Statistical Inference for Stochastic Processes (2005) 8:227–254
DOI 10.1007/s11203-005-6068-7

Exact Inference for Random Dirichlet Means

NILS LID HJORT[1] and ANDREA ONGARO[2,*]
[1]*Department of Mathematics, University of Oslo*
[2]*Department of Statistics, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy*

Abstract. Two characterisations of a random mean from a Dirichlet process, as a limit of finite sums of a simple symmetric form and as a solution of a certain stochastic equation, are developed and investigated. These are used to reach results on and new insights into the distributions of such random means. In particular, identities involving functional transforms and recursive moment formulae are established. Furthermore, characterisations for several choices of the Dirichlet process parameter (leading to symmetric, unimodal, stable, and finite mixture distributions) are provided. Our methods lead to exact simulation recipes for prior and posterior random means, an approximation algorithm for the exact densities of these means, and limiting normality theorems for posterior distributions. The theory also extends to mixtures of Dirichlet processes and to the case of several random means simultaneously.

AMS Mathematics Subject Classifications (1991): 62G99, 62F15.

Key words: Bernshteĭn–von Mises theorems, Dirichlet process, Hilbert and Stieltjes transforms, mixtures, nonparametric Bayes, random probability measure, stochastic equation.

1. Introduction and Summary

The Dirichlet process was brought to the attention of the statistical community as the first genuine nonparametric prior for use in Bayesian statistics by Ferguson (1973, 1974). It is still a cornerstone in nonparametric Bayesian methodology, where it is often used separately or as an ingredient in more complicated priors. It is also a special case of larger classes of nonparametric priors, like neutral to the right and tailfree processes, Beta processes and Pólya trees; see Walker et al. (1999) Hjort (2003) for recent reviews and discussion. This paper is concerned with distributional aspects for important functionals of the Dirichlet process. In Bayesian contexts one is more interested in the posterior distributions of such functionals, i.e. given a set of

*Author for correspondence: Tel.: +39-2-64487375; Fax: +39-2-6473312; e-mail: andrea.ongaro@unimib.it

observations, but since the random distribution underlying the data continues to be a Dirichlet process given the data, only with updated parameters, the distributional results we reach here continue to be relevant.

Let P be a Dirichlet process on some sample space Ω, with parameter aP_0, in terms of a probability distribution P_0 and a positive strength parameter a. It is characterised by the property that for any measurable partition $(A_1, \dots, A_k)$, the vector $(P(A_1), \dots, P(A_k))$ is Dirichlet distributed with parameter vector $(aP_0(A_1), \dots, aP_0(A_k))$, and we write $P \sim \mathrm{Dir}(aP_0)$. We shall be interested in the random mean $\theta = \theta(P) = \int g \, \mathrm{d}P$. Here g is in principle any measurable function making the integral finite almost surely (a.s.). Note that the mean can be represented as

$$\theta = \int_\Omega g(x) \, \mathrm{d}P(x) = \int_{-\infty}^{\infty} y \; \mathrm{d}Q(y), \tag{1.1}$$

where $Q = Pg^{-1}$ is the transformed Dirichlet process with parameter $aQ_0 = aP_0 g^{-1}$ on the real line.

There is a growing literature on formulae for and numerical approximations to distributions of random Dirichlet means. After some earlier and partial results of Hannum et al. (1981) and Yamato (1984), Cifarelli and Regazzini (1990, 1994) reached more general results, giving in particular (admittedly somewhat complicated) formulae in terms of limits of integrals in regions of the complex plane, following inversion of certain transforms. Diaconis and Kemperman (1996) provided new tools and surprising connections to other areas of probability and mathematics, and also gave some explicit formulae for the special case of $a = 1$, following the lead of Cifarelli and Regazzini; see also Kerov (1998). The recent paper of Regazzini et al. (2002) sums up earlier work and also discusses successes and difficulties with attempts at the quite nontrivial computer implementation of the mathematical results.

