chosen a priori, depending on the requirements of the statistician. In this case the parametric model would be selected if

$$H(X^n) = E^{\pi}[d(f, \mathcal{F})|X^n] \leq \epsilon,$$

where ϵ is fixed before the experiment. This is the Bayesian decision under the loss given by (2). As was mentioned in the introduction, situations, where ϵ can be fixed a priori, are not always possible and we also believe that to some extent ϵ should depend on the number of observations. In this respect, one could compare $H(X^n)$ with an approximation of its (frequentist) distribution under H_0. However, since θ is unknown, such an approximation is not available. We feel that a better approach consists in computing a reference distribution, characterizing the null hypothesis and being conditional on the observations. The most commonly used methods are then to compute either the plug-in p-value or the posterior predictive p-value. However both methods use twice the data in ways that can badly affect the result, as pointed out by Bayarri and Berger (2002). We therefore propose, here, a conditional predictive p-value, based on the maximum likelihood estimator. (In cases where its computation is intractable we can only suggest to use either a plug-in p value or a posterior predictive p-value.)

Let $\hat{\theta}_x$ be the maximum likelihood estimator in the parametric model, associated with the observations $x_1, ..., x_n$.

$$\begin{aligned}
\theta_1 &\sim \pi_0(\theta|\hat{\theta}_x)\,, & Y_1^n &\sim f(Y^n|\hat{\theta}_x; \theta_1), \\
\theta_2 &\sim \pi_0(\theta|\hat{\theta}_x)\,, & Y_2^n &\sim f(Y^n|\hat{\theta}_x; \theta_2), \\
&\cdots & &\cdots \\
\theta_N &\sim \pi_0(\theta|\hat{\theta}_x)\,, & Y_N^n &\sim f(Y^n|\hat{\theta}_x; \theta_N),
\end{aligned}$$

be iid copies from the conditional predictive distribution (conditional meaning conditional on the maximum likelihood estimator). Then

$$Y_i^n \sim m_0(Y^n|\hat{\theta}_x),$$

as defined by (3). For each Y_i^n we can calculate $H(Y^n)$ and thus get a sample from the predictive distribution of $H(Y)$ under the parametric model. If the null model is quite wrong then Y^n is very different from X^n and therefore, $H(Y^n)$ will be very different from $H(X^n)$. If the null model is correct, then Y^n is similar to X^n and so would be $H(Y^n)$ to $H(X^n)$. We then compare $H(X^n)$ with the quantiles of the predictive distribution of $H(Y)$. This predictive distribution is calculated under the parametric model. To make these statements more rigorous we now give the following asymptotic results.

First, we note that the general problem of simulating a sample conditional on a fixed maximum likelihood estimator $\hat{\theta}$ is quite interesting: within exponential families, since the maximum likelihood estimator is a sufficient statistic, the conditional distribution usually is straightforward. For instance, a normal $\mathcal{N}(\theta, 1)$ sample Y^n can be simulated as

$$(y_1, \ldots, y_{n-1}) \sim \mathcal{N}_{n-1}(\bar{x}\mathbf{1}, I - \frac{1}{n+1}\mathbf{1}\mathbf{1}^T)\,.$$

(See Section 3.3 for another illustration in the case of the $\mathcal{E}xp(\lambda)$ distribution.) Outside the exponential families, the conditional distribution depends on both the maximum likelihood estimator and the true parameter and its simulation requires more complex tools, like reversible jump MCMC and tempering (see Robert and Rousseau, 2003, for details).

We assume that $\theta^{\perp} = \arg\min_{\theta} \mathcal{J}(f_0, f_\theta)$ is unique (which is obvious when $f_0 \in \mathcal{F}$) and that the following three conditions hold, namely

$$\sup_{|\theta' - \theta^{\perp}| < 2\delta_n} E_0[|D^3 \log f_{\theta'}(X)|^4] < \infty, \tag{12}$$

$$\sup_{|\theta' - \theta^{\perp}| < 2\delta_n} E_0[|D^2 \log f_{\theta'}(X)|^4] < \infty, \tag{13}$$

and for all $H > 0$, there exists K large enough, such that

$$P_0^n \left[|\hat{\theta}_x - \theta^{\perp}| > K \log n/\sqrt{n} \right] \leq n^{-H}, \tag{14}$$

We do not really need (14) as written: it is enough to assume the condition with $H = 2$. However, this condition is generally satisfied under usual regularity conditions on the parametric model. In particular, (14) is almost a consequence of (12) and (13). In the case $f_0 \in \mathcal{F}$ (12), (13) and (14) are quite standard. If $f_0 \notin \mathcal{F}$, see Arcones (2002) for simple conditions to obtain (14).

Theorem 4. *Under the above conditions,*

$$p_{cpred} = P_\theta^n \left[H(Y^n) > H(x^n) | \hat{\theta}_y = \hat{\theta}_x ; x^n \right] + R_n,$$

where the residual R_n is such that, for some $p, p' \in \mathbb{N}$,

$$P_\theta^n [R_n > M \log n^p / \sqrt{n}] \leq M \log n^{p'} / \sqrt{n}\,.$$

If $f_0 \notin \mathcal{F}$, then $\theta = \theta^{\perp}$ and if $g_0(\hat{\theta})$, the density of the maximum likelihood estimate, allows for an Edgeworth expansion to the order $n^{-3/2}$, then

$$P_0^n [R_n > M\sqrt{n} \log n^p] \leq M \log n^{p'} [n^{-3c/2} + n^{-1/2}]\,,$$

for some constant $0 < c \leq 1$ depending on f_0, and some $p, p' \in \mathbb{N}$.

This result implies in particular that if the true density $f_0 \notin \mathcal{F}$, for all $\varepsilon > 0$,

$$\lim_{n \to \infty} P_0^n \left[p_{cpred} > \varepsilon \right] \to 0\,.$$

Also, under H_0, i.e. if $f_0 = f_\theta$,

$$P_\theta^n \left[p_{cpred} \leq \alpha \right] = \alpha + O(n^{-1/2}).$$

The assumption on the Edgeworth expansion of the maximum likelihood estimate, when $f_0 \notin \mathcal{F}$, is not a strong assumption, regularity conditions on the parametric model associated with moment conditions on the derivatives of the log-likelihood under f_0 (such as (12) and (13)) would typically imply such a result. Note also that this result is true for any test statistic H under the above regularity conditions on the parametric model.

The test is therefore asymptotically equivalent to using a conditional p-value, which has a uniform distribution under H_0. Thus, asymptotically, p_{cpred} is uniformly distributed on $[0, 1]$. Note also that marginally, p_{cpred} is exactly uniformly distributed on $[0, 1]$.

