Consistency of Bernstein Polynomial Posteriors

Sonia Petrone and Larry Wasserman
Università dell'Insubria and Carnegie Mellon University

Abstract

A Bernstein prior is a probability measure on the space of all the distribution functions on $[0,1]$. Under very general assumptions, it selects absolutely continuous distribution functions, whose densities are mixtures of known beta densities. The Bernstein prior is of interest in Bayesian nonparametric inference with continuous data.

In this paper we study the consistency of the posterior from a Bernstein prior. We first show that, under mild assumptions, the posterior is weakly consistent for any distribution function P_0 on $[0,1]$ with continuous and bounded Lebesgue density.

With slightly stronger assumptions on the prior, the posterior is also Hellinger-consistent. This implies that the predictive density from a Bernstein prior, which is a Bayesian density estimate, converges in Hellinger sense to the true density (assuming it continuous and bounded). We also study a sieve maximum likelihood version of the density estimator, and show that it is also Hellinger consistent under weak assumption.

When the order of the Bernstein polynomial, i.e. the number of components in the beta-mixture, is truncated, we show that under mild restrictions, the posterior concentrates on the set of so called pseudo-true densities. Finally, we study the behavior of the predictive density numerically and we also study a hybrid Bayes-maximum likelihood density estimator.

1 Introduction

A *Bernstein prior* (Petrone, 1999*a*) is a probability measure on the space of all the distribution functions on the unit interval, based on Bernstein polynomials. Under very general assumptions, it selects absolutely continuous distribution functions, with a continuous and smooth density; for these properties, the Bernstein prior can be an appropriate prior for Bayesian nonparametric inference with continuous data. Let $\{X_1, X_2, \ldots\}$ be a sequence of exchangeable random variables with values in $[0,1]$. Roughly speaking, assuming a Bernstein prior is equivalent to the following hierarchical model

(a) for any $n = 1, 2, \ldots, (X_1, \ldots X_n)$ are conditionally independent and identically distributed, given $\tilde{k} = k, \tilde{w}_k = w_k$, with common density

$$b(x; k, w_k) = \sum_{j=1}^{k} w_{j,k} \beta(x; j, k-j+1),$$

a mixture of beta densities with parameters $(j, k-j+1)$, with k components and mixing weights $w_k = (w_{1,k}, \ldots, w_{k,k})$, $w_{j,k} \geq 0$, $\sum_{j=1}^{k} w_{j,k} = 1$. The density $b(\cdot; k, w_k)$ is called a Bernstein density with parameters k and w_k;

(b) conditionally on $\tilde{k} = k$, the random vector $\tilde{w}_k$ has distribution function $H_k(\cdot)$ on the $(k-1)$-dimensional simplex;

(c) $\tilde{k}$ is a random variable with probability mass function p on the integers.

The attraction of this model is that it is non parametric yet it is fairly simple since, for each k, the model is a mixture model with no unknown parameters except the mixing weights, so that it is easy to do Bayesian estimation (by Gibbs sampling) and maximum likelihood estimation (using the EM algorithm). Furthermore, results about this model are the first step for studying more general mixture models of the kind $\sum_{j=1}^{k} w_{j,k} f_{j,k}(x)$, where $f_{j,k}$ is a given density (univariate or multivariate; see e.g. Epifani, 1999; Petrone and Veronese, 1999). The number of components of the mixture and the mixing weights are unknown and are given a prior distribution. Priors of this form are sometimes called sieve priors. In the context of regression, a sieve prior was used by Diaconis and Freedman (1993).

In this paper, we will adopt a Bayesian approach. Our main aim is to study consistency of the posterior from a Bernstein prior. In particular, we are interested in the predictive distribution function of the next observation X_{n+1}, given $(X_1 = x_1, \ldots, X_n = x_n)$:

$$\hat{F}(x) = P(X_{n+1} \le x \mid X_1 = x_1, \ldots, X_n = x_n) .$$

It is easy to show that the predictive distribution function from a Bernstein prior is absolutely continuous, with density

$$\hat{f}(x) = \sum_{k=1}^{\infty} b(x; k, E(\tilde{w}_k \mid k, x_1, \ldots, x_n))\, p(k \mid x_1, \ldots, x_n), \tag{1}$$

a mixture of Bernstein densities with parameters k and $E(\tilde{w}_k \mid k, x_1, \ldots, x_n)$ (the vector of conditional expectations of $\tilde{w}_k \mid \tilde{k} = k, X_1 = x_1, \ldots, X_n = x_n$), and mixing weights $p(k \mid x_1, \ldots, x_n)$ (the posterior probability function of $\tilde{k}$). The predictive distribution function is of interest for predicting X_{n+1}, given the data $(x_1, \ldots, x_n)$, and it also provides the Bayesian estimate of the unknown distribution function of the data, with quadratic loss. Analogously, the predictive density is of interest for predictive purposes, or as a Bayesian density estimate.

Our first results are about consistency of the posterior and of the predictive distribution from a Bernstein prior. In what follows, we denote by the same symbol a d.f. and the corresponding probability measure. All statements about convergence are a.s. with respect to the infinite product measure under the true distribution P_0, assuming independent sampling from P_0. $\pi(\cdot|X^n)$ denotes the natural version of the posterior probability measure from the Bernstein prior π, given $X^n = (X_1, \ldots, X_n)$. We show that the posterior is consistent under mild conditions. Specifically, we prove that if the data are i.i.d. with a distribution function P_0 on $[0,1]$ which has a bounded, continuous density f_0, then:

(i) under very weak conditions on the prior weights $p(k)$ and on the conditional distribution of $\tilde{w}_k \mid k$, the posterior is weakly consistent at P_0, i.e. $\pi(W|X^n) \to 1$ a.s. for every weak neighborhood W of P_0. It follows that the sequence of predictive distribution functions converges weakly to P_0 (see e.g. Ghosal, Ghosh and Ramamoorthi, 1998).

(ii) With slightly stronger conditions on the prior weights $p(k)$, the posterior is strongly consistent, i.e. $\pi(H|X^n) \to 1$ a.s. for every Hellinger neighborhood H of P_0; it follows that the sequence of predictive densities converges to the true density f_0 in Hellinger sense (Barron, Schervish and Wasserman, 1999, Corollary 1).

We also study maximum likelihood estimation for the Bernstein model. More specifically, we consider the sieve estimator $\hat{f}_n(\cdot) \equiv b(\cdot; k_n, \hat{w}_{k_n})$ where k_n is an increasing sequence of integers and $\hat{w}_k$ is the maximum likelihood estimate of w_k. A similar estimator has been proposed by Vitale (1975) and Tenbusch (1995), where $w_{j,k}$ is estimated by the number of data points falling in the interval $[\frac{j-1}{k}, \frac{j-1}{k})$. In our numerical examples, we consider choosing k using the data as well as a model-averaged version of the sieve estimator. We show that

(iii) if the true density f_0 is continuous and bounded away from zero and infinity, the sieve maximum likelihood estimator $\hat{f}_n(\cdot) = b(\cdot; k_n, \hat{w}_{k_n})$ is Hellinger consistent if $k = k_n$ grows at an appropriate rate with the sample size n.

The results about consistency of the posterior are contained in sections 2 and 3. Consistency of the sieve estimator is in section 5. We also consider (section 4) the case when k is truncated and we show that the posterior asymptotically concentrates on the set of so called pseudo-true densities (see Bunke and Milhaud, 1998).

In sections 6 and 7, we discuss computational issues and study the numerical behavior of the estimators for finite sample size. The predictive density from the Bernstein prior is estimated by the MCMC algorithm discussed in Petrone (1999a,b). We also propose a different approximation, which consists in two steps. First, we approximate $E(\tilde{w}_k \mid k, x_1, \ldots, x_n)$ with the maximum likelihood estimate of $\tilde{w}_k$, which can be easily computed by the EM algorithm. Then, we compute the BIC (or AIC) values for each value of k, and normalize them for obtaining an approximation of $p(k \mid x_1, \ldots, x_n)$. This proposal can be also regarded as a Bayes-maximum likelihood hybrid estimator in which the maximum likelihood estimates from each model are averaged with respect to data based weights, i.e. as a frequentist method that averages over models rather than selecting one model. We study the performance of the proposed hybrid estimator numerically, and we also compare our results with the frequentist Bernstein estimator proposed by Tenbusch (1995).

The final section contains a discussion of possible extensions of our results to more general sample spaces.

2 Weak consistency of the Bernstein posterior

Let us briefly recall the construction of the Bernstein prior. For a given integer k and a function F on $[0,1]$ define the Bernstein polynomial by

$$B(x;k,F) \equiv \begin{cases} 0 & x<0 \\ \sum_{j=0}^{k} \; F(\frac{j}{k}) \; \binom{k}{j} x^j (1-x)^{k-j} & 0 \le x \le 1 \\ 1 & x>1 \end{cases} \tag{2}$$

It can be easily shown that if F is a probability distribution function (d.f.) with support $[0,1]$, then $B(\cdot;k,F)$ is a d.f. too. Precisely, the d.f. $B(\cdot;k,F)$ is a mixture of a point mass $F(0)$ on $\{0\}$ and an absolutely continuous d.f. with density

$$b(x;k,w_k) = \sum_{j=1}^{k} w_{j,k} \beta(x;j,k-j+1)$$

where $w_{j,k} = F(\frac{j}{k}) - F(\frac{j-1}{k})$ and $\beta(\cdot;a,b)$ denotes a beta density with parameters (a,b). Let $\mathbf{N} = \{1,2,\ldots\}, \mathcal{P}(\mathbf{N})$ be the power set of $\mathbf{N}$ and Δ be the class of all the distribution functions on the unit interval, equipped with the σ-field $\mathcal{S}$ generated by weak topology. For defining the Bernstein prior, let $\tilde{k}$ be a random variable with values in $1,2,\ldots$ and $\tilde{F}$ be a random d.f. with support $[0,1]$. Suppose that $(\tilde{k},\tilde{F})$ have joint probability law $\mathbf{P}$ on $(\mathbf{N}\times\Delta, \mathcal{P}(\mathbf{N})\times\mathcal{S})$. Then, the Bernstein polynomial $B(\cdot;\tilde{k},\tilde{F})$ is a random d.f. on $[0,1]$, whose probability law is induced by $\mathbf{P}$ and is called *Bernstein prior* with parameter $\mathbf{P}$. In particular, when $\tilde{k}$ has probability mass function $p(k)$ on $\mathbf{N}$ and $\tilde{F}$ is a Dirichlet process $D(MF_0)$, where M is a positive constant and F_0 is a d.f. on $[0,1]$, we say that $B(\cdot;\tilde{k},\tilde{F})$ has a *Bernstein-Dirichlet prior*, with parameters $(p(\cdot);MF_0)$.

If $\mathbf{P}(\tilde{F}(0)=0)=1$, then the random d.f. $B(\cdot;\tilde{k},\tilde{F})$ is a.s. absolutely continuous, with density $b(x;\tilde{k},\tilde{w}_k)$ where $\tilde{w}_k = (\tilde{w}_{1,k},\ldots,\tilde{w}_{k,k})$ are the increments $\tilde{w}_{j,k} = \tilde{F}(\frac{j}{k}) - \tilde{F}(\frac{j-1}{k}), j=1,\ldots,k$ of the random d.f. $\tilde{F}$. In this paper, we restrict attention to this model.