Random Dirichlet means can be used in several statistical contexts in addition to the most natural one, which is to make inference on such mean parameters in a framework of nonparametric Bayesian statistics. Ferguson (1983) and Lo (1984) were early papers working on Bayesian density estimation where the prior involves a kernel smoothed integral of a Dirichlet process (see also Hjort, 1996). In a Bayesian regression framework one would like to model noise with mean zero around the main structure, and this can be accomplished with a Dirichlet process having its mean subtracted. This calls for somewhat complicated calculations involving simultaneous aspects of the process and its mean, where results and moment formulae of this paper are relevant. Finally, in Hjort (2003) a model for random shapes is being discussed, involving a smoothed normalised gamma process to represent a process of random radii. The random shape thus takes the form $(R(s)\cos(2\pi s), R(s)\sin(2\pi s))$, with random radius

$$R(s) = \int h^{-1} K(h^{-1}(s-u))\,\mathrm{d}P(u) \quad \text{for } 0 \leqslant s \leqslant 1$$

in terms of a Dirichlet process P on the unit interval and a kernel function K with support $[-1/2, 1/2]$, involving a window width parameter h; the integral is modulo the circle, so that in particular $R(0) = R(1)$. The random radii are accordingly of the type (1.1).

We shall investigate here two representations of a random Dirichlet mean, with the aim of demonstrating their potential both for reaching new results and for giving independent and simpler proofs of earlier results. The lay-out of our article is as follows. In Sections 2 and 3 the two representations are introduced and applied to prove general results for random Dirichlet means. Specifically, in Section 2 Dirichlet means are constructed as limits of finite sums of simple symmetric random variables. This enables us to obtain necessary and sufficient conditions for finiteness of a Dirichlet mean and a characterisation in terms of characteristic functions. This also leads to explicit expressions for the so-called generalised Stieltjes transform (also called the generalised Hilbert transform) of a random Dirichlet mean. Furthermore, we develop a strategy for simulating from the exact prior or posterior distribution of a Dirichlet mean, and describe the asymptotic behaviour of θ as a grows. This leads to asymptotic normality of the posterior distribution of θ as the sample size grows, as explained in detail in Section 6.2.

In Section 3 Dirichlet means are characterised as solutions to certain stochastic equations. For any fixed value of the parameter a, such stochastic equations are shown to establish a one-to-one correspondence between the law of a Dirichlet mean and the distribution $Q_0 = P_0 g^{-1}$. This is also related to the generalised Stieltjes transform. As a further direct consequence of such equations, simple recursive formulae are derived for direct and centralised moments, together with necessary and sufficient conditions for their existence.

The two representations are then applied in Section 4 to obtain useful information for Dirichlet means associated with distributions Q_0 belonging to various general classes of interest. In particular we consider symmetric, unimodal, and stable distributions. In Section 5 we deal with both mixtures of Dirichlet means and means of Dirichlet mixtures, with a number of illustrations. Finally in Section 6 we briefly address extensions of our results to the multi-dimensional case, give an approximation algorithm for computing the exact density of a random Dirichlet mean, and use our methods to derive a general Bernshteĭn–von Mises theorem for Dirichlet process priors.

We should make clear that these two approaches to studying Dirichlet means are not entirely new, as other recent articles have presented related techniques and results, partly independently of the efforts of the present authors. The relation to other literature is discussed in more detail in

Sections 2 and 3. Our contribution is to develop the techniques further and to reach various new results, and to re-prove some existing results in new and sometimes simpler ways. In particular, our exact simulation strategies for prior and posterior distributions are new (avoiding Markov chain Monte Carlo methods that need convergence diagnostics), as are our results for mixtures of means and means of mixtures.

