Ann Inst Stat Math (2011) 63:481–496
DOI 10.1007/s10463-009-0227-3

Extended Bernstein prior via reinforced urn processes

Lorenzo Trippa · Paolo Bulla · Sonia Petrone

Received: 26 July 2007 / Revised: 10 November 2008 / Published online: 25 April 2009

Abstract A reinforced urn process, which induces a prior on the space of mixtures of Bernstein distributions is introduced. A nonparametric Bayesian model based on this prior is presented: the elicitation is treated and some connections with Dirichlet mixtures are given. In the last part of the article, an MCMC algorithm to compute the predictive distribution is discussed.

Keywords Bayesian nonparametrics · Bernstein polynomials · Polya urn schemes

1 Introduction

The Bernstein polynomial of degree k associated with a bounded function F on [0, 1] is defined as

$$B(x; k, F) = \sum_{j=0}^{k} F\left(\frac{j}{k}\right)\binom{k}{j} x^j (1-x)^{k-j}, \quad x \in [0, 1]. \tag{1}$$

It is known that Bernstein polynomials well approximate F under general assumptions. If x is a continuity point of F: $\lim_{k\to\infty} B(x; k, F) = F(x)$. When F is a distribution function (df) the Bernstein approximation is also a df on [0, 1] and,

L. Trippa · P. Bulla (✉) · S. Petrone
Università Bocconi, Milan, Italy
e-mail: paolo.bulla@unibocconi.it

L. Trippa
e-mail: lorenzo.trippa@phd.unibocconi.it

S. Petrone
e-mail: sonia.petrone@unibocconi.it

if $F(0) = 0$, the polynomial $B(x; k, F)$ can be expressed as a mixture of beta distributions. Under these hypotheses the derivative of the Bernstein polynomial is $B'(\cdot\,; k, F) = \sum_{j=1}^{k} w_{jk}\beta(\cdot, j, k-j+1)$, where $\beta(\cdot, a, b)$ denotes the beta density with parameters (a, b) and $w_{jk} = F(\frac{j}{k}) - F(\frac{j-1}{k})$.

Petrone (1999) applied Bernstein polynomials in Bayesian inference: a nonparametric prior on the set of absolutely continuous distributions on the unit interval is expressed by specifying a probability measure on the set of df's which belong to the Bernstein polynomials' space. Therein is studied the random df $B(x; K, F) = I_{(1,\infty)}(x) + \sum_{j=0}^{K} F(\frac{j}{K})\binom{K}{j}x^j(1-x)^{K-j}I_{[0,1]}(x)$, where F is a Dirichlet process (Ferguson 1973) on the unit interval parameterized with a finite measure α and K has an independent distribution p on the integers. Provided that $\alpha(\{0\}) = 0$, $F(0) = 0$ *a.s.*, and the random Bernstein polynomial is *a.s.* a mixture of beta distributions. In practice, the df F_0 obtained by the normalization of the measure α expresses the initial guess of an unknown df, although $B(x)$ is centered on $\sum_k B(x; k, F_0)p(k)$, which is generally different from F_0. Computational procedures to estimate an unknown df through the described Bayesian model have been proposed by Petrone (1999) and Petrone and Wasserman (2002); however, they assume that the random variable (rv) K has finite support, i.e., the order of the random polynomial is a priori bounded.

We propose an extension of the Bernstein prior that can be easily centered on whichever continuous df on a real interval. Posterior inference can be fairly simply approximated through simulation techniques; in particular, computing the predictive distribution does not require to truncate the order of the polynomial.

The first part of the article is devoted to defining a prior probability measure on the beta mixtures space

$$\mathbf{B} = \left\{ \sum_{\tilde{S}} w_{jk}\beta(j, k-j+1) : \sum_{\tilde{S}} w_{jk} = 1;\ w_{jk} \geq 0,\ \forall (j,k) \in \tilde{S} \right\}, \tag{2}$$

where $\tilde{S} = \{(j,k) : j, k \in \mathbb{N}^+;\ j \leq k\}$. To this aim, we first construct a prior for the mixing weights $W = \{W_{j,k}\}_{(j,k)\in\tilde{S}}$; note that W is a random probability measure on $\tilde{S}$. Then, the prior on the beta mixtures space is obtained by mapping W to $\mathbf{B}$: $W \longrightarrow \sum_{\tilde{S}} W_{j,k}\beta(j, k-j+1)$. The peculiarity of our construction is to define the random mixing distribution W by means of an auxiliary reinforced urn process (RUP). Muliere et al. (2000) give a detailed account of RUP's probabilistic properties. We show that a random probability measure W on $\tilde{S}$, which generalizes the Dirichlet process, can be characterized through a RUP $\{X_n\}_{n\geq 0}$. Indeed, the RUP generates an exchangeable sequence of rv's $\{Y_n\}_{n\geq 1}$ with values in $\tilde{S}$, whose de Finetti measure gives the required probability law of the random measure W.

The above construction defines a prior with large support on the space of absolutely continuous df's on [0, 1]. In the second part of the article, we extend the Bernstein model to construct a prior for a random df on a general real interval. The main features of the extension are (i) the possibility of easily centering the prior on any a priori guess F_0, (ii) the availability of simulation techniques for implementing posterior inference, and (iii) the ability of the model to combine mixture components with remarkable heterogeneous variances. The latter property differentiates the proposed extension from the Bernstein model discussed in Petrone (1999); we illustrate that such peculiarity can determine relevant differences between the predictive distributions corresponding to the two models. Such differences have an intuitive explanation. The beta components of a Bernstein density B' act as kernels, with k having the role of a smoothing parameter. Intuitively, one can think of B' as a smoothing of a histogram with bins of equal length $1/k$, where the rectangular kernels of the histogram are replaced by the beta kernels $\beta(j, k - j + 1)$. For a better local behavior, it is natural to think of histograms with bins of different length, smoothed through the extended Bernstein mixtures in $\mathbf{B}$, where the mixture is taken with respect to the *joint* distribution W of j and k.

The outline of the article is the following. In Sect. 2 some definitions and results about RUPs from Muliere et al. (2000) are reviewed. In Sect. 3, a class of RUPs $\{X_n\}_{n\geq 0}$ which generates a random probability measure W on $\tilde{S}$ is introduced. The extended Bernstein model is presented in Sect. 4. In Sect. 5, an MCMC algorithm for posterior inference and some applications are illustrated. Final remarks and open issues are briefly discussed in the last section.

