

# On the role of mixtures in Bayesian nonparametrics

## *Sul ruolo delle misture nell'inferenza bayesiana nonparametrica*<sup>1</sup>

Sonia Petrone<sup>2</sup>

Istituto di Metodi Quantitativi  
Università L. Bocconi  
sonia.petrone@unibocconi.it

**Riassunto:** Combinazioni lineari convesse di distribuzioni (“modelli mistura”) sono ampiamente utilizzate, in diversi contesti applicativi, per modellare l’eterogeneità dei dati, o come strumento per ottenere modelli flessibili. In queste note si discute il ruolo dei modelli mistura nell’inferenza Bayesiana nonparametrica. Basandoci su risultati di Feller, presentiamo una procedura costruttiva di approssimazione di una funzione di ripartizione mediante misture. Ciò fornisce un quadro di riferimento generale per studiare iniziali basate su misture per l’inferenza bayesiana nonparametrica. Si riprendono alcuni risultati nel caso unidimensionale, in particolare relativi alla proprietà di consistenza della distribuzione finale, e si suggeriscono estensioni al caso di dati multidimensionali.

**Keywords:** Approximation of random curves; natural exponential family; nonparametric priors; random scheme; weak consistency.

## 1. Introduction

In the last years, the Bayesian nonparametric literature has grown extensively, with applications from basic problems with exchangeable sequences to problems with more complex structures of dependence among the data; for an overview we refer to Walker *et al.* (1999). The most popular nonparametric prior is the Dirichlet process, whose statistical implications have been illustrated in the celebrated paper of Ferguson (1973); since then, the Dirichlet process has been extended in several ways; see Regazzini *et al.* (2003) and the references therein.

In fact, another way of obtaining flexible models is provided by mixtures of distribution functions (d.f.’s). Lo (1984) proposed mixture models with a Dirichlet process mixing distribution for Bayesian nonparametric inference (see also Ferguson, 1983) and this approach has become more popular after the development of MCMC methods for approximating the resulting posterior and predictive distributions (see e.g. Escobar and West (1995), Ishwaran and James (2001)). However, the properties of mixture priors have not been as carefully studied as for the Dirichlet process and its generalizations. Theoretical properties, such as consistency of the posterior distribution on the random density, have been studied for special cases (see e.g. a recent work of Ghosal and van der Vaart (2003) and the reference therein, for mixtures of Gaussians; for mixtures of betas, see Petrone and Wasserman (2002) and Ghosal (2001)). It is still debatable whether finite mix-

---

<sup>1</sup>This work has been partially founded by MIUR, research project PRIN 2002 “Costruzione ed analisi di misure di probabilità aleatorie e loro applicazione nell’ambito dell’inferenza bayesiana nonparametrica”.

<sup>2</sup>Mailing address: Istituto di Metodi Quantitativi - Università Bocconi - Viale Isonzo 25 - 20135 Milano (Italy)

tures with a random number of components can be considered as *nonparametric*, and the problem of determining the number of components of the mixture is usually treated in the Bayesian literature as a model selection problem with a changing dimension parameter space more than as a problem related to nonparametric inference on the unknown d.f. of the data.

These notes aim to be a contribution to the discussion of the role of mixtures in Bayesian nonparametric inference, in the direction of proposing a framework for studying their general features. There are of course several kinds of mixture models. One might consider finite, countable or continuous mixtures; the mixture components might have a physical interpretation, as distributions in subgroups of the population, or they might be used more formally as basis functions in order to construct flexible models (in this case, possible clusters are individuated by the number of modes of the distribution). Our focus in this paper is in the later case.

## 2. Priors for Bayesian nonparametric inference

Let us start with the basic situation of an exchangeable sequence of random variables  $\{X_1, X_2, \dots\}$  with values in a set  $\mathcal{X} \subseteq \mathbb{R}$ . By de Finetti's representation theorem, for any  $n \geq 1$  the joint d.f. of  $(X_1, \dots, X_n)$  can be represented as

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = \int_{\mathcal{F}} \prod_{i=1}^n F(x_i) d\pi_F(F) ,$$

where  $\pi_F$  is the limit law of the sequence of the empirical d.f.'s  $F_n(x) = (\sum_{i=1}^n \delta_{X_i}(x))/n$ ,  $\delta_X$  denoting a d.f. degenerate on  $X$ . A *parametric prior*  $\pi_F$  restricts the set of the realizations of the limit of the empirical d.f.'s to a parametric family  $\mathcal{P}_\Theta = \{F_\theta, \theta \in \Theta\}$  included in the class  $\mathcal{F}$  of all the d.f.'s on the sample space. A *nonparametric prior*, instead, selects d.f.'s in a family  $\mathcal{B}$  which is weakly dense in  $\mathcal{F}$  (in the sense that any d.f. in  $\mathcal{F}$  can be weakly approximated by a sequence of d.f.'s in  $\mathcal{B}$ ). For example, the Dirichlet process and its generalizations select d.f.'s in the family of discrete d.f.'s  $\mathcal{B} = \{F(\cdot) = \sum_{j=1}^k p_j \delta_{\lambda_j}(\cdot), k \leq \infty\}$ , and for this reason these priors are also called *discrete nonparametric priors*.