Let $\mathcal{B}_k$ be the class of all absolutely continuous Bernstein d.f.'s of order k. With the restriction $\mathbf{P}(\tilde{F}(0)=0)=1$, the Bernstein prior is supported on the space of all absolutely continuous Bernstein d.f.'s, $\cup_{k=1}^{\infty}\mathcal{B}_k$, and can be written as

$$\pi(B) = \sum_{k=1}^{\infty} p(k)\pi_k(B) \quad , \; B \in \mathcal{S} \tag{3}$$

where $p(k) = \mathbf{P}(k\times\Delta)$ and π_k is the restriction of π on $\mathcal{B}_k$. We note that, while having large support, the Bernstein prior selects d.f.'s which have a parametric expression, being mixtures of given beta distributions. Therefore we can talk of an unknown parameter in the Bernstein model, which is given by $\tilde{\theta} = (\tilde{k},\tilde{F})$, the parameter space being $\Theta = \mathbf{N}\times\Delta$. Note however that, conditionally on $\tilde{k}=k$, the prior is restricted on $\mathcal{B}_k$, and the relevant parameter, given k, is a function of $\tilde{F}$, namely the vector $\tilde{w}_k$ of its increments; therefore, conditionally on k the parameter space reduces to the $(k-1)$ dimensional simplex S_{k-1}. One might be interested in establishing weak consistency (in the sense of Diaconis and Freedman, 1986) of the posterior of $(\tilde{k}, \tilde{F})$. However, this is a difficult task, due to the complicated expression of the likelihood in the mixture model. Moreover, there are identifiability issues that have to be handled carefully. For example, suppose that the true density is a mixture of two beta densities of the form $\beta(\cdot; j, k-j+1)$, for some values of j and k, say (j_1, k_1) and (j_2, k_2). Asymptotically, the posterior of k might concentrate on k_1 and k_2 and the posterior of $w_{k_i}|\tilde{k} = k_i$, for $i = 1, 2$, might concentrate on the vector w_{k_i} whose j_i-th component is 1. However, since the true density can be uniformly approximated by a single Bernstein density with $k \to \infty$ (see theorem 3 below), the posterior of k might also/alternatively concentrate towards infinity as the sample size increases.

On the other hand, in many problems we are not interested in $(\tilde{k}, \tilde{F})$ themselves, but rather in some functionals, in particular, in the unknown distribution $\tilde{Q}$ of the data. Therefore the parameter of interest is $\tilde{Q}$. Still, it can be difficult to study the behavior of the posterior distribution of $\tilde{Q}$ analytically, but one can use sufficient conditions for consistency. A result by Schwartz (1965, Theorem 6.1; see also Barron (1988) and Ghosal, Ghosh and Ramamoorthi, 1998) provides a sufficient condition for weak consistency. We recall Schwartz's theorem for the reader's convenience.

Theorem 1 *Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ denote the sample space, $(\Theta, \mathcal{B}(\Theta))$ denote the parameter space and let π be the prior and $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ be the class of all statistical models. Suppose that $\mathcal{P}$ is dominated i.e. there exists a σ-finite measure ν such that each $P_\theta \in \mathcal{P}$ has density p_θ with respect to ν, and that the densities $p_\theta(x)$ can be chosen $\mathcal{B}(\Theta) \times \mathcal{B}(\mathcal{X})$-measurable.*

Let $P_{\theta_0} \in \mathcal{P}$. Under regularity conditions, if $\pi(U_\delta) > 0$ for every Kullback-Leibler neighborhood U_δ of P_{θ_0}, then the posterior is weakly consistent at P_{θ_0}.

The regularity conditions are assumptions A.1 to A.4 of Schwartz (1965). In our case, the sample space is $([0,1], \mathcal{B}([0,1])$ and we consider $(\Theta, \mathcal{B}(\Theta)) = (\Delta^*, \mathcal{S}^*)$ where Δ^* is the class of all the absolutely continuous distribution functions on $[0,1]$, and $\mathcal{S}^*$ is the restriction of $\mathcal{S}$ to Δ^*. The class of all statistical models is $\mathcal{P} = \Delta^*$ and is therefore dominated by Lebesgue measure λ. It is easy to verify that the regularity conditions are satisfied; in fact, it is not obvious to show that assumption (A2) hold, namely that Δ^* is a Borel set in $\mathcal{S}$, but this is proved by Theorem 4.1 in Gaudart and Hadwin (1989). Furthermore, using the sufficient condition (V.5.58) in Dellacherie and Meyer (1982), it can be shown that the densities $f_G(x) = dG(x)/d\lambda(x)$ for $G \in \Delta^*$ can be chosen so that the map $\psi : (\Delta^* \times [0,1]) \to \mathbf{R}$ defined by $\psi(G, x) = f_G(x)$ is $(\mathcal{S} \times \mathcal{B}([0,1]))$-measurable.

In what follows, we denote by f_0 the density of the (true) d.f. $P_0 \in \Delta^*$, and by f_P the density of a d.f. $P \in \Delta^*$, and we let

$$U_\delta = \left\{ P \in \Delta^* : \int f_0(x) \log \frac{f_0(x)}{f_P(x)} dx \leq \delta \right\}$$

be a δ Kullback-Leibler neighborhood of P_0. Note that for any δ, $U_\delta \in \mathcal{S}^*$ since the map $K : P \rightarrow \int f_0(x) log \frac{f_0(x)}{f_P(x)} dx$ from $\Delta^* \rightarrow \Re$ is $\mathcal{S}^*$-measurable (see the Appendix).

Next we state the result about weak consistency of Bernstein posteriors.

Theorem 2 *Suppose that P_0 has a bounded, continuous density f_0. If for all $k = 1, 2, \ldots, p(k) > 0$ and $\tilde{w}_k | \tilde{k} = k$ has an absolutely continuous d.f., with density $h_k(\cdot)$ which is positive on the $(k-1)$-dimensional simplex, then the posterior from a Bernstein prior is weakly consistent at P_0.*

To prove the theorem, we make use of the following known result (see for example Altomare and Campiti, 1994, p.333), which in particular implies that Bernstein densities can uniformly approximate any density continuous on $[0, 1]$.

Theorem 3 *Let f be a continuous function on $[0, 1]$. Let $w_k(f) = (w_{1,k}(f), \ldots, w_{k,k}(f))$ where $w_{j,k}(f) = \int_{\frac{j-1}{k}}^{\frac{j}{k}} f(x)\, dx$. Then,*

$$lim_{k\rightarrow\infty} sup_{x\in[0,1]} |b(x; k, w_k(f)) - f(x)| = 0.$$

Proof of Theorem 2. According to theorem 1 we need only show that $\pi(U_\delta) > 0$ for every $\delta > 0$.

We first suppose that f_0 is bounded away from zero, i.e. $inf_{x\in[0,1]} f_0(x) = a > 0$. Let $K(f, g)$ denote the Kullback-Leibler divergence between f and g. Now, we claim that for any $\delta > 0$ there exists $b_0(\cdot) \equiv b(\cdot; k_0, w_{k_0}(f_0))$ such that $K(f_0, b_0) < \delta$. Fix $c > 0$. From Theorem 3 we may choose $k_0 = k_0(c)$ such that $\sup_x |f_0(x) - b(x; k, w_k(f_0))| < c$, for any $k \geq k_0$. For c sufficiently small, $b(\cdot; k, w_k(f_0))$ will be also bounded away from zero, and bounded, for any $k \geq k_0$; therefore,

$$|log \frac{f_0(x)}{b(x; k, w_k(f_0))}| < M,$$

say, for any $x \in [0, 1]$ and $k \geq k_0$. From the dominated convergence theorem,

$$lim_{k\rightarrow\infty} \int log \frac{f_0(x)}{b(x; k, w_k(f_0))} f_0(x) dx = \int lim_{k\rightarrow\infty} log \frac{f_0(x)}{b(x; k, w_k(f_0))} f_0(x) dx = 0$$

Therefore, we have shown that for any $\delta > 0$ there exists k_0 such that $K(f_0, b(\cdot; k_0, w_{k_0}(f_0))) < \delta$.

Now we show that there is a Kullback-Leibler neighborhood of $b_0(\cdot) = b(\cdot; k_0, w_{k_0}(f_0))$ with positive prior probability. For $\epsilon > 0$ let

$$N_\epsilon = \{w_{k_0} : sup_{j=1,\ldots,k_0} |w_{j,k_0} - w_{j,k_0}(f_0)| < \epsilon\}.$$

Remember that b_0 is bounded and bounded away from zero. Therefore, being

$$sup_{x\in[0,1]} |b(x; k_0, w_{k_0}) - b(x; k_0, w_{k_0}(f_0))| \leq k_0\, sup_{j=1,\ldots,k_0} |w_{j,k_0} - w_{j,k_0}(f_0)|,$$

by choosing ϵ small enough we can make $b(\cdot; k_0, w_{k_0})$ bounded and bounded away from zero too, for any $w_{k_0} \in N_\epsilon$. It follows that

$$|log \frac{f_0(x)}{b(x; k_0, w_{k_0})}| < M,$$

say, for any $x \in [0,1]$ and $w_{k_0} \in N_\epsilon$; and from the dominated convergence theorem,

$$lim_{w_{k_0} \to w_{k_0}(f_0)} K(f_0, b(\cdot; k_0, w_{k_0})) = K(f_0, b(\cdot; k_0, w_{k_0}(f_0)))$$

i.e. $K(f_0, b(\cdot; k_0, w_{k_0}))$ is continuous at $w_{k_0}(f_0)$. Hence, for any $\delta > 0$ we can choose ϵ such that $K(f_0, b(\cdot; k_0, w_{k_0})) < \delta$ for any $w_{k_0} \in N_\epsilon$. Therefore

$$\pi(U_\delta) = \pi(\{b : K(f_0, b) \leq \delta\}) \geq \pi(\{b(\cdot; k_0, w_{k_0}) : w_{k_0} \in N_\epsilon\})$$
$$= p(k_0) \int_{N_\epsilon} h_{k_0}(w) dw > 0,$$

since N_ϵ has positive Lebesgue measure and by assumption h_k is positive, so that the integral above is positive, and $p(k) > 0$ for any k by assumption.

Now consider the case when $\inf_x f_0(x) = 0$. We make use of the following result by Ghosal, Ghosh and Ramamoorthi (1999a; Lemma 5.1). If f_0 and f_1 are densities so that $f_0 \leq C f_1$ for a positive constant C then for any f

$$K(f_0, f) \leq (C+1) logC + C[K(f_1, f) + \sqrt{K(f_1, f)}]. \tag{4}$$

Now choose $a > 0$ and define

$$f_1(x) = \frac{f_0(x) \vee a}{\int_{[0,1]} (f_0(x) \vee a) dx}$$

where $(a \vee b)$ is the maximum between a and b. The density f_1 is bounded away from zero and infinity, therefore it is covered by the case above, so that for any $\epsilon > 0$ we can choose a set $U_\epsilon(f_1)$ with positive prior probability such that $K(f_1, b) < \epsilon$ for any $b \in U_\epsilon(f_1)$. Now, $f_0 \leq C f_1$, where $C = \int_{[0,1]} (f_0(x) \vee a) dx$; for the inequality (4), we have

$$K(f_0, b_0) \leq (C+1) logC + C[K(f_1, b_0) + \sqrt{K(f_1, b_0)}]$$

which can be made $< \delta$ by choosing ϵ and a small enough. Therefore

$$\pi(U_\delta) \geq \pi(\{b : K(f_1, b) < \epsilon\}) > 0.$$

This concludes the proof. $\diamond$

3 Strong Consistency

Weak consistency of the posterior implies that the predictive d.f., which is the Bayesian estimate of the distribution function with quadratic loss, converges weakly to the true distribution function. However, in general it does not imply that the predictive density (when it exists) converges to the true density, and stronger forms of convergence are of interest. In this section we study consistency in Hellinger sense. For Q_1 and Q_2 in Δ^*, the Hellinger metric is defined by

$$d_H(Q_1, Q_2) = \{\int (\sqrt{f_{Q_1}(x)} - \sqrt{f_{Q_2}(x)})^2 d\lambda(x)\}^{1/2}.$$

where λ is Lebesgue measure on $[0, 1]$ and $f_{Q_i} = dQ_i/d\lambda$. The posterior is Hellinger consistent at P_0 if for every $\epsilon > 0$,

$$\pi(H_\epsilon | X^n) \to 1 \tag{5}$$

where $H_\epsilon = \{Q \in \Delta^* : d_H(P_0, Q) \leq \epsilon\}$ is a Hellinger neighborhood of P_0. Note that $H_\epsilon \in \mathcal{S}$ since the map $d_H : Q \to d_H(P_0, Q)$, from Δ^* in $\Re$, is $\mathcal{S}^*$-measurable (see the Appendix).