2 Reinforced urn processes

Reinforced urn processes may be defined very briefly as reinforced random walks on a state space of urns. Indeed, they blend two basic probabilistic models: a Pólya urn scheme and a random walk in such a way as to describe the attitude of a random mover to repeat transitions that had already occurred in the past. Each RUP consists of four basic elements.

Definition 1 Let

1. S be a countable state space;
2. $C = \{c_1, \ldots, c_k\}$ be a finite set of colors of cardinality $k \geq 1$;
3. $U(s) = (a_s(c_1), \ldots, a_s(c_k))$ be an urn composition function which maps S into the set of k-tuples of nonnegative numbers whose sum is strictly positive;
4. $q : S \times C \to S$ be a law of motion such that for every $x, y \in S$, there is at most one color $c(x, y) \in C$ such that $q(x, c(x, y)) = y$.

Fixed $X_0 = x_0 \in S$, for $n \geq 1$ if $X_{n-1} = x \in S$, a ball is sampled from the urn associated with x and, given its color c, we set

$$X_n = q(x, c).$$

Finally, the ball is replaced in the urn along with one of the same color.

The sequence $X = \{X_n,\ n \geq 0\}$ is said to be a RUP with initial state x_0 and parameters S, C, U, q.

Muliere et al. (2000) show that partial exchangeability is a momentous feature enjoyed by the RUPs. Therefore, if a RUP $\{X_n\}_{n\geq 0}$ is also recurrent, a representation theorem of Diaconis and Freedman (1980) states that it is a mixture of Markov chains.

More formally, let $R^{(0)} = \{x_0\}$ and for all $x \in S$, let $R_x = \{y \in S : a_x(c(x, y)) > 0\}$ be the set of all the states attainable from the state x in one step. Then define, for $n \geq 1$, $R^{(n)} = \cup_{x \in R^{(n-1)}} R_x$ and $R = \cup_{n=0}^{\infty} R^{(n)}$, the set of states visited with strictly positive probability by the RUP. The above-cited result assures the existence of a probability measure μ on the set $\mathcal{P}$ of $R \times R$ transition matrices such that for every $n \geq 1$ and $(x_1, \ldots, x_n) \in R^n$,

$$P\,[X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n] = \int_{\mathcal{P}} \prod_{j=0}^{n-1} \pi(x_j, x_{j+1}) \mu(\mathrm{d}\pi). \tag{3}$$

Consider the random matrix Π with distribution μ. Let $\Pi(x)$ be the xth row of Π and $\alpha(x)$ the measure on R which assigns mass $a_x(c)$ to $q(x, c)$ for each $c \in C$ such that $a_x(c) > 0$ and null mass to all the other elements of R. The following theorem describes the measure μ making the point about the properties of the random matrix Π.

Theorem 1 (Muliere et al. 2000) *If the RUP $\{X_n\}_{n \geq 0}$ is recurrent, the rows of Π are mutually independent random probability distributions on R and, for all $x \in R$, the law of $\Pi(x)$ is that of a Dirichlet process with parameter $\alpha(x)$.*

Following Diaconis and Freedman (1980), for a process $\{X_n\}_{n \geq 0}$ on S, a x_0-block is defined to be a finite sequence of states which begins with x_0 and contains no further x_0. Let S^* be the countable space of all finite sequences of elements of S. Under the recurrence hypothesis of a RUP $\{X_n\}_{n \geq 0}$, the sequence of the successive x_0-blocks, which retraces the trajectory of the process, say $B_1, B_2, \ldots$ with $B_n \in S^*$ for every $n \geq 1$, is well defined. A simple example helps to clarify that the trajectory of a recurrent process can be decomposed in x_0-blocks. Let $S = \{0, 1, 2\}$ and $x_0 = 0$; a specific trajectory, say $\{0, 2, 1, 0, 0, 1, 0, \ldots\}$, is constituted by x_0-blocks: $\{B_1 = [0, 2, 1], B_2 = [0], B_3 = [0, 1], \ldots\}$. The block sequence $\{B_n\}_{n \geq 1}$ is well defined because of the recurrence hypothesis: the event $\{\cup_{m=1}^{\infty} \cap_{n \geq m} (X_n \neq x_0)\}$ has null probability.

As illustrated in the following paragraph, the sequence $\{B_n\}_{n \geq 1}$ is exchangeable and, for every measurable function φ from S^* to another space S', the same property characterizes the sequence $\{\varphi(B_n)\}_{n \geq 1}$. In the next section, this property is exploited to define a random probability measure on the countable space $\tilde{S} = \{(i, k) : i, k \in \mathbb{N}^+; i \leq k\}$.

A simple example shows that the exchangeability of the x_0-blocks follows from the representation (3). From the definition of x_0-block it follows that, $\{B_1 = [0, 2], B_2 = [0]\}$ and $\{X_1 = 0, X_2 = 2, X_3 = 0, X_4 = 0\}$ are two representations of the same event; as well, permuting B_1 and B_2, the events $\{B_1 = [0], B_2 = [0, 2]\}$ and $\{X_1 = 0, X_2 = 0, X_3 = 2, X_4 = 0\}$ are equivalent. Moreover, the count of the transitions between states in the two cases is identical: there is one transition from 0 to 0, one from 0 to 2 and one transition from 2 to 0. Thus, $Pr(B_1 = [0, 2], B_2 = [0]) = Pr(B_1 = [0], B_2 = [0, 2])$, since both are obtained from (3) and the integrand functions are the same. The same arguments can be adopted to verify that all the permutations of a finite sequence of x_0-blocks $B_1, \ldots, B_n$ are equally probable.

3 Probability measures on the beta mixtures space

The class of RUPs that we consider to define probability measures on the set of mixtures of beta distributions is parameterized as follows:

1. the state space: $\tilde{S} = \{(i, k) : i, k \in \mathbb{N}^+, i \leq k\}$,
2. the initial state of the process: $\tilde{x}_0 = (1, 1)$,
3. the color space: $\tilde{C} = \{\tilde{c}_1, \tilde{c}_2, \tilde{c}_3\}$,
4. the urn composition function $\tilde{U}(s) = (a_s(\tilde{c}_1), a_s(\tilde{c}_2), a_s(\tilde{c}_3))\ \forall s \in \tilde{S}$, and
5. the law of motion:

$$\tilde{q}((i,k), c) = \begin{cases} (i, k+1) & c = \tilde{c}_1 \\ (i+1, k+1) & c = \tilde{c}_2 \\ (1, 1) & c = \tilde{c}_3 \end{cases} \quad \forall (i,k) \in \tilde{S}.$$

The tilde symbol is used to distinguish the above-described class from the whole RUPs family.