However other classes of nonparametric priors can be obtained by assigning a probability measure on a family of d.f.'s represented in terms of expansions by basis functions:

$$\mathcal{B} = \{F \in \mathcal{F} : \phi(F(\cdot)) = \sum_{j=1}^k p_j H_j(\cdot), k \leq \infty\} , \quad (1)$$

where  $\phi(F(\cdot))$  is some transformation of  $F$  and  $H_j(\cdot)$  are basis function, possibly indexed by some parameters. Roughly speaking, this construction stems on the possibility of showing that any d.f. in  $\mathcal{F}$  can be weakly approximated by sequences of d.f.'s of the form  $\phi^{-1}(\sum_{j=1}^k p_j H_j(\cdot))$ . Of course there are many different expansions (by polynomials, splines, trigonometric functions, wavelets, etc.) that could be used to this aim, and it is an open question how to compare and choose among them.

Mixtures seem to be an attractive particular family of the kind (1), where  $\phi(\cdot)$  is the identity and the  $H_j$ 's are d.f.'s. Mixture are appealing since they have a more probabilistic nature and more naturally interpretable parameters than other expansions. Before

discussing more formally the construction of nonparametric mixture priors in the next section, we underline some general points.

There are many applications where one might look for a prior which is a smoothing of a nonparametric discrete prior, and mixtures can be useful to this aim. Perhaps the main reason of the popularity of the Dirichlet process is its conjugacy property; if  $F$  as a Dirichlet process prior, with parameters  $(a, F_0)$ , the posterior probability law of the random d.f., given  $(X_1 = x_1, \dots, X_n = x_n)$ , is still a Dirichlet process, with updated parameters. The predictive d.f. has the appealing simple form of a weighted average of the prior guess  $E(F(\cdot)) = F_0(\cdot)$  and the empirical d.f.  $F_n(\cdot)$ . This structure of the predictive d.f. (for  $n = 1, 2, \dots$ ) is indeed a characterizing property of the Dirichlet process (Regazzini (1978); a similar characterization holds more generally for neutral to the right priors (Walker and Muliere (1999)). It implies that the conditional probability of a result in a set  $A$  only depends on the number of previous outcomes in  $A$ . However, this property is sometimes a drawback, since one might want to spread the information from the observations in  $A$  on sets close to  $A$ , especially in the case of continuous data. In fact, with continuous data, one might want a prior which gives probability one to the set of absolutely continuous d.f.'s, not only because this is consistent with the prior opinion that the data are continuous, but also in order to include possibly available prior information on the shape of the unknown *density* of the data. For example, in studies on the income distribution, the economist might find more natural to elicit prior opinion on the shape of the income density (for example, she/he might expect the density to be similar to that of a previous period, with ticker tails for effect of an impoverishment of the middle class), rather than reasoning on discrete trajectories of a discrete random d.f. In these cases, a discrete nonparametric prior might be used as a first step in eliciting prior opinion, but then it should be smoothed in order to take into account prior information on the density of the data. The price to pay for that is a more complicated structure of the posterior and the predictive distributions, but this difficulty can be overcome by stochastic or numeric approximation techniques.

### 3. Mixture priors

In a parametric framework, where  $X_1, X_2, \dots$  are conditionally independent and identically distributed (i.i.d.), given  $\theta$ , with d.f.  $F_\theta(\cdot)$ , a common choice for the prior on the parameter  $\theta$  is a conjugate prior. However, this choice might be mainly due to mathematical convenience and might not honestly express the available prior information (for example in the case when the prior is expressed by a group of experts with discordant opinions). Diaconis and Ylvisaker (1985) argue that a honest elicitation of the prior can be always obtained by mixtures of conjugate priors; see also Dalal and Hall (1983). Somehow in the same spirit, when eliciting the de Finetti measure for a sequence of exchangeable r.v.'s, the choice of a parametric prior (or, in other words, the choice of a parametric model  $F_\theta(\cdot)$ ) might not express honest beliefs about the possible limits of the empirical d.f., and the question is whether a honest elicitation can always be obtained by a mixture of parametric models. This basic idea is developed by Petrone and Veronese (2002; 2003), starting from an approximation technique referred to Feller. Their construction is a step for a unified framework for studying mixture priors for Bayesian nonparametric inference, also providing a technique for smoothing discrete nonparametric priors by mixtures. Petrone and Veronese treats the case of a prior on a random univariate d.f. It is of interest to extend

their construction to the case of random multivariate d.f.'s, and here we provide a simple generalization.

**3.1 Multidimensional Feller operators.** In the approximation theory literature, a fundamental role is played by positive approximation processes. A relevant family of these processes are the so-called "Korovkin-type approximations"; an extensive monograph on this subject is provided by Altomare and Campiti (1994). In particular, in chapter 5 they discuss a probabilistic way of constructing Korovkin-type approximations and refer the result to Feller (1971). We call *Feller operators* the class of linear operators so defined. Feller operators include Bernstein polynomials as a special case. The standard definition of Feller operators is given for the univariate case. Here we focus on multidimensional Feller operators, which are based on what is called a multidimensional random scheme (see Altomare and Campiti (1994) p. 314).