Proof. The proof of Theorem 4 is given in Appendix D. The idea is to use the fact that $\hat\theta$ is asymptotically a sufficient statistic, so that to first order $f(y^n|\hat\theta_x;\theta)$ is independent of θ. Therefore writing p_{cpred} as

$$\begin{aligned} p_{cpred} &= \int \mathbb{I}_{H(y^n)>H_x} \frac{\int_\Theta f(y^n|\hat\theta_x;\theta)g(\hat\theta_x|\theta)\pi(\theta)d\theta}{\int_\Theta g(\hat\theta_x|\theta)\pi(\theta)d\theta} \\ &= \int_{A_n^c} \mathbb{I}_{H(y^n)>H_x} \frac{\int_{|\hat\theta_x-\theta|<\delta_n} f(y^n|\hat\theta_x;\theta)g(\hat\theta_x|\theta)\pi(\theta)d\theta}{\int_\Theta g(\hat\theta_x|\theta)\pi(\theta)d\theta} \\ &\quad + \int_{A_n} \mathbb{I}_{H(y^n)>H_x} \frac{\int_{|\hat\theta_x-\theta|<\delta_n} f(y^n|\hat\theta_x;\theta)g(\hat\theta_x|\theta)\pi(\theta)d\theta}{\int_\Theta g(\hat\theta_x|\theta)\pi(\theta)d\theta} \\ &\quad + \int \mathbb{I}_{H(y^n)>H_x} \frac{\int_{|\hat\theta_x-\theta|>\delta_n} f(y^n|\hat\theta_x;\theta)g(\hat\theta_x|\theta)\pi(\theta)d\theta}{\int_\Theta g(\hat\theta_x|\theta)\pi(\theta)d\theta}, \end{aligned}$$

where $\delta_n = K\log n/\sqrt{n}$ and where A_n is essentially a set on which $f(y^n|\hat\theta_x;\theta)$ is asymptotically independent of θ. We then use the fact that $P_\theta^n[A_n^c]$ is small and that $g(\hat\theta_x|\theta)$ is also small when $|\hat\theta_x-\theta| > K\log n/\sqrt{n}$, for K large enough. □

This test procedure is therefore asymptotically equivalent to using a conditional p-value, up to a constant. Note that the order of approximation here is $O(n^{-1/2})$ which is a lot smaller than the nonparametric rate of convergence of $E^\pi[d(f,f_0)|X^n]$.

To simplify the computation of the test procedure, which is quite heavy, we can use $G(X^n) = E^\pi[d(g_\psi,1)|X^n]$ (and $G(Y^n)$) instead of $H(X^n)$ (and $H(Y^n)$). Indeed $G(X^n)$ and $G(Y^n)$ have the same properties as $H(X^n)$ and $H(Y^n)$. We have

$$\begin{aligned} G(X^n) &= E^\pi[d(f_{\theta,\psi},f_\theta)|X^n] \\ &\geq d(f_0,\mathcal{F}) - E^\pi[d(f_0,f_{\theta,\psi})|X^n] \end{aligned} \tag{15}$$

and

$$G(X^n) \leq E^\pi[d(f_0,f_\theta)|X^n] + E^\pi[d(f_0,f_{\theta,\psi})|X^n] \tag{16}$$

Hence, if $d(f_0,\mathcal{F}) = \epsilon > 0$, i.e. under H_1, using (15) and the consistency of the posterior, we obtain that $G(X^n)$ is asymptotically almost surely greater than ϵ. If $f_0 \in \mathcal{F}$, then both $E^\pi[d(f_0,f_\theta)|X^n]$ and $E^\pi[d(f_0,f_{\theta,\psi})|X^n]$ go to 0 as n goes to infinity, almost surely. H and G therefore have the same asymptotic behaviour and we can build the same test procedure with G as with H.

Such a simplification is however not entirely satisfying, though, since it forgets the width of $\mathcal{F}$; it is somehow, like reducing $\mathcal{F}$ to the pluggin density $f_{\hat\theta}$, where $\hat\theta$ is for instance the maximum likelihood estimator.

3.3. Evaluation

Two representative cases are considered for the evaluation of our testing procedure: the parameterised family $\mathcal{F}$ is the exponential distribution $\mathcal{E}xp(\theta)$ and the data is simulated either from a $\mathcal{E}xp(1)$ distribution or from a Gamma $\mathcal{G}a(7,1/7)$ distribution, that is, when H_0 holds and when it does not.

In this case, the conditional distribution of a sample $Y_1, \ldots, Y_n$ given the maximum likelihood estimator $\hat{\theta}_Y = \hat{\theta}_x$ is independent of the true parameter value, since $\hat{\theta}_x$ is a sufficient statistics. For the computation of the conditional p-value, we thus simulate samples Y_j^n as uniform distributions over the simplex $\sum_{i=1}^n y_i = n\hat{\theta}_x$, which is equivalent to simulate from the Dirichlet distribution $\mathcal{D}_n(1, \ldots, 1)$, and multiply by $n\hat{\theta}_x$.

For our computation of $H(Y^n)$, we replaced the Hellinger distance with the L_1 distance, $\int |1 - g_\psi(x)| dx$, between the uniform distribution and the mixture of beta distributions, in order to simplify the numerical approximation (which uses a quadrature of 1000 points).

Using 1500 MCMC iterations and $N = 250$ Monte Carlo replications, we obtained the distributions represented on Figures 9 and 10 for the normal and gamma samples. In the first case, that is when H_0 holds, for $n = 50$, the predictive p-value is 0.82 and for $n = 80$, 0.75, thus within acceptable values. (This corresponds to $H(y)$ equal to 0.14 and 0.13. In the second case, that is when H_0 does not hold, for both $n = 40$, and $n = 90$, the predictive p-value is 0.00: no simulated $H(Y^n)$ has produced a value larger than the observed $H(x^n)$, equal to 0.38 and 0.44, respectively.

References

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

Barron, A., Schervish, M.J. and Wasserman, L. (1999) The consistency of posterior 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. American Statist. Assoc.* **95**, 1127–1142.

Bhattacharya, R. and Ghosh, J.K. (1978) On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6**, 434–451.

Cappé, O., Robert, C.P. and Rydén, T. (2003) Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *J. Roy. Statist. Soc. Ser. B* (to appear).

Celeux, G., Hurn, M. and Robert, C.P. (2000) Computational and inferential difficulties with mixtures posterior distribution *J. American Statistical Society* **95**, 957–979.

Diaconis, P. and Freedman, D. (1986) On the consistency of Bayes estimates. *Ann. Statist.*, **14**, 1-26.

Diebolt, J. and Robert, C.P. (1994) Estimation of finite mixture distributions by Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56**, 363–375.

Fraser, D.A.S and Reid, N. (1995) Ancillaries and third order significance, *Utilitas Mathematica*, **47**, 33-53.

Genovese, C. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** 1105–1127.

Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999) Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**(1), 143–158.

Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

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

Petrone, S. and Wasserman, L. (2002) Consistency of Bernstein polynomial posteriors. *J. Roy. Statist. Soc. Ser. B* **64**, 79–100.

Richardson, S. and Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 731–792.