As anticipated, if a RUP $\{\tilde{X}_n\}_{n \geq 0}$ is recurrent, a sequence of exchangeable rv's $\{Y_n\}_{n \geq 1}$ can be easily defined by mapping the x_0-blocks $B_1, B_2, \ldots$ to a measurable space. We consider the following map:

$$\varphi : \left[B_i = (\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_m)\right] \to [Y_i = \tilde{x}_m], \quad i \geq 1. \tag{4}$$

Let $(\Omega, \mathfrak{F}, \mathbf{P})$ be the probability space on which the process $\{\tilde{X}_n\}_{n \geq 0}$ and the random sequence $\{Y_n = \varphi(B_n)\}_{n \geq 1}$ is defined. By de Finetti's representation theorem, the limit of the empirical distributions of the latter sequence, $W = \{W_s = \lim \frac{1}{N} \sum_1^N I_{(s)}(Y_n)\}_{s \in \tilde{S}}$, is well defined on the same space; W is a random probability measure on $\tilde{S}$, with probability law given by the de Finetti measure of the exchangeable sequence $\{Y_n = \varphi(B_n)\}_{n \geq 1}$.

Let us now denote the class of probability measures on the unit interval with $\triangle$, the Borel σ-field generated by the topology of the weak convergence with $\mathcal{H}$ and its restriction to $\mathbf{B}$ with $\mathfrak{B}$. Given the RUP $\{\tilde{X}_n\}_{n \geq 0}$, and the associated random probability measure W, a probability measure on $(\mathbf{B}, \mathfrak{B})$ can be induced in a natural way via a measurable map from Ω to $\mathbf{B}$:

$$\omega \to \sum W_{i,k}(\omega) \beta(i, k - i + 1)\cdot$$

We will still denote such probability measure with $\mathbf{P}$.

From the properties of Bernstein polynomials, it follows easily that the prior $\mathbf{P}$ has full weak support Δ under mild assumptions.

Proposition 1 *If the urn composition function of the RUP $\{\tilde{X}_n\}_{n \geq 0}$ is such that $Pr(\varphi(B_1) = (i, k)) > 0\ \ \forall\, (i, k) \in \tilde{S}$, then the probability measure $\mathbf{P}$ defined above has full weak support Δ.*

The proofs of the results in this section are provided in the Appendix.

The following results show that the urn composition function $\tilde{U}(s)$ can be chosen in such a way that the random probability measure $W = \{W_s\}_{s\in\tilde{S}}$ is a Dirichlet process with parameter G, where G is an arbitrary finite measure on $\tilde{S}$. The following theorem holds for a general RUP.

Theorem 2 *Let $\{X_n\}_{n\geq 0}$ be a recurrent RUP with parameter (S, U, C, q) and initial state x_0. Let φ be the map from $\cup_{m\geq 1} S^m$ to S such that*

$$\varphi : (s_1, s_2, \ldots, s_m) \to s_m \quad \forall\, (s_1, s_2, \ldots, s_m) \in S^m, \quad \forall\, m \geq 1.$$

If

$$\sum_{c\,\in\, C} a_{s^*}(c) = \sum_{s\,\in\, S} a_s(c(s, s^*)) \quad \forall s^* \in S, \tag{5}$$

then the sequence $\{Y_n = \varphi(B_n)\}_{n\geq 1}$ is exchangeable and its associated de Finetti measure is a Dirichlet process with parameter α, where α is the finite measure on S with $\alpha(s) = a_s(c(s, x_0))$, $\forall s \in S$.

An elementary example of RUP that satisfies the condition (5) is the following. Let $S = \{0, 1, 2\}$ and $x_0 = 0$. Consider the urn composition such that the urn associated with the state s contains $(2-s)$ balls of color c_1 and one of color c_2. If $q(s, c_1) = s+1$ and $q(s, c_2) = 0$, then the described RUP parametrization implies that the de Finetti measure associated with $\{Y_n = \varphi(B_n)\}_{n\geq 1}$ is a Dirichlet distribution with parameters $(1, 1, 1)$.

The next proposition states how the urn compositions of the RUP $\{\tilde{X}_n\}_{n\geq 0}$ can be initialized to get a Dirichlet process prior for the sequence of the last states of the x_0-blocks.

Proposition 2 *Let G be a finite measure on $\tilde{S}$. Let $\{\mu_h\}_{h\geq 1}$ be a sequence of measures on $(0, 1]$ with $\mu_h((0, 1] \setminus \mathbb{Q}) = 0$ and*

$$\mu_h(r) = G((i, k) \in \tilde{S} : k \geq h, \frac{i}{k} = r), \quad \forall\, r \in \mathbb{Q}\cdot$$

Let $\{\tilde{X}_n\}_{n\geq 0}$ be the RUP with initial state $x_0 = (1, 1)$ and parameters $(\tilde{S}, \tilde{U}, \tilde{C}, \tilde{q})$ where $\tilde{U}(i, k) = (\mu_{k+1}(\frac{i-1}{k}, \frac{i}{k+1}], \mu_{k+1}(\frac{i}{k+1}, \frac{i}{k}], G(i, k))$. Then, the sequence $\{Y_n = \varphi(B_n)\}_{n\geq 1}$ is exchangeable and its de Finetti measure is a Dirichlet process with parameter G.

4 Extended Bernstein prior

In this section, a Bayesian nonparametric model whose prior distribution can be easily centered is specified through the probability measure **P** defined on the mixtures' space **B**.

We consider an exchangeable sequence $\{Z_i\}_{i\geq 1}$ of continuous real valued rv's; equivalently, $Z_i|F$ are i.i.d. according to F, where F is an absolutely continuous random df. Let F_0 be an arbitrarily chosen continuous df which represents the initial guess on the unknown df F of Z_i. To assign a nonparametric prior on F, centered on F_0, we transform the Z_i's into [0, 1] by letting $\zeta_i = F_0(Z_i)$, and assume that

$$\zeta_i \mid B \overset{i.i.d.}{\sim} B, \quad B \sim \mathbf{P}.$$

Therefore, the probability law of $\{Z_i\}_{i\geq 1}$ is uniquely defined by the sequence of df's

$$P\left(Z_1 \leq z_1, \ldots, Z_n \leq z_n\right) = \int_{\mathbf{B}} \prod_{j=1}^{n} B(F_0(z_i))\mathrm{d}\mathbf{P}(B), \quad n \geq 1.$$

In particular

$$P(Z_i \leq z \mid B) = \int_0^z b(F_0(t))\, F_0'(t)\mathrm{d}t, \tag{6}$$

where b denotes the beta mixture density of B. Thus, the prior models the unknown density as F_0' times a "distortion factor" $b(F_0(\cdot))$.