A *d-dimensional random scheme* is defined as a sequence of *d*-dimensional random vectors  $\{\mathbf{Z}_{k\mathbf{x}} = (Z_{1,k\mathbf{x}}, \dots, Z_{d,k\mathbf{x}}); k = 1, 2, \dots, \mathbf{x} \in Q \subseteq \mathbb{R}^d\}$  where  $\mathbf{Z}_{k\mathbf{x}}$  has d.f.  $P_{k\mathbf{x}}$  such that for  $\mathbf{x} = (x_1, \dots, x_d) \in Q$

$$\lim_{k \rightarrow \infty} E(Z_{i,k\mathbf{x}}) = x_i, i = 1, \dots, d \quad ; \quad \lim_{k \rightarrow \infty} \sum_{i=1}^d V(Z_{i,k\mathbf{x}}) = 0 .$$

Let  $U$  be a real function defined on  $\mathbb{R}^d$ . The *d*-dimensional Feller operator of order *d* for  $U$ , with random scheme  $\{\mathbf{Z}_{k\mathbf{x}} = (Z_{1,k\mathbf{x}}, \dots, Z_{d,k\mathbf{x}}) \sim P_{k,\mathbf{x}}, k = 1, 2, \dots, \mathbf{x} \in Q \subseteq \mathbb{R}^d\}$ , is defined as

$$B(\mathbf{x}; k, U) = E(U(\mathbf{Z}_{k\mathbf{x}})) = \int U(\mathbf{z}) dP_{k\mathbf{x}}(\mathbf{z}) . \quad (2)$$

The following theorem extends Lemma 1 in chapter VII of Feller (1971) to the multivariate case; even if simple, we include the proof for the reader's convenience.

**Theorem 1** *If  $U : \mathbb{R}^d \rightarrow \mathbb{R}$  is bounded, then for any  $\mathbf{x} \in Q$  of continuity of  $U$  we have*

$$\lim_{k \rightarrow \infty} B(\mathbf{x}; k, U) = U(\mathbf{x}) .$$

Proof. The proof uses a version of the multivariate Markov inequality: if  $(Y_1, \dots, Y_d)$  is a random vector with positive components, with  $E(Y_i) = \mu_i$  and  $V(Y_i) = \sigma_i^2$  finite, then

$$P\left(\max_{i=1, \dots, d} |Y_i - \alpha_i| > t\right) \leq \frac{\sum_{i=1}^d [\sigma_i^2 + (E(Y_i) - \alpha_i)^2]}{t^2} .$$

Now, we have that

$$\begin{aligned} |E(U(\mathbf{Z}_{k\mathbf{x}})) - U(\mathbf{x})| &\leq \int |U(\mathbf{z}) - U(\mathbf{x})| dP_{k\mathbf{x}}(\mathbf{z}) = \\ &= \int_{\{\mathbf{z}: \max_{i=1, \dots, d} |z_i - x_i| \leq \delta\}} |U(\mathbf{z}) - U(\mathbf{x})| dP_{k\mathbf{x}}(\mathbf{z}) + \int_{\{\mathbf{z}: \max_{i=1, \dots, d} |z_i - x_i| > \delta\}} |U(\mathbf{z}) - U(\mathbf{x})| dP_{k\mathbf{x}}(\mathbf{z}) . \end{aligned}$$

Being  $U(\cdot)$  continuous in  $\mathbf{x}$ , we can choose  $\delta$  such that  $|U(\mathbf{z}) - U(\mathbf{x})| < \epsilon/2$  for  $\mathbf{z}$  such that  $\max_{i=1, \dots, d} |z_i - x_i| \leq \delta$ , so the first addend above can be made smaller than  $\epsilon/2$ ; and being  $U(\cdot)$  bounded (smaller than  $M$ , say) we have

$$|E(U(\mathbf{Z}_{k\mathbf{x}})) - U(\mathbf{x})| \leq \frac{\epsilon}{2} + 2M P\left(\max_{i=1, \dots, d} |Z_{i,k\mathbf{x}} - x_i| > \delta\right) .$$

By the multivariate Markov inequality

$$|E(U(\mathbf{Z}_{kx}) - U(\mathbf{x}))| \leq \frac{\epsilon}{2} + 2M \frac{\sum_{i=1}^d [V(Z_{i,kx}) + (E(Z_{i,kx}) - x_i)^2]}{\delta^2},$$

which can be made arbitrarily small for large  $k$  by the properties of the random scheme.  
 $\diamond$

Despite its simplicity, the above result provides useful in many contexts. In the next section we discuss some implications in Bayesian nonparametric inference.

**3.2 Feller operators with exponential random scheme and mixture priors.** Petrone and Veronese (2003; in the sequel, [PV]) consider the case of  $d = 1$  and suggest an *exponential random scheme* where  $Z_{kx} = (1/k) \sum_{i=1}^k Y_i$  and  $Y_1, Y_2, \dots$  are i.i.d. random variables, with distribution in the natural exponential family (NEF) (in fact, [PV] consider a more general case, where  $Z_{kx}$  is obtained by power convolution, but for brevity here we restrict our attention to the simpler case). More precisely, let  $Y_1, Y_2, \dots$  be i.i.d. with distribution in a regular NEF, with density

$$p_\theta(y) = \exp(\theta y - M(\theta))$$

w.r.t. a  $\sigma$ -finite measure  $\nu$  on the Borel sets of  $\mathfrak{R}$ , and with natural parameter  $\theta \in \Theta$ . Denote by  $x$  the mean parameter, i.e.  $x = E_\theta(Y_i) = dM(\theta)/d\theta$ , and by  $\sigma^2 = \sigma^2(x)$  the variance of  $Y_i$ . The mean parameter space  $\Omega$  corresponds to the interior of the convex hull of the support of the measure  $\nu$ . The correspondence between the natural parameter  $\theta$  and the mean parameter  $x$  is one-to-one, so we can also write  $\theta = \theta(x)$ . Let us now consider the average  $Z_{kx} = (1/k) \sum_{i=1}^k Y_i$  of  $(Y_1, \dots, Y_k)$ .  $Z_{kx}$  has d.f.  $P_{k,\theta(x)}$  which is still in the NEF, with mean  $E(Z_{kx}) = x$  and variance  $V(Z_{kx}) = \sigma^2/k$  converging to zero as  $k \rightarrow \infty$ . Therefore  $\{Z_{kx} \sim P_{k,\theta(x)}, k = 1, 2, \dots, x \in \Omega\}$  is a random scheme, which is called *exponential random scheme*.