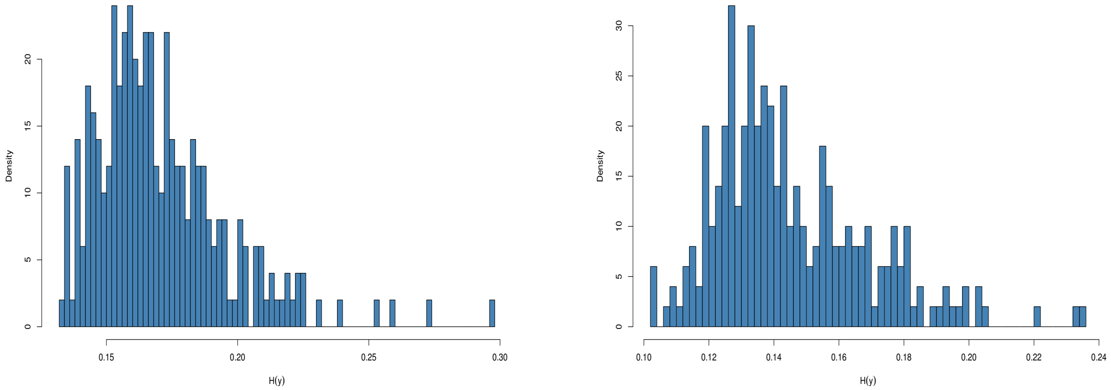

Fig. 9. Predictive distribution of the distance $H(Y^n)$ for $n = 50$ and $n = 80$, based on a sample x from a $\mathcal{E}xp(\theta)$ distribution. The observed $H(x^n)$ are equal to 0.14 and 0.13, respectively.

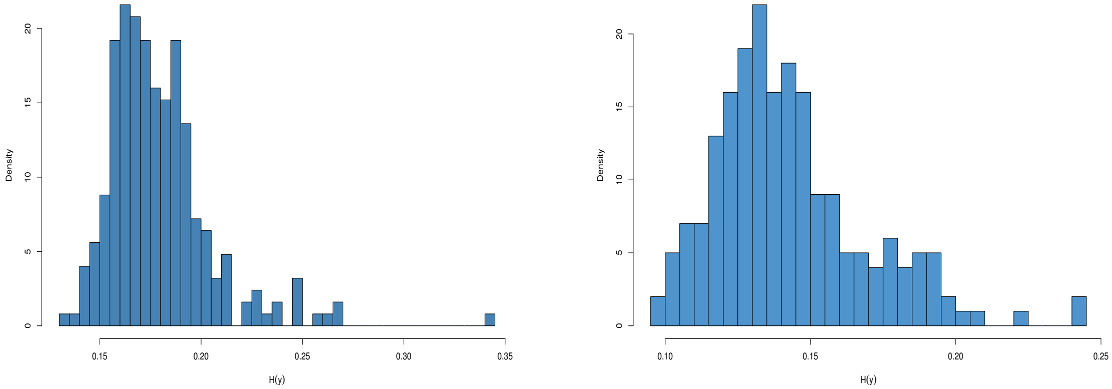

Fig. 10. Predictive distribution of the distance $H(Y^n)$ for $n = 40$ and $n = 90$, based on a sample x from a $\mathcal{G}a(7, 1/7)$ distribution.

Robert, C.P. (2001) *The Bayesian Choice* (second edition). Springer-Verlag, New York.

Robert, C.P. and Rousseau, J. (2003) Simulation of samples under maximum likelihood constraints. *In preparation.*

Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *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.

A. A theorem of Barron, Schervish and Wasserman

Let $\mathcal{P}$ be the set of probabilities on $\mathcal{X}$. For $\varepsilon > 0$ and $\mathcal{C} \subseteq \mathcal{P}$, define $\mathcal{L}(\mathcal{C}, \varepsilon)$ to be the logarithm of the infimum of the set of all k such that there exist nonnegative functions $f_1^U, \ldots, f_k^U$ such that

(a) $\int f_i^U(x) d\mu(x) \leq 1 + \varepsilon$ for all i,

(b) for each $P \in \mathcal{C}$ there exists i such that $f_P \leq f_i^U$ μ–a.s.

We now recall Theorem 1 of Barron *et al.* (1999), which enables us to prove the strong consistency of the posterior distribution. To do so, we first state the two conditions that have to be checked in their theorem:

A1 For every $\varepsilon > 0$, $\pi(N_\varepsilon) > 0$.

A2 For every $e > 0$, there exist a sequence $(\mathcal{F}_n)_{n=1}^\infty$ of subsets of $\mathcal{P}$, and positive, real numbers c_1, c_2, c_3 and ε, with

$$c_3 < ([e - \sqrt{\varepsilon}]^2 - \varepsilon)/2, \quad \varepsilon < e^2/4,$$

such that

(i) $\pi(\mathcal{F}_n^c) \leq c_1 \exp(-nc_2)$ for all but finitely many n;

(ii) $\mathcal{E}(\mathcal{F}_n, \varepsilon) \leq nc_3$ for all but finitely many n.

Barron *et al.* (1999) prove the consistency of the posterior distribution under these two hypotheses

Theorem 1 of Barron *et al.* **(1999)**: *Let A_ε be a Hellinger neighbourhood of f_0 the true density, which is defined in Section 2.2. Under conditions* **A1** *and* **A2**, *for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} \pi(A_\varepsilon | X^{(n)}) = 1 \quad P \text{ a.s.}$$

B. Proof of Theorem 2

The proof of this theorem is obtained using Theorem 1 of Barron, Schervish and Wasserman (1998) *[hereafter BSW]*, recalled above. To begin with, we prove condition **A1** in Theorem 1 of BSW. Let $g_0 = g_\psi$, where $\psi = (p_0, K, \omega_i, \alpha_i, \epsilon_i, i \leq K)$. So we first consider a finite mixture of Beta distributions. Then $\pi(N_\varepsilon) > \pi(N_\varepsilon^K)$, where N_ε^K is the set of densities in N_ε, that are mixtures of K Beta distributions.

To obtain condition [A1] in Theorem 1 of BSW, we prove that there exists $\delta > 0$ such that $|\psi - \psi'| < \delta$ implies that $f_{\psi'} \in N_\epsilon$, and thus $\pi(N_\epsilon) > \pi[\{|\psi - \psi'| < \delta\}]$. The prior density being strictly positive, the above probability will then be strictly positive.

When K is fixed, the model is a parametric model. A Taylor expansion of $\log g_{\psi'}$ around ψ leads to

$$|\log g_{\psi'}(u) - \log g_\psi(u)| \quad \leq \quad M(1 + \log u)|\psi - \psi'|,$$

and thus

$$\mathcal{I}(g_\psi, g_{\psi'}) \leq M|\psi - \psi'| \int_0^1 [1 + \log u]u^{-t}(1-u)^{-t}du \leq M'|\psi - \psi'|.$$

Let us consider some density $g_0 \in \Omega$ on $[0,1]$. Theorem 1 implies that $\forall \varepsilon > 0$, there exists $\psi = (p_0, K, \omega_j, a_j, b_j, j = 1, ..., K)$ such that $\mathcal{I}(g_0, g_\psi) \leq \varepsilon/2$. Using the above calculations, we deduce that for any $g_0 \in \Omega$, condition [A1] is satisfied.

We now consider condition **A2**. We construct $\mathcal{F}_n$ in the following way:

$$\mathcal{F}_n = \{g_\psi; K \leq t_n / \log n, t_n < a_j, b_j < T_n, j = 1, ..., K\},$$

with $T_n = n^l$, with $l \geq 1/c_0$ and $t_n = 2e^{-T_n}$, where $c_0 > 0$ is defined by (7). Then simple calculations imply that $\pi(\mathcal{F}_n^c) \leq e^{-nr}$, for some $r > 0$.