The law of the sequence $\{Z_i\}_{i\geq 1}$ can be alternatively represented conditionally on the RUP $\tilde{X} = \{\tilde{X}_i\}_{n\geq 0}$ used in the construction of $\mathbf{P}$. In this case:

$$\begin{aligned} P(Z_1 \leq z_1, \ldots, Z_n \leq z_n \mid \tilde{X}) &= \prod_{i=1}^{n} P\left(Z_i \leq z_i \mid \varphi(B_i)\right) \\ P\left(Z_i \leq z \mid \varphi(B_i)\right) &= \int_0^{F_0(z)} \beta\left(x, \varphi^{(1)}(B_i), \varphi^{(2)}(B_i) - \varphi^{(1)}(B_i) + 1\right)\mathrm{d}x, \end{aligned} \tag{7}$$

where $\varphi^{(1)}$ and $\varphi^{(2)}$ are the two components of the φ function. In this perspective, every observation Z_i has an associated latent rv $\varphi(B_i)$.

The introduced prior is centered on F_0 if and only if $E(B(x)) = x$ for every x in [0, 1], so that $P(Z_1 \leq z) = F_0(z)$. The simplest strategy to achieve this equality is constraining the parameter of the RUP $\tilde{X}$ as follows:

$$\frac{a_{i,k}(\tilde{c}_2)}{a_{i,k}(\tilde{c}_1) + a_{i,k}(\tilde{c}_2)} = \frac{i}{k+1} \quad \forall (i,k) \in \tilde{S}. \tag{8}$$

If these conditions are satisfied, the first x_0-block evolves exactly as a Pólya urn with initially two balls of two different colors; more formally $\forall\,(i,k) \in \tilde{S}$:

$$\begin{aligned} P(\tilde{X}_k = (i+1, k+1) | \tilde{X}_{k-1} = (i,k), \tilde{X}_k \neq (1,1)) &= \frac{i}{k+1} \\ P(\tilde{X}_k = (i, k+1) | \tilde{X}_{k-1} = (i,k), \tilde{X}_k \neq (1,1)) &= \frac{k-i+1}{k+1}. \end{aligned}$$

These are the transition probabilities of a Pólya urn containing i white balls and $k-i+1$ black. On the other hand, conditionally on $\{\tilde{X}_{k-1} = (i,k)\}$ and $\{\tilde{X}_k = (1,1)\}$, from

(7) we have that $F_0(Z_1)$ has a beta distribution with parameters $(i, k-i+1)$, which can be viewed as the random limit proportion of white balls in a Pólya urn initially containing i white balls and $k-i+1$ black. In a unified perspective, the law of $F_0(Z_1)$ can be represented as the random limit proportion of white balls in a Pólya urn initially containing 1 white and 1 black, which is well known to be uniformly distributed.

It can be easily shown, through these arguments, that for example if the RUP is parameterized as in Proposition 2, with

$$G(i,k) = M\frac{e^{-\lambda}\lambda^{k-1}}{k!} \quad \forall\,(i,k) \in \tilde{S},\ M > 0,\ \lambda > 0,$$

then the random probability measure $B\,(F_0(\cdot))$ is centered on F_0. More generally, if $P\,(\varphi(B_1) = (i,k))$ is a function of k, the equality $P(Z_1 \leq z) = F_0(z)$ is achieved. If the support of F_0 is a closed interval $[\gamma, \eta]$, it follows from Proposition (1) that the support of the random distribution $B\,(F_0(\cdot))$ is the set of the probability measures on the interval. In this example, the expected values of the random weights $W_{i,k}$ are functions of λ: the higher is λ, the larger are the weights of the random mixture B which correspond to beta components characterized by small variances.

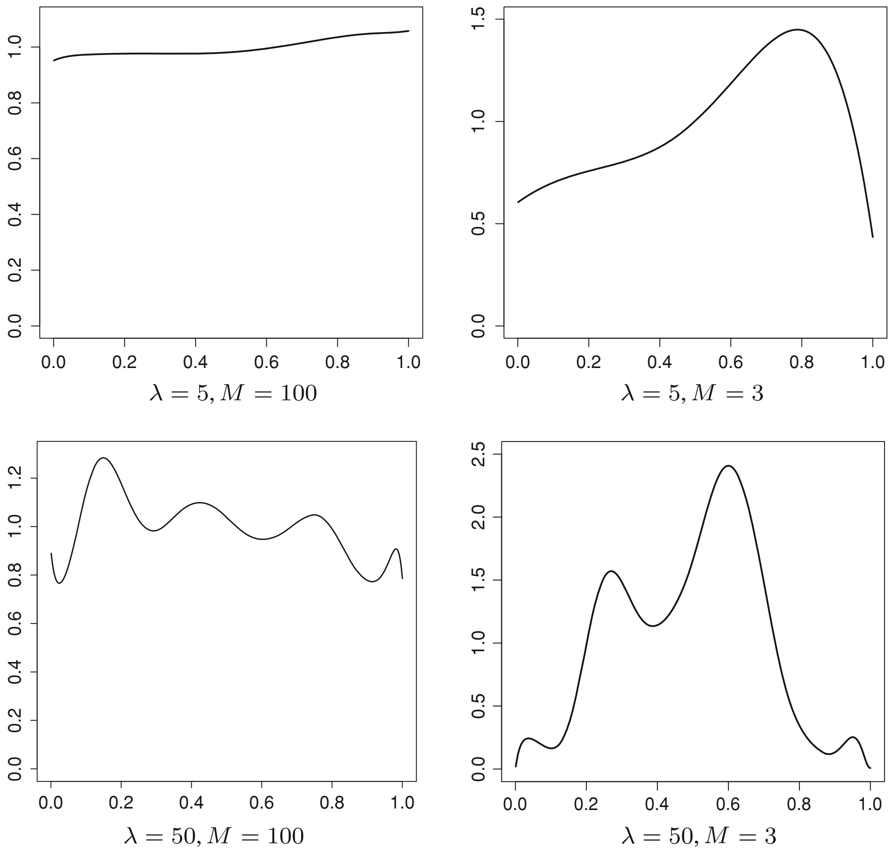
The constant λ can be interpreted as a smoothing parameter. For $\lambda \approx \infty$, given a finite partition $\{A_1, \ldots, A_m\}$ of [0, 1], the random vector $[B(A_1), \ldots, B(A_m)]$ will be approximately Dirichlet distributed, while if $\lambda \approx 0$ it will be approximately equal to the Lebesgue measure of the partition sets. For fixed values of λ, the parameter M is decisive to regulate the variances of the rv's $W_{i,k}$: the lower it is, the higher is the uncertainty on the random probability measure $B(F_0(\cdot))$. Figure 1 represents the density functions of four sampled probability measures B; the comparison of the four graphs emphasizes the interpretation of the parameters M and λ.