Let now  $U$  be a univariate d.f., and denote by  $\mathcal{X}$  the convex hull of its support (briefly, its convex support). Then  $U$  can be weakly approximated by a Feller operator  $B(\cdot; k, U)$  with exponential random scheme  $\{Z_{kx} \sim P_{k,\theta(x)}, k = 1, 2, \dots, x \in \mathcal{X}^\circ\}$ , appropriately extended on the real line. Theorem 3 in [PV] shows that

1.  $B(\cdot; k, U)$  is a d.f..
2. The derivative of  $B(x; k, U)$  for  $x \in \mathcal{X}^\circ$  is given by

$$b(x; k, U) = \int f_k(x; z) dU(z) \tag{3}$$

where

$$f_k(x; z) = \frac{k}{\sigma^2(x)} \int_{[z, \infty)} (t - x) dP_{k,\theta(x)}(t). \tag{4}$$

3. If  $\mathcal{X}$  is a real interval  $(a, b)$ ,  $B(\cdot; k, U)$  concentrates mass  $U(a)$  on  $a$  (if  $a$  is finite) and mass  $1 - U(b^-)$  on  $b$  if  $\nu(\{b\}) = 0$  (and  $b$  is finite). If  $U$  has no masses on the extremes of  $\mathcal{X}$ , then  $B(\cdot; k, U)$  is absolutely continuous and it can be written as a mixture of d.f.'s with density (3).

It is of interest to note that the kernel density  $f_k(\cdot; z)$  defined by (4) is the fiducial density of the mean parameter  $x$  of the NEF ([PV], remark 1).

We have thus provided a constructive procedure for the approximation of a d.f. by mixtures. The approximation holds for densities, too, in the sense that, if  $U$  is absolutely continuous, with a continuous and bounded density  $u$ , then for any  $x \in \mathcal{X}^\circ$   $\lim_{k \rightarrow \infty} b(x; k, U) = u(x)$  ([PV], theorem 4 and proposition 5).

These results show that the family  $\mathcal{B} = \{B(\cdot; k, U), k = 1, 2, \dots; U \in \mathcal{F}\}$  is weakly dense in the class  $\mathcal{F}$  of all the d.f.'s with convex support  $\mathcal{X}$ . Given a sequence of exchangeable random variables  $(X_1, X_2, \dots)$ , a nonparametric *mixture prior*  $\pi_B$  can consequently be defined by assigning a probability measure which select d.f.'s in  $\mathcal{B}$ , that is, by regarding  $(k, U)$  as random quantities and assigning them a joint probability law  $\pi$ . This means that, for any  $n \geq 1$

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = \int_{\mathcal{F}} \prod_{i=1}^n F(x_i) d\pi_B(F) = \int \prod_{i=1}^n B(x_i; k, U) d\pi(k, U). \quad (5)$$

Note that the parameter space is  $(\mathbf{N} \times \mathcal{F})$ , where  $\mathbf{N} = \{1, 2, \dots\}$ , even when  $B(\cdot; k, U)$  results to be a finite mixture with a random number of components, as appears clear in the following example.

*Example 1.* Let  $\{X_1, X_2, \dots\}$  be exchangeable r.v.'s with values in  $[0, 1]$ . Now, any d.f.  $U$  on  $[0, 1]$  can be approximated by a Feller operator with *binomial random scheme*, i.e.  $\{kZ_{kx} \sim \text{bin}(\cdot; k, x), k = 1, 2, \dots; x \in (0, 1)\}$ , where  $\text{bin}(\cdot; k, x)$  denotes the binomial probability function with parameters  $(k, x)$ . With this choice

$$B(x; k, U) = \sum_{j=0}^k U\left(\frac{j}{k}\right) \binom{k}{j} x^j (1-x)^{k-j}, \quad (6)$$

which corresponds to the Bernstein polynomial of order  $k$  for the approximation of  $U$ .  $B(\cdot; k, U)$  has a points mass  $p$  at zero if  $U(0) = p > 0$ . If  $U(0) = 0$ ,  $B(\cdot; k, U)$  is an absolutely continuous d.f., with density given by (3). It can be shown that, in this case, the kernel density is  $f_k(x; z) = be(x; [kz] + 1, k - [kz])$ , where  $be(x; \alpha, \beta)$  denotes a beta density with parameters  $(\alpha, \beta)$  and  $[kz]$  is the integer part of  $kz$ . Note that  $f_k(x; \cdot)$  is piecewise constant, therefore  $b(x; k, U)$  reduces to the finite mixture

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

where  $w_{j,k} = U(j/k) - U((j-1)/k), j = 1, \dots, k$ .

A nonparametric prior can be defined by assigning a probability law  $\pi$  on the random quantities  $(k, U)$  (see Petrone (1999a)), so that

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = \int \prod_{i=1}^n \left( \sum_{j=1}^k w_{j,k} \int_{(0,x]} be(t; j, k - j + 1) dt \right) d\pi(k, U).$$

A simple and interesting choice for  $\pi$  is to assume that  $k$  and  $U$  are independent, and  $U$  has a discrete nonparametric prior, for example a Dirichlet process; the mixture prior

provides in this case a smoothing of the discrete prior. Note that the introduction of a smoothing parameter  $k$  allows more flexibility in eliciting the prior on the unknown d.f. (see e.g. Petrone, 1999a).

