Statistica Sinica **20** (2010), 379-404

FELLER OPERATORS AND MIXTURE PRIORS IN BAYESIAN NONPARAMETRICS

Sonia Petrone and Piero Veronese

L. Bocconi University

Abstract: Priors for Bayesian nonparametric inference on a continuous curve are often defined through approximation techniques, e.g. basis-functions expansions with random coefficients. Using constructive approximations is particularly attractive, since it may facilitate the prior elicitation. With this motivation we study a class of operators, introduced by Feller, for the constructive approximation of a bounded real function. Feller operators have a simple, probabilistic structure. We prove that, when the random elements used in their construction are chosen in the natural exponential family, they have several properties of interest in statistical applications, and can be represented as mixtures of simple probability distribution functions. As a by-product, we give some new results on the natural exponential family. Our construction offers more insights on the role of mixtures in Bayesian nonparametrics. A fairly general class of mixture priors arises, which includes continuous, countable, or finite mixtures, with kernels suggested by the approximation scheme. This allows the study of theoretical properties in a unified setting; in particular, we give results on the Kullback-Leibler property for the proposed class of mixture priors, and on the consistency of the corresponding posterior, extending results known only for specific kernels.

Key words and phrases: Bernstein polynomials, fiducial densities, Kullback-Leibler support, mixture models, natural exponential family, weak consistency.

1. Introduction

Bayesian nonparametric inference is an active research area that has grown extensively over the last years. A vast literature on discrete nonparametric priors has been developed, with applications in a large range of fields. For continuous data, flexible (nonparametric) models are commonly obtained through mixtures of distributions. The use of mixtures in Bayesian nonparametrics goes back to Ferguson (1983) and Lo (1984), who introduced mixtures with a Dirichlet process mixing distribution as priors for continuous data. Mixture models are widely used for model-based clustering and classification, or as smoothing techniques, with components having the role of kernels. Mixtures are also applied in shape constrained density estimation: if the constrained set of densities is convex, Choquet's theorem provides a representation of the density as a mixture

of known extreme measures, so that a prior on the constrained density can be obtained more easily by giving a prior on the unconstrained mixing density (see Hoff (2003)). When an exact mixture representation is not available, mixture models are used as approximations of the unknown density of the data for defining a prior with rich support. Despite the richness of contexts of application, the theory on mixture-priors does not seem fully developed yet. Properties of mixtures have been mostly studied for special cases, in particular for Gaussian kernels (Ghosal and van der Vaart (2007) and references therein), or for beta kernels (e.g., Petrone and Wasserman (2002)), and many results are confined to density estimation by mixtures. Extending the analysis to more general kernels or different estimation contexts does not seem an easy task.

This paper offers a further contribution on the role of mixtures in Bayesian nonparametric inference. We present a general class of mixture models that originates from a constructive approximation technique, due to Feller, which is a generalization of Bernstein polynomials. In fact, another popular approach in Bayesian nonparametrics is to express a flexible model (in other words, a prior with large support) for an unknown density, or more generally a random curve, by using basis-function expansions with random coefficients. However, in many cases the expansion coefficients do not have a simple interpretation so that a honest prior elicitation is difficult. It is therefore attractive to use constructive approximations, where the coefficients are directly related to the function to be approximated. For bounded functions on the unit interval, a simple, constructive approximation is provided by Bernstein polynomials. Feller (1971, Chap. VII) defines a more general approximation scheme for a bounded curve on a possibly unbounded interval. Feller operators have the same probabilistic nature as Bernstein polynomials, and indeed they are known in the approximation theory literature as Feller's probabilistic way of constructing positive approximation processes of Korovkin-type (see Altomare and Campiti (1994)). Besides their attractive constructive nature, Bernstein polynomials have several properties that are of interest in statistical applications. A first one is monotonicity preservation, which is a basic requirement for ensuring that the approximation of a probability distribution function (d.f.) is still a d.f. Also, the derivative of the Bernstein polynomial turns out to be a mixture of beta densities, and this property nicely relates polynomial approximation to mixture models and kernel methods. In Bayesian parametric inference, this property was exploited by Dalal and Hall (1983) and Diaconis and Ylvisaker (1985) for showing that a honest prior on the parameter of a Bernoulli model can always be approximated by a mixture of beta densities, that is, of conjugate priors. They extend this result to more general models, but they do not refer explicitly to Feller's operators.

In Bayesian nonparametric inference, *random* Bernstein polynomials have been studied by Petrone (1999) and applied in several contexts (see e.g., Choudhuri, Ghosal and Roy (2004)). Here, the curve to be approximated is no more deterministic, but random. The so-called Bernstein prior can be extended to inference on a random curve on a possibly unbounded interval (see e.g. the implementation in the function BDPdensity of the R-package "DPpackage", Jarra (2007)). However, this requires a preliminary transformation of the sample space into the unit interval.

We use the extension of Bernstein polynomials, provided by Feller operators, to define a class of priors for Bayesian nonparametric inference on an unknown bounded curve on a general interval in $\mathbb{R}$. Moreover, we want to maintain some desirable properties of Bernstein polynomials, in particular their connection with mixture models. We show that this is possible if the probabilistic elements used in their construction (the *random scheme*, see Section 2) are chosen in the natural exponential family (NEF). As a by-product, we give some new results for the NEF, that are of independent interest. In inference on a random d.f., our construction leads to a class of mixture models where the choice of the kernel is automatically provided by the approximation scheme, which is appropriately chosen according to the sample space. This avoids difficulties such as the so-called boundary bias effect in density estimation. The proposed class of mixture models is fairly general, including continuous, countable, or finite mixtures; thus, we have a unified framework for studying theoretical properties of mixture models in Bayesian nonparametric inference. Since we show that any d.f. can be (constructively) approximated by Feller operators, we can define a class of "mixture priors" with rich support. In particular, we discuss the Kullback-Leibler support of the proposed priors. The Kullback-Leibler property is a sufficient condition for weak consistency in Bayesian density estimation, and it is usually maintained in more refined sets of conditions for stronger asymptotic results; see for example Walker, Lijoi and Prünster (2005). Our results are a first extension to a general class of mixture models of asymptotic properties usually proved for specific (Gaussian) kernels. We focus on Bayesian nonparametric inference on a random d.f. or density in the basic framework of exchangeable data, but our construction of the prior is attractive in more general statistical applications such as multiple shrinkage estimation, or inference on a bounded random curve, e.g. a regression function.

The paper is organized in two parts. In the first part (Sections 2 and 3) we define Feller operators with exponential random scheme and we study their properties, focussing on the approximation of a d.f.. In the second part, we discuss applications in Bayesian nonparametric inference. The proposed class of nonparametric priors based on Feller operators is in Section 4. In particular,

Kullback-Leibler properties are provided in Subsection 4.2. Further statistical applications and extensions are discussed in Section 5. More technical results and all the proofs are provided in the Appendix, available on line at `http://www.stat.sinica.edu.tw/statistica`. The Appendix is organized as follows. Section A.1 contains the proofs of the results stated in Section 3. Section A.2 shows some analytical properties of the mixture kernels arising in our scheme. These properties, together with continuity properties of Feller operators provided in Section A.3, are used in Section A.4 for proving the results on the Kullback-Leibler property for mixture priors.

2. Feller Operators for a Bounded Function

Feller (1971, Chap. VII) presents the following constructive approximation procedure for real bounded functions.

Let $U : \mathbb{R} \to \mathbb{R}$ be a bounded function. Consider a family of random variables (r.v.'s) $\{Z_{k,x},\ x \in (a,b) \subseteq \mathbb{R},\ k = 1, 2, \ldots\}$ with $Z_{k,x}$ having distribution function $P_{k,x}$, expected value $E(Z_{k,x}) = \mu_k(x)$ and finite variance $Var(Z_{k,x}) = \sigma_k^2(x)$. Define

$$B_{k,U}(x) = E(U(Z_{k,x})) = \int_{-\infty}^{\infty} U(z) dP_{k,x}(z)\ , \quad x \in (a,b)\ . \tag{2.1}$$

The following theorem, which is a slight generalization of the result of Feller (1971, Chap. VII, Lemma 1), shows that $B_{k,U}$ provides an approximation of U on the interval (a,b).

Theorem 1. *If $\{Z_{k,x}, x \in (a,b), k = 1, 2, \ldots\}$ is such that, for $k \to \infty$, $\mu_k(x) \to x$ and $\sigma_k^2(x) \to 0$, then $\lim_{k\to\infty} B_{k,U}(x) = U(x)$ at any continuity point x of U, $x \in (a,b)$. If U is continuous, the convergence is uniform in every closed interval in which $\mu_k(x) \to x$ and $\sigma_k^2(x) \to 0$ uniformly.*

Roughly speaking, the value of $U(x)$ at x is approximated by a weighted average of the values $U(z)$, with weights that are more and more concentrated around x as k increases. Despite its simplicity, Theorem 1 proves useful in a variety of problems. For example, from a statistical view point, if $\{Z_{k,x}\}$ is a sequence of statistics, then it says that $U(Z_{k,x})$ is an asymptotically unbiased estimator of $U(x)$. In the approximation theory literature, it is referred as Feller's probabilistic way of constructing Korovkin-type approximations of bounded functions (see Altomare and Campiti (1994, Sec. 5.2.)). The sequence of random variables $\{Z_{k,x}\}$ (or equivalently of d.f.'s $\{P_{k,x}\}$) satisfying the conditions of Theorem 1 is called a *random scheme* for the approximation. We call $B_{k,U}$ a *Feller operator* of order k for U, with random scheme $\{P_{k,x}\}$. A simple way for defining a random scheme is to consider independent and identically distributed (i.i.d.)

r.v.'s $Y_i, i = 1, 2, \ldots$, with mean $E(Y_i) = x$ and variance $\sigma^2(x)$. Then the random variables $\{Z_{k,x} = \sum_{i=1}^{k} Y_i/k,\ k = 1, 2, \ldots\}$ provide a random scheme since $E(Z_{k,x}) = x$ and $V(Z_{k,x}) = \sigma^2(x)/k$ converges to zero.