Let $\mathcal{A}_n = \{g_{a,b},\, 0 < t_n \leq a, b \leq T_n\}$, where $g_{a,b}$ is a Beta density with parameters a and b and $\eta = (a,b) \in [t_n, T_n]^2$.

Denote $\tau = (\tau_1, \tau_2)$, $\bar{\eta} = a + b$, $\bar{\tau} = \tau_1 + \tau_2$, and let $B(\eta)$ be the renormalising constant of the Beta density with parameter $\eta = (a,b)$ and C and ρ be generic positive constants. For all $\eta' = (\eta'_1, \eta'_2) \in [\eta - \tau, \eta + \tau]$

$$g_{\eta'}(u) \leq g_{\eta - \tau}(u) \frac{B(\eta - \tau)}{B(\eta + \tau)} = g^U(u).$$

We now determine conditions on τ_1 and τ_2 such that

$$\int g^U(u)du = \frac{B(\eta - \tau)}{B(\eta + \tau)} \leq 1 + \delta\,. \tag{17}$$

Using simple calculations on $\log \Gamma(x)$, we obtain that

(i) If $a, b < 2$, $i = 1, 2$

$$\log\left(\frac{\Gamma(a - \tau_1)\Gamma(b - \tau_2)}{\Gamma(a + \tau_1)\Gamma(b + \tau_2)}\right) + \log\left(\frac{\Gamma(\bar{\eta} + \bar{\tau})}{\Gamma(\bar{\eta} - \bar{\tau})}\right) \quad \leq \quad \frac{2\tau_1}{a - \tau_1} + \frac{2\tau_2}{b - \tau_2} - 2(\tau_1 + \tau_2)C.$$

Then the integral is bounded by $1 + \delta$ if $\tau_1 \leq \delta\rho a$ and $\tau_2 \leq \delta\rho b$, with $1/2 > \rho > 0$.

(ii) If $a < 2$, $b > 2$, then $\bar{\eta} > 2$ and

$$\log\left(\frac{\Gamma(a - \tau_1)\Gamma(b - \tau_2)}{\Gamma(a + \tau_1)\Gamma(b + \tau_2)}\right) + \log\left(\frac{\Gamma(\bar{\eta} + \bar{\tau})}{\Gamma(\bar{\eta} - \bar{\tau})}\right) \quad \leq \quad \frac{2\tau_1}{a - \tau_1} + \bar{\tau}[\log(\bar{\eta} + 1) - C].$$

Then the integral is bounded by $1 + \delta$ if $\tau_1 \leq \rho\delta a \left(1 + \log(\bar{\eta} + 1)\right)^{-1}$ and $\tau_2 \leq \rho\delta \left(1 + \log(\bar{\eta} + 1)\right)^{-1}$.

(iii) If $b < 2$, $a > 2$, then things are symmetrical to the previous case.
(iv) If $a, b > 2$, $i = 1, 2$, then

$$\log\left(\frac{\Gamma(a-\tau_1)\Gamma(b-\tau_2)}{\Gamma(a+\tau_1)\Gamma(b+\tau_2)}\right) + \log\left(\frac{\Gamma(\bar{\eta}+\bar{\tau})}{\Gamma(\bar{\eta}-\bar{\tau})}\right) \quad \leq \quad -2(\tau_1+\tau_2)[\rho - \log(\bar{\eta}+1)].$$

The integral is then bounded by $1+\delta$ if $\tau_i \leq \rho\delta[1+\log(\bar{\eta}+1)]^{-1}$, $i = 1, 2$.

We now count the number of upper bounds in $\mathcal{F}_n$:

(i) In the cube $[t_n, 2]^2$, $t\,(1+\rho\delta)^K \geq 2$ implies that the number of upper bounds in this cube is bounded by:

$$N_1 \leq \left(\log(2/t_n)\log(1+\rho\delta)^{-1}\right)^2.$$

(ii) In the cubes $[t_n, 2] \times [2, T_n]$ or $[2, T_n] \times [t_n, 2]$, in each column (for b fixed), the number of upper bounds is bounded by

$$K \leq \log(2/t_n)\log\left(1+\rho\delta\log(3+T_n)^{-1}\right)^{-1} \leq 2\log(2/t_n)\frac{\log(3+T_n)}{\rho\delta},$$

when T is large enough. The total number of bounds in the cube is then bounded by:

$$N_2 \leq \frac{\log(2/t_n)C}{\delta^2} T_n \log {T_n}^2.$$

(iii) In the cube $[2, T_n] \times [2, T_n]$, the number of upper bounds is bounded by:

$$N_3 \leq \frac{C}{\delta^2} T_n^2 \log {T_n}^2.$$

Finally, the total number of cubes is bounded by

$$\mathcal{N} = \frac{3C}{\delta^2} T_n^2 \log {T_n}^2, \qquad \text{since} \quad t_n = 2e^{-T_n}.$$

Using Genovese and Wasserman (2000), we obtain that the number of upper bounds for the elements of $\mathcal{F}_n$ can be bounded by

$$\begin{aligned} \mathcal{N}_n &= \sqrt{2(k_n+1)}\frac{B^{2k_n}}{\epsilon^{2k_n}}\mathcal{N}^{k_n} \\ &= \sqrt{2(tn/\log n+1)}\frac{(3MB)^{2tn/\log n}}{\delta^{4tn/\log n}}\left(T^2\log T^2\right)^{tn/\log n}. \end{aligned}$$

When $T_n = n^l$, with $l \geq 1/c_0$ and by choosing $t = c/6l$, we obtain $\log \mathcal{N}_n \leq nc$ and (6) is proved.

C. Proof of Theorem 3

As in Theorem 1, the proof is based on Theorem 1 of BSW. The hypothesis **H2** implies that $\forall\varepsilon > 0$, $\forall\theta \in \Theta$,

$$\pi[N_\varepsilon|\theta] > 0,$$

thus $\pi[N_\varepsilon] > 0$ and condition **A1** in BSW is satisfied. We now prove condition **A2**. Let

$$\bar{\mathcal{F}}_n = \{f_\theta(x)g_\psi(F_\theta(x)), \psi \in \mathcal{F}_n, \theta \in \Theta\},$$

and construct the upper bounds g_j^U as in the previous section, i.e. in the proof of Theorem 2, but with the constraint:

$$\int_0^1 g_j^U(u)du \leq 1 + \delta/2,$$

instead of δ. Since $f_{\theta,\psi}$ is a mixture of parametric densities, we first count the number of upper bounds for $g_\psi = g_{a,b}$, i.e. a Beta density with parameters $(a,b) \in \mathbb{R}_+^2$, as in the proof of Theorem 2. Then, $g_j(u)$ has the form of a Beta distribution with a larger renormalisation constant: it can be written as

$$g_j(u) = g_{a,b}(u)(1 + \delta/2)\,.$$

We can therefore work as if $g_j = g_{a,b}$. As in the proof of Theorem 2, let $t_n \leq a, b \leq T_n$, with $T_n = n^l$, $l \geq 1/c_0$ and $t_n = n^{-\alpha}$, for some $\alpha \geq c_1$ so that $\pi[\mathcal{F}_n^c] \leq e^{-nr}$ for some $r > 0$. Throughout the proof, C denotes a generic constant.