We show that with slightly stronger conditions on the prior weights $p(k)$, the posterior from a Bernstein prior is consistent in the Hellinger distance. We need a little more notation. Throughout, c_1, c_2, c_3, c_4 are fixed, universal positive constants; see Theorem 8. The Hellinger pseudo-metric (we denote it by the same symbol d_H) between two integrable functions f and g is defined by $d^2_H(f, g) = \int(\sqrt{f} - \sqrt{g})^2$. A set of pairs of functions $(\ell_1, u_1), \ldots, (\ell_r, u_r)$ is an ϵ-bracketing for a set of absolutely continuous d.f.'s $\mathcal{F}$ if $d_H(\ell_j, u_j) \leq \epsilon$ for $j = 1, \ldots, r$ and for every $Q \in \mathcal{F}$ there exists $1 \leq j \leq r$ such that $\ell_j(x) \leq f_Q(x) \leq u_j(x)$. The minimum number of such brackets required to cover $\mathcal{F}$ is denoted by $N_{[]}(\epsilon, \mathcal{F})$. The bracketing entropy is defined by $H_{[]}(\epsilon, \mathcal{F}) = \log N_{[]}(\epsilon, \mathcal{F})$. Define the bracketing integral by

$$J(\epsilon, \mathcal{F}) = \int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \sqrt{H_{[]}(u/c_3, \mathcal{F})} du.$$

We will use the following result from Wasserman (1998) which is a slight variant of the main result Barron, Schervish and Wasserman (1999).

Theorem 4 *Suppose the following two conditions hold.*

(A1) $\pi(U_\delta) > 0$ *for every Kullback-Leibler neighborhood* U_δ *of* P_0 *;*

(A2) For every $\epsilon > 0$ *there exists* $r > 0$ *and subsets* $\mathcal{F}_1, \mathcal{F}_2, \ldots$ *of* Δ^* *such that* $\pi(\mathcal{F}_n^c) < e^{-nr}$ *and* $J(\epsilon, \mathcal{F}_n) \leq c_4\sqrt{n}\epsilon^2$.

Then, the posterior is Hellinger consistent at P_0.

Theorem 5 *Suppose there exists $k_n \to \infty$ such that $k_n = o(n)$ and such that $\sum_{k \geq k_n} p(k) \leq e^{-nr}$ for some $r > 0$. Further suppose that the assumptions of theorem 2 hold and that P_0 has a bounded continuous density f_0. Then the posterior from the Bernstein prior is Hellinger consistent at P_0.*

For the proof, we use theorem 4, which requires to find a sequence $\{\mathcal{F}_n, n = 1, 2, \ldots\}$ (a sieve) such that the prior probability of $\mathcal{F}_n^C$ is exponentially small; at the same time, the set $\mathcal{F}_n$ should not be too large, since assumption (A2) requires that $J(\epsilon, \mathcal{F}_n) \leq c_4\sqrt{n}\epsilon^2$. In our case, the role of $\mathcal{F}_n$ is played by the set of Bernstein d.f.'s of order $\leq k_n$. For showing that it is not too large, we use the following result which follows from Theorem 2 of Genovese and Wasserman (1998).

Theorem 6 *There exists a positive constant $c > 0$ such that, for all $1 \geq \epsilon > 0$,*

$$N_{[]}(\epsilon, \mathcal{B}_k) \leq \left(\frac{c}{\epsilon}\right)^k .$$

Proof of Theorem 5. Note that (A1) of Theorem 4 was already shown to hold. We need only show that condition (A2) holds. Remember that $\mathcal{B}_k$ is the class of all absolutely continuous Bernstein d.f.'s of order k, and let $\mathcal{F}_n = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_{k_n}$. Then $\pi(\mathcal{F}_n^c) = \sum_{j \geq k_n} p(j) < e^{-nr}$ by assumption. Clearly, $N_{[]}(\epsilon, \mathcal{F}_n) \leq \sum_{j=1}^{k_n} N_{[]}(\epsilon, \mathcal{B}_j) \leq k_n(c/\epsilon)^{k_n} \leq (2c/\epsilon)^{k_n}$. Direct integration shows that for any $c' > 0$, $J(\epsilon, \mathcal{F}_n) \leq c'\sqrt{n}\epsilon^2$ for all large n. Hence, condition (A2) holds and the result then follows. $\diamond$

Remark. One can consider mixtures of Bernstein densities, of the form

$$\sum_{r=1}^{k} \xi_r b(x; r, w_r) ,$$

with $\xi_r \geq 0$, $\sum_r \xi_r = 1$, which can be also written as

$$b^*(x; k, W_k) = \sum_{r=1}^{k} \sum_{j=1}^{r} W_{j,r} \beta(x; j, r - j + 1)$$

where $W_k = (w_1, w_2, \ldots, w_k)$ and $w_r = (w_{1,r}, \ldots, w_{r,r})$ for $1 \leq r \leq k$. Thus, W_k is a vector of length $d_k = k(k+1)/2$. The elements of W_k are non-negative and sum to 1. This an "extended Bernstein density." It can be easily shown that the results about weak and strong convergence still hold for the extended Bernstein density.

4 Asymptotic behavior when k is truncated

In the previous sections, we assumed that $p(k) > 0$ for all integers $k = 1, 2, \ldots$; this assumption is necessary to guarantee that the Bernstein prior is supported on the set of all Bernstein d.f.'s $\cup_{k=1}^{\infty} \mathcal{B}_k$, and has been used for proving consistency.

However, for practical reasons the order k is often truncated to a maximum value k_{max}, so that the prior has support $\mathcal{B}^* = \cup_{k=1}^{k_{max}} \mathcal{B}_k$. In this section we study the asymptotic behavior of the posterior in this case. There are a few papers on the asymptotic behavior of Bayes estimators in the case of a possibly incorrect parametric model. We quote Berk (1966; 1970) and Bunke and Milhaud (1998). These results provide regularity conditions under which the posterior distribution a.s. concentrates on the subset Θ_0 of the parameter space Θ on which the Kullback-Leibler divergence of the true d.f. P_0 against the distributions P_θ in the model is minimal. The points of Θ_0 are called pseudo-true parameters. We prove a result of this kind for the truncated Bernstein prior. Our proof is based on the theory of empirical processes and in particular on the fact that $\mathcal{B}^*$ is a Glivenko-Cantelli class.

Theorem 7 *Suppose that the prior π has support $\mathcal{B}^* = \cup_{k=1}^{k_{max}} \mathcal{B}_k = \{$ all absolutely continuous Bernstein d.f.'s of order $k \le k_{max}\}$ and that for any $k = 1, \dots, k_{max}$, $\tilde{w}_k | \tilde{k} = k$ has an absolutely continuous d.f., with density $h_k(\cdot)$ which is positive on the $(k-1)$-dimensional simplex. Let $\eta = inf_{b \in \mathcal{B}^*} K(f_0, b)$. Fix $\epsilon > 0$ and let $A_\epsilon = \{b \in \mathcal{B}^* : K(f_0, b) < \eta + \epsilon\}$. If the true density f_0 satisfies $\int f_0(x) |log(f_0(x))| dx < \infty$, then for any $\epsilon > 0$, $\pi(A_\epsilon | X^n) \to 1, a.s.$*

Proof. For a Bernstein d.f. $B \in \mathcal{B}^*$, we denote by b_B the corresponding density. Let P_0^* denote the outer measure of P_0.
Now fix $\epsilon > 0$. Let $0 < \delta < \epsilon/6$. Let $A = \{B \in \mathcal{B}^*; K(f_0, b_B) < \eta + (\epsilon/2)\}$ and $E = \{B \in \mathcal{B}^*; K(f_0, b_B) > \eta + (\epsilon/2) + 3\delta\}$. Note that $E \subset A^c$. The proof has 3 steps.

(i) show that $\pi(A) > 0$;

(ii) show that the class $\mathcal{F} = \{f = log f_0/b_B, B \in \mathcal{B}^*\}$ is a Glivenko Cantelli class, i.e.

$$sup_{f \in \mathcal{F}} \left| \frac{\sum_{i=1}^n f(X_i)}{n} - \int f dP_0 \right| \to 0, \ a.s. P_0^{*\infty} ;$$

(iii) show that $\pi(A_\epsilon | X^n) \to 1$.

Assume that (i) and (ii) hold. Then let us consider

$$\frac{\pi(E|X^n)}{\pi(A|X^n)} = \frac{\int_E \prod_{i=1}^n b_B(X_i) \pi(dB)}{\int_A \prod_{i=1}^n b_B(X_i) \pi(dB)} = \frac{\int_E e^{-nD_n} \pi(dB)}{\int_A e^{-nD_n} \pi(dB)}$$

where

$$D_n = D_n(f_0, b_B) = \frac{1}{n} \sum_{i=1}^n \log \frac{f_0(X_i)}{b_B(X_i)}.$$

From (ii) it follows that

$$sup_{B \in \mathcal{B}^*} |D_n(f_0, b_B) - K(f_0, b_B)| \to 0$$

a.s. $P_0^{*\infty}$, therefore for any $\delta > 0$ we have

$$K(f_0, b_B) - \delta < D_n(f_0, b_B) < K(f_0, b_B) + \delta$$

for any $B \in \mathcal{B}^*$, for n sufficiently large. In particular

$$D_n(f_0, b_B) < \eta + \epsilon/2 + \delta$$

for any $B \in A$ and

$$D_n(f_0, b_B) > \eta + \epsilon/2 + 2\delta$$

for any $B \in E$. It follows from this and from $\pi(A) > 0$ that

$$\frac{\pi(E|X^n)}{\pi(A|X^n)} \leq \frac{e^{-n(\eta+(\epsilon/2)+2\delta)}\pi(E)}{e^{-n(\eta+(\epsilon/2)+\delta)}\pi(A)} = e^{-n\delta}\frac{\pi(E)}{\pi(A)},$$

and so

$$lim_{n\to\infty}\frac{\pi(E|X^n)}{\pi(A|X^n)} = 0, \quad \text{a.s.}$$

Now let $C = \{w : w \notin E, w \notin A\}$. Then,

$$\frac{\pi(E|X^n)}{\pi(E^c|X^n)} = \frac{\pi(E|X^n)}{\pi(A|X^n) + \pi(C|X^n)} \leq \frac{\pi(E|X^n)}{\pi(A|X^n)} \to 0 \ \text{ a.s.} P_0^{*\infty}.$$

Thus, $\pi(E^c|X^n) \to 1$ a.s. and since $\delta < \epsilon/6$, $E^c \subset A_\epsilon$, therefore we have that $\pi(A_\epsilon|X^n) \to 1$ as we wanted to prove.