The complicated structure of the likelihood in mixture models makes computing the posterior probability law of  $(k, U)$  analytically difficult, but sampling from the posterior is possible via simulation techniques. Note that the fact that the parameter space is  $(\mathbf{N} \times \mathcal{F})$  allows to implement MCMC algorithms without having to deal with a changing dimensions parameter space (see Petrone (1999b)).

**3.3. Consistency.** Consistency is a frequentist property of the Bayesian procedures, which, in a parametric setting, ensures that, if the data are i.i.d. with d.f.  $F_{\theta_0}$ , then the posterior distribution of the parameter  $\theta$  almost surely cumulates around the value  $\theta_0$  as the sample size increases to infinity. From a Bayesian viewpoint, consistency can be interpreted in terms of merging of opinions (Diaconis and Freedman, 1986). It is known that consistency might fail in a nonparametric framework, and in the past this drawback has limited the development of Bayesian nonparametric inference. In the more recent literature, however, there are general positive results about consistency also in a nonparametric framework. For mixture models, consistency can be studied with concern to the posterior of the mixing distribution, or to the posterior of the unknown d.f., or of the number of components (for finite mixtures). Many results in the literature deal with the asymptotic behavior of the posterior of the d.f. (or density) of the data. Let  $\pi_F$  be the de Finetti measure of the exchangeable sequence  $\{X_1, X_2, \dots\}$  and denote by  $\pi_{F,n}$  the posterior of the random d.f., given  $(X_1, \dots, X_n)$ . We say that  $\pi_{F,n}$  is consistent at  $F_0$  if

$$\lim_{n \rightarrow \infty} \pi_{F,n}(W_\delta(F_0) | X_1, \dots, X_n) = 1 \quad , \quad \text{a.s. } F_0^\infty \quad ,$$

for any  $W_\delta(F_0)$  which is a neighborhood of  $F_0$  in some topology (the topology of weak convergence for weak consistency; in Hellinger metric for strong consistency). Sharp results have been proved for specific families of mixture models (e.g. mixtures of Gaussians, or mixtures of beta distributions). These results exploit the particular features of the model under exam (e.g., the compactness of the sample space, for mixtures of betas; the behavior of the Gaussian tails when  $\mathcal{X} = \mathbb{R}$ ), and it is not easy to extend them to a general mixture model. For mixtures of Gaussians, consistency of the posterior distribution of the unknown density has been first proved under the assumption that the *true* d.f. is itself a mixture of Gaussians, of the form  $F_0(x) = \int \Phi(x; \mu, \sigma_0^2) dU_0(\mu)$  (where  $\Phi(\cdot; \mu, \sigma^2)$  denotes a Gaussian d.f. with parameters  $(\mu, \sigma^2)$ ), with mixing distribution  $U_0$  having compact support (Ghosal *et al.* (1999)). These results (and also convergence rates) have been extended to location-scale mixtures of Gaussians with variance bounded away from zero and infinity (Ghosal and van der Vaart (2001); see also Genovese and Wasserman (2000)) and only recently proved for more general cases (Scricciolo (2001), Ghosal and van der Vaart (2003)). More general results hold for mixtures of betas, for which consistency and convergence rates can be proved under the mild assumption that the true density is bounded and continuous (Ghosal (2001); Petrone and Wasserman (2002)). A rather general consistency result for (finite) mixtures is provided by Ishwaran *et al.* (2001), assuming however a particular choice for the prior on the random mixing distribution.

A consistency result for mixture models with a fairly general kernel is provided in [PV]. Consider a mixture prior  $\pi_B$  as defined in (5). More specifically, suppose that  $X_1, X_2, \dots$  are conditionally i.i.d., with common density  $b(x; k, U) = \int f_k(x; z) dU(z)$ ,  $x \in \mathcal{X}^\circ$ , arising from an exponential random scheme such that the measure  $\nu$  for the NEF

is absolutely continuous with respect to the Lebesgue or to the counting measure. Let  $U|k \sim \pi(\cdot|k)$  and  $k \sim p(k)$ . [PV] study some properties of the kernel densities  $f_k(\cdot; z)$ , from which they can prove the following consistency result.

**Theorem 2** ([PV], theorem 5). *Suppose that  $\pi(\cdot|k)$  has full weak support and  $p(k) > 0$ , for any  $k = 1, 2, \dots$ . Let  $F_0$  be an absolutely continuous d.f., with density  $f_0(x) = \int f_{k_0}(x; z) dU_0(z)$ , where  $k_0 \in \{1, 2, \dots\}$  and  $U_0$  has support  $[c_1, c_2] \subset \mathcal{X}^\circ$ . Then the mixture prior  $\pi_B$  defined in (5) is weakly consistent at  $F_0$ , i.e.*

$$\lim_{n \rightarrow \infty} \pi_{B,n}(W|X_1, \dots, X_n) = 1 \quad , \text{ a.s. } F_0^\infty$$

for any weak neighborhood  $W$  of  $F_0$ , where  $F_0^\infty$  is the probability measure of the sequence  $\{X_n, n \geq 1\}$  according to which  $X_1, X_2, \dots$  are i.i.d. with d.f.  $F_0$ .