We thus need to bound

$$\sup_{|\theta'-\theta|<d} f_{\theta'}(x)g_{a,b}(F_{\theta'}(x)).$$

First, let $a, b > 1$ and denote

$$h_\theta(x) = H_1^{-1}f_\theta(x)^{1-\tau}, \quad H_\theta(x) = \int_{-\infty}^x h_\theta(y)dy,$$

where H_1 is the renormalising constant. Note that the hypothesis **H5** implies that $H_1 < \infty$. We have

$$\begin{aligned} 1 - F_{\theta'}(x) &= \int_x^\infty f_{\theta'}(x)dx \leq m_2H_1(1 - H_\theta(x)), \\ F_{\theta'}(x) &\leq m_2H_1F_\theta(x), \end{aligned}$$

thus, $\forall|\theta' - \theta| < d$, with $d \leq d_0$,

$$f_{\theta'}g_{a,b}(F_{\theta'})(x) \leq m_2^{a+b-1}H_1^{a+b-1}h_\theta(x)g_{a,b}(H_\theta(x)) = \bar{h}(x).$$

So,

$$\int_{\mathcal{X}} \bar{h}(x)dx = m_2^{a+b-1}H_1^{a+b-1} \leq 1 + \delta/3$$

if $m_2H_1 \leq (1 + \delta/3)^{1/(a+b-1)}$. Replacing a, b by $T_n = n^l$, this is satisfied if

$$m_2H_1 \leq 1 + \delta/(8T_n). \tag{18}$$

Hypothesis **H5** implies that (18) is valid when $d \leq \delta n^{-l/\beta}/(8C)$. The number of such upper bounds, for fixed a, b is then bounded by $N(\Theta)_n^1 \leq C\delta^{-p}n^{-pl/\beta}$.

Let $a < 1$ and $b > 1$ (or similarly $a > 1$ and $b < 1$). Writing $h_\theta = h_\theta^\alpha h_\theta^{1-\alpha}$, with $\alpha = (1 + \tau)/(1 + 2\tau)$ and using Holder's inequality, we obtain,

$$H_\theta(x)^{1+2\tau} \leq \left(\int_{-\infty}^x h_\theta^{1+\tau}(y)dy\right)\left(\int_{-\infty}^x \sqrt{h_\theta}(y)dy\right)^{2\tau} \tag{19}$$

This is finite because of hypothesis **H4**. Hypothesis **H5** implies that

$$F_{\theta'}(x) \geq m_1 \int_{-\infty}^{x} h_\theta^{(1+\tau)/(1-\tau)}(y)dy,$$

we obtain, using $\tau' = (1+\tau)/(1-\tau) - 1$ instead of τ in equation (19),

$$\begin{aligned} F_{\theta'}(x) &\geq \frac{m_1}{\left(\int_{-\infty}^{\infty} \sqrt{h_\theta}(y)dy\right)^{2\tau'}} H_\theta(x)^{1+2\tau'} \\ &= m_1' H_\theta(x)^{1+2\tau'}. \end{aligned} \tag{20}$$

Note that m_1' goes to 1 and τ' goes to 0, as d goes to 0. We thus have

$$\begin{aligned} f_{\theta'}(x)g_{a,b}(F_{\theta'}(x)) &\leq B(a,b)^{-1}H_1 m_2 (m_1')^{a-1} h_\theta(x) H_\theta(x)^{(1+2\tau')(a-1)}(1 - m_1' H_\theta(x)^{(1+2\tau')})^{b-1}dx \\ &\leq H_1 m_2 (m_1')^{a-1}(1+\tau') h_\theta(x) H_\theta(x)^{(1+2\tau')(a-1)}(1 - m_1' H_\theta(x))^{(1+2\tau')(b-1)} \\ &= \bar{h}(x), \end{aligned}$$

which implies that

$$\begin{aligned} \int_{\mathcal{X}} \bar{h}(x)dx &\leq H_1 m_2 (m_1')^{a-1}(1+\tau') \int_0^1 u^{(1+2\tau')(a-1)}(1 - m_1' u)^{(1+2\tau')(b-1)} du \\ &\leq H_1 m_2 (m_1')^{-2}(1+\tau') \frac{B(a',b')}{B(a,b)}, \end{aligned}$$

where $a' = a + \tau'(a-1)$ and $b' = b + \tau'(b-1)$.

Therefore, $\int_{\mathcal{X}} \bar{h}(x)dx \leq 1 + \delta/3$ if

(i) $H_1 \leq 1 + \delta/15$.
(ii) $m_2 \leq 1 + \delta/15$.
(iii) $m_1 H_0^{-4\tau/(1-\tau)} \geq (1+\delta/15)^{-1}$, with $H_0 = \int_{\mathcal{X}} f_\theta(x)^{1-\tau_1}dx$, for some fixed $\tau_1 \geq \tau$.
(iv) $B(a',b')/B(a,b) \leq 1+\delta/15$. Using the calculations of Section 2.2, this will be satisfied if $2\tau(1-a)/(1-\tau) \leq \rho\delta a(1+\log{(b+2)})^{-1}$ and $2\tau(b-1)/(1-\tau) \leq \rho\delta(1+\log{(b+2)})^{-1}$, for some $\rho > 0$.

Note that τ depends on d the distance between θ' and θ. As a crude upper bound we can let $a = t_n$ and $b = T_n$, so that when n is large enough, the most constrictive condition is (iv). We thus need

$$\tau \quad \leq \quad \rho'\delta(1 + l\log n)^{-1}n^{-h} = \tau_n, \tag{21}$$

where $h = \max(l, \alpha)$. Let d_n be such that when $|\theta' - \theta| \leq d_n$,

$$m_1 f_\theta(x)^{1+\tau_n} \leq f_{\theta'}(x) \leq m_2 f_\theta(x)^{1-\tau_n},$$

as in hypothesis **H5**, then $d_n \geq \tau_n^{1/\beta}/\tau_1 \geq \rho'\delta^{1/\beta}n^{-h/\beta-1}$, where ρ' is some constant. The number of such upper bounds, for fixed a, b, is then bounded by $Cd_n^{-d} = O(n^T)$, for some $T > 0$.

Let $a, b < 1$. We then use the same calculations as above to obtain

$$f_{\theta'}(x)g_{a,b}(F_{\theta'}(x)) \leq H_1(m_2')^{b-1}(m_1')^{a-2}h_\theta(x)H_\theta(x)^{(1+2\tau')(a-1)}(1-H_\theta(x))^{(1+2\tau')(b-1)} = \bar{h}(x),$$

and

$$\int_{\mathcal{X}} \bar{h}(x)dx \quad \leq \quad H_1(m_2')^{b-1}(m_1')^{a-1}\frac{B(a',b')}{B(a,b)} \leq 1+\delta/3$$

To obtain the above inequality we therefore need the same conditions as previously, i.e. (i)–(iv), apart from (iv) which is now expressed as

$$2\tau(1-a)/(1-\tau) \leq \delta\rho a \quad \text{and} \quad 2\tau(1-b)/(1-\tau) \leq \delta\rho b$$

as in the proof of Theorem 2. This condition is again the most constrictive, when replacing a and b by $t_n = n^{-\alpha}$. The above inequality will therefore be satisfied when $\tau \leq \rho'\delta n^{-\alpha}$, for some $\rho' > 0$, when n is large enough.