Bernstein polynomials are a special case of this construction with $Y_i \sim$ Bernoulli with parameter x, and $kZ_{k,x}$ having a binomial distribution with parameters k and x, so that

$$B_{k,U}(x) = \sum_{j=0}^{k} U(\frac{j}{k}) \binom{k}{j} x^j (1-x)^{k-j}\ ,\ x \in (0,1). \tag{2.2}$$

It is natural to study the more general case where $P_{k,x}$ belongs to the natural exponential family. In particular, we want to explore if this generalization preserves the desirable properties of Bernstein polynomials discussed in the Introduction.

3. Feller Operators with Exponential Random Scheme

To introduce the notation, we recall some basic notions about the natural exponential family; for a more complete reference, see e.g., Brown (1986). The proofs of the results contained in this section are provided in Appendix A.1.

3.1 Preliminaries on the natural exponential family

Given a non degenerate σ-finite measure ν on the Borel sets of $\mathbb{R}$, denote with $M(\theta) = \ln \int \exp\{\theta\, x\}\nu(dx)$ the *cumulant transform* of ν, and let $\mathcal{N} = \{\theta \in \mathbb{R} : M(\theta) < \infty\}$. The real *Natural Exponential Family* (NEF), with natural parameter θ, is the family of probability measures $\mathcal{F}$ on $\mathbb{R}$ whose densities, with respect to (w.r.t.) ν, are of the form

$$p_\theta(y) = \exp\{\theta\, y - M(\theta)\}, \quad \theta \in \Theta, \tag{3.1}$$

where Θ is the interior of $\mathcal{N}$. The set Θ is supposed non-empty and open. The family is *regular* if $\mathcal{N}$ is open, i.e., $\mathcal{N} = \Theta$. In the sequel we only consider regular NEF's. A NEF can be alternatively parametrized in the mean parameter $\mu = \mu(\theta) = dM(\theta)/d\theta$, since $\mu(\cdot)$ is a one-to-one transformation from Θ onto $\Omega = \mu(\Theta)$. Notice that if a NEF is regular, then $\Omega = (a, b)$, where (a, b) is the interior of the convex support of ν (by convex support we mean the smallest closed interval containing the support of the measure ν). The variance of a NEF is given by $d^2M(\theta)/d\theta^2$. As a function of μ, it is called *variance function* and is denoted by $V(\mu)$. It is convenient to work with a NEF parametrized by the mean parameter μ, denoted as P_μ.

A NEF is closed under the sum operation: if $Y_1, \ldots, Y_k$ are random variables i.i.d. according to the density (3.1), then $S_{k,\mu} = \sum_{i=1}^{k} Y_i$ is distributed according

to a NEF with density, w.r.t. the convolution measure ν_k^*, given by

$$p^*_{k,\mu}(s) = \exp\{\theta(\mu)\, s - kM(\theta(\mu))\}, \quad \theta \in \Theta. \tag{3.2}$$

The average $Z_{k,\mu} = S_{k,\mu}/k$ has density belonging to a NEF, w.r.t. the appropriate measure, with mean parameter $E(Z_{k,\mu}) = \mu$ and variance $Var(Z_{k,\mu}) = V(\mu)/k$. In fact, a NEF is closed under the more general operation of *power convolution*. More precisely, given a NEF $\mathcal{F}$, let k be a positive real number and suppose there exists a non degenerate measure ν_k^* such that its cumulant transform is given by $kM(\theta)$ for each $\theta \in \Theta$. Then the NEF generated by ν_k^* is called the k-th *convolution power* of $\mathcal{F}$ and is denoted by $\mathcal{F}^{*k}$. The set Λ of all real positive k for which one can construct a convolution power of $\mathcal{F}$ is called a *Jorgensen set* (Jorgensen (1987)); clearly it always contains all positive integers. The density of a NEF $\mathcal{F}^{*k}$, with $k \in \Lambda$, is formally equivalent to (3.2). If $S_{k,\mu}$ has distribution in $\mathcal{F}^{*k}$, then $Z_{k,\mu} = S_{k,\mu}/k$ is still distributed according to a NEF, with density given by $p_{k,\,\mu}(z) = \exp\{k[\theta(\mu)z - M(\theta(\mu))]\}$ w.r.t. the appropriate measure ν_k. Clearly, $E(Z_{k,\mu}) = \mu$ and $Var(Z_{k,\mu}) = V(\mu)/k$. Notice that the convex support of $Z_{k,\mu}$ concides with that of the NEF $\mathcal{F}$.

3.2. Feller operators with ERS

We exploit the property of closeness under power convolution of a NEF to define a general *exponential random scheme* (ERS) for Feller operators. Let U be a real bounded function defined on $\mathbb{R}$. For the approximation of U on an interval (a, b), one can take a NEF $\mathcal{F}$ having mean x defined on the same interval (a, b), and consider the r.v.'s $k\, Z_{k,x}$ distributed according to the convolution power $\mathcal{F}^{*k}$ of $\mathcal{F}$, $k \in \Lambda$. The family $\{Z_{k,x}, x \in (a, b), k \in \Lambda\}$ provides a random scheme, since $E(Z_{k,x}) = x$ and $Var(Z_{k,x}) = V(x)/k$ converges to zero as $k \to \infty$.

Definition 1. Given a NEF $\mathcal{F}$ with mean parameter $x \in (a, b)$, let $kZ_{k,x}$ be a r.v. with distribution in the k-convolution power of $\mathcal{F}$, for $k \in \Lambda$. We call

$$B_{k,U}(x) = E(U(Z_{k,x})), \quad x \in (a, b)$$

a *Feller operator with ERS* $\{Z_{k,x},\, x \in (a, b),\, k \in \Lambda\}$, for the approximation of the function U on the interval (a, b). The ERS is said to be continuous if the r.v.'s $Z_{k,x}$ are absolutely continuous, and discrete if the $Z_{k,x}$'s are discrete.

If $Z_{k,x} \sim P_{k,x}$, we will also denote the ERS by $\{P_{k,x},\, x \in (a, b),\, k \in \Lambda\}$.

Example 1 (discrete ERS).

Binomial random scheme (Bernstein polynomials). If U is a bounded function on the unit interval $[0, 1]$, the Feller operator with ERS for U may be based on a NEF with mean parameter space $(0, 1)$. A natural choice is a Bernoulli distribution

with parameter x. In this case $\Lambda = \{1, 2, \ldots\}$ and $\mathcal{F}^{*k}$ is the binomial family with parameters (k, x), $\mathrm{bi}(k, x)$. As already noted, the Feller operator with Binomial random scheme $\{Z_{k,x},\ x \in (0,1),\ k = 0, 1, 2 \ldots\}$, where $kZ_{k,x} \sim \mathrm{bi}(k, x)$, is the Bernstein polynomial (2.2) of order k for U.

Poisson random scheme. If the function U to be approximated is defined on $(0, \infty)$, the ERS may be based on a NEF with mean parameter space $(0, \infty)$. One possible choice is a Poisson family with mean x, $\mathrm{Po}(x)$. In this case, $\Lambda = (0, \infty)$ and $\mathcal{F}^{*k}$ is the Poisson family with mean kx. Therefore a Feller operator with Poisson random scheme $\{Z_{k,x},\ x \in (0, \infty),\ k > 0\}$, where $kZ_{k,x} \sim \mathrm{Po}(kx)$, is defined as

$$B_{k,U}(x) = E(U(Z_{k,x})) = \sum_{j=0}^{\infty} U(\frac{j}{k})\,(kx)^j\,\frac{e^{-kx}}{j!}, \quad x > 0.$$

Example 2 (continuous ERS).

Gaussian random scheme. A bounded function U on $(-\infty, \infty)$ can be approximated by a Feller operator with ERS having mean in $(-\infty, \infty)$. A natural choice is a Gaussian family with mean x and fixed variance σ^2, $N(x, \sigma^2)$. In this case, $\Lambda = (0, \infty)$ and $\mathcal{F}^{*k}$ is the family $N(kx, k\sigma^2)$ with $k > 0$. We obtain a Gaussian random scheme $\{Z_{k,x} \sim \mathrm{N}(x, \sigma^2/k),\ x \in \mathbb{R},\ k > 0\}$, which gives

$$B_{k,U}(x) = E(U(Z_{k,x})) = \int_{-\infty}^{\infty} U(z)\ \mathrm{N}(z; x, \frac{\sigma^2}{k})\,dz, \tag{3.3}$$

where $\mathrm{N}(\cdot; \mu, \sigma^2)$ denotes the density function of a $\mathrm{N}(\mu, \sigma^2)$.

