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,

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

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 $E^\pi[d(f,\mathcal{F})|X^n]$, 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, MCMC algorithms, nonparametric estimation, 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.

†Work partially supported by EU TMR network ERB–FMRX–CT96–0095 on *'Computational and Statistical Methods for the Analysis of Spatial Data'*.

In regular testing problems, the 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. 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. In this setup, an acceptable resolution of the nonparametric problem is to use mixtures of Beta distributions, of the form

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

whose shapes are variate enough to allow for an approximation of an arbitrary distribution on $[0,1]$, at least in the sense of the Hellinger distance

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

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 we think that by allowing the parameters to be free, we do need less components to approximate a given density on $[0,1]$. Moreover, 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) can approximate unsmoothed densities as well as densities that diverge at 0 or 1.

The resolution being nonparametric, there is no hope to determine a subjective prior on the whole set of parameters. It is thus 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 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 is, the more relevant the parametric model might get.

The quantity of interest is then the distance between the true density and the proposed model, $d(f,\mathcal{F})$. We approximate this quantity using $E^\pi[d(f,\mathcal{F})|X^n]$. To test the parametric model, we must therefore compare the above posterior expectation with a quantity that would characterize its behaviour under the null hypothesis. To do so, we use the distribution of $E^\pi[d(f,\mathcal{F})|Y^n]$, when Y^n is distributed according to

$$m_0(y^n|X^n) = \int_\Theta f_\theta(Y^n)\, d\pi_0(\theta|X^n)\,,$$

which is its *predictive* distribution under the null hypothesis. The test consists in evaluating

$$P\left[E^\pi[d(f,\mathcal{F})|y^n] \geq E^\pi[d(f,\mathcal{F})|X^n]|\, X^n\right],$$

where the probability is calculated under $m_0(y^n|X^n)$.

We prove, in section 5 that such a test is consistent, in the sense that the above probability goes to zero as n goes to infinity under the alternative and that the distribution of $E^\pi[d(f,\mathcal{F})|y^n]$ is equivalent to the true distribution of $E^\pi[d(f,\mathcal{F})|X^n]$ under the null hypothesis, see Theorem 4. This test procedure is therefore also satisfying from a frequentist point of view, since it is equivalent to using a p-value.

The paper is organised as follows: in Section 2, we define the prior distribution associated with the specific mixture of Beta distributions (1); in Section 3, we show that the posterior distribution of the parameters of a mixture of Beta distributions is consistent; in Section 4, we explain the estimation procedure associated with this posterior distribution; in Section 5, we detail the test of goodness of fit; Section 6 concludes with a illustration of the Hellinger distances for some simulated samples.

2. Mixtures of Beta distributions

2.1. Representation

Given that any distribution on $[0,1]$ can be approximated as a infinite mixture of Beta distributions,

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

we define the general alternative to $X \sim \mathcal{U}([0,1])$ to be

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

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:

$$p_0\mathcal{U}(0,1) + (1-p_0)\sum_{k=1}^{K} p_k \mathcal{B}(\alpha_k\epsilon_k, \alpha_k(1-\epsilon_k)) \qquad \sum_{k\geq 1} p_k = 1\,, \tag{2}$$

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 (2) and $\mathcal{U}(0,1)$, bypassing parameterisatio problems, and by (b) selecting an appropriate prior distribution.

For simulation reasons discussed in Cappé *et al.* (2001), we also choose to replace the weights p_k with their unscaled version, ω_k, namely $(k = 1, \dots, K)$

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

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 $0 < \epsilon_k < 1$.

2.2. Testing priors for Beta mixtures

Although a regular conjugate prior could be used in this setting just as in Diebolt and Robert (1990) or Richardson and Green (1997), we now build a specific prior distribution in order to oppose the uniform component of the mixture (2) with the other components. So we choose a uniform $\{1, \dots, K_{\max}\}$ 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

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

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

$$\begin{aligned}(\alpha_k, \epsilon_k) \quad \sim \quad & \{1 - \exp\left[-\{\beta_1(\alpha_k - 2)^{c_3} + \beta_2(\epsilon_k - .5)^{c_4}\}\right]\} \\ & \exp\left[-\tau_0 \alpha_k^{c_0}/2 - \tau_1/\{\alpha_k^{2c_1}\epsilon_k^{c_1}(1-\epsilon_k)^{c_1}\}\right], \end{aligned} \tag{3}$$

on the (α_k, ϵ_k)'s, where $c_0, \dots, c_4, \tau_0, \tau_1, \beta_1, \beta_2$ are hyperparameters. This choice is purposedly 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 in the choice of that specific prior on the (α_k, ϵ_k)'s, 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

$$(\alpha_k, \epsilon_k) \sim \left\{1 - \exp\left[-\xi\left\{(\alpha_k - 2)^2 + (\epsilon_k - .5)^2\right\}\right]\right\} \exp\left[-\zeta/\{\alpha_k^2 \epsilon_k(1-\epsilon_k)\} - \kappa\alpha_k^2/2\right] \tag{4}$$

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.

As in many Bayesian nonparametric analyses, we now prove the consistency of the posterior, which validates the choice of the prior. These consistency results are also used to prove the consistency of the test procedure, in Section 3.2.