2. Construction as Limit of Symmetric Distributions

A basic tool used in the following to derive results for the random mean $\theta = \int g\,\mathrm{d}P$ is based on approximations via finite sums of certain symmetric distributions. This stems from the following simple approximation of a Dirichlet process. Let $\xi_1, \ldots, \xi_m$ be independent from P_0 and independent of $(\beta_1, \ldots, \beta_m)$, which we give a symmetric Dirichlet distribution with parameter $(a/m, \ldots, a/m)$. Then the random probability measure

$$P_m = \sum_{j=1}^{m} \beta_j \delta(\xi_j) \tag{2.1}$$

converges in distribution to a Dirichlet process with parameter aP_0 as $m \to \infty$; for a proof, see Hjort and Ongaro (2004), Hjort (2003). Here $\delta(\xi)$ denotes unit point mass at position ξ. As a direct consequence of this one obtains (see Theorem 4.2 in Kallenberg, 1986) convergence in distribution of

$$\theta_m = \int g\,\mathrm{d}P_m = \sum_{j=1}^{m} g(\xi_j)\beta_j \tag{2.2}$$

to θ when g is a continuous function with compact support. It is the symmetry of the Dirichlet distribution used in (2.1) and (2.2) that for some applications gives a simpler treatment than if working with other characterisations of the Dirichlet process.

The random probability (2.1) has been considered by several authors, independently and partly in quite unrelated contexts; an extensive list of references is given in Ishwaran and Zarepour (2002, Section 4). In particular, Ishwaran and Zarepour proved convergence of θ_m under the assumption that g is P_0-integrable.

In the following theorem we shall prove convergence of θ_m under completely general conditions, i.e. whenever θ exists and is finite. As a consequence of this approximation, we shall also derive necessary and sufficient conditions for finiteness of θ and an identity in terms of characteristic functions which fully determines its distribution. The latter identity produces, as a special case, the so-called Stieltjes transform of order a for θ. Below we denote by ξ a random variable with distribution P_0.

THEOREM 1. *Let* $P \sim \mathrm{Dir}(aP_0)$ *and consider* $\theta = \int g\,\mathrm{d}P$ *for a measurable* g *for which*

$$\mathrm{E}\log\{1+|g(\xi)|\} = \int \log(1+|g|)\,\mathrm{d}P_0 = \int_{-\infty}^{\infty} \log(1+|y|)\,\mathrm{d}Q_0(y) \tag{2.3}$$

is finite. Let furthermore $G_a \sim \mathrm{Gam}\,(a,1)$ *be independent of* θ. *Then* θ *is a.s. finite and its distribution is fully characterised by the characteristic function of* $(G_a\theta, G_a)$:

$$\mathrm{E}\exp\{i(tG_a\theta + sG_a)\} = \exp[-a\,\mathrm{E}\log\{1-i(tg(\xi)+s)\}] \quad \textit{for } t,s\in\mathcal{R}. \tag{2.4}$$

As a special case one obtains

$$\mathrm{E}\exp(it\theta G_a) = \mathrm{E}\Big(\frac{1}{1-it\theta}\Big)^a = \exp[-a\,\mathrm{E}\log\{1-itg(\xi)\}] \quad \textit{for } t\in\mathcal{R}. \tag{2.5}$$

Furthermore, θ_m *of* (2.2) *converges in distribution to* θ. *If on the other hand* $\mathrm{E}\log\{1+|g(\xi)|\}$ *is infinite, then* $\int|g|\,\mathrm{d}P$ *is a.s. infinite, so that* θ *does not exist finite.*

Proof. Write $\beta_j = G_j/S_m$ in terms of independent and identically distributed $G_j \sim \mathrm{Gam}\,(a/m, 1)$, with sum $S_m = \sum_{j=1}^m G_j$. This leads to $\theta_m = R_m/S_m$, with $R_m = \sum_{j=1}^m G_j g(\xi_j)$ being a random mixture of many independent small gammas. We shall first compute the characteristic function of (R_m, S_m) and its limit under condition (2.3).

Use first $\mathrm{E}\exp(itG_j) = (1-it)^{-a/m}$ to find

$$\begin{aligned}\mathrm{E}\big[\exp\{i(tR_m+sS_m)\}\,|\,\xi_1,\ldots,\xi_m\big] &= \prod_{j=1}^m \{1-i(tg(\xi_j)+s)\}^{-a/m} \\ &= \exp\Big[-a\frac{1}{m}\sum_{j=1}^m \log\{1-i(tg(\xi_j)+s)\}\Big].\end{aligned}$$

Note that if (2.3) holds, then automatically also $\mathrm{E}\log\{1-i(tg(\xi)+s)\}$ is finite for all $t,s\in\mathcal{R}$. Under this assumption,

$$\mathrm{E}\exp\{i(tR_m+sS_m)\} \to \exp[-a\,\mathrm{E}\log\{1-i(tg(\xi)+s)\}] \quad \text{as } m\to\infty \tag{2.6}$$

by the law of large numbers.