Gamma random scheme. For a bounded function U on $(0, \infty)$, we used in Example 1 a discrete ERS, based on the Poisson distribution. A continuous ERS can be obtained by considering the (negative) exponential distribution with mean x. Then $\Lambda = (0, \infty)$ and the convolution power is the gamma family with shape parameter $k > 0$ and mean kx, $\mathrm{Ga}(k, 1/x)$. Thus, we have a Gamma random scheme $\{Z_{k,x} \sim \mathrm{Ga}(k, k/x),\ x \in (0, \infty),\ k > 0\}$ and

$$B_{k,U}(x) = E(U(Z_{k,x})) = \int_0^{\infty} U(z)\mathrm{Ga}(z; k, \frac{k}{x})dz,\ \ x > 0,$$

where $\mathrm{Ga}(\cdot; \alpha, \beta)$ denotes the density function of a $\mathrm{Ga}(\alpha, \beta)$.

An attractive aspect of Feller operators with ERS is that they preserve some desirable properties of Bernstein polynomials, such as monotonicity and their connection with kernel or mixture approximation. For a NEF, $P_{k,x}(t)$ is a non-increasing function in x for any fixed t (see Lehmann (1959, Chap. 3, Lemma 2)), therefore $Z_{k,x'}$ stochastically dominates $Z_{k,x}$ if $x' > x$. It follows that, if U

is monotone (non-decreasing), $E(U(Z_{k,x'}) \geq E(U(Z_{k,x}))$, i.e., $B_{k,x}$ is monotone non-decreasing. Thus we have the following

Proposition 1. *A Feller operator with ERS for a monotone (not constant) bounded function U is monotone.*

Note that if U is monotone (non-decreasing), it can be written as $U(x) = \int_{-\infty}^{x} dU(t)$, for $a < x < b$, and we can re-express the Feller operator $B_{k,U}(x)$ in Theorem 1, interchanging the role of $P_{k,x}$ and U, as

$$\begin{aligned} B_{k,U}(x) &= \int_{-\infty}^{\infty} U(z) dP_{k,x}(z) = \int_{-\infty}^{\infty} \int_{(-\infty,z]} dU(t) dP_{k,x}(z) \\ &= \int_{-\infty}^{\infty} P_{k,x}([t,\infty)) dU(t). \end{aligned} \tag{3.4}$$

Roughly speaking, the right side of (3.4) gives a representation of Feller operators in terms of "kernels" $H_k(x;t) = P_{k,x}([t,\infty))$, weighted by the function U. However, if we do not impose restrictions on $P_{k,x}$, the functions $H_k(\cdot;t)$ are too general and not easily interpretable; moreover, monotonicity preservation may fail. Instead, if $P_{k,x}$ is chosen in the NEF, it can be shown that $H_k(\cdot;t)$ has several attractive properties. In particular, we will show that it can be seen as a probability distribution function. Therefore, when the function U of interest is a d.f., Feller operators with ERS have a representation as mixtures of d.f.'s, related to the NEF used in the ERS. Thus, properties proved for Feller operators with ERS hold for a fairly general class of mixture models, which is of particular interest in statistical applications. To these aims, we need some technical results on the NEF, that are given in the following subsections.

3.3. Some new results for the NEF

The properties of the NEF presented in this section, besides being useful for studying Feller operators with ERS, are of independent interest. To our knowledge, the following lemma is only partially known.

Lemma 1. *Let P_μ be the d.f. of a real NEF with mean parameter $\mu \in (a,b)$, $-\infty \leq a < b \leq \infty$, and dominating measure ν. Then*

(i) $\lim_{\mu \to a^+} P_\mu(y) = 1$ *and* $\lim_{\mu \to b^-} P_\mu(y) = 0, \quad \forall y \in (a,b)$;

(ii) *if $a > -\infty$ and $\nu\{a\} > 0$, then*
$\lim_{\mu \to a^+} p_\mu(a)\nu\{a\} = 1$ *and* $\lim_{\mu \to b^-} p_\mu(a)\nu\{a\} = 0$;

(iii) *if $b < \infty$ and $\nu\{b\} > 0$ then*
$\lim_{\mu \to a^+} p_\mu(b)\nu\{b\} = 0$ *and* $\lim_{\mu \to b^-} p_\mu(b)\nu\{b\} = 1$.

Although the lemma is stated in terms of the mean parameter of the NEF, it can be rephrased for the natural parameter, since there is a one-to-one positive monotone correspondence between the two parametrizations.

As shown in Subsection 3.1, if P_μ belongs to a NEF, so does its convolution power $P_{k,\mu}$. Motivated by (3.4), we study the properties of $P_{k,\mu}([z,\infty))$. The next theorem shows that $P_{k,\mu}([z,\infty))$, as a function of μ and appropriately completed on the real line, is a d.f.. Since we study the function $P_{k,(\cdot)}([z,\infty))$ on the real line and not only on the interval (a,b), we denote its argument with a generic symbol x rather than with μ. For $z \in \mathbb{R}$ and $k \in \Lambda$, define

$$H_k(x;z) = \begin{cases} 0 & x < a \\ \lim_{x \to a^+} P_{k,x}([z,\infty)) & x = a \\ P_{k,x}([z,\infty)) & a < x < b \\ 1 & x \geq b. \end{cases} \tag{3.5}$$

Theorem 2. *For any $z \in \mathbb{R}$ and $k \in \Lambda$, the function $H_k(\cdot;z)$ defined in* (3.5) *is a d.f.. In particular*:

for $a < z < b$, $H_k(\cdot;z)$ is an absolutely continuous d.f. with density

$$h_k(x;z) = \frac{k}{V(x)} \left(\int_{[z,\infty)} (t-x) dP_{k,x}(t) \right) I_{(a,b)}(x), \tag{3.6}$$

where $I_{(a,b)}(x)$ is the indicator function of (a,b);

for $z \leq a$ (with a finite), $H_k(\cdot;z)$ is a d.f. degenerate on a;

for $z = b$ (with b finite), $H_k(\cdot;z)$ is an absolutely continuous d.f. with support $[a,b]$ and density $(k/V(x))(b-x)p_{k,x}(b)\nu_k(b)$ if the d.f. $P_{k,x}$ is discrete, and it is degenerate on b if $P_{k,x}$ is continuous;

for $z > b$, $H_k(\cdot;z)$ is a d.f. degenerate on b.

The proof is based on the property of monotonicity of $P_{k,(\cdot)}(t)$; furthermore, it uses Lemma 1 to establish the appropriate behavior of $H_k(\cdot;z)$ at the extremes of the interval (a,b).

Remark. (*Fiducial distributions*). In our context, Theorem 2 is a technical result, but it is interesting in itself since the density $h_k(\cdot;z)$ is related to the *fiducial density* (Fisher (1973)) of the mean parameter of the NEF. The fiducial density for the parameter μ of a *continuous* family P_μ is

$$f(\mu;z) = -\frac{d}{d\mu} P_\mu((-\infty,z]), \tag{3.7}$$

provided $f(\mu; z)$ has the properties of a probability density function. If P_μ belongs to a continuous NEF,

$$f(\mu; z) = \frac{d}{d\mu}(1 - P_\mu((-\infty, z])) = \frac{d}{d\mu}P_\mu((z, \infty)) = \frac{d}{d\mu}P_\mu([z, \infty)) = h(\mu; z),$$

and Theorem 2 shows that $h(\cdot; z)$ is a density for any $z \in (a, b)$. Therefore, $h(\cdot; z)$ is interpretable as the fiducial density for μ. The notion of fiducial probability has generated a long controversy about its interpretation and possible extensions to *discrete* models; see e.g., Zabell (1992). Here we simply note that, using Theorem 2, f at (3.7) is still the fiducial density for a discrete NEF. In fact, for a discrete P_μ with support points $\{a = z_1 < z_2 < \cdots < z_N = b\}$, $N \leq \infty$, we have

$$f(\mu; z_i) = \frac{d}{d\mu}P_\mu((z_i, \infty)) = \frac{d}{d\mu}P_\mu([z_{i+1}, \infty)) = h(\mu; z_{i+1}), \quad i = 1, \ldots, N-1,$$

where $h(\mu; z_{i+1})$ is a probability density function for $i = 1, \ldots, N-2$. Clearly, since $P_\mu(b) = 1$ for all μ, a fiducial density cannot be defined for $z = b$. Notice that we can derive a fiducial density also for the natural parameter of the NEF, by simply restating Theorem 2 in terms of the natural parameter.