Now we prove claim (ii). We first note that $\mathcal{B}^*$ is a Vapnik-Cervonenkis (V-C) class, since it is a linear combination of finitely many functions (see e.g. van der Vaart, 1998, example 19.17). Then $\mathcal{F}$ is also a V-C class (by (viii) and (v) of lemma 2.6.18 in van der Vaart and Wellner, 1996). Then we use a sufficient condition for $\mathcal{F}$ being a Glivenko-Cantelli class, namely the fact that if $\mathcal{F}$ is a V-C class with a measurable envelope function F such that $\int F dP_0^* < \infty$, then $\mathcal{F}$ is a Glivenko-Cantelli class (see van der Vaart and Wellner,1996, p. 125). F is an envelope function for $\mathcal{F}$ if $|f(x)| < F(x)$ for any $f \in \mathcal{F}$). So it remains to show that there exists a measurable function F such that $|log\frac{f_0(x)}{b_B(x)}| < F(x)$, for any $B \in \mathcal{B}^*$, and $\int_{[0,1]} F(x)f_0(x)dx < \infty$. Now, we have that for any $B \in \mathcal{B}^*$, b_B is bounded (being a mixture of finitely many bounded functions), so $b_B(x) < M$, say, and $b_B(x) \geq a(x)$ where

$$a(x) = min_{j=1,\ldots,k}\beta(x; j, k-j+1) = \begin{cases} k(1-x)^{k-1} & 0 < x < 1/2 \\ kx^{k-1} & 1/2 \leq x < 1. \end{cases}$$

Now, let $C_1 = \{x \in [0,1] : f_0(x) \leq a(x)\}; C_2 = \{x \in [0,1] : f_0(x) \geq M)\}, C_3 = \{x \in [0,1] : a(x) < f_0(x) < M\}$. Then let

$$F(x) = \begin{cases} log(M/f_0(x)) & x \in C_1 \\ log(f_0(x)/a(x)) & x \in C_2 \\ log(M/a(x)) & x \in C_3. \end{cases}$$

Clearly, $|log(f_0(x)/b_B(x))| \leq F(x)$, for any $B \in \mathcal{B}^*$. It is easy to show that F is integrable; indeed,

$$\int_{[0,1]} F(x)f_0(x)dx = \int_{C_1} f_0(x)log(M/f_0(x))dx + \int_{C_2} f_0(x)log(f_0(x)/a(x))dx + \int_{C_3} f_0(x)log(M/a(x))dx;$$

the integrals on the right side are finite, since $\int f_0(x)loga(x)dx < \infty$, and $\int f_0(x)|logf_0(x)|dx < \infty$ by assumption. This proves (ii).

To show (i), note that being $|log\frac{f_0(x)}{b(x;k,w_k)}| \leq F(x)$ for any $x \in [0,1]$ and for any w_k, with $\int F_0 f_0 < \infty$, we can apply the dominated convergence theorem to show that for any k, $K(f_0, b(\cdot;k,w_k))$ is continuous at w_k. Now, from the definition of η it follows that there exists $(k_0, w^*_{k_0})$ such that $K(f_0, b(\cdot;k_0,w^*_{k_0})) < \eta + \epsilon/4$. Since K is continuous at $w^*_{k_0}$, there exists a neighborhood N_ϵ of $w^*_{k_0}$ such that $|K(f_0, b(\cdot;k_0,w_{k_0})) - K(f_0, b(\cdot;k_0,w^*_{k_0}))| < \epsilon/4$ for any $w_{k_0} \in N_\epsilon$. Therefore, $K(f_0, b(\cdot;k_0,w^*_{k_0})) < \eta + \epsilon/2$ on N_ϵ, and

$$\pi(A) \geq p(k_0) \int_{N_\epsilon} h_{k_0}(w)dw,$$

which is positive since by assumption $p(k_0) > 0$ and $h_{k_0}(w) > 0$ for any w in the simplex, and N_ϵ has positive Lebesgue measure. This concludes the proof. $\diamond$

5 Maximum Likelihood Sieve

Let k_n be an increasing sequence and let $\hat{f}_n(\cdot) = b(\cdot;k_n,\hat{w}_{k_n})$ where $\hat{w}_{k_n}$ is the maximum likelihood estimator for w_{k_n}. Then $\hat{f}_n$ is called a sieve maximum likelihood estimator. We now show that $\hat{f}_n$ is consistent for any density f_0 continuous on $[0,1]$ and bounded away from zero, i.e. $d_H(f_0,\hat{f}_n) = o_{P_0{}^*}(1)$.

Theorem 8 *Let $k_n = n^\alpha$ for some $\alpha \in (0,1)$. Let $\hat{f}_n(x) = f(x;k_n,\hat{w}_{k_n})$ where $\hat{w}_{k_n}$ is the maximum likelihood estimator of w_{k_n} (using either the Bernstein or extended Bernstein model). Suppose that the true density f_0 is continuous, bounded and bounded away from zero. Then, for every $\epsilon > 0$,*

$$P^*(d_H(f_0,\hat{f}_n) > \epsilon) < 4e^{-nc_2\epsilon^2}.$$

Thus, $d_H(f_0,\hat{f}_n) \to 0$ a.s.

Before proving this result, let us quote the following result due to Wong and Shen (1995).

Theorem 9 *Let $X_1,\ldots,X_n$ be an i.i.d. sample from a distribution P_0 with density f_0. Let $\mathcal{F}$ be a set of densities. There exist universal positive constants c_1,c_2,c_3,c_4 such that $J(\epsilon,\mathcal{F}) \leq c_4\sqrt{n}\epsilon^2$ implies that*

$$P^*_0\left(\sup_{f\in\mathcal{F};d_H(f_0,f)>\epsilon} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} > e^{-nc_1\epsilon^2}\right) \leq 4e^{-c_2n\epsilon^2}.$$

Proof of Theorem 8. Fix $\epsilon > 0$. From our calculations in Theorem 2, and from the fact that f_0 is bounded away from zero, we see that for all large k, there exists $B_k \in \mathcal{B}_k$ such that $D_k \equiv K(f_0, b_{B_k}) < c_1\epsilon^2/4$. In fact, those calculations show that $\sigma_k^2 = \int f_0(\log f_0/b_{B_k})^2$ is finite and that $\text{limsup}_k \sigma_k < \infty$. Let $A_n = \mathcal{B}_{k_n} \cap \{f; d_H(f_0, f) > \epsilon\}$. In what follows we write k in place of k_n. Now,

$$\begin{aligned}
P_0^*(d_H(f_0, \hat{f}) > \epsilon) &\leq P_0^*\left(\sup_{f \in A_n} \prod_i \frac{f(X_i)}{b_{B_k}(X_i)} \geq e^{-c_1 n\epsilon^2/2}\right) \\
&\leq P_0^*\left(\sup_{f \in A_n} \prod_i \frac{f(X_i)}{f_0(X_i)} \geq e^{-c_1 n\epsilon^2}\right) + P_0\left(\prod_i \frac{f_0(X_i)}{b_{B_k}(X_i)} \geq e^{c_1 n\epsilon^2/2}\right) \\
&\equiv P_1 + P_2, \quad \text{say.}
\end{aligned}$$

By Theorem 8 and the earlier entropy calculations in the proof of Theorem 5, $P_1 \leq 4e^{-c_2 n\epsilon^2}$ (in either the Bernstein or extended Bernstein case). To bound P_2 we use Chebyshev's inequality:

$$\begin{aligned}
P_0\left(\prod_i \frac{f_0(X_i)}{b_{B_k}(X_i)} \geq e^{c_1 n\epsilon^2/2}\right) &= P_0\left(\sum_i \log \frac{f_0(X_i)}{b_{B_k}(X_i)} \geq c_1 n\epsilon^2/2\right) \\
&= P_0\left(\sum_i \left[\log \frac{f_0(X_i)}{b_{B_k}(X_i)} - D_k\right] \geq n\left[c_1\epsilon^2/2 - D_k\right]\right) \\
&\leq \frac{n\sigma_k^2}{n^2[c_1\epsilon^2/2 - D_k]^2} \\
&\leq \frac{1}{n}\frac{16\sigma_k^2}{c_1^2\epsilon^4} \to 0. \quad \diamond
\end{aligned}$$

Tenbusch (1995) proposed a Bernstein density estimator defined by $\hat{f}_n^*(\cdot) = b(\cdot; k_n, r_{k_n}/n)$, where $r_k = (r_{1,k}, \ldots, r_{k,k})$ and r_{j,k_n} is the number of data points in the interval $[\frac{j-1}{k}, \frac{j}{k})$, $j = 1, \ldots, k$, and investigates its L_1 convergence properties (in fact, Tenbusch considers the more general case of data in $[0,1]^d$). We compare our density estimates with Tenbusch estimates for fixed sample size in some examples of section 7.

6 Computing the predictive density

In this section we discuss computational issues. Since the Bernstein density is a mixture of beta densities, with unknown number of components, the likelihood, and therefore the posterior, have a complicated analytical expression, and approximation techniques are needed.

Petrone (1999*a,b*) proposed an MCMC algorithm for the Bernstein-Dirichlet prior. Here we remind the basic features of the algorithm and will use it in the next section for approximating the predictive density (1). The results are satisfactory, however the algorithm can be slow, especially if the sample size is large. Therefore, we will also consider a different asymptotic approximation, in which the maximum likelihood density estimates from each model, i.e. for each k, are averaged with respect to data based weights, computed from BIC (or AIC). This can be regarded either as an approximate Bayes estimate or as a frequentist method that averages over models rather than selecting one model. A different approximation of the Bernstein density estimate, based on a histogram approximation of the likelihood, has been recently proposed by Wu and Wells (1999).

The MCMC algorithm proposed by Petrone (1999*a, b*) makes use of auxiliary random variables which have the role of hidden labels. The way in which hidden labels are introduced is not standard, and is based instead on the possibility of regarding the Bernstein model, conditionally on k, as a mixture of kernel densities with a random mixing distribution. For simplicity, we will restrict our attention to the Bernstein-Dirichlet prior. It can be shown that the Bernstein-Dirichlet prior $\pi(p(k), \mathcal{D}(MF_0))$, with F_0 absolutely continuous, is equivalent to the following hierarchical model

- $\tilde{k} \sim p(k)$;
- conditionally on k, $\tilde{F}$ is a Dirichlet process with parameter (MF_0), $\tilde{F} \mid k \sim \mathcal{D}(MF_0))$;
- $(Y_1, \ldots, Y_n)$ are auxiliary random variables with values in $(0, 1]$; conditionally on F, $(Y_1, \ldots, Y_n)$ are i.i.d. according to F;
- $X_1, \ldots, X_n \mid k, F, y_1, \ldots, y_n \;\sim \prod_{i=1}^n \phi(x_i|k, y_i)$, where $\phi(x|k, y) = \beta(x; j, k-j+1)$ if $y \in (\frac{j-1}{k}, \frac{j}{k}], j = 1, \ldots, k$.

It follows that

$$X_1, \ldots, X_n \mid k, F \;\sim\; \prod_{i=1}^n \int \phi(x_i|k, y) dF(y)$$

i.e., conditionally on k the model is a mixture of the kernel densities $\phi(\cdot|k, y)$, with a Dirichlet process mixing distribution. Since $\phi(\cdot|k, y)$ is a piece-wise constant function of y, we have

$$\int \phi(x_i \mid k, y) dF(y) = \sum_{j=1}^k w_{j,k} \beta(x_i; \theta(k, y_i), k - \theta(k, y_i) + 1) = b(x_i; k, w_k) \tag{6}$$

where $\theta(k, y) = j$ if $y \in (\frac{j-1}{k}, \frac{j}{k}], j = 1, \ldots, k$. Note that this formulation allows a comparison with other Bayesian density estimates, based on a Dirichlet process convolution of kernel densities (see e.g. Lo, 1984). Recently, Ghosal, Ghosh and Ramamoorthi (1999*b*) discussed consistency of Dirichlet process mixtures with a Gaussian kernel, $N(\cdot; \theta, \sigma^2)$, where θ is a sample from a Dirichlet process and the variance σ^2 (which regulates the smoothing) is given a prior density. Expression (6) writes the Bernstein model, conditionally on k, as a Dirichlet mixture with a piece-wise constant kernel; the role of smoothing parameter is played by k.