5 Inference

The proposed extension of the Bernstein prior allows to define a Gibbs sampling algorithm in order to approximate the predictive distribution of Z_{n+1} conditionally on $Z_1, Z_2, \ldots, Z_n$

$$P\left(Z_{n+1} \leq z | Z_1, \ldots, Z_n\right) = \sum_{\tilde{S}} E\left(W_{i,k} | Z_1, \ldots, Z_n\right) \int_0^{F_0(z)} \beta(x, i, k-i+1)\mathrm{d}x. \tag{9}$$

The algorithm is based on the updating structure proper of the Pólya scheme. If the prior is suitably parameterized, the computational procedure does not require to truncate the number of components of the random mixture $B(F_0(\cdot))$. This is an advantage with respect to the algorithms in literature for the Bernstein model, which bound the order of the random polynomial, see for example Petrone (1999).



$\lambda = 5, M = 100$

$\lambda = 5, M = 3$

$\lambda = 50, M = 100$

$\lambda = 50, M = 3$

Fig. 1 Samples of density functions from differently parameterized prior distributions

An essential description of the algorithm is given as follows. Since, using (7), we have

$$E(W_{i,k} \mid Z_1, \ldots, Z_n) = \int E(W_{i,k} \mid B_1, \ldots, B_n)\, \mathrm{d}P(B_1, \ldots, B_n \mid Z_1, \ldots, Z_n),$$

to compute the predictive distribution (9) it suffices to sample from the posterior distribution of $(B_1, \ldots, B_n | Z_1, \ldots, Z_n)$. To this aim, fix a sequence of x_0-blocks $B_1^1, \ldots, B_n^1$ such that $P([B_1, \ldots, B_n] = [B_1^1, \ldots, B_n^1]) > 0$ and sample iteratively B_l^{j+1}, from the conditional distribution of B_l given $(Z_1, \ldots, Z_n)$ and $(B_1 = B_1^{j+1}, \ldots, B_{l-1} = B_{l-1}^{j+1}, B_{l+1} = B_{l+1}^{j}, \ldots, B_n{=}B_n^j)$. The iterations simulate a Markov chain $\{B_1^j, \ldots, B_n^j\}_{j \geq 1}$ whose stationary distribution is the conditional law of $(B_1, \ldots, B_n)$ given $(Z_1, \ldots, Z_n)$. So, for every couple (i, k) in $\tilde{S}$, $E(W_{i,k} | Z_1, Z_2, \ldots, Z_n)$ can be approximated through $\frac{1}{N} \sum_{j=1}^{N} E(W_{i,k} | B_1{=}B_1^j, \ldots, B_n = B_n^j)$. These conditional expectations can be computed by means of the following equalities:

$$
\begin{aligned}
&E\left(W_{i,k}|B_1, B_2, \ldots, B_n\right)\\
&\quad= \sum_{\varphi(B)=(i,k)} P\left(B_{n+1}=B|B_1, B_2, \ldots, B_n\right)\\
&\qquad \times P\left(\tilde{X}_{m+1}=\tilde{x}_{m+1}, \ldots, \tilde{X}_{m+k}=\tilde{x}_{m+k}|\tilde{X}_0=\tilde{x}_0, \ldots, \tilde{X}_m=\tilde{x}_m\right)\\
&\quad= \prod_{j=0}^{k-1} \frac{a_{\tilde{x}_{m+j}}\left(c\left(\tilde{x}_{m+j}, \tilde{x}_{m+j+1}\right)\right)+\sum_{i=0}^{m+j-1} I_{(\tilde{x}_i,\tilde{x}_{i+1})}\left(\tilde{x}_{m+j}, \tilde{x}_{m+j+1}\right)}{a_{\tilde{x}_{m+j}}(\tilde{c}_1)+a_{\tilde{x}_{m+j}}(\tilde{c}_2)+a_{\tilde{x}_{m+j}}(\tilde{c}_3)+\sum_{i=0}^{m+j-1} I_{(\tilde{x}_i)}(\tilde{x}_{m+j})}. \qquad (10)
\end{aligned}
$$

In two relevant cases it is easy to sample (B_l) conditionally on $(Z_1, \ldots, Z_n)$ and $(B_1, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n)$: when the condition (8) is satisfied and when the random probability measure associated with $\{\varphi(B_n)\}_{n\geq 1}$ is a Dirichlet process.

In the first case, let $\tilde{k}=\max(\varphi^{(2)}(B_j), j \in \{1, \ldots, l-1, l+1, \ldots, n\})$; the same trick adopted to illustrate that condition (8) allows to center the prior distribution on F_0 points out the equality:

$$
\begin{aligned}
&\sum_{\tilde{\mathcal{S}}} E(W_{i,k}|B_1, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n)\beta(\cdot, i, k-i+1)\\
&\quad= \sum_{k<\tilde{k}} \sum_{i=1}^{k} P\left(\varphi(B_l)=(i,k)|B_1, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n\right)\beta\left(\cdot, i, k-i+1\right)\\
&\qquad+\sum_{i=1}^{\tilde{k}} P\left(\varphi(B_l^{\tilde{k}})=(i,\tilde{k})|B_1, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n\right)\beta(\cdot, i, \tilde{k}-i+1), \qquad (11)
\end{aligned}
$$

where $B_l^{\tilde{k}}$ denotes the x_0-block truncated at the $\tilde{k}$th component. The equality (11) allows to sample from the conditional distributions of $\varphi(B_l)$ and B_l given $(Z_1, \ldots, Z_n)$ and $(B_1, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n)$:

$$
\begin{aligned}
&P\left(\varphi(B_l)=(i^*, k^*)|Z_1, \ldots, Z_n, B_1, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n\right)\\
&\quad= \frac{E\left(W_{i^*,k^*}|B_1, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n\right)\beta\left(F_0(Z_l), i^*, k^*-i^*+1\right)}{\sum_{\mathcal{S}} E\left(W_{i,k}|B_1, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n\right)\beta\left(F_0(Z_l), i, k-i+1\right)}\\
&\qquad\times P\left(B_l=B^*|\varphi(B_l)=(i^*, k^*), Z_1, \ldots, Z_n, B_1, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n\right)\\
&\quad= \frac{P\left(B_l=B^*|B_1, B_2, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n\right)I_{(i^*,k^*)}(\varphi(B^*))}{\sum_{\varphi(B)=(i^*,k^*)} P\left(B_l=B|B_1, B_2, \ldots, B_{l-1}, B_{l+1}, \ldots, B_n\right)}. \qquad (12)
\end{aligned}
$$

Finally, adapting the expression (11), the density function of Z_{n+1} conditionally on $(Z_1, \ldots, Z_n)$ and $(B_1, \ldots, B_n)$ in z

$$
F_0'(z)\sum_{\tilde{S}} E\left(W_{i,k}|B_1, B_2, \ldots, B_n\right)\beta(F_0(z), i, k-i+1)
$$

can be exactly computed.