3. Convergence of the posterior distributions for Beta mixtures

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

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

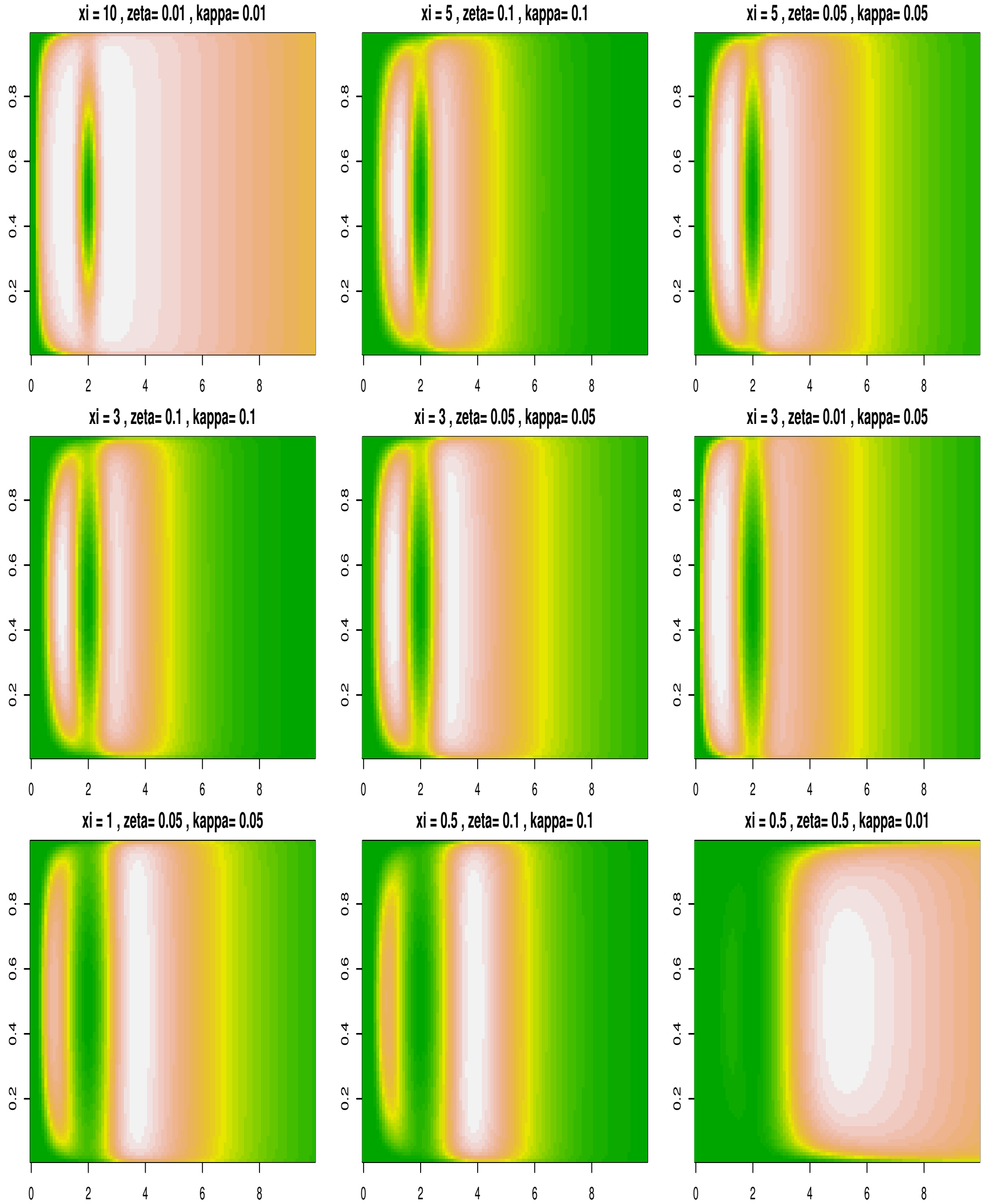

Fig. 1. R's `filled.contour` representation of the prior distribution (3) for various values of (ξ, ζ, κ)

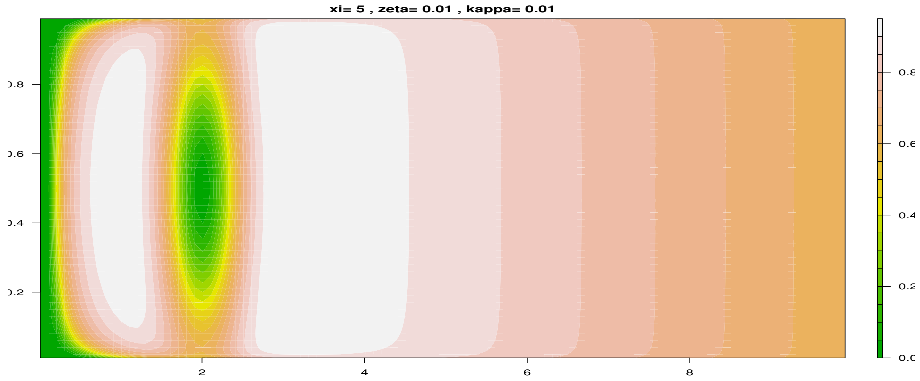

Fig. 2. R's `filled.contour` representation of the prior distribution (3) for $(\xi, \zeta, \kappa) = (5, 0.01, 0.01)$

where F_θ is a cdf indexed by a parameter $\theta \in \Theta$, for a given sample $x_1, , \ldots, x_n$, as a test of *uniformity* for the transforms $u_1 = F_\theta(x_1), \ldots, u_n = F_\theta(x_n)$, for a certain value of θ. 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 *[but not restricted to integer valued parameters as in Petrone and Wasserman (2002)]* .

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

$$\mathcal{H} = \{f = f_\theta(x) g_\psi(F_\theta(x)), \theta \in \Theta, \psi \in S\}, \tag{5}$$

where g_ψ is the density of a mixture of Beta distributions,

$$g_\psi(u) = p_0 + (1 - p_0) \sum_{j=1}^{K} p_j g_j(u), \tag{6}$$

with g_j the density of the Beta distribution $\mathcal{B}e(a_j, b_j)$, $K \in \mathbb{N}$, $p_0 \in [0, 1]$, $p_j = \omega_j / \sum_{l=1}^{K} \omega_l$, $\omega_l > 0$. The parameters for this general model are then

$$\psi = (K, p_0, \{\omega_j, a_j, b_j, j = 1, \ldots K\}) \quad \text{and} \quad \theta \in \Theta.$$

In a more general framework than Section 2.2, we consider priors of the form $\pi(\psi, \theta) = \pi_1(\psi)\pi_2(\theta|\psi)$, where 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
$$P(K \geq tn/\log n) \leq e^{-rn}, \tag{7}$$
(b) $p_0 \sim \pi(p_0)$ a.c. wrt Lebesgue and with support is $[0, 1]$.
(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 denote $p_K(a_1, b_1, ..., a_K, b_K)$ the prior density on the parameters of the Beta densities. We impose the following form on p_K:
$$p_K(a_1, b_1, ..., a_K, b_K) = \prod_{j=1}^{K} p(a_j, b_j)$$

where $a_j = \alpha_j \epsilon_j$, $b = \alpha_j(1-\epsilon_j)$, with $\alpha_j > 0$ and $\epsilon_j \in (0,1)$, and (α_j, ϵ_j) is distributed from (3).

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 (7) is satisfied in particular by the Poisson distribution.

We first consider the consistency of the plain model, i.e. $\{g_\psi, \psi \in S\}$, without the additional level of estimating the parameter θ.

3.1. Mixtures of Beta distributions

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

Let $A_\varepsilon(g_0) = \{g : d(g_0, g) \le \varepsilon\}$ and $N_\varepsilon = \{g : \mathcal{I}(g_0, g) \le \varepsilon\}$, 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 that have at most a countable set of discontinuities and that satisfy s

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

It is in fact well-known (Petrone and Wasserman, 2002) that any continuous density on $[0,1]$ can be approximated by Bernstein polynomials, which constitute a subset of Ω. Obviously, a general mixture of Beta distributions can, at least, also approximate a density that explodes around 0 at a polynomial rate x^{-d}, $d < 1$ (or around 1 at a rate $(1-x)^{-d}$).