Suppose now that g is bounded and write $\theta = \int y \, \mathrm{d}Q(y)$ with $Q \sim \mathrm{Dir}(aP_0g^{-1})$. Convergence of θ_m to θ is then a consequence, by Theorem 4.2 in Kallenberg (1986), of convergence in distribution of P_m in (2.1). This in turn implies $(R_m, S_m) \sim (G_a\theta_m, G_a) \to_d (G_a\theta, G_a)$, where $G_a \sim \mathrm{Gam}\,(a, 1)$ is independent of θ_m and θ. It follows then by (2.6) that relation (2.4) must hold when g is bounded.

Consider next a general measurable g and let us see when θ exists and is finite, i.e. when $\theta^+ = \int |g| \, \mathrm{d}P < \infty$ a.s.. Let $h_k(y) = yI\{|y| \leqslant k\} + kI\{y > k\} - kI\{y < -k\}$. Then as $|h_k(y)| \uparrow |y|$, by the theorem on monotone convergence we have $\theta_k^+ = \int |h_k(g)| \, \mathrm{d}P \uparrow \theta^+$ a.s., where θ^+ is an extended nonnegative random variable. This implies $G_a\theta_k^+ \uparrow G_a\theta^+$, which in its turn leads to convergence of the Laplace transform $\mathrm{E}\exp(-tG_a\theta_k^+)$ to $\mathrm{E}\exp(-tG_a\theta^+)$ for $t > 0$. Note that this convergence takes place even if $G_a\theta^+$ is not finite with probability one. Noticing that $|h_k(g)| \leqslant k$, the same argument used to prove (2.6) gives

$$\mathrm{E}\exp(-tG_a\theta_k^+) = \exp\bigl[-a\,\mathrm{E}\log\{1 + t|h_k(g(\xi))|\}\bigr].$$

The latter, again by monotone convergence, tends to $\exp[-a\,\mathrm{E}\log\{1 + t|g(\xi)|\}]$, which is then the Laplace transform of $G_a\theta^+$. Under condition (2.3) $\mathrm{E}\log\{1 + t|g(\xi)|\}$ is finite for any positive t, so that such a Laplace transform tends to 1 as $t \uparrow 0^+$. This implies that $G_a\theta^+$ and therefore θ^+ are finite a.s.

If on the other hand $\mathrm{E}\log\{1 + |g(\xi)|\} = \infty$, so that $\mathrm{E}\log\{1 + t|g(\xi)|\} = \infty$ for any positive t, then

$$\Pr\{G_a\theta^+ < \infty\} = \lim_{t \to 0^+} \mathrm{E}\exp(-tG_a\theta^+) = 0.$$

Consequently, $\theta^+ = \infty$ a.s.

Let us prove that (2.4) holds under condition (2.3). Consider $\theta_k = \int h_k(g) \, \mathrm{d}P$. As $h_k(g) \to g$ and $|h_k(g)| \leqslant |g|$ with $|g|$ integrable, we have by the theorem on dominated convergence that $\theta_k \to \theta$ a.s.. This implies convergence in distribution of $(G_a\theta_k, G_a)$ to $(G_a\theta, G_a)$ and therefore convergence of the corresponding characteristic functions. As h_k is bounded, the characteristic function $\mathrm{E}\exp\{i(tG_a\theta_k + sG_a)\}$ is equal to $\exp[-a\,\mathrm{E}\log\{1 - i(th_k(g(\xi)) + s)\}]$. Furthermore, under condition (2.3), $\mathrm{E}\log\{1 - i(tg(\xi) + s)\}$ is finite, and it is possible to apply the theorem on dominated convergence to show that this characteristic function converges to $\exp[-a\,\mathrm{E}\log\{1 - i(tg(\xi) + s)\}]$. This proves relation (2.4).