It can be easily shown that

$$\tilde{w}_k \mid k, y_1, \ldots, y_n, x_1, \ldots, x_n \ \sim D(\alpha_k + N_k)$$

where $D(a)$ denotes the Dirichlet distribution with parameter a, α_k depends on the prior specification of the Dirichlet process $\mathcal{D}(MF_0)$, being $\alpha_k = (M\xi_{1,k}, \ldots, M\xi_{k,k})$ with $\xi_{j,k} = E(W_{j,k}) = F_0(\frac{j}{k}) - F_0(\frac{j-1}{k})$ and $N_k = (N_{1,k}, \ldots, N_{k,k})$ is defined as a function of the auxiliary r.v.'s $(Y_1, \ldots, Y_n)$, as $N_{j,k}$ =number of $(Y_1, \ldots, Y_n) \in (\frac{j-1}{k}, \frac{j}{k}]$. It follows that

$$E(\tilde{w}_{j,k} \mid k, x_1, \ldots, x_n) = \frac{M}{M+n}\xi_{j,k} + \frac{1}{M+n}E(N_{j,k} \mid k, x_1, \ldots, x_n) \ \ j = 1, \ldots, k\,, \tag{7}$$

and the predictive density (1) can be written as

$$\sum_{k=1}^{\infty} [\frac{M}{M+n}b(x; k, \xi_k) + \frac{n}{M+n}b(x; k, E(N_k/n \mid k, x_1, \ldots, x_n))]\ p(k \mid x_1, \ldots, x_n) \tag{8}$$

where $\xi_k = (\xi_{1,k}, \ldots, \xi_{k,k})$.

A Monte Carlo sample from the posterior of $(\tilde{k}, \tilde{F}, Y_1, \ldots, Y_n)$ can be obtained by sampling from the posterior distribution of $(\tilde{k}, Y_1, \ldots, Y_n) \mid x_1, \ldots, x_n$ and then sampling from the conditional distribution of $\tilde{F} \mid k, Y_1, \ldots, Y_n, x_1, \ldots, x_n$. The latter step is easy; in fact, conditionally on k, it suffices to sample the increments w_k of F, and their conditional distribution is the Dirichlet distribution $D(\alpha_k + N_k)$ above. The first step is more complicated, but it can be obtained by Gibbs sampling. Details are provided in Petrone (1999*a*,*b*). An MCMC sample $(k^{(s)}, \tilde{w}^{(s)}_{k^{(s)}}, s = 1, 2, \ldots, S)$ from the posterior distribution of $(\tilde{k}, \tilde{w}_k)$ is then obtained by disregarding the sampled y values, and the predictive density (1) can be estimated by

$$\hat{b}(x) = \frac{1}{S}\ \sum_{s=1}^{S}\sum_{j=1}^{k^{(s)}} w^{(s)}_{j,k^{(s)}}\beta(x; j, k^{(s)} - j + 1). \tag{9}$$

A posterior confidence band can be approximated by considering empirical quantiles.

We also study a different approximation of the Bernstein estimate (1), which consists in two steps. First, we compute the maximum likelihood estimate $\hat{w}_k$ of the weights $\tilde{w}_k$ for each k; these can be very easily obtained by the EM algorithm, since the Bernstein density is a mixture with no unknown parameters except the mixing weights. Then the estimator is defined to be

$$f^*(\cdot) = \sum_{k=1}^{\infty} \gamma_k\, b(\cdot\ ;\ k, \hat{w}_k) \tag{10}$$

where $\gamma_k = a_k / \sum_{j=1}^{\infty} a_j$ and

$$\log a_k = \sum_{i=1}^{n} \log b(x_i; k, \hat{w}_k) - \ell_k d_k. \tag{11}$$

Here, d_k is the dimension of the parameter space ($d_k = (k-1)$) and ℓ_k is a sequence of constants. Choosing $\ell_k = 1$ gives the familiar AIC and $\ell_k = (\log n)/2$ gives BIC. The motivation for such an estimator is two-fold. First, in finite dimensional models, it is known (see Kass and Wasserman, 1995) that BIC gives an approximation to the posterior. However, for non parametric cases, it is not known what ℓ_k gives a valid approximation to the posterior.

From the frequentist point of view, this estimator is motivated as follows. Let $\hat{k}$ maximize (11) and define $\hat{f}(\cdot) = b(\cdot; \hat{k}, \hat{w}_{\hat{k}})$. Barron and Yang (1995) and Barron, Birge and Masart (1998) have shown that, if the ℓ_k's are sufficiently large, then

$$E({d_H}^2(f_0, \hat{f})) \leq c \inf_k \left\{ \inf_{f \in \mathcal{B}_k} K(f_0, f) + \frac{\ell_k d_k}{n} \right\}$$

for some $c > 0$. In the above expression, $K(\cdot,\cdot)$ is Kullback-Leibler divergence, $d_H(\cdot,\cdot)$ is Hellinger distance and $\mathcal{B}_k$ is (in our case) the set of Bernstein densities of order k. The right hand side is known to give optimal or near optimal risks in some cases.

However, there are pitfalls to choosing a single k. First, there may be values of k different than $\hat{k}$ which come close to maximizing (11) but which lead to very different estimators. Second, the true density might be spatially inhomogeneous in which case we would need to choose a very large k to fit the density. One possible solution to both problems is to average the estimators and this motivates the hybrid estimator. No theory exists yet for such estimators so we study the estimator numerically.

7 Examples

We present four examples, with simulated data from the following densities (which are similar, respectively, to densities 1, 8, 3 and 4 in Marron and Wand, 1992). The prior we use for $\tilde{k}$, in all examples, is truncated to a maximum value k_{max}.

Example 1. Simulated data from a beta density with parameters $(8, 8)$, $\beta(\cdot; 8, 8)$. This is an easy example; the beta density $\beta(\cdot; 8, 8)$ is a Bernstein density with parameters $k = 15$ and $w_{j,k} = 1$ for $(j = 8, k = 15)$, and zero otherwise. In general, it is difficult to study analytically the asymptotic behavior of the posterior of $(\tilde{k}, \tilde{w}_k)$, due to its complicated expression, but also to the possible lack of identifiability; indeed, our results on consistency concern the posterior distribution of the random density $b(\cdot; \tilde{k}, \tilde{w}_k)$, not of $(\tilde{k}, \tilde{w}_k)$ themselves. In this simple example, however, we expect that the posterior of $\tilde{k}$ cumulates on the true value $k = 15$ and the distribution of the weights, conditionally on $k = 15$ and the data, asymptotically selects only the 8-th beta component of the mixture.

Example 2. Simulated data from a mixture of two beta densities, $0.5\beta(\cdot; 5, 6) + 0.5\beta(\cdot; 40, 10)$. This is a bimodal density. The Bernstein estimate should be able to adapt to the inhomogeneous data, since it averages on different values of k, therefore on varying levels of resolution.

Example 3. Data from an exponential density with parameter 8, truncated at 1. This is not a Bernstein density, and it is strongly asymmetric, so generally difficult to estimate. However, the first of the beta basis functions, $\beta(\cdot; 1, k)$, have a similar shape.

Example 4. Simulated data from a highly peaked density, obtained as the mixture $0.5\beta(\cdot; 4, 4) + 0.5\beta(\cdot; 150, 150)$.

We simulated two data sets from each of the above densities, with sample size $n = 50$ and $n = 300$ respectively. For each example, we show the MCMC approximation (9) of the Bernstein estimate, together with the 0.05 and 0.95 pointwise empirical quantiles, for giving account of posterior uncertainty. The starting point of the Gibbs sampling algorithm in all examples is $(y_1^{(0)}, \ldots, y_n^{(0)}) = (x_1, \ldots, x_n)$. Roughly speaking, this choice corresponds to estimate $E(N_{j,k} \mid k, x_1, \ldots, x_n)$ in (8) with the number $r_{j,k}$ of data points $(x_1, \ldots, x_n)$ which fall in the jth interval $(\frac{j-1}{k}, \frac{j}{k}]$. In a similar fashion, we fixed $(r_{j,k}/n, j = 1, \ldots, k)$ as the starting value of the EM algorithm for computing the maximum likelihood estimate of w_k.

As prior parameters, we chose a uniform prior on $\{1, 2, \ldots, k_{max}\}$ for k, ($k_{max} = 100$ in the first three examples, $= 300$ in the last example), a prior guess F_0 uniform on $[0, 1]$ and $M = 1$, in all examples. From expression (8) of the Bernstein density estimate, as a weighted average of the prior guess $b(x; k, \underline{\xi}_k)$ and the empirical component $b(x; k, E(\underline{N}_k \mid k, x_1, \ldots, x_n))$, it follows that M has the role of weight of prior opinion, therefore the choice $M = 1$ can be considered as vaguely informative. However, it must be noted that the choice of M also has effect on the posterior of k, small values of M making the posterior more concentrated on small values of k; sensitivity of the posterior to the value of M and to the prior on k is also investigated in Petrone (1999a).

The results for the four examples are shown in figures 1 to 7. The dotted line in all plots is the true density from which the data have been generated. The first plot in each figure shows the MCMC estimate of the predictive density; below (first plot in the second row) it is shown a posterior quantile band. The Bernstein estimates are fairly satisfactory, and the posterior quantile band shrinks around the true density as the sample size increases, as expected from consistency.

The second and third plot in the first row of figures 1 to 7 shows the hybrid estimators. We show two curves, the maximum BIC (respectively AIC) estimator and the averaged hybrid estimator. The AIC estimator seems to perform better than the BIC estimator; however the hybrid estimate results less satisfactory than the MCMC Bernstein estimate even with moderately large sample size ($n = 300$).

The last three plots in figures 1 to 7 show the MCMC estimate of the posterior of k (the empirical histogram of the MCMC sample of k values), the normalized BIC and the normalized AIC weights. The examples suggest that BIC and AIC underestimate the posterior probability of k.

In the first example, the posterior of k cumulates around the true value $k = 15$ as the sample size increases, as conjectured. Figure 8 shows four densities sampled from the posterior; the dotted lines are the weighted beta components of the Bernstein mixture. As expected, there are many zero coefficients $w_{j,k}$, and $w_{j,k} > 0$ only for j around the value 8.

In the third example, the posterior of k cumulates around the value $k = 8$ and the posterior of $w_{k=8}$ concentrates on vectors whose first component is close to 1.

The last example is more difficult. We only show the results for $n = 300$; the Bernstein estimate capture the peak nicely, however it is too wiggly outside the mode.

For a rough comparison, the last plot in the first row of figures 1 to 7 shows the kernel estimate, with Gaussian kernel and bandwidth given by $1.06n^{-1/5}min(S, Q/1.34)$, where S is the data standard deviation and Q is the interquartile range (see e.g. Silverman, 1986).

Finally, we provide a comparison of our Bernstein density estimates with the frequentist Bernstein density estimator of Vitale (1975) and Tenbusch (1995), in some of the previous examples. For a given k, the Vitale-Tenbusch estimator is defined as $b(x; k, r_k/n)$, where $r_k/n = (r_{1,k}/n, \ldots, r_{k,k}/n)$ is the vector of the relative frequencies $r_{j,k}/n$ of data points $(x_1, \ldots, x_n)$ which fall in the jth interval $(\frac{j-1}{k}, \frac{j}{k}]$. We consider the hybrid estimator

$$f^{**}(\cdot) = \sum_{k=1}^{\infty} \gamma_k \, b(\cdot\ ;\ k, r_k/n)$$

where the weights γ_k are defined by BIC or AIC as in (10). We compare the MLE hybrid estimator (10) with the Vitale-Tenbusch estimator f^{**} in examples 2 and 4 (sample size = 300), with the weights γ_k computed by the normalized AIC values. The results are shown in figures 9 and 10; the second row of the figures shows the normalized BIC and AIC values, when $w_{j,k}$ is estimates with the frequency $r_{j,k}$ rather than the MLE. As expected for the fairly large sample size, the results are quite similar, however the Vitale-Tenbusch estimator has the advantage of being very quick to compute.