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

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

The idea of the proof, is the following. Since $g \in \Omega$, we can approximate

$$\int_0^1 g(x)\log g(x) dx \quad \text{by} \quad \int_0^1 \tilde{g}(x)\log\tilde{g}(x)dx\,,$$

where $\tilde{g}$ is piecewise constant, and since g has a countable set of discontinuities, the pieces where $\tilde{g}$ is constant can be considered as intervals in $(0,1)$. We then approach $\tilde{g}$ in terms of Kulback-Leibler divergence, using a mixture of Beta distributions. By considering a mixture, we can work on each sub-interval separately, and this simplifies the calculations.

Proof. Let $g \in \Omega$. There exists a piecewise constant function $\tilde{g}$ such that $\int |g\log g + \tilde{g}\log\tilde{g}|dx < \epsilon$. We can moreover impose wlog that $\tilde{g}(x) = \sum_{i=1}^N \mathbb{I}_{\Delta_i} w_i$, with $\sum_{i=1}^N w_i|\Delta_i| = 1$ and $\Delta_i = (\tau_i - \delta_i, \tau_i + \delta_i)$.

Now consider $g_\psi(x) = \sum_{i=1}^N w_i(2\delta_i)h_i(x)$, where $h_i(x)$ is a Beta density. We have

$$\begin{aligned}
&\sum_{i=1}^N \omega_i \left[(2\delta_i)\log\omega_i - \int_{\Delta_i} \log\left\{\sum_{l=1}^N w_l(2\delta_l)h_l(x)\right\} dx\right] \\
&\quad \le \sum_{i=1}^N \omega_i \left[(2\delta_i)\log\omega_i - \int_{\Delta_i} \log\left[w_i(2\delta_i)h_i(x)\right]dx\right] \\
&\quad = -\sum_{i=1}^N \omega_i \int_{\Delta_i} \log\left[(2\delta_i)h_i(x)\right]dx.
\end{aligned}$$

Since we can work on each Δ_i separately, we drop the subscript i for simplicity's sake. We prove that, by choosing carefully the α's and β's (the parameters of the Beta densities), we obtain

$$\int_{\Delta_i} \log\left[|\Delta_i| h_i(x)\right]dx = |\Delta_i| \times o(1). \tag{8}$$

For each $\varepsilon \in (0,1)$ (fixed), the normalising coefficient of the Beta distribution satisfies

$$\begin{aligned}
B(\alpha\varepsilon, \alpha(1-\varepsilon)) &= \frac{\Gamma(\alpha)}{\Gamma(\alpha\varepsilon)\Gamma(\alpha(1-\varepsilon))} \\
&= \frac{\alpha^\alpha\, e^{-\alpha}\sqrt{2\pi/\alpha}}{(\alpha\varepsilon)^{\alpha\varepsilon}\, e^{-\alpha\varepsilon}\sqrt{2\pi/(\alpha\varepsilon)}(\alpha(1-\varepsilon))^{\alpha(1-\varepsilon)}\, e^{-\alpha(1-\varepsilon)}\sqrt{2\pi/(\alpha(1-\varepsilon))}}(1+O((\alpha\varepsilon(1-\varepsilon))^{-1})) \\
&= \frac{[\varepsilon^\varepsilon(1-\varepsilon)^{1-\varepsilon}]^{-\alpha}\sqrt{\alpha\varepsilon(1-\varepsilon)}}{\sqrt{2\pi}}(1+o(1)),
\end{aligned}$$

as $\alpha, \alpha\varepsilon, \alpha(1-\varepsilon) \to \infty$, by Stirling's formula. Moreover, let $\delta/\tau = o(1)$, $\delta/(1-\tau) = o(1)$, and $\varepsilon = \tau$, then

$$\begin{aligned}
&\int_{\tau-\delta}^{\tau+\delta} \log\left[x^{\alpha\tau-1}(1-x)^{\alpha(1-\tau)-1}\right]dx \\
&\quad = (\alpha\tau - 1)\left((\tau+\delta)\log(\tau+\delta) - (\tau-\delta)\log(\tau-\delta) - 2\delta\right) \\
&\qquad +(\alpha(1-\tau)-1)\left((1-\tau+\delta)\log(1-\tau+\delta) - (1-\tau-\delta)\log(1-\tau-\delta) - 2\delta\right) \\
&\quad = (\alpha\tau - 1)\left[2\delta\log\tau + \frac{5\delta^3}{3\tau^2} + O\left(\frac{\delta^4}{\tau^3}\right)\right] \\
&\qquad +(\alpha(1-\tau)-1)\left[2\delta\log(1-\tau) + \frac{5\delta^3}{3(1-\tau)^2} + O\left(\frac{\delta^4}{(1-\tau)^3}\right)\right].
\end{aligned}$$

Therefore, when $\alpha = o(\delta^{-4}/(\tau(1-\tau))^2)$, $\alpha\tau, \alpha(1-\tau) \to \infty$,

$$\int_\Delta \log\left[(2\delta)h(x)\right]dx = 2\delta\left[\frac{5\delta^3\alpha}{3\tau(1-\tau)} - \log(\tau(1-\tau)) + \log\sqrt{\alpha} - \frac{\log 2\pi}{2} + \log(2\delta) + o(1)\right]$$

and (8) is satisfied as soon as

$$\frac{5\delta^3\alpha}{3\tau(1-\tau)} + \log\sqrt{\alpha} = \log(\tau(1-\tau) - \log(2\delta) + \frac{\log(2\pi)}{2}.$$

There exists a solution to this equation, which is essentially of order $O(\delta^{-3}\tau(1-\tau)|\log\delta/(\tau(1-\tau))|)$ and Theorem 1 is proved. □

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:* $\forall \varepsilon > 0$,

$$\pi[A_\varepsilon(g_0)|U_1, \ldots, U_n] \to 1, \quad g_0 \ a.s. \tag{9}$$

The consistency of the posterior mean of the density (which is a standard Bayesian estimate for the density) follows from (9) in terms of the Hellinger distance d.

Proof. The proof of this theorem is obtained using Theorem 1 of Barron, Schervish and Wasserman (1998) *[hereafter BSW]*, which we recall in Appendix A.

To begin with, we prove condition [A1] in Theorem 1 of BSW. Let $g_0 = g_\psi$, where $\psi = (p_0, K, \omega_i, \alpha_i, \epsilon_i, i \le K)$. 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

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

and thus,

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

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 $\mathcal{I}(g_0, g_\psi) \le \varepsilon/2$. Using the above calculations, we deduce that for any such $g_0 \in \Omega$, condition [A1] is satisfied.