Note now that the characteristic function of $(G_a\theta_m, G_a)$ converges to the characteristic function of $(G_a\theta, G_a)$, by (2.6); this implies

$$(G_a\theta_m, G_a) \to_d (G_a\theta, G_a)$$

and by the continuous mapping theorem $\theta_m \to_d \theta$. Finally, relation (2.5) is obtained by setting $s=0$ in (2.4) and then computing the characteristic function of $G_a\theta$ conditionally on θ. □

Remarks. Several comments and illustrations are now in order.

(1) We term the (2.5) transform the generalised Stieltjes transform, of order a, since the original Stieltjes transform corresponds to $a=1$. It is sometimes also called the generalised Hilbert transform; see Henrici (1993), Widder (1971), and Guglielmi (1998).

(2) Condition (2.3) for finiteness of θ and relation (2.5) are known (see Regazzini et al., 2002, and references therein). Their proof is based on properties of the Gamma process and its relation with the Dirichlet process. The above proof, albeit somewhat similar, uses only elementary tools and well-known properties of the Dirichlet distribution.

(3) The characteristic function in (2.4) is related to the Laplace transform of a Gamma process, see e.g. Hjort (1990) and Vershik et al. (2001), where similar results are discussed.

(4) Attempts at inverting expression (2.4) to exhibit the underlying density $f_a(r,s)$ for $(R,S) \equiv (G_a\theta, G_a)$ lead to complexities resembling those encountered in Cifarelli and Regazzini (1990) and Regazzini et al. (2002). A relatively simple formula emerges only for $a=1$, where the density d_1 of θ is equal to

$$d_1(t) = \pi^{-1}\sin(\pi F_0(t))\exp\left\{-\int_{-\infty}^{\infty}\log|y-t|\,\mathrm{d}Q_0(y)\right\}, \tag{2.7}$$

where $F_0(t)$ denotes the distribution function associated with $Q_0 = P_0 g^{-1}$; see also Diaconis and Kemperman (1996). For this case (θ, G_a) has density $d_1(t)\exp(-s)$, implying a density for (R,S) of the form

$$f_1(r,s) = d_1(r/s)\exp(-s)s^{-1}. \tag{2.8}$$

(5) It is also clear that the density $f_a(r,s)$ can be expressed as the convolution $f_1 \star \cdots \star f_1$. This generally produces analytically rather complicated expressions, however, even for special cases of P_0.

(6) These results also lead to a simulation recipe for a an arbitrary integer, as follows. First notice that (R,S) is an infinitely divisible random vector, as the right hand side of (2.4) is a characteristic function for any positive a. Let then $(\bar{R}_i, \bar{S}_i)$ for $i=1,\ldots,a$ be independently drawn from $f_1(r,s)$. This is accomplished by drawing a unit exponential $\bar{S}_i$ and independently a θ_i from $d_1(t)$ using e.g. a rejection-acceptance routine, and then setting $\bar{R}_i = \theta_i \bar{S}_i$. Then (R,S), with $R = \sum_{i=1}^{a} \bar{R}_i$ and $S = \sum_{i=1}^{a} \bar{S}_i$, has the density $f_a(r,s)$ of (R,S), and $\theta = R/S$ is a simulated outcome from the exact density $d_a(t)$ of θ. This, in particular, allows exact simulation from

the posterior distribution of θ, whenever the total mass parameter of the prior is an integer. This contrasts with the Markov chain Monte Carlo method described in conjunction with Proposition 2 in Section 3.