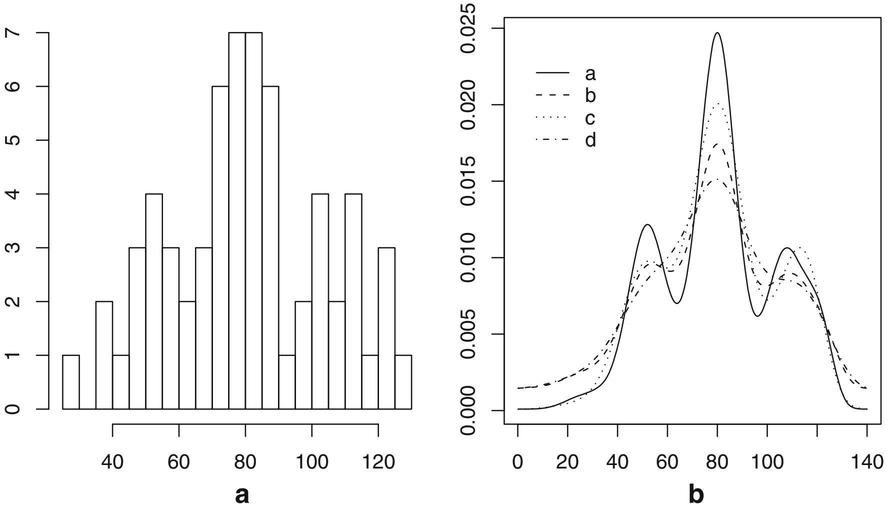


Fig. 2 **a** Histogram of the Buffalo snowfall data. **b** Predictive density functions; *a, b, c* and *d* correspond to the parametrizations ($\lambda = 100, M = 1$), ($\lambda = 100, M = 20$), ($\lambda = 50, M = 1$) and ($\lambda = 50, M = 20$)

The sampling procedure in the second case is much the same as in the first case; the only difference consists in the computation of the left member of the equality (11) which, exploiting the fact that the Dirichlet prior is conjugate, is equal to:

$$\frac{G(\tilde{S})f(\cdot)}{G(\tilde{S})+n-1}+\frac{\sum_{j\neq l}\beta\left(\cdot,\varphi^{(1)}(B_j),\varphi^{(2)}(B_j)-\varphi^{(1)}(B_j)+1\right)}{G(\tilde{S})+n-1},$$

where f is the density function of the rv $F_0(Z_1)$ and G is the parameter of the Dirichlet process.

If the specification of the prior distribution is different from those in the two discussed cases the Gibbs sampling exploits the approximations

$$E(W_{i,k}|Z_1,\ldots,Z_n)\approx E(W_{i,k}|Z_1,\ldots,Z_n,\max\{\varphi^{(2)}(B_1),\ldots,\varphi^{(2)}(B_n)\}<\tilde{k}),$$

where $\tilde{k}$ is a fixed large integer, and sample from the conditional distribution of B_l given $(Z_1,\ldots,Z_n)$, $(B_1,\ldots,B_{l-1},B_{l+1},\ldots,B_n)$ and $\varphi^{(2)}(B_l)<\tilde{k}$.

We have applied the described algorithm to the Buffalo snowfall data. This dataset consists of 63 observations and it has been extensively used in the density estimation literature, see for example Silverman (1986). We recall that the predictive df is the optimal Bayesian estimate of the unknown df under a quadratic loss function.

Four different prior distributions are considered. F_0 is set equal to a truncated normal distribution with mean $\theta = 70$ and standard deviation $\sigma = 100$ having support [0, 140]. The RUP $\tilde{X}$ is parameterized as in Proposition 1 with $G(i,k) = M\frac{e^{-\lambda}\lambda^{k-1}}{k!}$ and the predictive distribution is computed for four different values of the couple (M,λ) (Fig. 2).

The predictive densities suggest a trimodal distribution. The different shapes of the four estimates agree with the interpretation of the parameters M and λ given in Sect. 4: the higher is the value of M the more the predictive distribution is similar to the initial guess F_0, while different values of λ correspond to predictive densities with different degrees of smoothness.

In some cases, the proposed model outperforms the Dirichlet–Bernstein prior, as illustrated in the following example. The random distributions in the Dirichlet–Bernstein model are the mixtures of a random number K of beta distributions with Dirichlet distributed random weights. The posterior distribution from a Bernstein–Dirichlet prior, when the unknown density is very picked, typically concentrates on distributions that overfit the data on the tails: the densities generated from the posterior have peaks corresponding to single observations on the tails. This behavior of the Dirichlet–Bernstein model has an intuitive explanation. If a relevant portion of observations lies in a small subinterval of the support, the posterior concentrates on mixtures with a high number of components and, conditionally on a large value of K, the Dirichlet–Bernstein model inherits the peculiarities of the Dirichlet process; in particular, the predictive distribution closely follows that of the Dirichlet process which has point masses on the observations.

The extended Bernstein prior solves the issue: in the Dirichlet–Bernstein model K determines how the mixture components concentrate around their means, while the proposed prior is more flexible in combining components with different shapes. The underlined difference is similar to the improvement that can result if a data-set is fitted through a location-scale mixture of Gaussians (see for example Escobar and West 1995) rather than by a location mixture.

Figure 3 represents the density estimates obtained fitting a sample generated from a mixture of a truncated normal density and an uniform distribution, through the Bernstein model and the proposed extension. We have kept the parameterizations of the two priors as similar as possible. In the first case, $P(K = k) \propto 0.9^k$ and the distribution function F in (1) is modeled through a Dirichlet process centered on the uniform distribution $U[0, 1]$, in the second case, the prior is centered on the Uniform distribution and the expected values of the mixing weights $\{W_{j,k}\}_{(j,k)\in\tilde{S}}$ depend only on k: $E(W_{j,k}) = \frac{0.9^k}{9k}$.