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

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

with $T_n = n^l$, with $l \ge 1/c_0$ and $t_n = 2e^{-T_n}$, where $c_0 > 0$ is defined by (3). Then simple calculations imply that $\pi(\mathcal{F}_n^c) \le e^{-nr}$, for some $r > 0$.

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

Denote $\tau = (\tau_1, \tau_2)$, $\bar\eta = a + b$, $\bar\tau = \tau_1 + \tau_2$, $B(\eta)$ be the renormalising constant of the Beta density with parameter $\eta = (a, b)$ and C and ρ be generic positive constants. For all $\eta' = (\eta_1', \eta_2') \in [\eta - \tau, \eta + \tau]$

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

We now determine conditions on τ_1 and τ_2 such that

$$\int g^U(u) du = \frac{B(\eta - \tau)}{B(\eta + \tau)} \le 1 + \delta, \tag{10}$$

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 $\tau_1 \leq \rho\delta a\,(1+\log(\bar\eta+1))^{-1}$ and $\tau_2 \leq \rho\delta\,(1+\log(\bar\eta+1))^{-1}$.

(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 (9) is proved. □

3.2. General goodness of fit model

We now consider the general parametric model

$$\mathcal{H} = \{f(x) = f_\theta(x) g_\psi(F_\theta(x)), \theta \in \Theta, \psi \in S\} \tag{11}$$

where g_ψ is defined as in the previous section. We establish the strong consistency of the posterior distribution, under some regularity conditions on f_θ.

Let $X_1, ..., X_n$ be n iid observations from a distribution with density f_0 against Lebesgue measure. Let $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ be the parametric model, with $\Theta \subset \mathbb{R}^p$ compact. Denote $\pi(\theta)$ 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

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

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

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

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

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

Although the expression of the hypotheses **H4** and **H5** is rather unusual, they are in essence fairly general and are satisfied for most known models, when the parameter space is compact. For instance, if $f_\theta(x) = e^{-\theta x}\theta$, with $\theta \in [\epsilon, E]$, $0 < \epsilon < E$, then if $|\theta' - \theta| \leq d\theta$,

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

Heavy tail distributions can also 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**, $\forall \epsilon > 0$,

$$\pi[A_\epsilon | X^n] \to 1, \quad as \quad n \to \infty, \quad f_0\, a.s. \tag{12}$$

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 ∞.

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

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

therefore $\pi[N_\varepsilon] > 0$ 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), 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

$$\begin{aligned} 1 - F_{\theta'}(x) &= \int_x^\infty f_{\theta'}(x)dx \leq m_2 H_1(1 - H_\theta(x)), \\ F_{\theta'}(x) &\leq m_2 H_1 F_\theta(x), \end{aligned}$$

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 $m_2 H_1 \leq (1 + \delta/3)^{1/(a+b-1)}$. Replacing a, b by $T_n = n^l$, this is satisfied if

$$m_2 H_1 \leq 1 + \delta/(8T_n). \tag{13}$$

Hypothesis **H5** implies that (13) 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,

$$H_\theta(x)^{1+2\tau} \leq \left(\int_{-\infty}^{x} h_\theta^{1+\tau}(y)dy\right)\left(\int_{-\infty}^{x} \sqrt{h_\theta}(y)dy\right)^{2\tau} \tag{14}$$

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 (14),

$$\begin{aligned} F_{\theta'}(x) &\geq \frac{m_1}{\left(\int_{-\infty}^{\infty}\sqrt{h_\theta}(y)dy\right)^{2\tau'}} H_\theta(x)^{1+2\tau'} \\ &= m_1' H_\theta(x)^{1+2\tau'}. \end{aligned} \tag{15}$$

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 3.1, 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

$$\tau \leq \rho'\delta(1+l\log n)^{-1}n^{-h} = \tau_n, \tag{16}$$

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 3.1, in the case of mixtures of betas densities, and C is a positive constant. It is thus bounded by cn, for n large enough. □

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.

4. Estimation of Beta mixtures

4.1. 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

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

with a fixed number of components K. For instance, a Gibbs sampling strategy as in Diebolt and Robert (1990) 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 algorithm 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.* (2001), 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.* (2001) 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.* (2001). 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.*, 2001).

4.2. Performances of the sampler

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 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

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

5. 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 what the statistician wants. In this case the parametric model would be selected if

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

where ϵ is fixed a priori. These situations 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, besides this would be a purely frequentist test, and thus not so entirely satisfying, since it would bypass the prior Bayesian

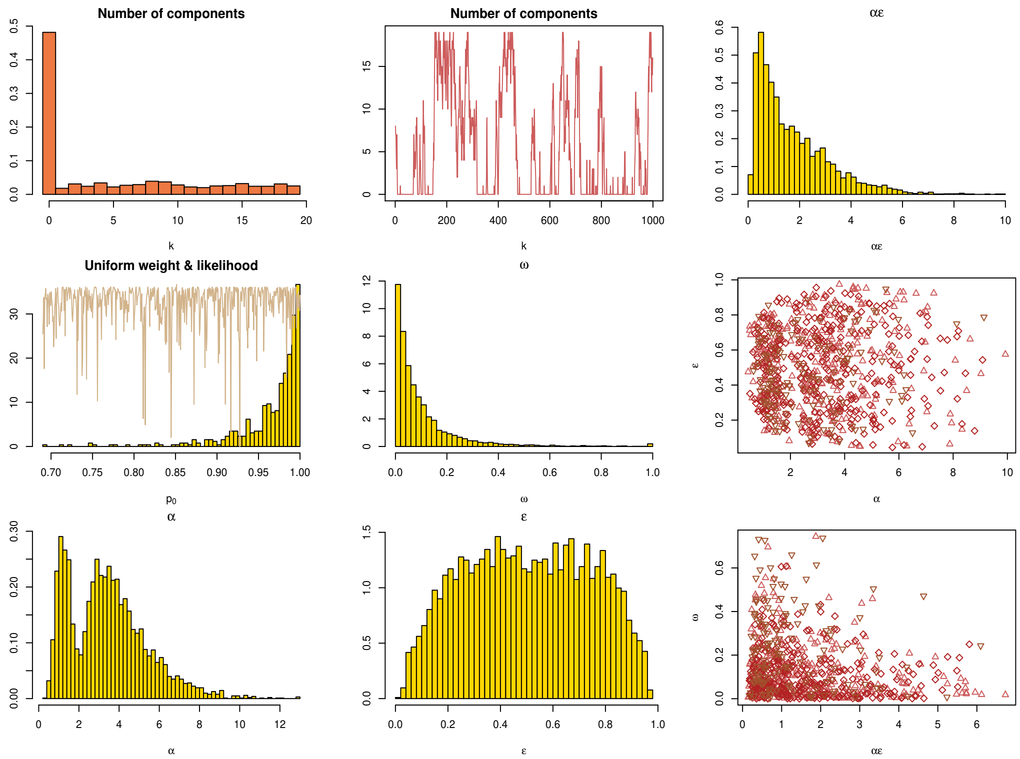

Fig. 3. Monitorings of the convergence of the birth and death sampler for an equidistributed sequence of 1000 points: *(upper left)* histogram of K; *(upper center)* sequence of the simulated K's; *(upper right)* histogram of the $\alpha_k\epsilon_k$'s; *(central left)* histogram of p_0 for $K > 0$ and 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 $K > 0$.