Example 1 ctd.
Binomial random scheme. If $kZ_{k,x} \sim \text{bi}(k, x)$, $k = 1, 2, \ldots$, by the known relationship between the retro-cumulative sum of binomial weights and the incomplete beta function, we have

$$P_{k,x}([z, \infty)) = \int_0^x \frac{k!}{(j-1)!(k-j)!}\, t^{j-1}(1-t)^{k-j}dt, \ \frac{j-1}{k} < z \leq \frac{j}{k}, \ j = 1, \ldots, k.$$

Thus, $h_k(x; z)$ is a beta density with parameters $(j, k - j + 1)$ for $(j-1)/k < z \leq j/k$. It can be interpreted as a fiducial density for the mean parameter of a Bernoulli distribution, when $kz = j - 1$ "successes" have been observed, $j = 1, \ldots, k$. Interestingly, this density was found by Heike, Târcolea, Demetrescu and Târcolea (2003) from a completely different approach.
Poisson random scheme. If $kZ_{k,x} \sim \text{Po}(kx)$, $k > 0$, then (3.6) gives

$$h_k(x; z) = \frac{k}{x}\sum_{t=(j+1)/k}^{\infty}(t - x)\frac{(kx)^{kt}\exp\{-kx\}}{kt!} = \frac{1}{x}\sum_{u=0}^{j}(kx - u)\frac{(kx)^u\exp\{-kx\}}{u!}$$
$$= k^{j+1}x^j\frac{\exp\{-kx\}}{j!} \qquad \frac{j}{k} < z \leq \frac{j+1}{k}, \quad j = 0, 1, \ldots.$$

Thus, $h_k(x; z)$ is a Gamma density with parameters $(j + 1, k)$ for $j/k < z \leq (j+1)/k$. It can be regarded as a fiducial density for the mean of the Poisson distribution given $kz = j$ events.

Example 2 ctd.
Gaussian random scheme. In this case $Z_{k,x} \sim N(x, \sigma^2/k)$, $k > 0$. Then $H_k(x;z) = 1 - P_{k,x}(z) = 1 - \Phi(\sqrt{k}(z-x)/\sigma) = \Phi(\sqrt{k}(x-z)/\sigma)$, where Φ denotes the standard normal d.f.. Thus $h_k(x;z)$ is $N(x; z, \sigma^2/k)$, which is the well-known fiducial density for the mean parameter of the Gaussian distribution.
Gamma random scheme. For $Z_{k,x} \sim Ga(k, k/x), k > 0$ we have

$$H_{k,x}(z) = 1 - P_{k,x}(z) = 1 - (\frac{k}{x})^k \frac{1}{\Gamma(k)} \int_0^z t^{k-1} \exp\{-(\frac{k}{x})t\} dt \, .$$

Deriving $H_{k,x}$ w.r.t. x we obtain

$$h_k(x;z) = \frac{d}{dx} H_k(x;z) = \frac{(kz)^k}{\Gamma(k)} x^{-k-1} \exp\left\{-kz\frac{1}{x}\right\}, \tag{3.8}$$

an inverse-gamma density with parameters (k, kz). This is the fiducial density for the mean parameter of the Gamma distribution.

In the above examples, the density $h_k(x;z)$ (i.e., the fiducial density) belongs to the conjugate family of the NEF used in the ERS. However, this is not always the case. A counter example is obtained starting from the regular NEF in Example 1.1 of Consonni and Veronese (1992).

3.4. Approximation of a distribution function

An interesting case in statistical applications occurs when the function to be approximated is a probability distribution function. Let U be a d.f. with convex support $E \subseteq \mathbb{R}$. Even if U is defined on the whole real line, the interval on which it is interesting to study its approximation is clearly E. Therefore, it seems natural to consider Feller operators $B_{k,U}$ based on a random scheme $P_{k,x}$ with parameter $x \in (a,b) = E$. The choice $(a,b) \supset E$ will not be treated explicitly, but the following results can be extended to that case, too.

Complete $B_{k,U}$ on the whole real line as

$$B_{k,U}(x) = \begin{cases} 0 & x < a \\ U(a) & x = a \\ \int U(z)\, dP_{k,x}(z) & a < x < b \\ 1 & x \geq b, \end{cases} \tag{3.9}$$

with the obvious changes when $a = -\infty$, with $U(-\infty) = \lim_{x \to -\infty} U(x)$, and when $b = \infty$. For a general random scheme, $B_{k,U}$ does not preserve the shape properties of a d.f.. However, with an ERS, monotonicity is preserved by Proposition 1; furthermore, $B_{k,U}$ becomes a mixture of d.f.'s with kernel related to the ERS, as shown in the following theorem.

Theorem 3. *If U is a d.f. with convex support $[a, b]$, the Feller operator $B_{k,U}$ defined by* (3.9) *is still a d.f. with support $[a, b]$. More precisely, it is a mixture of the d.f.'s $\{H_k(\cdot; z), -\infty < z < \infty\}$, with mixing distribution U, i.e.,*

$$B_{k,U}(x) = \int_{-\infty}^{\infty} H_k(x; z) dU(z). \tag{3.10}$$

The d.f. $B_{k,U}$ has mass $U(a)$ concentrated on a (if a is finite) and it has mass $U(\{b\}) = 1 - \lim_{t \to b^-} U(t)$ concentrated on b (if b is finite) only if $\nu\{b\} = 0$. For $x \in (a, b)$, $B_{k,U}(x)$ is continuous, with derivative

$$b_{k,U}(x) = \int h_k(x; z) dU(z), \tag{3.11}$$

where the function $h_k(x; z)$ is defined in (3.6).

By Theorem 1, the d.f. $B_{k,U}$ converges weakly to the d.f. U as $k \to \infty$. Thus, we have shown that any d.f. on $\mathbb{R}$ can be constructively weakly approximated by mixtures of d.f.'s, arising as Feller operators with ERS. If $U(a) = 0$ and $U(\{b\}) = 0$, the derivative $b_{k,U}$ of $B_{k,U}$ is a probability density function, represented as a mixture of the kernels $h_k(x; z)$. As a function of z, the kernel $h_k(x; z)$ is continuous if the ERS is continuous, while it is piecewise constant for a discrete ERS (see Example 1). Thus, when the ERS is discrete with support points $\{z_{1,k}, z_{2,k}, \ldots, z_{N_k,k}\}$, $b_{k,U}$ reduces to a finite or countable mixture

$$b_{k,U}(x) = \sum_{i=1}^{N_k - 1} w_{i,k}\, h_k(x; z_{i,k}),$$

with simple components having k as the only unknown parameter, and mixing weights $w_{i,k} = U(z_{i+1,k}) - U(z_{i,k})$. In this case one can construct the approximation $B_{k,U}$ even if U is known only at the points $U(z_{j,k})$.

3.5. Approximation of a density

We have shown that any d.f. U can be constructively approximated, in the sense of weak convergence, by Feller operators. With further assumptions, stronger convergence results can be proved. Interestingly, if U is absolutely continuous with a continuous and bounded derivative u, the density $b_{k,U}$ of $B_{k,U}$ turns out to be a Feller-type approximation of the density u. This result makes use of the following lemma.

Lemma 2. *Let $\{P_{k,x},\ x \in (a, b),\ k \in \Lambda\}$ be a continuous or discrete ERS. Then, for each $x \in (a, b)$, the following hold.*

1. *The function* $h_k(x;z)$ *defined by* (3.6) *is a probability density in* z *w.r.t. Lebesgue measure, with support* $[a,b]$.
2. *If* $Z^*_{k,x}$ *has density* $h_k(x;\cdot)$, *then*

$$E(Z^*_{k,x}) = \frac{1}{2k}V'(x) + x \tag{3.12}$$

$$Var(Z^*_{k,x}) = \frac{1}{12\,k^2}V'(x)^2 + \frac{1}{3\,k^2}V''(x)V(x) + \frac{1}{k}V(x). \tag{3.13}$$

3. *If the ERS is continuous, then* $h_k(x;z)$ *is a unimodal continuous density in* z, *with a maximum at* $z = x$. *If the ERS is discrete, with support points* $\{z_{1,k}, z_{2,k}, \ldots, z_{N_k,k}\}$, *then* $h_k(x;z)$ *is a piecewise constant density function in* z, *with modal interval* $(z_{i,k}, z_{i+1,k}]$ *if* $x \in (z_{i,k}, z_{i+1,k}]$.

Lemma 2 shows that $h_k(x;\cdot)$ is a probability density function, and that the sequence of r.v.'s $Z^*_{k,x} \sim h_k(x;\cdot)$ has the properties of a random scheme, since $E(Z^*_{k,x}) \to x$ and $Var(Z^*_{k,x}) \to 0$ as $k \to \infty$. Therefore, $b_{k,U}$ can be written as

$$b_{k,U}(x) = \int_{-\infty}^{\infty} u(z)\; h_k(x;z)dz = E(u(Z^*_{k,x})),$$

that is, as a Feller-type approximation of the density u of U. Thus, if u is continuous and bounded, $b_{k,U}$ converges pointwise to u by Theorem 1. By Scheffé Theorem, this implies that $B_{k,U}$ converges to U in total variation. Thus we have proved the following.