8 Final remarks

We have proved, under weak assumptions on the prior parameters, that the posterior from a Bernstein prior is consistent for continuous and bounded densities on the unit interval, in weak and strong sense; it follows that the predictive density, which is a natural Bayesian density estimator, converges in Hellinger sense to the true density.

We also studied consistency for the maximum likelihood sieve estimator $\hat{f}_n(\cdot) = b(\cdot; k_n, \hat{w}_{k_n})$ (section 5), under assumptions on the sequence $\{k_n\}$. Besides the choice of the MLE $\hat{w}_{k_n}$ rather than the conditional posterior expectation of the weights $\tilde{w}_k$, the difference between $\hat{f}_n(\cdot)$ and the Bayesian estimator given by the predictive density is that, in the first case, the smoothing parameter k_n has to be chosen according so some criteria, while in the latter case, k_n is regulated by its posterior probability. Unfortunately, we have no results on the asymptotic behavior of the posterior probability of $\tilde{k}$, that would allow a clearer comparison between the Bayesian and the maximum likelihood estimators.

The Bernstein prior is defined for data on $[0, 1]$, but extensions to more general sample spaces are possible. The idea is to define a prior as $\pi = \sum_{k=1}^{\infty} \lambda_k \pi_k$, see (3). For the Bernstein prior, π_k is supported on the set of Bernstein densities with parameter k (mixtures of k beta densities); for a general sample space $\mathcal{X} \subseteq \mathbf{R}^d$, π_k can be supported on a set $\mathcal{F}_k$ of densities of the form $f(x) = \sum_{j=1}^{k} w_{j,k} f_{j,k}(x)$, where $f_{j,k}$ are known densities on $\mathcal{X}$. An example of such a prior is the generalized Bernstein prior for data in the d-dimensional hypercube $[0, 1]^d$ (Epifani, 1999), where $f_{j,k}$ is the product of known beta densities; other examples are in Petrone and Veronese (1999). The mixture densities considered by these authors maintain many of the approximation properties of Bernstein densities. This suggests that our study on consistency for the Bernstein prior might be useful for exploring consistency for these more general sieve priors. Techniques of the kind used by Shen and Wasserman (1998) and Ghosal, Ghosh and Van der Vaart (1998) might be used for establishing convergence rates.

A direction of further study is about reformulating our asymptotic results as finite-sample ones, such as in Diaconis and Freedman (1990), where inequalities replace the asymptotics and observed frequencies stand for the true parameter. For example, it would be interesting to establish an inequality for the posterior probability around the empirical d.f., or showing that a normalized version of the random d.f. $B(\cdot; \tilde{k}, \tilde{F})$ converges to a Gaussian process.

APPENDIX

Kullback-Leibler divergence is $\mathcal{S}^*$-measurable. We want to show that the map

$$K : P \to \int f_0(x) log \frac{f_0(x)}{f_P(x)} dx$$

from Δ^* to $\Re$ is $\mathcal{S}^*$-measurable, where the integral is defined to be infinity if f_P is zero on a set of positive f_0 probability, and the integrand is defined to be zero whenever f_0 is zero.

First, it is easy to show that the sufficient conditions (V.5.58) in Dellacherie and Meyer (1982, p.52) hold, so that for any $P \in \Delta^*$, the density f_P can be chosen $(\mathcal{S} \times \mathcal{B}([0, 1]))$-measurable. Therefore, the function h defined by

$$h(P, x) = \begin{cases} 0 & f_0(x) = 0 \\ +\infty & f_P(x) = 0, f_0(x) > 0 \\ f_0(x) log \frac{f_0(x)}{f_P(x)} & otherwise \end{cases}$$

is $(\mathcal{S} \times \mathcal{B}([0,1]))$-measurable, too. Now, we have

$$\int f_0(x) log \frac{f_0(x)}{f_P(x)} dx = \int_{x: f_0(x)>0} f_0(x) log \frac{f_0(x)}{f_P(x)} dx = \int h(P,x) dx$$

and the map $K: P \to \int h(P,x)dx$ is $\mathcal{S}^*$-measurable, from Fubini Theorem. $\diamond$

Analogously, it can be shown that the Hellinger distance $d_H(P_0, \cdot)$ is $\mathcal{S}^*$-measurable.

Acknowledgements. We wish to thank Patrizia Berti, R.V. Ramamoorthi and Pietro Rigo for very helpful suggestions, and Christian Prinoth for computing support. Sonia Petrone was in part supported by grants from M.U.R.S.T., Roma, and F.A.R., Università degli Studi dell'Insubria. Larry Wasserman was supported by NSF grant 9803433. This work was partially written when Sonia Petrone was visiting the Department of Statistics at Cornell University.

References

[1] Altomare, F. and Campiti, M. (1994). *Korovkin-type approximation theory and its application.* W. de Gruyter, Berlin.

[2] Barron, A. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Unpublished manuscript.

[3] Barron, A. and Yang, Y. An asymptotic property of Model Selection criteria. *IEEE Transactions on Information Theory*, **44**, 95-116.

[4] Barron, A., Birge, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, **113**, 301-413.

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

[6] Berk, R.H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Ann. Math. Statist.*, **37**, 51-58.

[7] Berk, R.H. (1970). Consistency of a posteriori. *Ann. Math. Statist.*, **41**, 894-906.

[8] Bunke, O. and Milhaud, X. (1998). Asymptotic behavior of Bayes estimates under posibly incorrect models. *Ann. Statist.*, **26**, 617-644.

[9] Dellacherie, C. and Meyer, P.-A. (1982).*Probabilities and potentials (B)*, North Holland, Amsterdam.

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

[11] Diaconis, P. and Freedman, D. A. (1990). On the uniform consistency of Bayes estimates for multinomial probabilities *Ann. Statist.*, **18**, 1317-1327.

[12] Diaconis, P. and Freedman, D. A. (1993). Nonparametric binary regression: a Bayesian approach *Ann. Statist.*, **21**, 2108-2137.

[13] Epifani, I. (1999). Some results on the probability law of random Bernstein polynomials and applications to Bayesian Statistics (in italian). *PhD Dissertation*, Dottorato di ricerca in Statistica Metodologica - XI ciclo, Trento.

[14] Gaudart, M. and Hadwin (1989). Sigma algebras on spaces of probability measures. *Scand. J. Statist.*, **16**, 167-175.

[15] Genovese, C. and Wasserman, L. (1998). Convergence Rates for the Gaussian Mixture Sieve *Tech. report*, Dept of Statistics, Carnegie Mellon University.

[16] Ghosal, S. Ghosh, J.K. and Ramamoorthi, R.V. (1998). Consistency issues in Bayesian nonparametrics. *Asymptotics, Nonparametrics and Times Series: A Tribute to Madan Lal Puri* (eds: Subir Gosh). Marcel Dekker.

[17] Ghosal, S. Ghosh, J.K. and Ramamoorthi, R.V. (1999*a*). Consistent semi-parametric Bayesian inference about a location parameter. *Journal of Statistical Planning and Inference*,**77**, 181-193.

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

[19] Ghosal, S. Ghosh, J.K. and van der Vaart, A.W. (1998) Convergence rates of posterior distributions. *Tech. Rep. WS-504*, Vrije Universiteit, Amsterdam.

[20] Kass, R.E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Stat. Assoc.*, **90**, 928-934.

[21] Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. Density estimates. *Ann. Statist.*, **12**, 351-357.

[22] Marron, J.S. and Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.*, **20**, 712-736.

[23] Petrone, S. (1999a). Random Bernstein polynomials. *Scand. J. Statist.*, **26**, 373-393.

[24] Petrone, S. (1999b). Bayesian density estimation using Bernstein polynomials. *Cand. J. Statist.*, **27**, 105-126.

[25] Petrone, S. and Veronese, P. (1999). An exponential family random scheme for nonparametric priors. (preliminary draft).

[26] Schwartz, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4**, 10-26.

[27] Shen, X. and Wasserman, L. (1998). Rates of convergence of posterior distributions.*Tech. report*, Dept of Statistics, Carnegie Mellon University.

[28] Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* Chapman and Hall, London.

[29] Tenbusch, A. (1995). *Nonparametric curve estimation with Bernstein estimates*, Universitätsverlag Rasch. Osnabrück, Germany.

[30] Van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge University Press.

[31] Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak convergence and Empirical processes.* Springer, New York.

[32] Vitale, R. A. (1975). A Bernstein polynomial approach to density estimation. *Statistical Inference and Related Topics*, Madan Lal Puri (ed.), Vol 2, Academic Press Inc. New York, 87-100.

[33] Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In: *Practical Nonparametric and Semiparametric Bayesian Statistics.* (D. Dey, P. Müller and D. Sinha eds.) Springer: New York.

[34] Wong, W.H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLE's. *Ann. Statist.*,**23**,339-362.

[35] Wu, S.S. and Wells, M.T. (1999). Nonparametric Bayesian density and hazard rate estimation using Bernstein polynomials (preliminary draft).

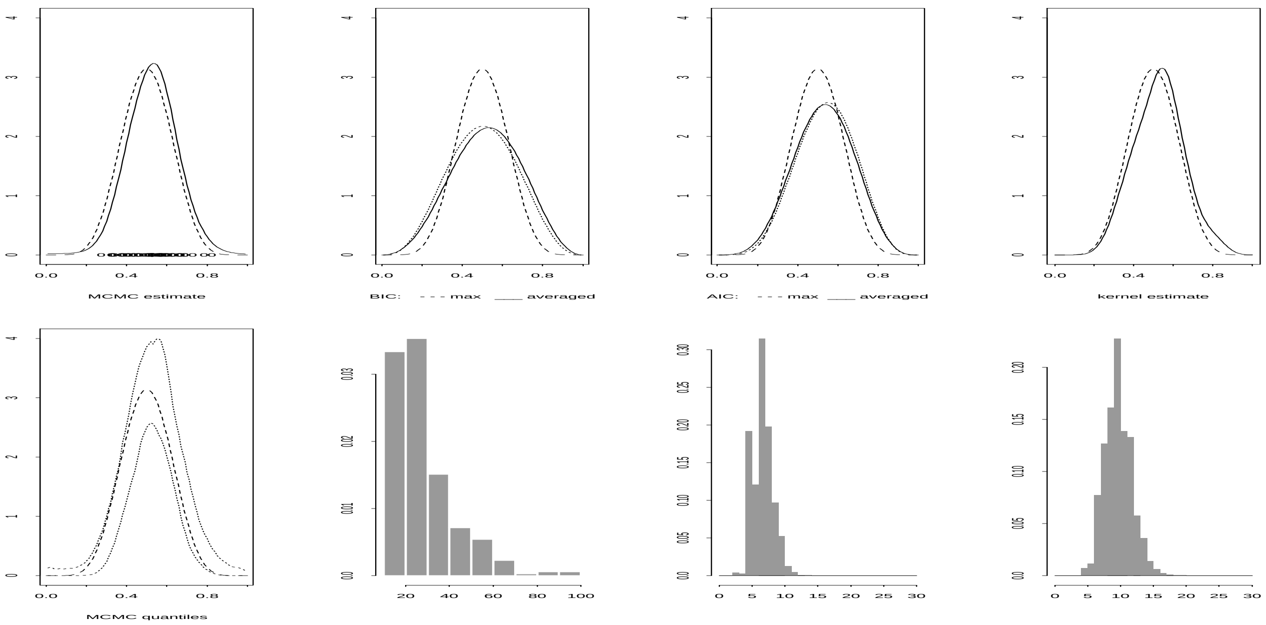

Figure 1: Example 1: simulated data from a $beta(8,8), n = 50$.