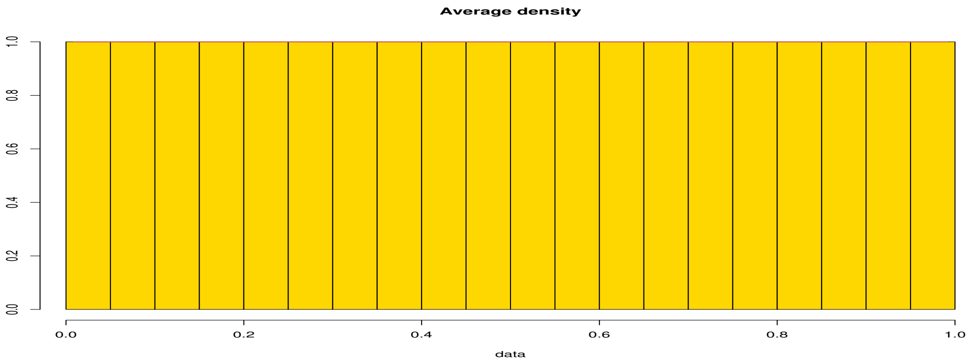

Fig. 4. Histogram of the equidistributed sequence of 1000 points and averaged density estimator.

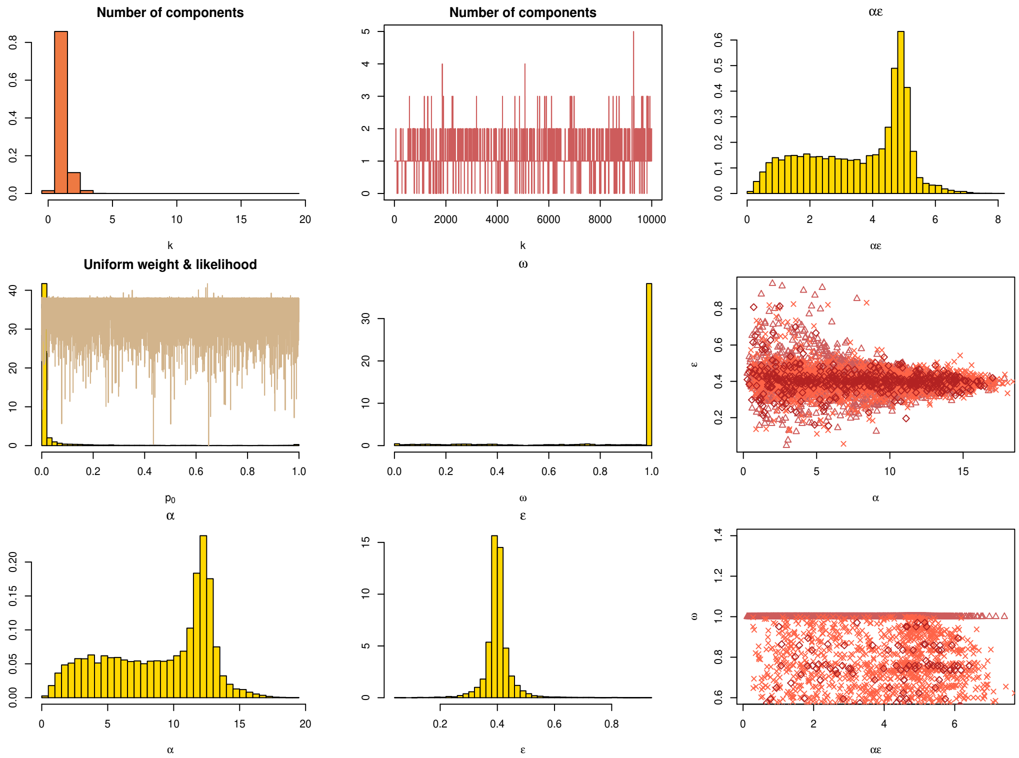

Fig. 5. Monitorings of the convergence of the birth and death sampler for a random sample of 1500 points (same legend as Figure 3).

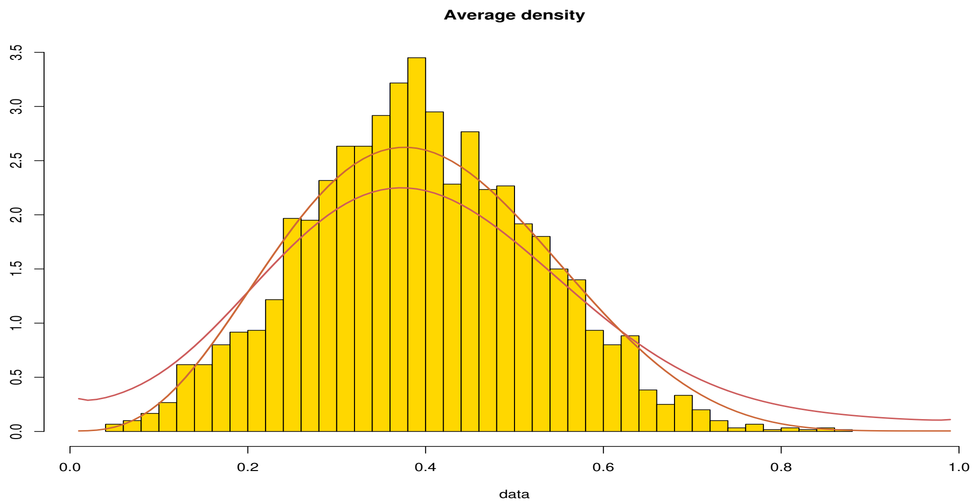

Fig. 6. Histogram of a random sample of 1500 points and averaged density estimators. The curve with the fatter tails corresponds to the average of the densities over the MCMC simulations and the curve with the thinner tails corresponds to the plugg-in estimate of the density where the parameters $(p_k,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