Theorem 4. *Let* $B_{k,U}$ *be a Feller operator with continuous or discrete ERS for an absolutely continuous d.f.* U *with support* $[a,b]$, *having bounded and continuous density* u. *Then, as* $k \to \infty$, $b_{k,U}(x)$ *converges pointwise to* $u(x)$ *and* $B_{k,U}$ *converges to* U *in total variation.*

Example 1 ctd.
Mixture of Beta distributions. For approximating a d.f. U on $[0,1]$, one can use a completed Bernstein polynomial of order k, $(k = 1, 2, \ldots)$

$$B_{k,U}(x) = \begin{cases} 0 & x \le 0 \\ U(0) & x = 0 \\ E(U(Z_{k,x})) = \sum_{j=0}^{k} U(\frac{j}{k})\binom{k}{j}x^j(1-x)^{k-j} & 0 < x < 1 \\ 1 & x \ge 1. \end{cases}$$

The d.f. $B_{k,U}$ has mass $U(0)$ concentrated on zero, and it is continuous for $0 < x < 1$, with derivative

$$b_{k,U}(x) = \int h_k(x;z)dU(z) = \sum_{j=1}^{k} w_{j,k}\ \mathrm{be}(x; j, k-j+1),$$

a linear combination of beta densities with parameters $(j, k-j+1)$ and weights $w_{j,k} = U(j/k) - U((j-1)/k)$. If $U(0) = 0$, then $\sum_{j=1}^{k} w_{j,k} = 1$ and $b_{k,U}$ is a mixture of beta densities. Note that $h_k(x; z)$ can also be read as a piecewise constant density in z, according to Lemma 2.

Mixture of Gamma distributions. For approximating a d.f. U with support $[0, \infty)$, one can use a completed Feller operator with Poisson random scheme

$$B_{k,U}(x) = \begin{cases} 0 & x \le 0 \\ U(0) & x = 0 \\ E(U(Z_{k,x})) = \sum_{j=0}^{\infty} U(\frac{j}{k})(kx)^j \frac{e^{-kx}}{j!} & x > 0. \end{cases}$$

The d.f. $B_{k,U}$ has mass $U(0)$ concentrated on zero, and it is continuous for $x > 0$, with derivative

$$b_{k,U}(x) = \sum_{j=1}^{\infty} w_{j,k} \mathrm{Ga}(x; j, k),$$

a countable linear combination of gamma densities with parameters (j, k) and mixing weights $w_{j,k} = U(j/k) - U((j-1)/k)$.

Example 2 ctd.

Mixtures of Gaussians. If U is a d.f. with support $(-\infty, \infty)$, the Feller operator with Gaussian random scheme (3.3) is an absolutely continuous d.f. with density

$$b_{k,U}(x) = \int_{-\infty}^{\infty} h_k(x; z) dU(z) = \int N(x; z, \frac{\sigma^2}{k}) dU(z),$$

a location mixture of Gaussian densities. In this case it is immediate to see (according to Lemma 2) that $h_k(x; z)$ is a density in z, namely a $N(z; x, \sigma^2/k)$. If U has a continuous and bounded density u, then

$$b_{k,U}(x) = E(u(Z^*_{k,x})) = \int_{-\infty}^{\infty} u(z)\ \mathrm{N}(z; x, \frac{\sigma^2}{k}) dz$$

converges to $u(x)$ for $k \to \infty$.

Mixtures of Inverse-gamma. If U is a d.f. with support $[0, \infty)$, we can use a completed Feller operator with Gamma random scheme

$$B_{k,U}(x) = \begin{cases} 0 & x \le 0 \\ E(U(Z_{k,x})) = \int_0^{\infty} U(z) \mathrm{Ga}(z; k, \frac{k}{x}) dz, & x > 0 \end{cases}$$

The derivative of $B_{k,U}$ is

$$b_{k,U}(x) = \int_0^{\infty} \frac{(kz)^k}{\Gamma(k)} x^{-k-1} e^{-kz/x} dU(z)\ , \tag{3.14}$$

a mixture of Inverse-gamma densities. Notice that $h_k(x,z)$, as a function of z, is a $\mathrm{Ga}(z; k+1, k/x)$ density. If U has continuous and bounded density u, then

$$b_{k,U}(x) = E(Z^*_{k,x}) = \int_0^\infty u(z)\frac{(k/x)^{k+1}}{\Gamma(k+1)} z^k e^{-(k/x)z} dz\ , \quad x > 0,$$

converges to $u(x)$ for $k \to \infty$.

Remark. (*Kernel interpretation*). One of the advantages of our procedure stems from the automatic choice of the kernel which is suitable for the specific problem under study. In statistical applications of kernel methods, typical kernels are in fact Gaussian densities or, more generally, densities of the form $q_k(x-\ell)$, where ℓ is a location parameter and k a dispersion or smoothing parameter. However, if the support of the density to be approximated is restricted to a subset (a,b) of $\mathbb{R}$, the choice of a kernel defined on $\mathbb{R}$ requires boundary correction methods, such as "reflection techniques", for improving the quality of the approximation at the frontier (see e.g., Jones and Foster (1996)). Clearly, these difficulties are attenuated if k is large, since the kernels become very concentrated. Our kernel $h_k(x;z)$ does not suffer from boundary bias since, by construction, it has the same support of the density to be approximated. In fact, it is quite close to a Gaussian density for large values of k. Roughly speaking, $h_k(x;z)$ is the density corresponding to the d.f. $H_k(x;z) = 1 - P_{k,x}(z)$ (see Theorem 2). Since $P_{k,x}$ is the distribution of the "average" $Z_{k,x}$ of i.i.d. random variables with mean x, the Central Limit Theorem implies that $P_{k,x}(z)$ is close, for large k, to the Gaussian d.f. $\Phi((z-x)/\sqrt{V(x)/k})$, which in turn can be approximated by $\Phi((z-x)/\sqrt{V(z)/k})$, being $Z_{k,x} \approx x$ when k is large. Therefore $H_k(x;z) = 1 - P_{k,x}(z) \approx \Phi((x-z)/\sqrt{V(z)/k})$. As $k \to \infty$, the kernel $h_k(x;z)$ becomes more and more concentrated around x and, in this sense k has the role of a smoothing parameter. In fact, $h_k(x;z)$ can be quite different from a Gaussian density for small values of k. Consider for example the approximation of a density on $(0,\infty)$. In Example 2, we used Feller operators with a Gamma random scheme, leading to the Inverse-gamma kernel Inv-$\mathrm{Ga}(k, kz)$ expressed by (3.8). In this case, z represents a scale parameter and the Feller operator (3.14) is a scale mixture of Inverse-gamma densities. However, if we make the usual logarithmic change of variables to transform the kernel from $\mathbb{R}^+$ to $\mathbb{R}$, it becomes a location density. More specifically, let $X \sim$ Inv-$\mathrm{Ga}(k, kz)$ and consider $W = log(X)$. Letting $\ell = \log(z)$, we obtain

$$q_k(w-\ell) = \frac{k^k}{\Gamma(k)} \exp\left\{-k\left[(w-\ell) + e^{-(w-\ell)}\right]\right\}, \quad w \in \mathbb{R},\ \ell \in \mathbb{R}.$$

The kernel $q_k(t)$ is plotted in Figure 3.1. for various choices of k. It shows a slight asymmetry (with a longer right tail) which rapidly disappears as k increases, as expected.

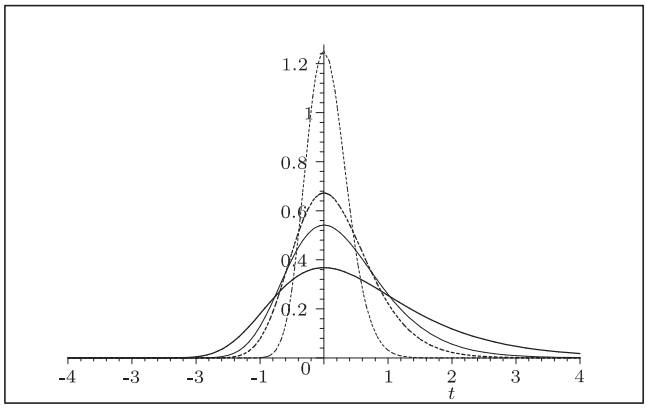

Figure 3.1. Kernels (transformed on $\mathbb{R}$) for the Gamma random scheme, for varying choices of the smoothing parameter k ($k = 1$: solid thick line, $k = 2$: solid thin line, $k = 3$: dotted thick line, $k = 10$: dotted thick line).

4. Applications in Bayesian Nonparametric Inference

In the previous sections, the curve U to be approximated was deterministic. A natural extension is to consider the approximation of a *random* curve by Feller operators. If U is random, then $B_{k,U}$ will be a random curve whose probability law can be of interest as a prior in Bayesian nonparametric inference. Again, even if Feller operators can be used for Bayesian inference on a general bounded random curve (a regression curve, a hazard function etc.), we focus on inference on a random d.f..

More precisely, let U be a random d.f. defined on a measurable space $(\Omega, \sigma(\Omega))$ with values in the space $\Delta = \Delta(E)$ of all the d.f.'s with convex support $E \subseteq \mathbb{R}$, equipped with the σ-field $\mathcal{F}$ generated by the topology of weak convergence. Consider the completed Feller operator $B_{k,U}$ with ERS $\{P_{k,x},\ x \in (a,b) \supseteq E,\ k \in \Lambda\}$. By Theorem 3, $B_{k,U}$ is a d.f. in Δ, and one can easily show that the map $: U \mapsto B_{k,U} = \int H_k(x;z)dU(z)$ is $\mathcal{F}$-measurable. Therefore, $B_{k,U}$ is a random d.f., with probability law induced by the probability law of U. An extension of Feller's Theorem 1 to random d.f.'s shows that any random d.f. can be constructively weakly approximated by random Feller operators with ERS. The proof is similar to that of Theorem 5 in Petrone (1999).