The comparison between the predictive densities in Fig. 3 emphasizes the previously mentioned difference. In both cases, the mode of the density from which the data have been generated is well approximated. The predictive densities approach in a similar manner the peak of the unknown density and concurrently appear rather different corresponding to the two tails. The predictive density obtained updating the Bernstein prior is characterized by several peaks while in the second case the predictive density has approximately flat tails.

6 Final remarks

In this article, a random mixing measure for mixtures of beta kernels, which is a generalization of the Dirichlet process, is characterized through a RUP and some advantages with respect to the Bernstein prior are illustrated.

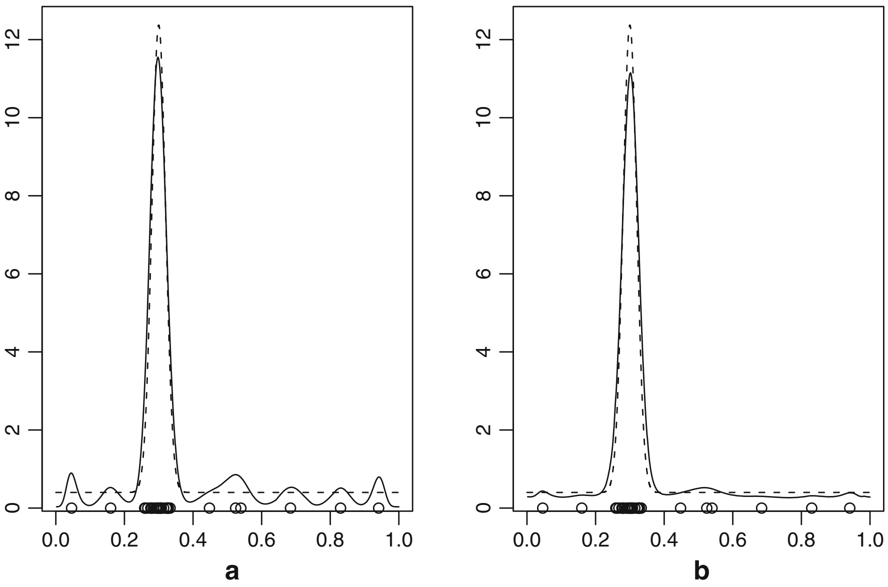


Fig. 3 **a** Bernstein prior. **b** Extended Bernstein prior. *Solid lines* Predictive densities. *Dotted line* The density from which the data have been generated. Sample size = 30

The theoretical properties of the Bernstein prior and of other Bayesian mixture models based on the Dirichlet process have been investigated by several authors. Among these, Ghosal (2001), Ghosal et al. (2008) and Kruijer and van der Vaart (2008) give results on the asymptotic properties of the Bernstein model. The general results of Ghosal et al. (2008) can be applied to the Bernstein model, for studying the effect of the a priori distribution of the unknown polynomial degree K on the convergence rates of the posterior distribution. The recent work of Kruijer and van der Vaart (2008) underlines the relevance of investigating random mixing measures alternative to the Dirichlet process; they show that a slight modification of the random mixing measure of the Bernstein model, under specific assumptions, improves the posterior rate of convergence.

As illustrated in the previous section, the proposed mixture model may provide a better small sample behavior than the Bernstein model. An open problem, in the above direction of research, is to investigate if it can also improve the convergence rates in density estimation.

Appendix

Proof (Proposition 1) Under the assumptions, it can be easily verified that, given a finite subset $\{(i_1, k_1), (i_2, k_2), \ldots, (i_m, k_m)\}$ of $\tilde{S}$ and a vector of positive numbers $(w_1, w_2, \ldots, w_m)$ such that $\sum w_j \leq 1$, we have

$$P\left(\bigcap_{j=1}^{m}[w_j - \delta \leq W_{i_j,k_j} \leq w_j + \delta]\right) > 0, \quad \forall \delta > 0.$$

Then, given a finite set $\{g_1, g_2, \ldots, g_l\}$ of continuous functions and a probability measure Q on $[0, 1]$, for every strictly positive ϵ

$$\begin{aligned}&\mathbf{P}\left[B \in \mathbf{B} : \max_{j=1,\ldots,l} \left[\left|\int g_j \mathrm{d}b - \int g_j \mathrm{d}Q\right|\right] < \epsilon\right]\\ &\quad \geq \mathbf{P}\left[B \in \mathbf{B} : \max_{j=1,\ldots,l} \left[\left|\int g_j \mathrm{d}b - \int g_j \mathrm{d}B_k^Q\right|\right] < \frac{\epsilon}{2}\right] > 0, \end{aligned} \tag{13}$$

where B_k^Q is a Bernstein distribution satisfying the inequalities

$$\left|\int g_j \mathrm{d}B_k^Q - \int g_j \mathrm{d}Q\right| < \frac{\epsilon}{2} \quad \forall j = 1, 2, \ldots, l,$$

whose existence is ensured by the fact that Bernstein-distributions are dense in $\triangle$. □

Proof (Theorem 2) Without loss of generality consider $S = \mathbb{N}$ and $x_0 = 0$. The finite-dimensional laws of the process $\{X_n\}_{n\geq 0}$ are described in (3) where μ is the distribution of the random transition matrix Π such that the rows are independent and the ith row $\Pi(i)$ is a Dirichlet processes with parameter $\beta_i = \{\beta_{i0}, \beta_{i1}, \ldots\}$, where $\beta_{ij} = a_i(c(i, j))$.

Equality (5) implies $\sum_j \beta_{ij} = \sum_j \beta_{ji}$. As the probability measure μ is conjugate, conditionally on $[X_0 = x_0, X_1 = x_1, \ldots, X_m = x_m]$, the rows $\Pi(i)$ are independent Dirichlet processes an the parameter of the ith raw is

$$\left\{\beta_{i0} + \sum_{l=0}^{m-1} I_{(i,0)}(x_l, x_{l+1}), \quad \beta_{i1} + \sum_{l=0}^{m-1} I_{(i,1)}(x_l, x_{l+1}), \ldots\right\}.$$

If $x_m = x_0$,

$$\sum_j \left(\beta_{ij} + \sum_{l=0}^{m-1} I_{(i,j)}(x_l, x_{l+1})\right) = \sum_j \left(\beta_{ji} + \sum_{l=0}^{m-1} I_{(j,i)}(x_l, x_{l+1})\right) \tag{14}$$

and the condition (5) still holds.
Let us denote $A^{[b]} = \Pi_{i=1}^{b}(A + i - 1),\ \forall b \in \mathbb{N}^+,\ \forall A \in \mathbb{R}^+$ and $A^{[0]} = 1$.