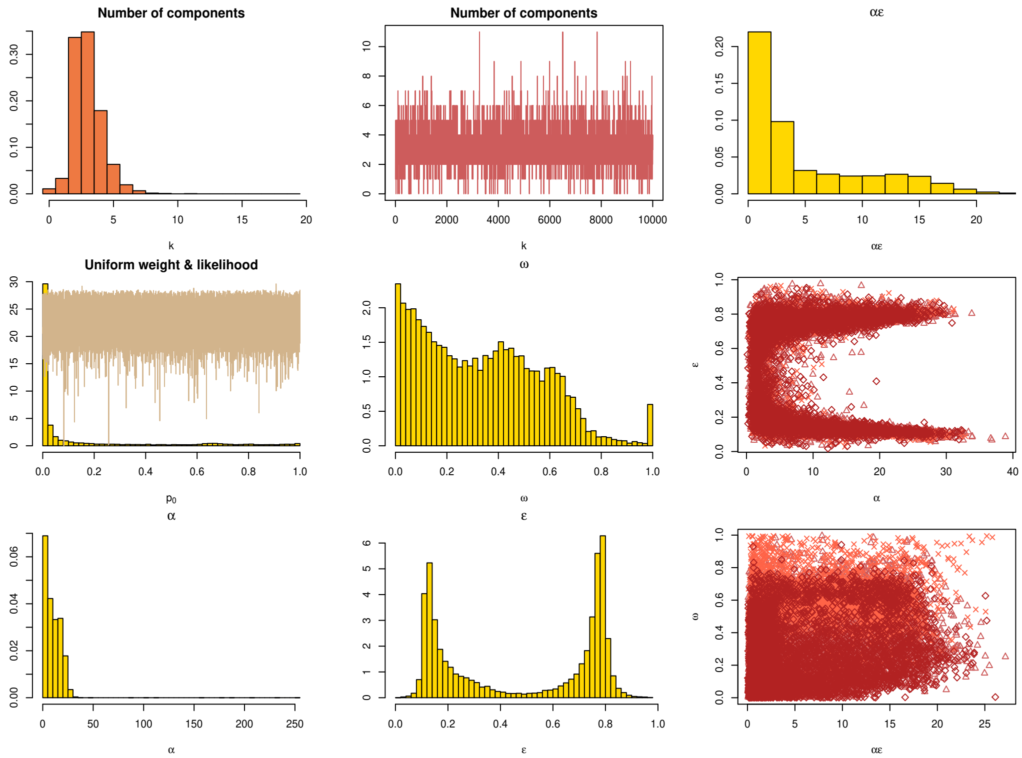

Fig. 7. Monitorings of the convergence of the birth and death sampler for a random sample of 1250 points (same legend as Figure 3).

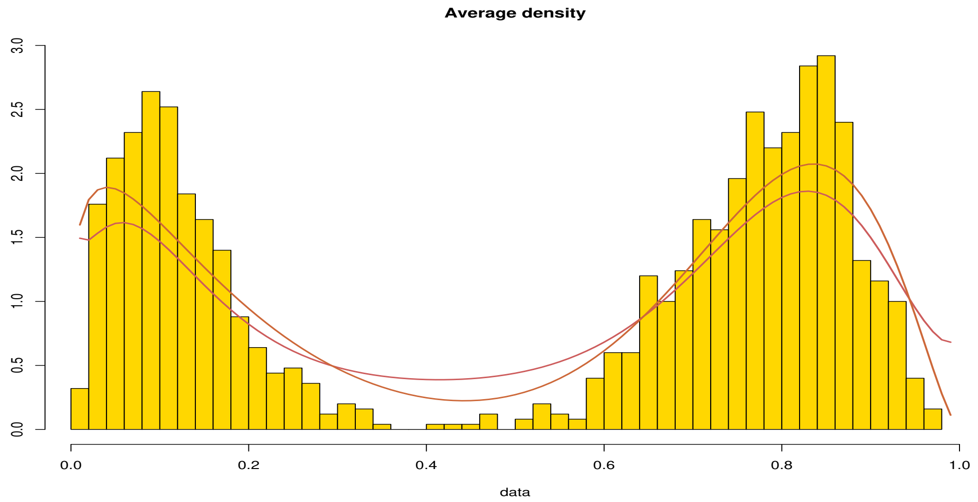

Fig. 8. Histogram of a random sample of 1250 points and averaged density estimators conditional on $K = 2$ (same legend as Figure 6).

analysis of the model. We feel that a better approach consists in computing a reference distribution, characterizing the null hypothesis and being conditional on the observations.

The goodness of fit test is based upon a quantity that is intrinsically difficult to scale. We thus resort to a bootstrap approximation of its distribution. Let

$$\begin{array}{ll} \theta_1 \sim \pi_0(\theta|X^n)\,, & Y_1^n \sim f(Y^n|\theta_1) \\ \theta_2 \sim \pi_0(\theta|X^n)\,, & Y_2^n \sim f(Y^n|\theta_2) \\ & \ldots \quad \ldots \\ \theta_N \sim \pi_0(\theta|X^n)\,, & Y_N^n \sim f(Y^n|\theta_N) \end{array}$$

be iid copies from the posterior×predictive distribution. Then

$$Y_i^n \sim m_0(Y^n|X^n) = \int f(Y^n|\theta)\pi(d\theta|X^n).$$

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.

We assume that f_θ satisfies the usual regularity conditions to obtain a first order Laplace expansion of the posterior density (under the parametric model), see for instance Johnson (1970).

THEOREM 4. *Under H_1, $\exists B \in \mathcal{X}^\infty$ such that $P_0(B) = 1$ and $\forall X \in B$, $\exists N_X$ such that $\forall n \geq N_x$ $H(X^n) \geq d(f_0, \mathcal{F})/2$ and*

$$P[H(Y^n) > H(X^n)|X^n] \leq P[H(Y^n) > d(f_0, \mathcal{F})/2|X^n] \to 0.$$

Under H_0, $P[H(Y^n) > H(X^n)|X^n]$ does not converge to zero. Up to a relative error of order $n^{-1/2}$, we have:

$$2^{-k/2} P_{\theta_0}^n[H(Y^n) > H(X^n)|X^n] \leq P[H(Y^n) > H(X^n)|X^n] \leq P_{\theta_0}^n[H(Y^n) > H(X^n)|X^n].$$

Proof. We proved that $H(X^n) \to d(f_0, \mathcal{F})$, as n goes to infinity, f_0 a.s. We now have the following lemma.