Proposition 2. *If U is a random d.f., $(B_{k,U})$ provides a sequence of random d.f.'s that converges in distribution to U for $k \to \infty$.*

This construction can be extended to the case of a random order k of the approximation. Then $B_{k,U}$ has a probability law, which we denote by π_B, induced by the joint distribution π of (k, U). Since, as shown in the previous section, $B_{k,U}$

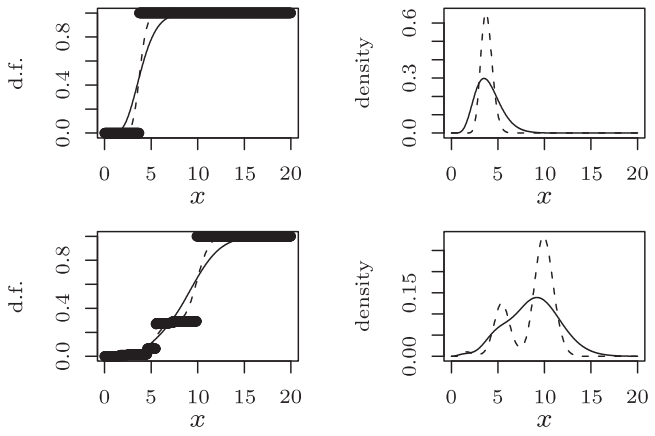

Figure 4.2. Left panels: one realization U (dark) from a $DP(\alpha U_0)$, with U_0=Gamma$(2, 0.5)$ and $\alpha = 0.05$ (in the first row) or $\alpha = 2$ (second row), together with its approximation $B_{k,U}$ with Poisson random scheme, plotted for varying values of k ($k = 2$ (solid line) and $k = 10$ (dashed)). The right panels show the corresponding densities $b_{k,U}$.

can be represented as a mixture, we refer to π_B as a "Feller prior", or "mixture prior", with ERS $\{P_{k,x},\, x \in (a, b),\, k \in \Lambda\}$ and parameters (k, U).

4.1. Inference for exchangeable sequences

The probability law of $B_{k,U}$ can play the role of prior in Bayesian nonparametric inference on a bounded random function. We focus on the basic case of inference for exchangeable data, but the proposed class of priors is attractive in more general contexts, as we will discuss in Section 5.

A Feller prior may be used as a constructive smoothing of a discrete, nonparametric prior. Consider a sequence of continuous, exchangeable r.v.'s $(X_i, i \geq 1)$, and suppose we choose a Feller prior as their de Finetti measure. This is equivalent to assuming that, given k and U, the X_i's are i.i.d. according to $B_{k,U}$, with a prior on (k, U). That is, the prior is assigned hierarchically, by first giving a prior on U and then modeling the distribution of the data as $B_{k,U}$, with the prior on k controlling how close the shape of $B_{k,U}$ is to that of U. This "two stage" specification should simplify prior elicitation. One may find it easier to start with a simple, discrete sketch U of the unknown d.f. (using on U one of the familiar discrete nonparametric priors available in the literature), and subsequently transform it into a prior on the continuous d.f. of the data, using some form of smoothing of U. A smoothing via Feller operators is attractive since their constructive nature allows one to directly relate the shape of U to that of the model $B_{k,U}$. Suppose for example that U is given a Dirichlet process prior

with parameters α, U_0, denoted as $U \sim DP(\alpha U_0)$. Then jumps of U reflect into modes of $B_{k,U}$, unless k is so small to smooth them. The joint effects of k and of discontinuities in U are illustrated in Figure 1. We plot one realization of U and its approximation $B_{k,U}$ with the corresponding density, for varying values of α and k; U_0 is Gamma$(2, 0.5)$ and $B_{k,U}$ is based on a Poisson random scheme. For the Dirichlet process, the parameter α regulates both the prior uncertainty and the number of support points with non-negligible probability mass. If α is close to zero ($\alpha = 0.05$ in the first row), U concentrates most of the unit mass on a single point, and consequently its smoothing $B_{k,U}$ is unimodal. The smoothing parameter k gives a further degree of freedom. A more vague prior belief on a unimodal density can be expressed by choosing a less extreme value of α ($\alpha = 2$ in the second row) with a prior on k concentrated on small values; while large values of k imply a multimodal density.

A further motivation for the two-stage prior specification arises from a predictive approach. We illustrate this issue for the Dirichlet process, but similar remarks hold for more general discrete priors. The Dirichlet process prior $DP(\alpha U_0)$ is characterized by the predictive rules: $P(X_1 \le x) = U_0(x)$ and

$$P(X_{n+1} \le x \mid x_1, \dots, x_n) = \frac{\alpha}{\alpha + n} U_0(x) + \frac{n}{\alpha + n} F_n(x), \quad n > 1, \tag{4.1}$$

where $F_n(\cdot) = \sum_{i=1}^n \delta_{x_i}(\cdot)/n$ is the empirical d.f. of $x_1, \dots, x_n$. Therefore, for any measurable set A, $P(X_{n+1} \in A \mid x_1, \dots, x_n)$ depends on the data only through n and the number of observations in A. However, with continuous data, it is natural to also exploit the information provided by the observations in sets close to A, thus looking at some form of smoothing of (4.1); in other words, to spread the mass concentrated by the empirical d.f. at a data point x_i to the neighborhood sets. A smooth version on (4.1) can be obtained by starting from a Feller smoothing of the DP. Let X_i, $i = 1, 2, \dots$ be i.i.d., conditionally on (k, U), with distribution $B_{k,U}$, and assume for simplicity that k and U are independent with $k \sim p(k)$ and $U \sim DP(\alpha U_0)$. Then we easily get that $X_i \mid k \sim b_{k,U_0}$. Furthermore, exploiting the mixture representation of $B_{k,U}$ given in Theorem 3, i.e., $B_{k,U}(x) = \int H_k(x; z) dU(z)$, one can reformulate the model hierarchically by means of hidden variables Z_i,

$$X_i \mid z_i, k, U \overset{ind}{\sim} H_k(\cdot; z_i); \qquad Z_i \mid k, U \overset{i.i.d.}{\sim} U,$$

with k and U as before. It follows that

$$\begin{aligned} Pr(X_{n+1} \le x \mid k, x_1, \dots, x_n, z_1, \dots, z_n) &= E(B_{k,U}(x) \mid k, z_1, \dots, z_n) \\ &= B_{k,E(U|k,z_1,\dots,z_n)}(x) = \frac{\alpha}{\alpha + n} B_{k,U_0}(x) + \frac{1}{\alpha + n} B_{k,\tilde{F}_n}(x), \end{aligned}$$

since $U \mid k, z_1, \ldots, z_n, x_1, \ldots, x_n \sim DP(\alpha U_0 + n\tilde{F}_n)$, where $\tilde{F}_n$ is the empirical d.f. of the latent variables $Z_1, \ldots, Z_n$. Therefore,

$$P(X_{n+1} \leq x \mid x_1, \ldots, x_n, k) = \frac{\alpha}{\alpha + n} B_{k,U_0}(x) + \frac{1}{\alpha + n} E(B_{k,\tilde{F}_n}(x) \mid x_1, \ldots, x_n, k)$$

which is a smooth version of (4.1) where, however, the empirical d.f. of the data is replaced by $\tilde{F}_n$. Since $B_{k,\tilde{F}_n}(x) = (1/n) \sum_{i=1}^{n} H_k(x; z_i)$, the predictive density (conditionally on k) is

$$f_n(x \mid k) = \frac{\alpha}{\alpha + n} b_{k,U_0}(x) + \frac{n}{\alpha + n} E\left(\sum_{i=1}^{n} \frac{1}{n} h_k(x; Z_i) \mid x_1, \ldots, x_n, k \right), \quad (4.2)$$

a weighted average between the prior guess b_{k,U_0} and the conditional expectation of a "Bayesian kernel estimate" $\sum_{i=1}^{n} h_k(x; Z_i)/n$. The choice of the smoothing parameter k is guided by the data through its posterior distribution. The predictive density is the Bayesian density estimate with quadratic loss. Note that in the second term of (4.2), the kernels are centered on the latent variables Z_i, while in a frequentist density estimate we would have $\sum_{i=1}^{n} h_k(x; x_i)/n$. Roughly speaking, this operates a form of shrinkage of the estimates x_i of the Z_i through the posterior of $(Z_1, \ldots, Z_n)$. Therefore, one can expect that (4.2) has better small sample properties (e.g. admissibility) than a frequentist kernel estimate centered on the sample points x_i. Furthermore, from the properties of the DP, ties may appear in the sample $(Z_1, \ldots, Z_n)$ with positive probability, so that the kernels are centered on the distinct values of the latent variables, inducing a kind of multiple shrinkage and dimension reduction; a number of kernels much fewer than n is used for reconstructing the unknown density of the data.

The form (4.2) of the predictive density (given k) arises from the representation of the model as a DP-mixture, and in this sense it is well known. DP-mixtures are indeed popular priors in Bayesian nonparametric inference for continuous data. The novelty of our approach is in starting from the Feller smoothing of the DP, then obtaining the connection with mixtures from the results in Section 3. Our construction also underlines one aspect of DP-mixtures that is sometimes understated in applications. DP-mixtures are widely used for their capability of modeling the clustering structure of the data; however, they often suggest more clusters than expected. The nature of the mixture model, as a smoothing of the discrete d.f., is one possible explanation; in this case the components play the role of building kernels, and population clusters are rather described by the modes of the predictive density.