If $P(\varphi(B_1) = i) = \dfrac{\beta_{i0}}{\sum_l \beta_{l0}}$ for every $i \in \{0, 1, \ldots\}$, then

$$P\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m\right) = \frac{\prod_i \beta_{i0}^{[n_i]}}{\left(\sum_i \beta_{i0}\right)^{[m]}}, \tag{15}$$

where $n_i = \sum_{l=1}^{m} I_{(i)}(y_l)$, and the de Finetti measure associated with the sequence $\{Y_n = \varphi(B_n)\}_{n\geq 1}$ is a Dirichlet process with parameter α.

Consider the process $\{X_n^*\}_{n\geq 0}$ with the same state space of $\{X_n\}_{n\geq 0}$ such that $P\left[X_0^* = x_0, \ldots, X_n^* = x_n\right] = \int_{\mathcal{P}} \Pi_{j=0}^{n-1}\pi(x_j, x_{j+1})\mu^*(d\pi)$, where μ^* is the distribution of a random transition matrix Π^* such that the rows are independent random measure and the row $\Pi^*(i)$ is a Dirichlet process with parameter $\beta_i^* = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \ldots)$.

Observe that

$$\begin{aligned} &P\left(X_0 = x_0, X_1 = x_1, X_2 = x_2, \ldots, X_{m-1} = x_{m-1}, X_m = x_0\right) \\ &\quad = \prod_l \left(\frac{\prod_j \beta_{lj}^{[n_{lj}]}}{\left(\sum_j \beta_{lj}\right)^{[\sum_j n_{lj}]}} \right) \\ &\quad = P\left(X_0^* = x_0, X_1^* = x_{m-1}, X_2^* = x_{m-2}, \ldots, X_{m-1}^* = x_1, X_m^* = x_0\right), \quad (16) \end{aligned}$$

where $n_{lj} = \sum_{i=0}^{m-1} I_{(l,j)}(x_i, x_{i+1})$, thus $P(\varphi(B_1) = i) = P(X_1^* = i) = \frac{\beta_{i0}}{\sum_l \beta_{l0}}$ and equality (15) is verified. □

Proof (Proposition 2) The proposition is a direct application of Theorem 2. The first condition (5) is straightforwardly verified. Thus we need to verify the recurrence of the process.

The identity

$$P(\tilde{X}_{k-1} = (i, k)) = \frac{\mu_k\left(\frac{i-1}{k}, \frac{i}{k}\right]}{G(\tilde{S})} \quad \forall\, (i, k) \in \tilde{S}. \quad (17)$$

holds by definition for $k = 1$ and can be recursively verified for every $k \in \mathbb{N}^+$ exploiting the following equality:

$$\begin{aligned} P(\tilde{X}_k = (i, k+1)) &= P(\tilde{X}_k = (i, k+1)|\tilde{X}_{k-1} = (i, k))P(\tilde{X}_{k-1} = (i, k)) \\ &\quad + P(\tilde{X}_k = (i, k+1)|\tilde{X}_{k-1} = (i-1, k))P(\tilde{X}_{k-1} = (i-1, k))\cdot \quad (18) \end{aligned}$$

Indeed, the first term of the sum in (18) is equal to $\frac{\mu_{k+1}(\frac{i-1}{k}, \frac{i}{k+1}]}{\mu_k(\frac{i-1}{k}, \frac{i}{k}]} \frac{\mu_k(\frac{i-1}{k}, \frac{i}{k}]}{G(\tilde{S})}$ if $k \geq i$ and 0 otherwise, while the second is equal to $\frac{\mu_{k+1}(\frac{i-1}{k+1}, \frac{i-1}{k}]}{\mu_k(\frac{i-2}{k}, \frac{i-1}{k}]} \frac{\mu_k(\frac{i-2}{k}, \frac{i-1}{k}]}{G(\tilde{S})}$ if $i \geq 2$ and 0 otherwise. It follows that

$$P\left(\varphi(B_1) = (i, k)\right) = \frac{G(i, k)}{G(\tilde{S})}, \quad \lim_{k\to\infty} P(\varphi^{(2)}(B_1) > k) = 0 \quad \text{and}$$

$$\lim_{k\to\infty} P(\varphi^{(2)}(B_1) > k|\tilde{X}_{h-1} = (i, h)) = 0 \quad \forall (i, h) \in \tilde{S}$$

Then the equality

$$\begin{aligned}&\lim_{k\to\infty} P(\varphi^{(2)}(B_i) > k|\varphi^{(2)}(B_j) = k_j;\, j \in \{1,\ldots,i-1\})\\ &\quad= \lim_{k\to\infty}\sum_i \Big[P\left(\tilde{X}_{\hat{k}+\sum_j k_j} = (i,\hat{k}+1)|\varphi^{(2)}(B_j) = k_j;\, j\in\{1,\ldots,i-1\}\right)\\ &\qquad\times P(\varphi^{(2)}(B_1) > k|\tilde{X}_{\hat{k}} = (i,\hat{k}+1))\Big] = 0 \qquad (19)\end{aligned}$$

where $\hat{k} = \max_{j\in\{1,\ldots,i-1\}}\{k_j\}$, proves the recurrence of the process. □

Acknowledgments The authors thank the referees for useful comments which led to improvements of the manuscript.

References

Diaconis, P., Freedman, D. (1980). de Finetti's theorem for Markov chains. *The Annals of Probability, 8*(1), 115–130.

Escobar, M. D., West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association, 90*(1), 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric. problems. *Annals of Statistics, 1*(2), 209–230.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Annals of Statistics, 29*(5), 1264–1280.

Ghosal, S., Lember, J., van der Vaart, A. (2008). Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics, 2*, 63–89.

Kruijer, W., van der Vaart, A. (2008). Posterior convergence rates for Dirichlet mixtures of beta densities. *Journal of Statistical Planning and Inference, 138*, 1981–1992.

Muliere, P., Secchi, P., Walker, S. G. (2000). Urn schemes and reinforced random walks. *Stochastic Processes and Their Applications, 88*, 59–78.

Petrone, S. (1999). Random Bernstein polynomials. *Scandinavian Journal of Statistics, 26*, 373–393.

Petrone, S., Wasserman, L. (2002). Consistency of Bernstein Polynomial Posteriors. *Journal of the Royal Statistical Society. Series B, 64*(1), 79–100.

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