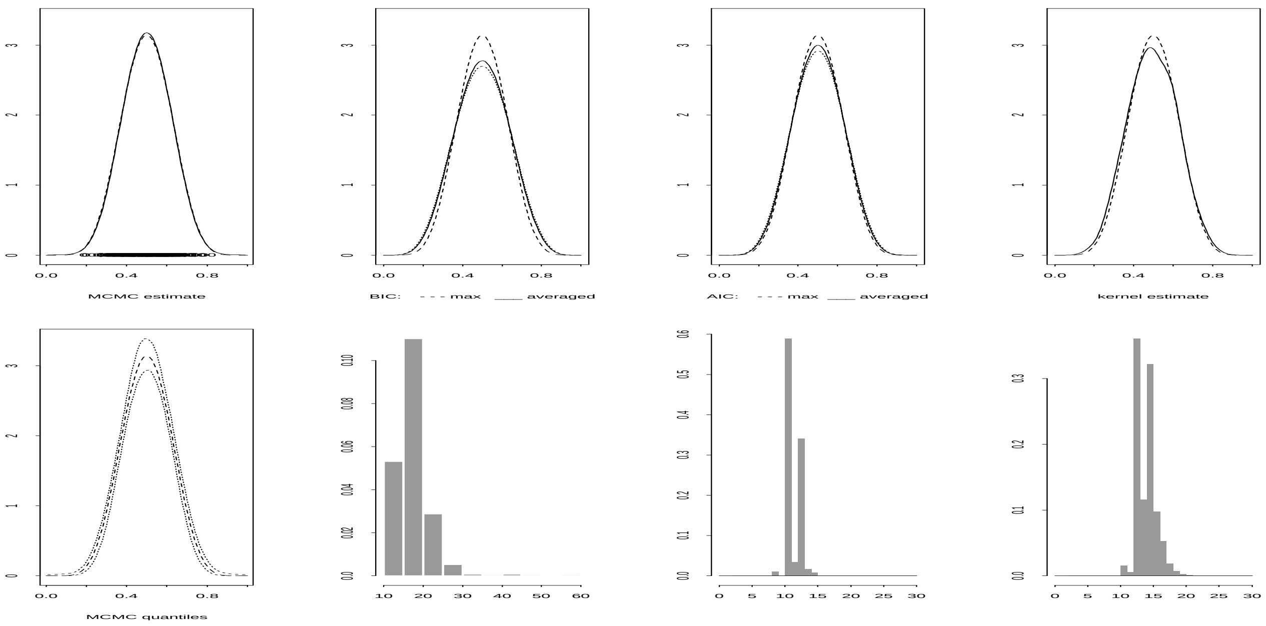

Figure 2: Example 1: simulated data from a $beta(8,8), n = 300$.

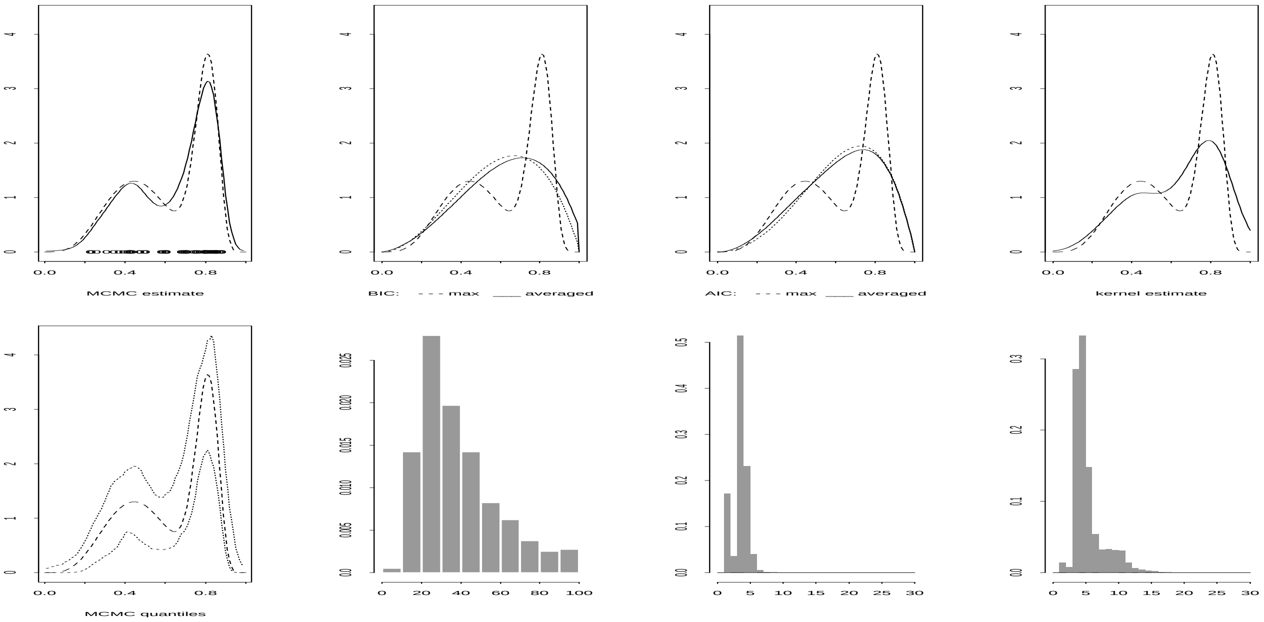

Figure 3: Example 2: simulated data from $0.5beta(5,6) + 0.5beta(40,10), n = 50$.

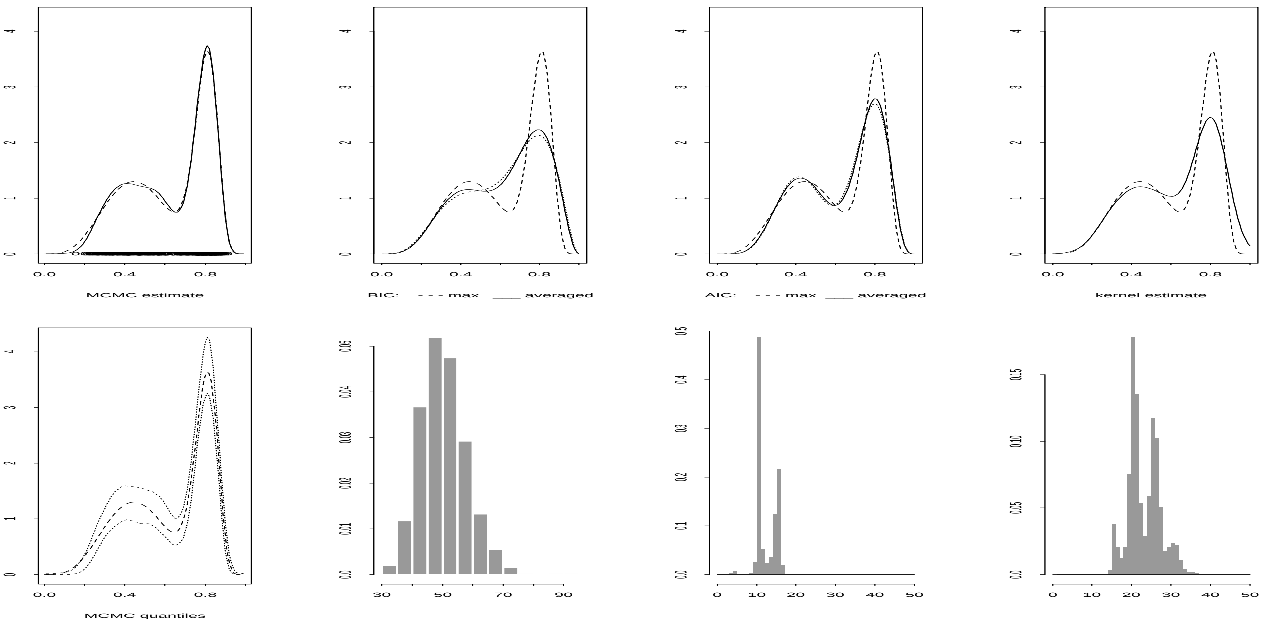

Figure 4: Example 2: simulated data from $0.5beta(5,6) + 0.5beta(40,10), n = 300$.

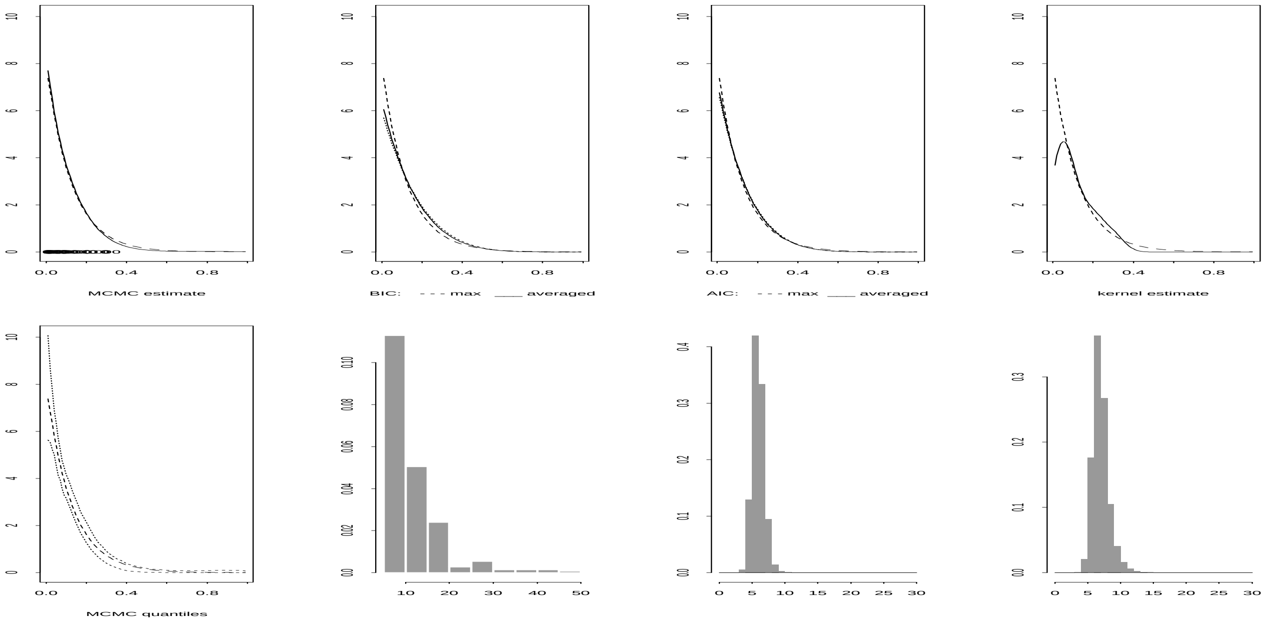

Figure 5: Example 3: simulated data from an $exponential(8), n = 50$.