Computation of the posterior distribution of the quantities of interest, namely $(k, U, Y_1, \ldots, Y_n)$, can be carried over fairly easily using MCMC algorithms that are now quite standard for mixture models. For the specific case of mixture

models arising from Feller operators, we refer to Petrone and Veronese (2002), who illustrate MCMC computations for both the cases of discrete and continuous ERS.

4.2. Support of mixture priors

In the previous section, we have referred to the mixture prior π_B as *nonparametric* since it selects d.f.'s in a class (the class of Feller d.f.'s $\{B_{k,U}, k \in \Lambda, U \in \Delta\}$) that is dense in Δ so that, roughly speaking, π_B has large support. In this section we state more formal results, that are proved in Appendix A.4. We say that a d.f. F_0 is in the support of π_B w.r.t. some topology if $\pi_B(W_\epsilon(F_0)) > 0$ for any neighborhood $W_\epsilon(F_0)$ of F_0 in that topology. In particular, we consider neighborhoods in the topology of weak convergence (weak support), in total variation (strong support), and in Kullback-Leibler.

Clearly, the support of π_B depends on the prior on (k, U). We say that the marginal distribution p of k is "positive" if it gives positive probability to any isolated point in Λ and is diffuse with a strictly positive density on the intervals in Λ. We say that the conditional probability law $\pi_U(\cdot \mid k)$ of U, given k, has full weak support if for any $U \in \Delta$, $\pi_U(W_\epsilon(U) \mid k) > 0$ for any weak neighborhood $W_\epsilon(U)$ of U.

Theorem 5. Suppose that $p(k)$ is positive.

(i) *If F_0 is a continuous d.f. in the weak support of $\pi_U(\cdot \mid k)$ for any k, then F_0 is in the weak support of π_B.*

(ii) *If F_0 is an absolutely continuous d.f. with continuous and bounded density f_0 and F_0 is in the strong support of $\pi_U(\cdot \mid k)$ for any k, then F_0 is in the strong support of π_B.*

It is easy to verify that if the prior π_B is based on a continuous ERS, then the assumption of continuity of F_0 in part (i) of the above theorem is not needed.

We now provide some results on the Kullback-Leibler support of the prior π_B. Since π_B can be regarded as a measure on the space $(\Delta_c, \mathcal{F}_c)$ of absolutely continuous d.f.'s, the so-called Kullback-Leibler property is sufficient, under regularity conditions, for weak consistency of the corresponding posterior (Schwartz (1965) and Ghosh and Ramamoorthi (2003)). Schwartz's regularity conditions can be shown to hold in our case; note that the parameter space here is $(\Delta_c, \mathcal{F}_c)$.

Many results have been recently established on frequentist consistency and convergence rates of Bayesian density estimators (Ghosal and van der Vaart (2007), Walker, Lijoi and Prünster (2007), and references therein). However, they are mainly focussed on convolutions of a discrete prior with specific kernels, often Gaussian kernels. Extending these results to more general classes of kernels

seems a difficult task. Our aim is to provide some results in this direction, showing Kullback-Leibler properties for the fairly general class of mixtures arising from Feller operators with ERS.

The first case to consider is when the *true* d.f. F_0 of the data is itself a mixture, $F_0 = B_{k_0,U_0}$. The following theorem extends an analogous result obtained for location mixtures of Gaussian distributions by Ghosal, Ghosh and Ramamoorthi (1999). From Schwartz's theorem, it implies that the mixture prior π_B on $(\Delta_c, \mathcal{F}_c)$ is weakly consistent at $F_0 = B_{k_0,U_0}$ if U_0 has support included in (a, b).

In the sequel we consider Feller operators with an ERS that is either continuous or discrete. In particular, for the discrete case, we assume that the carrier measure ν of the NEF in the random scheme has support $\{0, \pm h, \pm 2h, \ldots\}$ or $\{q+mh, q+(m+1)h, q+(m+2)h, \ldots\}$, where h and q are positive real constants and m a positive integer.

Theorem 6. *Let π_B be a mixture prior on $\Delta([a, b])$, with a continuous or discrete ERS and parameters (k, U) such that $U(\{a\}) = U(\{b\}) = 0$ a.s.. Suppose that $p(k)$ is positive and that for any k, $\pi_U(\cdot \mid k)$ has full weak support. Then $F_0 = B_{k_0,U_0}$, with U_0 having support strictly included in (a, b), is in the Kullback-Leibler support of π_B.*

The theorem relates the Kullback-Leibler support of π_B to the weak support of the prior of U given k. Often, k and U are assumed to be independent, with $U \sim DP(\alpha U_0)$. In fact, the theorem holds for more general priors on U as long as they have full weak support. This is the case for many nonparametric priors in the Bayesian literature, under mild conditions; for example, Ongaro and Cattaneo (2004, Corollary 2) study the weak support of a fairly general class of discrete priors on Δ, that includes the Dirichlet process, the Poisson-Dirichlet process, the beta two-parameter process, stick-breaking priors, and other generalizations of the Dirichlet process.

The basic steps of the proof of Theorem 6 are as follows. First, we write the Kullback-Leibler divergence between $f_0 = b_{k_0,U_0}$ and $b_{k,U}$ as

$$KL(f_0, b_{k,U}) = \int \log \frac{f_0(x)}{b_{k,U_0}(x)} f_0(x)dx + \int \log \frac{b_{k,U_0}(x)}{b_{k,U}(x)} f_0(x)dx. \tag{4.3}$$

Then, we prove some properties of the kernel $h_k(x; z)$ which are useful for showing that the integrals on the right hand side can be made arbitrarily small for (k, U) in a set with positive prior probability π_B. Properties of the kernel include unimodality of $h_k(x; \cdot)$, as shown in part 3) of Lemma 2 in Section 3.5; the other properties are collected in Lemmas 3, 4, and 5 in Appendix A.2. In particular, Lemma 3 in A.2 shows that the kernels $h_k(x; z)$ have a tail behavior which extends some tail properties of Gaussian kernels. Furthermore, we need some continuity

results of $B_{k,U}$ in k and U, that are provided in Section A.3 of the Appendix. These results exploit the properties of the kernels $h_k(x;z)$ and the construction of $B_{k,U}$ as a Feller operator with ERS, $B_{k,U}(x) = E(U(Z_{k,x}))$. Note that the first addend in (4.3) is the Kullback-Leibler divergence $KL(b_{k_0,U_0}, b_{k,U_0})$; its continuity in k is included as Lemma 7 in Appendix A.3.

Theorem 6 requires that F_0 is itself a mixture, $F_0 = B_{k_0,U_0}$. This assumption can be avoided if F_0 has finite support included in (a,b).

Theorem 7. *Let π_B be a mixture prior on $\Delta([a,b])$ with a continuous or discrete ERS, and parameters (k,U) satisfying the assumptions in Theorem 6. Let F_0 be a d.f. with a continuous and bounded density f_0 and finite support included in $[a,b]$. Then F_0 is in the Kullback-Leibler support of π_B.*

Theorems 6 and 7 give sufficient conditions for weak consistency of the mixture prior π_B for some classes of d.f.'s F_0. For stronger forms of consistency, or for establishing rates of convergence, further conditions are required which, however, usually include the Kullback-Leibler property. Therefore, results like those provided in the previous theorems are a basic step also for stronger results. A discussion about the reasons why a weakly consistent posterior might fail to cumulate on Hellinger or total variation neighborhoods of the *true* density is given by Walker, Lijoi and Prünster (2005). Roughly speaking, this might happen if there are d.f.'s which are weakly close to F_0 but far away from F_0 in strong sense. This behavior is related to the identifiability of the mixture model and the possibility of characterizing weak neighborhoods of $B_{k,U}$ in terms of the mixing distribution and the smoothing parameter. We do not pursue this issue here; the special case considered below, where k is fixed, illustrates some basic points and difficulties.

Let $\mathcal{B}_k$ the class of mixtures $\{B_{k,U} = \int H_k(x;z)dU(z),\ U \in \Delta([a,b])\}$, for a fixed $k \in \Lambda$. Note that if $B_{k,U}$ is based on a discrete ERS with support points $\{z_{1,k}, z_{2,k}, \ldots, z_{N_k,k}\}$, $N_k, k \leq \infty$, then $B_{k,U} = \sum_{j=1}^{N_k} w_{j,k}^U H_k(x; z_{j,k})$ is a finite or countable mixture, with $w_{j,k}^U = U((z_{j,k}, z_{j+1,k}])$, $j = 1, \ldots, N_k - 1$. We say that the class of mixtures $\mathcal{B}_k$ is *identifiable* (or, equivalently, the mixing distribution U is identifiable in $\mathcal{B}_k$) if $B_{k,U}(x) = B_{k,Q}(x)$ for any x implies: *i)* $U(z) = Q(z)$ for any z and any $Q \in \Delta$, if the ERS is continuous; *ii)* $w_{j,k}^U = w_{j,k}^Q$ for $j = 1, \ldots, N_k - 1$ and any $Q \in \Delta$, if the ERS is discrete.