The above result only proves weak consistency; however, weak consistency is the basic step, since sufficient conditions for stronger results require weak consistency as a prerequisite. In particular, the result in the above theorem might be extended by using the sufficient conditions for strong consistency recently proved by Walker (2003). The restrictions on  $F_0$  might be relaxed following the lines in Ghosal and van der Vaart (2003).

#### 4. Multivariate extensions

We provide now a simple way of extending Feller operators with exponential random scheme for the approximation of  $d$ -dimensional d.f.'s. For simplicity, we restrict our attention to d.f.'s whose convex support is a product space  $\mathcal{X} = (\mathcal{X}_1 \times \dots \times \mathcal{X}_d) \subseteq \mathbb{R}^d$ . Let  $\mathbf{x} = (x_1, \dots, x_d) \in \mathcal{X}^\circ$ , and consider a random vector  $\mathbf{Z}_{k\mathbf{x}} = (Z_{1,k\mathbf{x}}, \dots, Z_{d,k\mathbf{x}})$  whose components are independent, and defined as  $Z_{i,k\mathbf{x}} = (1/k) \sum_{j=1}^k Y_{i,j}$ , where for any  $i = 1, \dots, d$ ,  $(Y_{i,1}, Y_{i,2}, \dots)$  are i.i.d., with a distribution  $P_{\theta(x_i)}$  in the NEF such that

$$E(Y_{i,j}) = x_i \quad , \quad V(Y_{i,j}) = \sigma^2(x_i) \quad .$$

Then  $Z_{i,k\mathbf{x}}$  has d.f.  $P_{k,\theta(x_i)}$  in the NEF, with  $E(Z_{i,k\mathbf{x}}) = x_i$  and  $V(Z_{i,k\mathbf{x}}) = \sigma^2(x_i)/k$  converging to zero as  $k$  goes to infinity. Therefore, the sequence of random vectors  $\{\mathbf{Z}_{k\mathbf{x}}, k = 1, 2, \dots, \mathbf{x} \in \mathcal{X}^\circ\}$  is a  $d$ -dimensional exponential random scheme.

Let now  $U$  be a  $d$ -dimensional d.f. with convex support  $\mathcal{X}$ . According to (1), a  $d$ -dimensional Feller operator with exponential random scheme for  $U$  is defined as

$$B(\mathbf{x}; k, U) = E(U(\mathbf{Z}_{k\mathbf{x}})) = \int \dots \int U(z_1, \dots, z_d) dP_{k,\theta(x_1)}(z_1) \dots dP_{k,\theta(x_d)}(z_d) \quad , \quad (7)$$

for  $\mathbf{x} \in \mathcal{X}^\circ$ . The derivative  $b(\mathbf{x}; k, U)$  of  $B(\mathbf{x}; k, U)$ , for  $\mathbf{x} \in \mathcal{X}^\circ$ , is given by

$$b(\mathbf{x}; k, U) = \frac{\partial}{\partial x_1} \dots \frac{\partial}{\partial x_n} B(x_1, \dots, x_n; k, U) = \int \dots \int \prod_{i=1}^d f_k(x_i, z_i) dU(z_1, \dots, z_d) \quad (8)$$

(the computation follows the lines of the proof of theorem 3 in [PV]).

If we assume for simplicity that  $U$  has no mass concentrated on the frontier of  $\mathcal{X}$ , then  $B(\mathbf{x}; k, U)$ , appropriately extended on  $\mathfrak{R}^d$ , is an absolutely continuous d.f. with density given by the mixture (8). By theorem 1,  $B(\mathbf{x}; k, U)$  converges to  $U(\mathbf{x})$  as  $k$  goes to infinity, for any continuity point  $\mathbf{x}$  of  $U$ .

*Example 2.* Let  $U$  be a  $d$ -dimensional d.f. with convex support  $\mathcal{X} = [0, 1]^d$ , the hypercube of  $\mathfrak{R}^d$ . Then for any  $\mathbf{x}$  in the interior of  $\mathcal{X}$  and  $i = 1, 2, \dots$ , let  $(Y_{i,1}, Y_{i,2}, \dots)$  be independent Bernoulli random variables, with parameter  $x_i$ . Then

$${}_k Z_{i,kx_i} = \sum_{j=1}^k Y_{i,j} \sim \text{bin}(\cdot; k, x_i) \quad , \quad i = 1, \dots, d .$$

The Feller operator with *d-dimensional binomial random scheme* is defined by

$$B(\mathbf{x}; k, U) = \sum_{j_1=0}^k \cdots \sum_{j_d=0}^k U\left(\frac{j_1}{k}, \dots, \frac{j_d}{k}\right) \prod_{i=1}^d \binom{k}{j_i} x_i^{j_i} (1 - x_i)^{k-j_i} , \quad \mathbf{x} \in (0, 1)^d ,$$

and corresponds to the Bernstein operator of order  $k$  on the  $d$ -dimensional hypercube (Lorentz (1953)). The derivative for  $\mathbf{x} \in (0, 1)^d$  is

$$b(\mathbf{x}; k, U) = \sum_{j_1=1}^k \cdots \sum_{j_d=1}^k w_{j_1, \dots, j_d} \prod_{i=1}^d \text{be}(x_i; j_i, k - j_i + 1) ,$$

where

$$w_{j_1, \dots, j_d} = P_U\left(\left(\frac{j_1-1}{k}, \frac{j_1}{k}\right] \times \cdots \times \left(\frac{j_d-1}{k}, \frac{j_d}{k}\right]\right)$$

and  $P_U$  denotes the probability measure defined by the d.f.  $U$ . Bernstein polynomials on the hypercube and their implications for Bayesian nonparametric inference have been studied by Epifani (1999).