(7) For nonnegative random variables it is more convenient to work in terms of Laplace transforms, rather than with characteristic functions. The parallel of (2.5) says for such cases that

$$\mathrm{E}\exp\{-a\log(1+u\theta)\}=\exp\{-a\,\mathrm{E}\log(1+uY)\}\quad\text{for } u\geqslant 0,$$

where $Y=g(\xi)$. As an illustration, take P_0 to be the Beta$(1/2, 1/2)$ distribution on the unit interval and let $\theta=\int_0^1 x\,\mathrm{d}P$ be the random mean when $P\sim\mathrm{Dir}(aP_0)$. Then some calculations give

$$\int_0^1 \log(1+ux)f_0(x)\,\mathrm{d}x=2\log\{\tfrac{1}{2}+\tfrac{1}{2}(1+u)^{1/2}\}$$

for f_0 the Beta$(1/2, 1/2)$ density, and this leads to

$$\mathrm{E}(1+u\theta)^{-a}=\{\tfrac{1}{2}+\tfrac{1}{2}(1+u)^{1/2}\}^{-2a}\quad\text{for } u\geqslant 0.$$

In this case we are actually able to invert the transform explicitly, as one may show that

$$\int_0^1 (1+u\theta)^{-a}\frac{\Gamma(1+2a)}{\Gamma(1/2+a)^2}\{\theta(1-\theta)\}^{a-1/2}\,\mathrm{d}\theta=\{\tfrac{1}{2}+\tfrac{1}{2}(1+u)^{1/2}\}^{-2a}.$$

Thus $\theta\sim\mathrm{Beta}(a+1/2, a+1/2)$, as has also been discovered through different methods in Cifarelli and Melilli (2000, p. 1393). See also our discussion of formula (3.4) below.

A similar example lets P_0 have density $f_0(x)=\pi^{-1}x^{-1/2}(1+x)^{-1}$ on $(0,\infty)$, for which one finds

$$\int_0^\infty \log(1+ux)f_0(x)\,\mathrm{d}x=2\log(1+u^{1/2}),$$

which then leads to $\mathrm{E}(1+u\theta)^{-a}=(1+u^{1/2})^{-2a}$ for the random Dirichlet mean. But one may demonstrate that

$$\int_0^\infty (1+u\theta)^{-a}\frac{\Gamma(a+1)}{\Gamma(a+\frac{1}{2})\sqrt{\pi}}\frac{\theta^{a-1/2}}{(1+\theta)^{a+1}}\,\mathrm{d}\theta=\frac{1}{(1+\sqrt{u})^{2a}}$$

for $u\geqslant 0$ and $a>0$. The density for θ is accordingly

$$d_a(\theta)=k(a)\frac{\theta^{a-1/2}}{(1+\theta)^{a+1}},\quad\text{with}\quad k(a)=\frac{\Gamma(a+1)}{\Gamma(a+\frac{1}{2})\sqrt{\pi}}.$$

Note that the random mean has a density even though P_0 has infinite mean. Cifarelli and Melilli (2000, p. 1394) also lists this result, but with an incorrect constant for the density, and as being valid only for $a \geqslant 1$. Our method shows that $d_a(\theta)$ is the correct density, also for $a \in (0, 1)$.

As a further application of relation (2.4) and, in particular, of the infinite divisibility property of (R, S), we may now rather easily determine the behaviour of θ for a large.

PROPOSITION 1. *Assume that θ_0 and σ_0, the mean and standard deviation of $Y = g(\xi)$, where $\xi \sim P_0$, are finite. Then $(a+1)^{1/2}(\theta - \theta_0)$ tends to a normal $(0, \sigma_0^2)$ as a grows.*

Proof. Take a as an integer, for simplicity, and write θ as $\bar{R}(a)/\bar{S}(a)$, where $\bar{R}(a) = \sum_{i=1}^{a} \bar{R}_i$ and $\bar{S}(a) = \sum_{i=1}^{a} \bar{S}_i$, with $(\bar{R}_i, \bar{S}_i)$ being independent from the same distribution, namely the one of (R, S) for the case of $a = 1$ (see (2.8)). These have mean vector and covariance matrix equal to

$$\begin{pmatrix} \theta_0 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \sigma_0^2 + \theta_0^2 & \theta_0 \\ \theta_0 & 1 \end{pmatrix},$$

respectively, as found by a little analysis. The conclusion follows from the central limit theorem and the delta method. □

We note that the result of course must hold with scaling $a^{1/2}$ too; we use $(a+1)^{1/2}$ here to match the exact variance. It may also be pointed out that the representation of θ as a ratio between i.i.d. averages may be used to give suitable modifications to the first order asymptotic result above, for example via saddlepoint approximations. Some such computational schemes could rely on aspects of the distribution of (R, S) for the special case of $a = 1$, for which we have formula (2.8).