Proposition 3. *The class of mixtures $\mathcal{B}_k$ is identifiable.*

From the continuity of $B_{k,U}$ in U (Proposition 4 in Appendix A.3), it follows that U_n converges weakly to U_0 if and only if B_{k,U_n} converges weakly to B_{k,U_0}. Therefore, if the posterior on the random d.f. cumulates on weak neighborhoods of $F_0 = B_{k,U_0}$, the posterior on U cumulates on weak neighborhoods of U_0 (it is

weakly consistent at U_0). In turn, by part (a) of Proposition 4 in Appendix A.3, this implies strong consistency of the posterior of the d.f. at F_0.

5. Extensions and Final Remarks

In the paper we have studied a fairly general class of mixture models arising from Feller operators with ERS, focussing on applications in Bayesian nonparametric inference for exchangeable, continuous data. However, the proposed construction is attractive in more general contexts. Consider a sequence of independent r.v.'s $(X_i, i \geq 1)$ such that $X_i \mid \mu_i \sim p_{\mu_i}(x)$, with p_{μ_i} belonging to a NEF with mean parameter μ_i. In this context, shrinkage estimators of the individual parameters $\mu_1, \ldots, \mu_n$ are obtained, in a Bayesian nonparametric approach, by assuming that they are a sample from a random d.f. G (i.e., they are i.i.d. conditionally on G). The popular choice of a Dirichlet process prior on the latent distribution G implies that G is a.s. discrete, so that ties can be observed with positive probability among the μ_i's. This property is attractive as a clustering procedure of the individual parameters, but it is not appropriate if, for example, one wants to model random effects inside clusters. In this case the μ_i's are a.s. distinct, i.e., G is absolutely continuous, and clusters in the μ_i's are rather individuated by the modes of the latent distribution G. Using a Feller prior on G is particularly attractive in this context, since not only does it allow one to model G as a.s. absolutely continuous with a flexible multimodal form, but it also gives a representation of G as a mixture of kernel distributions that can be chosen quite naturally according to the distribution $p_{\mu_i}(x)$ of the data. That is, if p_{μ_i} belongs to a NEF (e.g. Poisson), it is natural to choose a Feller prior with ERS based on the same NEF (a Poisson random scheme). This implies that the kernel of the resulting mixture model has, in many cases, the form of a conjugate prior, as illustrated in the examples of Section 3.3, resulting in a fairly simple structure that facilitates posterior computations. For example, it allows one to obtain a fairly natural expression for the posterior expectation $E(\mu_i \mid x_1, \ldots, x_n)$ that represents the Bayesian multiple shrinkage estimator of μ_i, with an automatic choice of the shrinking guided by the estimated modes of the latent distribution G. Results of this kind were obtained by Petrone and Veronese (2002) in the case where $X_i \mid \mu_i$ are Binomial with mean parameter μ_i, using a Bernstein prior on the latent distribution G on $(0, 1)$. We do not enter in further details here, but since the Bernstein prior is a special case of a Feller prior based on a Binomial random scheme, one can easily envisage that their results can be extended to more general problems of combining several experiments.

The properties of Feller operators have been used for extending theoretical properties of nonparametric mixture priors, usually discussed for specific kernels. In particular, we have extended results on weak consistency of Bayesian density

estimators to fairly general, non Gaussian kernels. While completing a revision on this paper, we became aware of a recent work by Wu and Ghosal (2008) that addresses the latter problem. They give sufficient conditions on the kernels for weak consistency in mixture models, under some assumptions on f_0. Their results move from a decomposition of the Kullback-Leibler divergence, as in (4.3). The interesting aspect is that they allow a general form of f_0 by introducing a further condition that would control, in our case, the extra addend $\int \log(f_0/b_{k_0,U_0})f_0$ in (4.3). However, Wu and Ghosal (2008) also need properties of the mixture kernel for ensuring that the steps in their general Theorem 1 are satisfied. These properties have to be checked for the specific kernels of interest; giving results for general classes of kernels requires considerable additional work, and restrictions on f_0; see the discussion of location-scale kernels in their Section 3. We also give properties of the mixture kernel for weak consistency; furthermore, we verify that these properties actually hold for the fairly general family of kernels arising from Feller operators with ERS. The restriction on f_0 in our Theorem 6 is partially removed in Theorem 7, and further extensions could be possible through their approach.

The mixture-kernels arising from Feller operators with ERS are quite simple, having no unknown parameter besides the smoothing parameter k. This has computational advantages, for example in avoiding identifiability or label-switching difficulties, but many components might be needed for approximating irregular curves. When prior information is available on the mixture components, one might prefer to use kernels having a physical interpretation and free parameters. Our scheme could be generalized to this case.

We did not study the approximation rates of Feller operators (results can be obtained by Theorem 5.2.4 in Altomare and Campiti (1994)) but, in some cases, they are known to be slow. Thus, in applications to Bayesian density estimation, further extensions to achieve faster convergence rates for the posterior on the unknown density might be needed, as shown for Bernstein polynomials e.g. by Ghosal (2001) and Kruijer and van der Vaart (2008). In particular, one might expect that, being constructed in order to preserve specific properties of the density to be estimated, Bayesian density estimators via Feller's operators may give better rates than estimators based on Gaussian kernels. However, a comparison is not easy; for large k, the kernel resulting from Feller's operators is close to a Gaussian kernel, but this is not true for small values of k and rates remain a delicate issue. However, density estimation is only one possible application of the scheme presented here. It can be used more generally for Bayesian nonparametric inference on a random, bounded curve, and it looks particularly attractive when prior information is available on some general properties of the curve. This could be fairly naturally incorporated in the model due to the simple, constructive nature of the approximation. An example, on which we plan to return in further

research, is discussed in Petrone and Corielli (2005), where Bernstein polynomials are used in nonparametric dynamic regression for including prior information on the temporal evolution of the regression curve. Extensions to multivariate data are also of interest.

Acknowledgement

This research was partially supported by grants from MIUR and from L. Bocconi University. We are grateful to P. Diaconis for helpful discussions. We thank the referees for their suggestions and stimulating comments which led to a much improved presentation.

References

Altomare, F. and Campiti, M. (1994). *Korovkin-type Approximation Theory and its Application*. W. de Gruyter, Berlin.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families, with Applications in Statistical Decision Theory*. Lectures Notes **9**. Institute of Mathematical Statistics, Hayward.

Choudhuri, N., Ghosal, S. and Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *J. Amer. Statist. Assoc.* **99**, 1050-1059.

Consonni, G. and Veronese, P. (1992). Conjugate priors for exponential families having quadratic variance functions. *J. Amer. Statist. Assoc.* **87**, 1123-1127.

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 Ylvisaker, D. (1985). Quantifying prior opinion. In *Bayesian statistics* **2** (Edited by J. M. Bernardo, M. H. deGroot, D. V. Lindley and A. F. M. Smith), 133-156. Elsevier Science Publishers B.V., North Holland.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. II. Wiley, New York (First Edition: 1966).

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics* (Edited by H. Rizvi and J. Rustagi), 287-302. Academic Press, New York.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. 3rd edition. Hafner Press, New York.

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. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143-158.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer, New York.

Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35**, 697-723.

Heike, H., Târcolea, C., Demetrescu, M. and Târcolea, A. (2003). Fiducial inference for discrete and continuous distributions. *BSG Proceedings* **8**, 69-80. Geometry Balkan Press, Bucharest.

Hoff, D. P. (2003). Nonparametric estimation of convex models via mixtures. *Ann. Statist.* **31**, 174-200.

Jarra, A. (2007). Dirichlet Process Package. *R-package* available at `www.Rproject.org`.

Jones, M. C., and Foster, P. J. (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statist. Sinica* **6**, 1005-1013.

Jorgensen, B. (1987). Exponential dispersion models. *J. Roy. Statist. Soc. Ser. B* **49**, 127-162.

Kruijer, W. and van der Vaart, A. (2008). Posterior convergence rates for Dirichlet mixtures of Beta densities. *J. Statist. Plann. Inference* **138**, 1981-1992

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

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

Ongaro, A. and Cattaneo, C. (2004). Discrete random probability measures: a general framework for Bayesian inference. *Statist. Probab. Lett.* **67**, 33-45.

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

Petrone, S. and Corielli, F. (2005). No-arbitrage issues in dynamic models for estimating the yield curve. Tech. Rep., IMQ, Bocconi University, Milano.

Petrone, S. and Veronese, P. (2002). Non parametric mixture priors based on an exponential random scheme. *Stat. Methods Appl.* **11**, 1-20.

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

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

Walker, S., Lijoi, A. and Prünster, I. (2005). Data tracking and the understanding of Bayesian consistency. *Biometrika* **92**, 765-778.

Walker, S., Lijoi, A. and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* **35**, 738-746.

Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.* **2**, 298-331.

Zabell, S. L. (1992). R. A. Fisher and the Fiducial Argument. *Statist. Sci.* **7**, 369-387.

Department of Decision Sciences, L. Bocconi University, 20136 Milan. Italy.

E-mail: sonia.petrone@unibocconi.it

Department of Decision Sciences, L. Bocconi University, 20136 Milan. Italy.

E-mail: piero.veronese@unibocconi.it

(Received December 2007; accepted October 2008)