Finally the logarithm of the total number of upper bounds for the densities $f_\theta g_\psi(F_\theta)$, with $\theta \in \Theta$ and $\psi \in \mathcal{F}_n$ is bounded by

$$\log \mathcal{N}_n + C\log n,$$

where $\mathcal{N}_n$ is the number of upper bounds defined in Section 2.2, in the case of mixtures of Beta densities, and C is a positive constant. It is thus bounded by cn, for n large enough.

D. Proof of Theorem 4

Recall that $H_x = E^\pi\left[d(f,\mathcal{F})|X^n\right]$ the observed value of the test statistic and

$$p_{cpred} = Pr^{m_0(.|\hat{\theta}_x)}\left[H(Y^n) > H_x\right], \tag{22}$$

where $m_0(y^n|\hat{\theta}_x)$ is given by (3). We now prove that under the null hypothesis, i.e. if there exists $\theta_0 \in \Theta$ such that $f_0 = f_{\theta_0}$, then p_n is equivalent to the conditional p-value :

$$P_{\theta_0}[H(Y^n) > H_x|\hat{\theta} = \hat{\theta}_x] = p_0(\hat{\theta^x}),$$

and if $f_0 \notin \mathcal{F}$, and if $\theta^\perp = \arg\min_\theta \mathcal{I}(f_0, f_\theta)$ is unique, then p_n is equivalent to

$$P_{\theta^\perp}[H(Y^n) > H_x|\hat{\theta} = \hat{\theta}_x] = P_{\theta^\perp}(\hat{\theta^x}).$$

To do so, we first prove that when θ is closed to $\hat{\theta}_x$, $f(y^n|\hat{\theta}_x;\theta)$ is asymptotically independent of θ. Denote $S_n^{\hat{\theta}_x} = \{y^n; \hat{\theta}(y^n) = \hat{\theta}_x\}$ and $d\lambda_n$ the Lebesgue measure on $S_n^{\hat{\theta}_x}$. In the following M denotes a generic constant. Let $y^n \in S_n^{\hat{\theta}_x}$, then using Fraser and Reid's (1995) result we have that

$$f(y^n|\hat{\theta}_x;\theta) \quad = \quad \frac{f(y^n|\theta)|\hat{j}||l_{\theta;y}(\hat{\theta}_x,y)|^{-1}}{g(\hat{\theta}_x|\theta)}$$

where $\hat{j} = -D^2 l_n(\hat{\theta})/n$ is the observed Fisher information matrix and

$$|l_{\theta;y}(\hat{\theta}_x,y)| = |(\partial^2 l_n(\hat{\theta}_x)/(\partial\theta\partial y))(\partial^2 l_n(\hat{\theta}_x)/(\partial\theta\partial y))^t|^{1/2}$$

is therefore the determinant of a $k \times k$ matrix. Then

$$\begin{aligned}
\frac{f(y^n|\hat{\theta}_x;\theta)}{f(y^n|\hat{\theta}_x;\theta)} &= \frac{e^{l_n(\theta)-l_n(\hat{\theta})}}{e^{l_n(\theta_0)-l_n(\hat{\theta})}} \frac{g(\hat{\theta_x}|\theta_0)}{g(\hat{\theta_x}|\theta)} \\
&= \frac{1+R_n(\theta)/\sqrt{n}}{1+R_n(\theta)/\sqrt{n}} \frac{\int_{S_n^{\hat{\theta}_x}} f(y^n|\hat{\theta}_x)|\hat{j}||l_{\theta;y}(\hat{\theta}_x,y)|^{-1}(1+R_n(\theta_0)/\sqrt{n})d\lambda_n(y^n)}{\int_{S_n^{\hat{\theta}_x}} f(y^n|\hat{\theta}_x)|\hat{j}||l_{\theta;y}(\hat{\theta}_x,y)|^{-1}(1+R_n(\theta)/\sqrt{n})d\lambda_n(y^n)},
\end{aligned}$$

where

$$1+R_n(\theta) = \exp\{-n(\hat{\theta}_x-\theta)^t Z_{n,2}(\hat{\theta}_x)(\hat{\theta}_x-\theta)/2\sqrt{n}\} \exp\{n^{3/2}(\hat{\theta}_x-\theta)^{(3)}\mu_n^{(3)}(\bar{\theta})/\sqrt{n}\},$$

$$Z_{n,2}(\theta) = \sqrt{n}\left(-\frac{D^2 l_n(\theta)}{n} - I(\theta)\right),$$

$I(\theta) = E_\theta\left[-D^2 \log f(X|\theta)\right]$, the differential being wrt θ, $\bar{\theta} \in (\theta, \hat{\theta})$, and

$$\mu_n^{(3)}(\theta) = \frac{D^3 l_n(\theta)}{n}.$$

The notation $(\hat{\theta}_x - \theta)^{(3)}\mu_n^{(3)}$ means the sum over the components of these two terms.

Let $|\theta - \hat{\theta}_x| < K \log n/\sqrt{n} = \delta_n$, with K large enough, and denote A_n the set of y^n such that

$$\sup_{|\theta'-\hat{\theta}|<\delta_n} \left|\frac{D^3 l_n(\theta')}{n}\right| \leq M, \quad |Z_{n,2}(\hat{\theta}_x)| \leq \sqrt{n}/(4K^2 \log n^2).$$

Then, on A_n,

$$|R_n(\theta)| \leq 2|n(\hat{\theta}_x-\theta)^t Z_{n,2}(\hat{\theta}_x)(\hat{\theta}_x-\theta)/2| + M|n^{3/2}(\hat{\theta}_x-\theta)^{(3)}\mu_n^{(3)}(\bar{\theta})|.$$

We have

$$\begin{aligned}
p_{cpred} &= \int_{S_n^{\hat{\theta}_x}} \mathbb{I}_{H(y^n)>H_x} \frac{\int_\Theta f(y^n|\hat{\theta}_x;\theta)g(\hat{\theta}_x|\theta)\pi(\theta)d\theta}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta} d\lambda_n(y^n) \\
&= \int_{S_n^{\hat{\theta}_x}\cap A_n^c} \mathbb{I}_{H(y^n)>H_x} \frac{\int_{|\hat{\theta}_x-\theta|<\delta_n} f(y^n|\hat{\theta}_x;\theta)g(\hat{\theta}_x|\theta)\pi(\theta)d\theta}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta} d\lambda_n(y^n) \\
&\quad + \int_{S_n^{\hat{\theta}_x}\cap A_n} \mathbb{I}_{H(y^n)>H_x} \frac{\int_{|\hat{\theta}_x-\theta|<\delta_n} f(y^n|\hat{\theta}_x;\theta)g(\hat{\theta}_x|\theta)\pi(\theta)d\theta}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta} d\lambda_n(y^n) \qquad (23) \\
&\quad + \int_{S_n^{\hat{\theta}_x}} \mathbb{I}_{H(y^n)>H_x} \frac{\int_{|\hat{\theta}_x-\theta|>\delta_n} f(y^n|\hat{\theta}_x;\theta)g(\hat{\theta}_x|\theta)\pi(\theta)d\theta}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta} d\lambda_n(y^n).
\end{aligned}$$