We now give some basic properties of  $B(\cdot; k, U)$ . The proof of properties 1, 2 and 4 is simple and therefore omitted.

*Property 1.* It is easy to check that the marginal d.f.'s of  $B(\cdot; k, U)$  are

$$B_i(x_i; k, U_i) = \int f_k(x_i; z) dU_i(z) ,$$

where  $U_1(x_1), \dots, U_d(x_d)$  are the marginal d.f.'s of  $U(x_1, \dots, x_d)$ . From the previous results,  $B_i(\cdot; k, U)$  is a Feller-type approximation of the marginal d.f.  $U_i$ ,  $i = 1, \dots, d$ .

*Property 2.* The approximation preserves independence, that is, if  $U(x_1, \dots, x_d) = \prod_{i=1}^d U(x_i)$  for any  $(x_1, \dots, x_d) \in \mathfrak{R}^d$ , then

$$B(x_1, \dots, x_d; k, U) = \prod_{i=1}^d B_i(x_i; k, U_i) .$$

*Property 3.* If  $U$  is absolutely continuous, with density  $u(x_1, \dots, x_d)$  continuous and bounded in each argument, then the density  $b(\mathbf{x}; k, U)$  of  $B(\mathbf{x}; k, U)$  approximates the density  $u(\mathbf{x})$  of  $U$ , that is

$$\lim_{k \rightarrow \infty} b(\mathbf{x}; k, U) = u(\mathbf{x}) \quad , \quad \text{for } \mathbf{x} \in \mathcal{X}^\circ . \quad (9)$$

*Proof.* If  $U$  is absolutely continuous, with density  $u$ , then by (8) we have

$$b(x_1, \dots, x_d; k, U) = \int \cdots \int u(z_1, \dots, z_d) \left[ \prod_{i=1}^d f_k(x_i; z_i) \right] dz_1 \cdots dz_d .$$

Now, it can be proved ([PV], theorem 4) that for any  $x_i \in \mathcal{X}_i^\circ$ ,  $f_k(x_i; \cdot)$  is a probability density; in particular, if  $Z_{i,kx_i}^* \sim f_k(x_i; \cdot)$ , then  $E(Z_{i,kx_i}^*) \rightarrow x_i$  and  $V(Z_{i,kx_i}^*) \rightarrow 0$  if  $k \rightarrow \infty$ . It follows that  $\{\mathbf{Z}_{k\mathbf{x}}^* = (Z_{1,kx_1}^*, \dots, Z_{d,kx_d}^*), k = 1, 2, \dots, \mathbf{x} \in \mathcal{X}^\circ\}$  is a  $d$ -dimensional random scheme, so that the thesis follows by theorem 1.  $\diamond$

*Property 4.* Property 3 implies that, under the same assumptions, the approximation holds for the conditional densities, too; that is, letting for brevity  $d = 2$ ,

$$\lim_{k \rightarrow \infty} b_{2|1}(x_2; k, U|x_1) = \lim_{k \rightarrow \infty} \frac{b(x_1, x_2; k, U)}{b_1(x_1; k, U)} = \frac{u(x_1, x_2)}{u(x_2)} = u_{2|1}(x_2|x_1) .$$

The proposed construction shows that, for  $d$ -dimensional data in a product space  $\mathcal{X}$ , one can always define a nonparametric mixture prior which selects d.f.'s of the form (7), by selecting the random quantities  $(k, U)$  according to a probability law  $\pi$ . Again, an interesting case is when  $k$  and  $U$  are independent. The prior for  $U$  could be a Dirichlet process on the space of the  $d$ -dimensional d.f.'s with convex support  $\mathcal{X}$ ; other choices could be of interest, for expressing different structures of dependence among the marginal d.f.'s (see e.g. Muliere and Walker (2003)). Natural applications are to density estimation and to nonparametric stochastic regression. Indeed, the literature on Bayesian nonparametric regression is mainly focussed to the case of deterministic explanatory variables (see e.g. Graziani and Holmes (2003) and the references therein) and stochastic nonparametric regression seems less developed. The proposed procedure might be useful to this aim, of course requiring MCMC algorithms which are efficient also in the case of high-dimensional data.

The multivariate extension of Feller operators that we have discussed is fairly simple, but clearly more general  $d$ -dimensional random schemes might be constructed by sequences of random vectors  $\mathbf{Z}_x$  with dependent components. This should allow to extend the procedure to the approximation of d.f.'s whose convex support is not a product space, as illustrated in the following example.

*Example 3.* Let  $U$  be a d.f. on the  $d$ -dimensional simplex  $Q_d$ . Given  $\mathbf{x} = (x_1, \dots, x_d) \in Q_d$ , let  $k$   $Z_{k\mathbf{x}}$  be a random vector with a multinomial distribution with parameters  $(k, x_1, \dots, x_d)$ . The Feller operator with a *multinomial random scheme* is defined as

$$B(\mathbf{x}; k, U) = \sum_{\{j_1, \dots, j_d \geq 0, \sum_{i=1}^d j_i \leq k\}} U\left(\frac{j_1}{k}, \dots, \frac{j_d}{k}\right) p(j_1, \dots, j_d; k, \mathbf{x}) ,$$

where

$$p(j_1, \dots, j_d; k, \mathbf{x}) = \frac{k!}{j_1! \cdots j_d! (k - j_1 - \dots - j_d)!} x_1^{j_1} \cdots x_d^{j_d} (1 - x_1 - \dots - x_d)^{k - j_1 - \dots - j_d} .$$

The operator  $B(\cdot; k, U)$  so defined is called the Bernstein operator of order  $k$  on the  $d$ -dimensional simplex. The derivative of  $B(\mathbf{x}; k, U)$  for  $\mathbf{x}$  in the interior of  $Q_d$  is a mixture of Dirichlet distributions, as studied in the context of Bayesian inference by Regazzini and Sazanov (1999).