LEMMA 1. *For all X^n, $H(Y^n) \to 0$, as n goes to infinity, $m(.|X^n)$ a.s.*

Proof. Let $\epsilon > 0$, $\forall\theta$

$$d(f_\theta g_\psi(F_\theta), \mathcal{F}) \leq d(g_\psi, 1).$$

Conditionally on θ, Y^n is independent of X^n and is distributed from f_θ.

$$\begin{aligned} H(Y^n) &= \int_\Theta E^\pi[d(f, \mathcal{F})\mathbb{I}_{(\mathcal{F}_n^\theta)^c}|Y^n, \theta]d\pi(\theta|Y^n) \\ &\quad + \int_\Theta E^\pi[d(f, \mathcal{F})\mathbb{I}_{(\mathcal{F}_n^\theta)\cap N_\epsilon}|Y^n, \theta]d\pi(\theta|Y^n) \\ &\quad + \int_\Theta E^\pi[d(f, \mathcal{F})\mathbb{I}_{(\mathcal{F}_n^\theta)\cap N_\epsilon^c}|Y^n, \theta]d\pi(\theta|Y^n) \\ &\leq e^{-nr} + \epsilon + P^\pi[N_\epsilon^c|Y^n]. \end{aligned}$$

Therefore, conditionally on θ, $H(Y^n) \to 0$ P_θ a.s. Moreover, $m(Y^n|X^n)$ is a.c. wrt $f(Y^n|\theta)$, $\pi(d\theta|X^n)$ a.s., for all X^n. This implies that $H(Y^n) \to 0$, $m(Y^n|X^n)$ a.s. for all X^n. □

Under H_1 i.e. when $f_0 \notin \mathcal{F}$, we have: $H(X^n) \to d(f_0, \mathcal{F}) > 0$. Therefore, Lemma 1 implies the first part of theorem 4.

Under H_0,

$$\begin{aligned} P[H(Y^n) > H(X^n)|X^n] &= \int P[H(Y^n) > H(X^n)|\theta, X^n]\pi_0(\theta|X^n)d\theta \\ &= \int_{|u|\le n^\alpha} \frac{e^{-u'J(x)u/2}}{(2\pi)^{k/2}|J(x)|^{-1/2}} \times \\ &\qquad P[H(Y^n) > H(X^n)|\theta_0 + u/\sqrt{n}, X^n]du(1 + O(n^{-1})), \end{aligned}$$

for any $h > 0$, where $J(x)$ is the empirical Fisher information matrix associated with X^n. $J(x)$ converges to $I(\theta_0)$.

Let $Z_1^y = n^{-1/2} D \log f_{\theta_0}(Y^n)$. The first term of the rhs can then be expressed as:

$$\begin{aligned} &\int_{|u|\le n^\alpha} \frac{e^{-u'J(x)u/2}}{(2\pi)^{k/2}|J(x)|^{-1/2}} P[H(Y^n) > H(X^n)|\theta_0 + u/\sqrt{n}, X^n]du \\ &= \int \mathbb{I}_{H(Y^n)>H(X^n)} \int_{|u|\le n^\alpha} \frac{e^{-u'J(x)u/2}}{(2\pi)^{k/2}|J(x)|^{-1/2}} e^{u'Z_1^Y - u'J(y)u/2 + R_n(y,u)} du f(Y^n|\theta_0)dY^n \\ &= \int \mathbb{I}_{H(Y^n)>H(X^n)} \frac{e^{(Z_1^y)'(J(x)+J(y))^{-1}Z_1^y/2}|J(x) + J(y)|^{-1/2}}{|J(x)|^{-1/2}} f(Y^n|\theta_0)dy^n(1 + O_P(n^{-1/2})) \\ &\ge \int \mathbb{I}_{H(Y^n)>H(X^n)} \frac{|J(x) + J(y)|^{-1/2}}{|J(x)|^{-1/2}} f(Y^n|\theta_0)dY^n(1 + O_P(n^{-1/2})). \end{aligned}$$

Conditionally on θ_0 Y^n and X^n are iid, so $|J(x)+J(y)|^{-1/2} = |I(\theta_0)|^{-1/2}/2^{k/2} + O_P(n^{-1/2})$ and $|J(y)|^{-1/2} = |I(\theta_0)|^{-1/2} + O_P(n^{-1/2})$. This implies that

$$P[H(Y^n) > H(X^n)|X^n] \ \ge \ 2^{-k/2} P^n_{\theta_0}[H(Y^n) > H(X^n)|X^n](1 + O_P(n^{-1/2})).$$

Using the Schwarz inequality, we also have

$$P[H(Y^n) > H(X^n)|X^n] \ \le \ P^n_{\theta_0}[H(Y^n) > H(X^n)|X^n](1 + O_P(n^{-1/2})).$$

In other words the distribution of $H(Y^n)$ gives us, asymptotically a lower bound and an upper bound, up to a constant, to the frequentist distribution of $H(X^n)$. This test procedure is therefore asymptotically equivalent to using a p-value, up to a constant. Note that the order of approximation here is a $O(n^{-1/2})$ 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,

$$\begin{aligned} G(X^n) &= E^\pi[d(f_{\theta,\psi}, f_\theta)|X^n] \\ &\ge d(f_0, \mathcal{F}) - E^\pi[d(f_0, f_{\theta,\psi})|X^n] \qquad (17) \end{aligned}$$

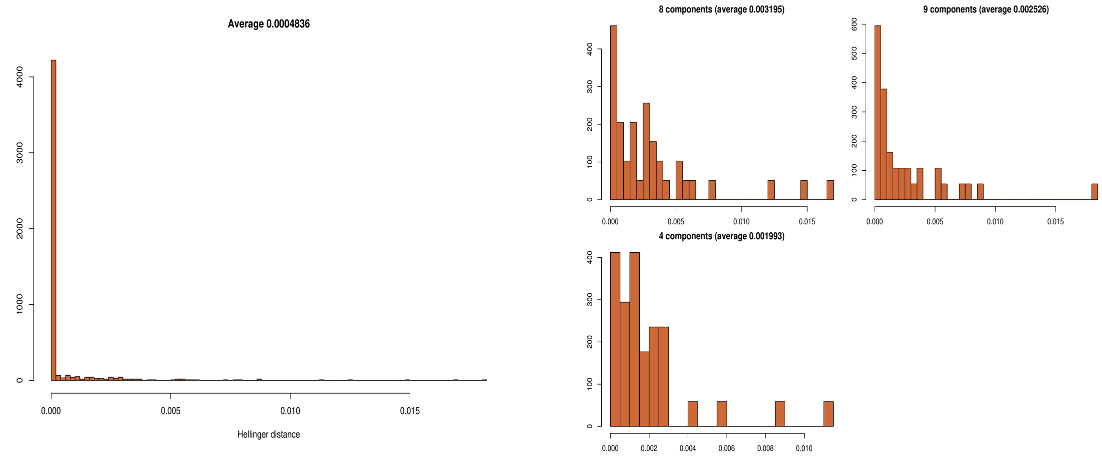

Fig. 9. Distribution of the Hellinger distance $d(g, 1)$ for an equidistributed sequence of 1000 points *(left)* and corresponding distributions conditional on $K > 0$ *(right)*.