We first consider the first term of the right hand side of (23) and we denote $p_{n,1}$ this term. Then

$$\begin{aligned}
p_{n,1} &\leq \frac{\int_{|\hat{\theta}_x-\theta|<\delta_n} \pi(\theta) \int_{S_n^{\hat{\theta}_x}\cap A_n^c} f(y^n|\hat{\theta}_x;\theta)d\lambda_n(y^n)g(\hat{\theta}_x|\theta)d\theta}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta} \\
&= \frac{\int_{|\hat{\theta}_x-\theta|<\delta_n} \pi(\theta)P_\theta^n[A_n^c|\hat{\theta}_x]g(\hat{\theta}_x|\theta)d\theta}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta}.
\end{aligned}$$

Using the hypotheses (12) and (13) we obtain that uniformly in θ,

$$P_\theta^n\left[A_n^c\right] \le \frac{M\log n^4}{n^2}.$$

Since

$$P_\theta^n\left[A_n^c\right] \quad \ge \quad \int_{|\hat\theta-\theta|<\delta_n} P_\theta^n\left[A_n^c|\hat\theta\right] g(\hat\theta|\theta)d\hat\theta,$$

we have

$$\int_{|\hat\theta-\theta|<\delta_n} P_\theta^n\left[A_n^c|\hat\theta\right] g(\hat\theta|\theta)d\hat\theta \le \frac{M\log n^4}{n^2}.$$

1. We first consider the case $f_0 = f_{\theta_0}$.Then

$$\int_\Theta g(\hat\theta_x|\theta)\pi(\theta)d\theta > c_0 \tag{24}$$

for some c_0 small enough and for n large enough. Indeed, the approximation of the density $g(\hat\theta|\theta)$ can be uniformly approximated by

$$g(\hat\theta_x|\theta_0) = n^{k/2}|I(\theta_0)|^{1/2}\varphi(I^{1/2}\sqrt{n}(\hat\theta_x-\theta_0))\left[1+\frac{P(\sqrt{n}(\hat\theta_x-\theta_0))}{\sqrt{n}}\right] + O(n^{-1/2}), \tag{25}$$

where $\varphi(u)$ is the density of a standard Gaussian random variable and $P(u)$ is a polynomial function with degree 3, see Bhattacharya and Ghosh (1978). We then obtain that

$$\begin{aligned}
\int_\Theta g(\hat\theta|\theta)\pi(\theta)d\theta \quad &\ge \quad \pi(\hat\theta)\int_{|u|<\sqrt{n}\delta_n}\varphi_{I^{-1/2}}(u)du - M\log n/\sqrt{n} \\
&\ge \quad \pi(\hat\theta) - C\log n/\sqrt{n} \\
&\ge \quad c_0
\end{aligned}$$

for $c_0 < \inf_\Theta \pi(\theta)$, when n is large enough. Hence,

$$\begin{aligned}
P_{\theta_0}^n\left[p_{n,1} > M/\sqrt{n}\right] \quad &\le \quad P_{\theta_0}^n\left[\int_{|\hat\theta_x-\theta|<\delta_n}\pi(\theta)P_\theta^n[A_n^c|\hat\theta_x]g(\hat\theta_x|\theta)d\theta > Mc_0/\sqrt{n}\right] \\
&\le \quad \frac{M\sqrt{n}}{c_0}\int_{|\hat\theta_x-\theta_0|<\delta_n}\int_{|\hat\theta_x-\theta|<\delta_n} g(\hat\theta_x|\theta_0)g(\hat\theta_x|\theta)P_\theta^n[A_n^c|\hat\theta_x]\pi(\theta)d\theta d\hat\theta_x + n^{-1}
\end{aligned}$$

Therefore,

$$\begin{aligned}
P_{\theta_0}^n\left[p_{n,1} > M/\sqrt{n}\right] \quad &\le \quad M\sqrt{n}n^{k/2}\int_{|\hat\theta_x-\theta_0|<\delta_n}\int_{|\hat\theta_x-\theta|<\delta_n} g(\hat\theta_x|\theta)P_\theta^n[A_n^c|\hat\theta_x]\pi(\theta)d\theta d\hat\theta_x + Mn^{-1/2} \\
&\le \quad M\sqrt{n}n^{k/2}\int_{|\hat\theta_0-\theta|<2\delta_n} P_\theta[A_n^c]\pi(\theta)d\theta + Mn^{-1/2} \\
&\le \quad \frac{M\log n^4 n^{k/2}}{\sqrt{n}}\pi[|\hat\theta_0-\theta| < 2\delta_n] \\
&\le \quad M\frac{\log n^{4+k/2}}{\sqrt{n}}.
\end{aligned}$$

We now consider the third term of (23), namely $p_{n,3}$.

$$p_{n,3} = \int_{S_n^{\hat{\theta}_x}} \mathbb{I}_{H(y^n)>H_x} \frac{\int_{|\hat{\theta}_x-\theta|>\delta_n} f(y^n|\hat{\theta}_x;\theta)g(\hat{\theta}_x|\theta)\pi(\theta)d\theta}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta}.$$

Using (24), we have :

$$\begin{aligned} p_{n,3} &\leq c_0^{-1}\int_{S_n^{\hat{\theta}_x}} \mathbb{I}_{H(y^n)>H_x} \int_{|\hat{\theta}_x-\theta|>\delta_n} f(y^n|\hat{\theta}_x;\theta)g(\hat{\theta}_x|\theta)\pi(\theta)d\theta d\lambda_n(y^n) \\ &\leq c_0^{-1}\int_{|\hat{\theta}_x-\theta|>\delta_n} g(\hat{\theta}_x|\theta)\pi(\theta)d\theta. \end{aligned}$$

Moreover, by choosing K large enough in the definition of δ_n and using (25) ,

$$p_{n,3} \quad \leq \quad Mn^{-1/2}. \tag{26}$$

We now consider the second term of (23), namely $p_{n,2}$. We prove that $p_{n,2}$ is equal to the conditional p-value, to the order $n^{-1/2}$.