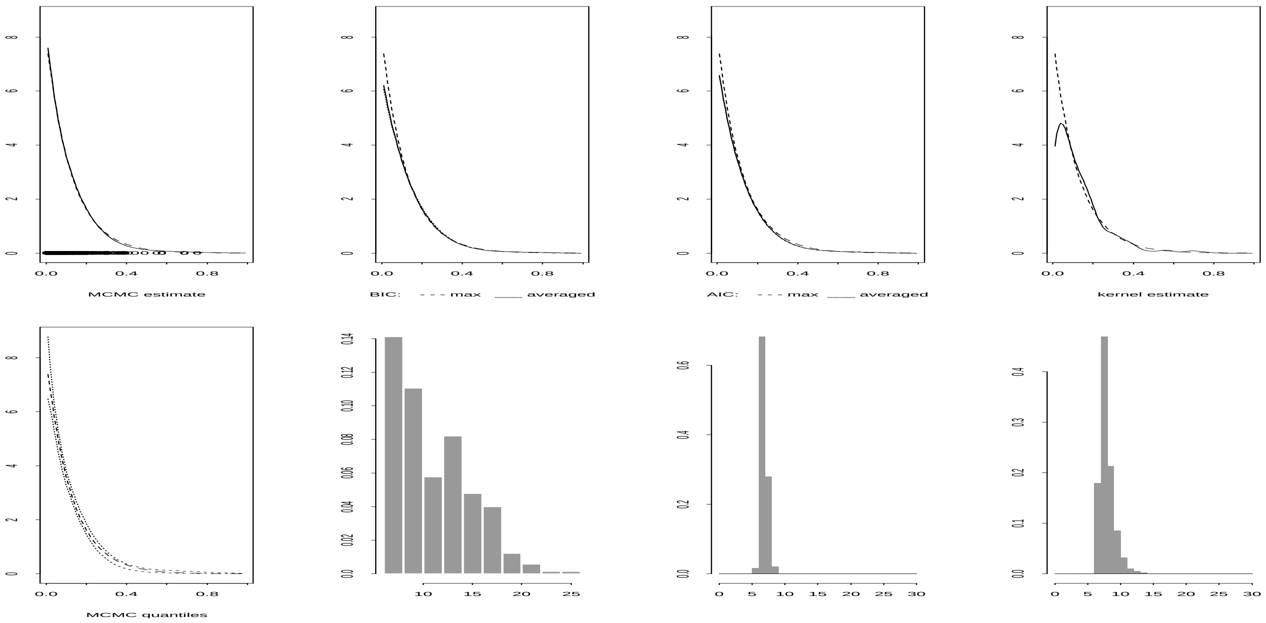

Figure 6: Example 3: simulated data from $exponential(8), n = 300$.

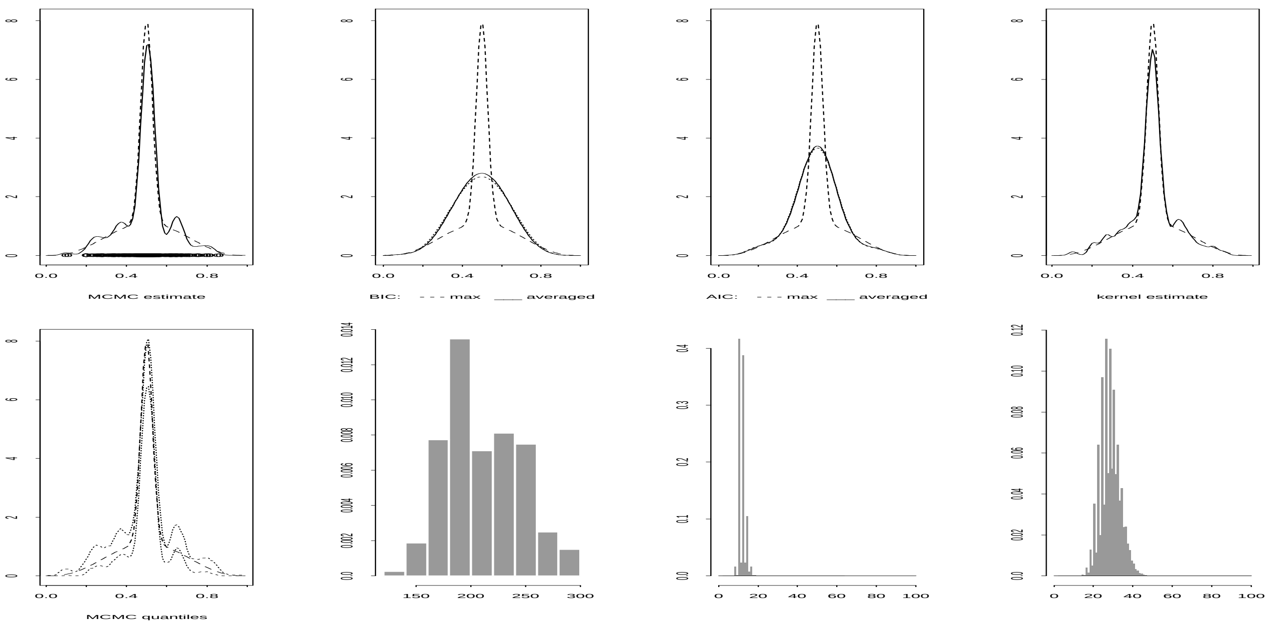

Figure 7: Example 4: simulated data from $0.5beta(4,4) + 0.5beta(150,150), n = 300$.

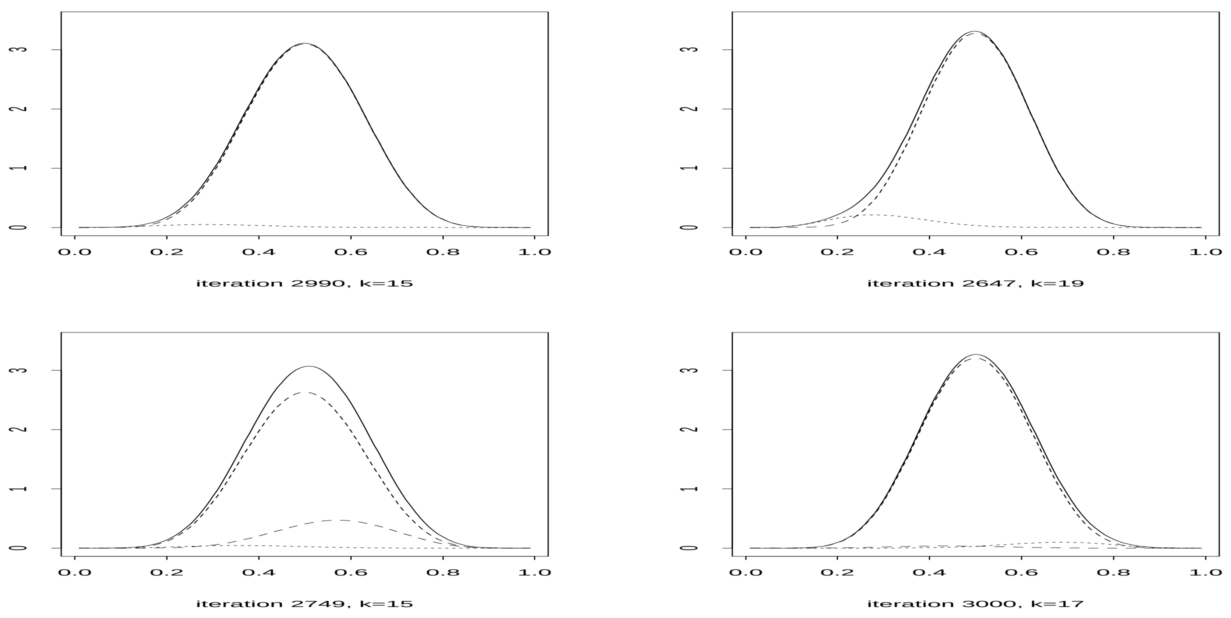

Figure 8: Example 1: sampled Bernstein densities from the posterior. The dotted lines are the weighted beta components of the mixture.

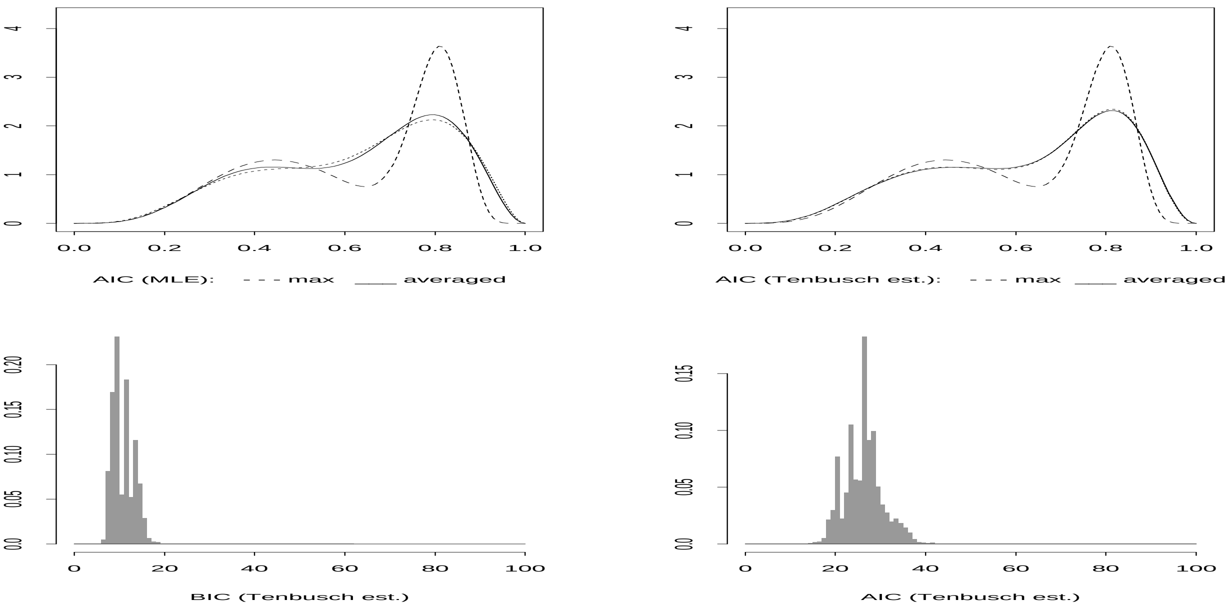

Figure 9: Example 2 (sample size= 300). Hybrid MLE and Vitale-Tenbusch estimates.

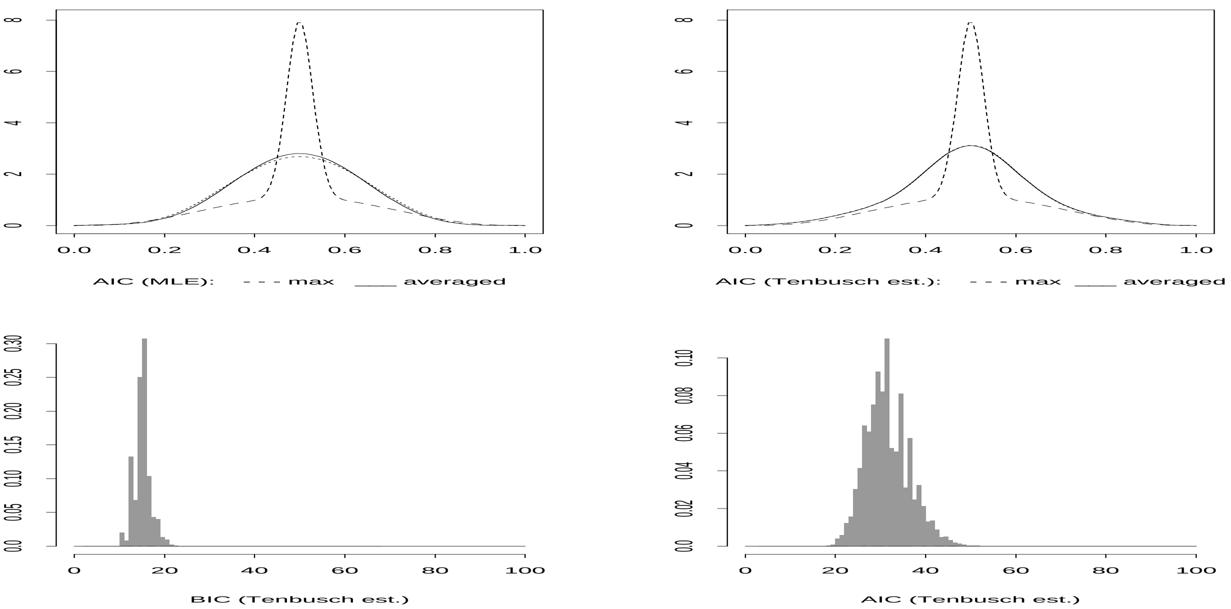

Figure 10: Example 4 (sample size= 300). Hybrid MLE and Vitale-Tenbusch estimates.