and

$$G(X^n) \leq E^{\pi}[d(f_0, f_\theta)|X^n] + E^{\pi}[d(f_0, f_{\theta,\psi})|X^n] \tag{18}$$

Hence, if $d(f_0, \mathcal{F}) = \epsilon > 0$, i.e. under H_1, using (17) 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 have therefore 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 since it forgets the width of $\mathcal{F}$, it is somehow, like reducing $\mathcal{F}$ to $f_{\hat\theta}$, where $\hat\theta$ is for instance the maximum likelihood estimator.

6. Discussion

If we go back to the three samples introduced in Section 4.2, we can see on Figure 9 that, in the first case of the equidistributed sequence, the Hellinger distance $d(g, 1) = 2(1 - \int \sqrt{g})$ is well-concentrated around zero. (Since the computation of $\int \sqrt{g}$ is not possible in closed form, this integral was replaced by a simple trapezoidal approximation.) This property remains true when conditioning on values of K larger than 0, as shown by the histograms on the right. Figures 10 and 11 exhibit much larger values for the two other samples. Note once more on Figure 11 that conditioning on the value of K does not significantly modify the distribution of the distance $d(g, 1)$, a fact we can relate to the weak identifiability of the mixture parameters and which argues in favour of a parameter free approach as the one we favour. On the opposite, the Bayes factor will suffer to some extent from this weak identifiability and will be more sensitive to the choice of the prior.

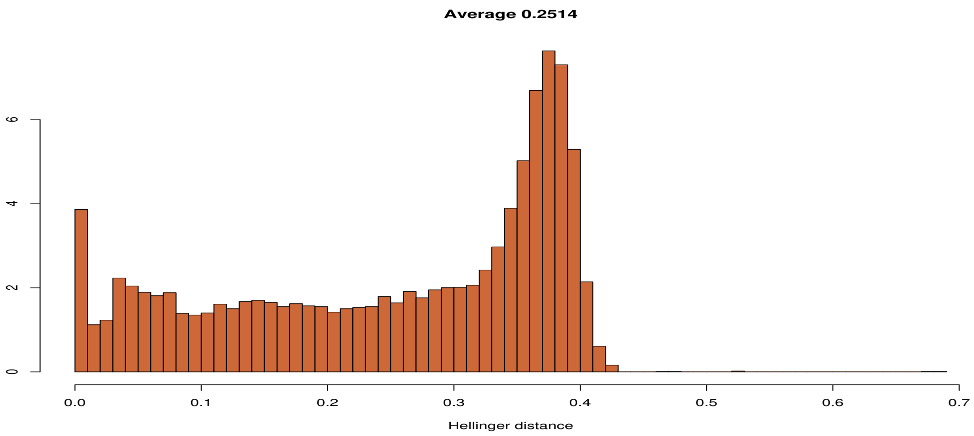

Fig. 10. Distribution of the Hellinger distance $d(g, 1)$ for a sequence of 1500 random points corresponding to the histogram of Figure 5.

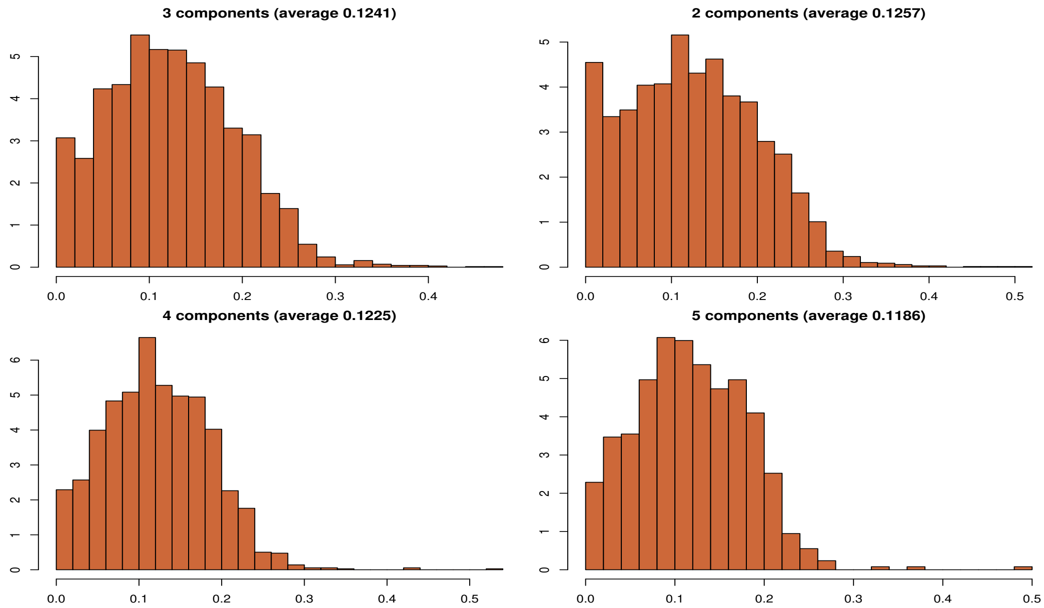

Fig. 11. Distribution of the Hellinger distance $d(g, 1)$ for a sequence of 1250 random points corresponding to the histogram of Figure 7, conditional on the most likely values of K.

References

Barron, A., Schervish, M.J. and Wasserman, L. (1999) The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27**, 536–561.
Cappé, O., Robert, C.P. and Rydén, T. (2002) 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. (1990) Estimation des paramètres d'un mélange par échantillonnage bayésien. *Notes aux Comptes–Rendus de l'Académie des Sciences I* **311**, 653–658.
Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.
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.
Robert, C. (2001) *The Bayesian Choice* (second edition). Springer-Verlag, New York.
Stephens, M. (2000a) 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) $\int f_i^U(x) d\mu(x) \le 1 + \varepsilon$ for all i,

(b) for each $P \in \mathcal{C}$ there exists i such that $f_P \le f_i^U$ 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 such that

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

and such that

(i) $\pi(\mathcal{F}_n^c) \le c_1 \exp(-nc_2)$ 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 (3.1). Under conditions* **A1** *and* **A2**, *for every $\varepsilon > 0$,*

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