$$p_{n,2} \quad = \quad \int_{S_n^{\hat{\theta}_x}\cap A_n} \mathbb{I}_{H(y^n)>H_x} \frac{\int_{|\hat{\theta}_x-\theta|<\delta_n} f(y^n|\hat{\theta}_x;\theta)g(\hat{\theta}_x|\theta)\pi(\theta)d\theta}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta},$$

we have proved that on A_n,

$$|R_n(\theta)| \leq 2|n(\hat{\theta}_x-\theta)^t Z_{n,2}(\hat{\theta}_x)(\hat{\theta}_x-\theta)/2| + M|n^{3/2}(\hat{\theta}_x-\theta)^{(3)}\mu_n^{(3)}(\bar{\theta})| \leq \sqrt{n}/2$$

whenever $|\theta-\hat{\theta}_x|<\delta_n$. Therefore,

$$\left|\frac{f(y^n|\hat{\theta}_x;\theta)}{f(y^n|\hat{\theta}_x;\theta_0)} - 1\right| \leq M\frac{|R_n(\theta)|+|R_n(\theta_0)|}{\sqrt{n}}$$

and

$$\begin{aligned} \left|p_{n,2}-p_0(\hat{\theta_x})\right| &\leq M\int_{S_n^{\hat{\theta}_x}\cap A_n} \mathbb{I}_{H(y^n)>H_x} f(y^n|\hat{\theta}_x;\theta_0)\int_{|\hat{\theta}_x-\theta|<\delta_n} \frac{|R_n(\theta)|+|R_n(\theta_0)|}{\sqrt{n}} g(\hat{\theta}_x|\theta)\pi(\theta)d\theta d\lambda_n \\ &+ \int_{S_n^{\hat{\theta}_x}\cap A_n^c} f(y^n|\hat{\theta}_x;\theta_0)\mathbb{I}_{H(y^n)>H_x} \frac{\int_{|\hat{\theta}_x-\theta|<\delta_n} g(\hat{\theta}_x|\theta)\pi(\theta)d\theta d\lambda_n}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta} \\ &+ \int_{S_n^{\hat{\theta}_x}} f(y^n|\hat{\theta}_x;\theta_0)\mathbb{I}_{H(y^n)>H_x} \frac{\int_{|\hat{\theta}_x-\theta|>\delta_n} g(\hat{\theta}_x|\theta)\pi(\theta)d\theta d\lambda_n}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta} \\ &\leq \frac{M}{\sqrt{n}} E_{\theta_0}\left[|Z_n(\hat{\theta^x})||\hat{\theta}_x\right]\left(1+\log n^3\right) + P_{\theta_0}\left[A_n^c|\hat{\theta}_x\right] + \frac{\int_{|\hat{\theta}_x-\theta|>\delta_n} g(\hat{\theta}_x|\theta)\pi(\theta)d\theta d\lambda_n}{\int_\Theta g(\hat{\theta}_x|\theta)\pi(\theta)d\theta} \\ &\leq Mn^{-1/2}, \end{aligned}$$

except on a set of probability $(P_{\theta_0}^n)$ less than $n^{-1/2}$, using the above calculations.

2. We now consider the case $f_0 \notin \mathcal{F}$.

To begin with, using condition (14), (25) and (24) remain unchanged.

We need to consider the asymptotic distributions of the mle $\hat{\theta}_x$ under f_0 ; we denote $g_0(\hat{\theta}_x)$ its density. We have, as usual

$$\sqrt{n}(\hat{\theta}_x - \theta^{\perp}) = I_0(\theta^{\perp})^{-1} Z_n(\theta^{\perp}) + R_n/\sqrt{n},$$

under assumptions (12), (13) and (14), which implies that

$$|g_0(\hat{\theta}) - n^{k/2} \varphi_{I_0^{-1}}(\sqrt{n}(\hat{\theta}_x - \theta^{\perp})| \leq M n^{-1/2}. \tag{27}$$

In other words, $g_0(\hat{\theta})$ behaves similarly to $g(\hat{\theta}|\theta)$ and we can then apply exactly the same argument as in the upper bound of $P_{\theta_0}^n\,[p_{n,1} > M/\sqrt{n}]$ to obtain

$$P_0^n\left[p_{n,1} > M/\sqrt{n}\right] \leq \frac{M \log n^{4+k/2}}{\sqrt{n}}. \tag{28}$$

We also have that

$$p_{n,3} \quad \leq \quad c_0^{-1} \int_{|\hat{\theta}_x - \theta| > \delta_n} g(\hat{\theta}_x|\theta)\pi(\theta) d\theta \leq M n^{-1/2}. \tag{29}$$

Moreover, we have the same decomposition for $p_{n,2}$ leading to

$$\begin{aligned} \left|p_{n,2} - p_0(\hat{\theta_x})\right| \quad &\leq \quad \frac{M}{\sqrt{n}} E_{\theta^{\perp}}\left[\left|Z_n(\hat{\theta^x})\right| \hat{\theta}_x\right]\left(1 + \log n^3\right) + P_{\theta^{\perp}}\left[A_n^c|\hat{\theta}_x\right] \\ &\qquad + \frac{\int_{|\hat{\theta}_x - \theta| > \delta_n} g(\hat{\theta}_x|\theta)\pi(\theta) d\theta d\lambda_n}{\int_{\Theta} g(\hat{\theta}_x|\theta)\pi(\theta) d\theta}. \end{aligned}$$

Replacing $g_0(\hat{\theta})$ and $g(\hat{\theta}|\theta^{\perp})$ by their Edgeworth expansion to the order $n^{-3/2}$ in our calculations, we obtain that $\tilde{g}_0(u) \leq \tilde{K}\tilde{g}(u|\theta^{\perp})^c$, where $u = \sqrt{n}(\hat{\theta} - \theta^{\perp})$, $c \in (0, 1]$ and where $\tilde{g}$ denotes the corresponding Edgeworth expansion. Hence, if

$$B_n = \{P_{\theta^{\perp}}\left[A_n^c|\hat{\theta}_x\right] \geq K \log n)^p/\sqrt{n}\},$$

up to a term of order $n^{-3/2}$,

$$P_0^n[B_n] \quad \leq \quad \tilde{K} \int_{|u| < \delta_n} \tilde{g}^c(u|\theta^{\perp}) \mathbb{I}_{B_n}(u) du$$

for some $1 \geq c > 0$ and with $\delta_n = K \log n$. Using Jensen's inequality we obtain that

$$\begin{aligned} P_0^n[B_n] \quad &\leq \quad \tilde{K}\left(\int_{|u| \leq \delta_n} \tilde{g}(u|\theta^{\perp})(u) \mathbb{I}_{B_n}(u) du\right)^{2-c} \left(\int_{|u| \leq \delta_n} \tilde{g}^2(u|\theta^{\perp})(u) \mathbb{I}_{B_n}(u) du\right)^{-(1-c)} \\ &\leq \quad \tilde{K}\left(\int_{|u| \leq \delta_n} \tilde{g}(u|\theta^{\perp})(u) \mathbb{I}_{B_n}(u) du\right)^{2-c} \left(\int_{|u| \leq \delta_n} \tilde{g}(u|\theta^{\perp}) \mathbb{I}_{B_n}(u) du\right)^{-(1-c)} \\ &\qquad \times \left(\int_{|u| \leq \delta_n} \mathbb{I}_{B_n}(u) du\right)^{1-c} \\ &\leq \quad K\left(P_{\theta^{\perp}}[B_n] + M n^{-3/2}\right)^c \log n^{1-c} \leq K' n^{-3c/2} \log n^{1-c(1+p)}, \end{aligned}$$

and the theorem is proved.