## 5. Final remarks

We have discussed a constructive procedure, based on Feller operators, for approximating a univariate or multivariate d.f. by mixtures. This scheme provides a general framework for studying nonparametric mixture priors for Bayesian nonparametric inference. We underline that the mixture components in our construction have in general no physical interpretation, acting in fact as basis-functions for the approximation. Of course, when prior information is available on possible subgroups of the population under study, one might prefer to use a mixture model whose components have a physical meaning, representing the distributions of sub-populations. In fact, this case might be formally included in the proposed general framework, by modifying the random scheme.

We focussed on the expansion by mixtures of random d.f.'s. In fact, the proposed approximation scheme can have application, more generally, in Bayesian nonparametric inference on a random curve. Curves of interest include, among many others, regression functions, hazard rates, transition densities of a Markov process, spectral densities of a time series. Some work in this direction is in Choudhuri *et al.* (2004).

## References

- Altomare, F. and Campiti, M. (1994) *Korovkin-type Approximation Theory and its Application*. W. de Gruyter, Berlin.
- Choudhuri, N., Ghosal, S. and A. Roy (2004). Bayesian estimation of the spectral density of a time series. (Tentatively accepted by *J. Amer. Statist. Assoc.*).
- Dalal, S.R. and Hall, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *J. Roy. Statist. Soc. Ser. B* **45**, 278-286.
- Diaconis, P. and Freedman, D. A. (1986). On the consistency of Bayes estimates. *Ann. Statist.*, **14**, 1-26.
- Diaconis, P. and Ylvisaker, D. (1985) Quantifying prior opinion, *Bayesian statistics 2*, J. M. Bernardo, M. H. deGroot, D. V. Lindley, A. F. M. Smith (Eds), Elsevier Science Publishers B.V. (North Holland), 133-156.
- Epifani, I. (1999). Alcuni risultati sulle leggi di probabilità dei polinomi aleatori di Bernstein e loro applicazioni alla statistica bayesiana. *Tesi del Dottorato di Ricerca in Statistica Metodologica*, Università di Trento.
- Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, **90**, 577-588.
- Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. II. Wiley, New York (First Edition: 1966).
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209-230.
- Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*, H. Rizvi and J. Rustagi eds. New York: Academic Press, 287-302.
- Genovese, C.R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.*, **28**, 1105-1127.
- Ghosal, S. (2001) Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.* **29**, 1264-1280.
- Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999b). Posterior consistency of

- Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143-158.
- Ghosal, S. and van der Vaart, A.W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of Normal densities. *Ann. Statist.* **29**, 1233-1263.
- Ghosal, S. and van der Vaart, A. (2003). Posterior convergence rates of Dirichlet mixtures of normal distributions for smooth densities. *Techn. Rep. 2003-12*, Mathematical Dept., Vrije Universiteit.
- Graziani, R. and Holmes, C. (2003). Bayesian free knot polynomial splines of random order. *Studi Statistici N.75*, Istituto di Metodi Quantitativi, Università Bocconi, Milano (to appear in *Statistics and Computing*)
- Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.*, **96**, 161-173.
- Ishwaran, H., James, L.F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.*, **96**, 1316-1332.
- Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates: I Density estimation. *Ann. Statist.*, **12**, 351-357.
- Lorentz, G.G. (1953). *Bernstein polynomials*. University of Toronto Press, Toronto.
- Muliere, P. and Walker, S. (2003). A bivariate Dirichlet process. *Statist. Prob. Lett.*, **64**, 1-7.
- Petrone, S. (1999a). Random Bernstein polynomials. *Scand. J. Statist.* **26**, 373-393.
- Petrone, S. (1999b). Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.* **27**, 105-126.
- Petrone, S. and Veronese, P. (2002). Non parametric mixture priors based on an exponential random scheme. *Statistical Methods and Applications* **11**, 1-20.
- Petrone, S. and Veronese, P. (2003). Approximation of distribution functions: a constructive scheme with application in Bayesian nonparametrics. *Studi Statistici 81*, Istituto di Metodi Quantitativi, Università Bocconi, Milano.
- Petrone, S. and Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *J. Roy. Statist. Soc. B* **64**, 79-100.
- Regazzini, E. (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilità. *Giornale dell'Istituto Italiano degli Attuari*, **1-2**, 77-89.
- Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.*, **31**, 560-585.
- Regazzini, E. and Sazonov, V. V. (1999). Approximation of laws of multinomial parameters by mixtures of Dirichlet distributions with applications to Bayesian inference. *Acta Applicandae Mathematicae*, **58**, 247-264.
- Walker, S. (2004). New approaches to Bayesian consistency. *Techn. Report*, University of Bath.
- Walker, S. and Muliere, P. (1999). A characterization of a neutral to the right prior via an extension of Johnson's sufficientness postulate, *Ann. Statist.*, **27**, 589-599.
- Walker, S., Damien, P., Laud, P.W. and Smith, A.F.M. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *J. R. Stat. Soc. Ser. B*, **61**, 485-527. with independent increments. *Ann. Statist.*, **31**, 560-585.