We also point out that a multidimensional version of Proposition 1 may be exhibited and proved, see Section 6.1. These results also imply approximate normality of the posterior distribution of functions of random means when the prior is kept fixed and the number of observations grows, as also explained in Section 6.2.

3. Stochastic Equation Characterisations

The second basic tool used to reach results for the random mean θ is a representation as a solution to a certain stochastic equation. Such stochastic equations can be derived from the following representation of the Dirichlet process (see Sethuraman and Tiwari, 1982; Sethuraman, 1994). Let

$$P = \sum_{j=1}^{\infty} \gamma_j \delta(\xi_j), \tag{3.1}$$

where the ξ_js are independent from P_0 and independent of the weights, which are constructed in terms of a sequence $\{B_j\}$ of independent $\mathrm{Beta}(1, a)$ variables as follows: $\gamma_1 = B_1$, $\gamma_2 = (1 - B_1)B_2$, $\gamma_3 = (1 - B_1)(1 - B_2)B_3$, and so on. Then P is a.s. a $\mathrm{Dir}(aP_0)$ process.

Among the consequences of (3.1) we have the following distributional equation. Similar versions have appeared earlier in the literature, cf. Guglielmi (1998), Epifani (1999), Hjort (2000, 2003), Guglielmi and Tweedie (2001), Ishwaran and Zarepour (2002), but we include it for easy reference and since we will use similar techniques to reach other results later. We use '$=_d$' to mean equality in distribution.

PROPOSITION 2. *If* $Y = g(\xi)$ *is such that* $\mathrm{E}\log(1 + |Y|)$ *is finite, then* $\theta = \int g\,\mathrm{d}P$ *satisfies the stochastic equation*

$$\theta =_d BY + (1 - B)\theta, \quad \textit{where } B \sim \mathrm{Beta}(1, a), \tag{3.2}$$

and where on the right hand side B, Y *and* θ *are independent.*

Proof. The infinite series representation (3.1) allows θ a $\sum_{j=1}^{\infty} \gamma_j Y_j$ representation, the random sum being a.s. convergent exactly under condition (2.3). This leads to

$$\begin{aligned}\theta &= B_1Y_1 + (1 - B_1)B_2Y_2 + (1 - B_1)(1 - B_2)B_3Y_3 + \cdots \\ &= B_1Y_1 + (1 - B_1)\{B_2Y_2 + (1 - B_2)B_3Y_3 \\ &\qquad + (1 - B_2)(1 - B_3)B_4Y_4 + \cdots\},\end{aligned}$$

which can be written $B_1Y_1 + (1 - B_1)\theta'$ with $\theta' =_d \theta$. □

Such stochastic or distributional equations imply identities involving characteristic functions or Laplace transforms. Specifically, for g nonnegative, write L_0 and L for the Laplace transforms of, respectively, Y and θ. Then conditioning first on (B, Y) in (3.2), and then taking the mean value w.r.t. Y, leads to

$$L(u) = \int_0^1 L_0(ub)L(u(1 - b))\beta_a(b)\,\mathrm{d}b,$$

where $\beta_a(b)$ is the $\mathrm{Beta}(1, a)$ density. Similarly an identity might be put up using convolution, involving the density f for θ in terms of the density f_0 for Y. These identities determine in principle L for given L_0 and f for given f_0, although the exact solution might be hard to come by.

The stochastic equation (3.2) is used in the bounded case by Guglielmi (1998) to derive an expression for the generalised Stieltjes transform of θ. Furthermore, related work has led to simulation strategies for θ by exploiting Markov chains of the form $\theta_i = B_iY_i + (1 - B_i)\theta_{i-1}$ for $i \geqslant 2$, with B_i and Y_i independently drawn from respectively the $\mathrm{Beta}(1, a)$ and $Q_0 = P_0g^{-1}$. The Markov chain has the required equilibrium distribution. Guglielmi and

Tweedie (2001) and Guglielmi et al. (2002) work with such chains, following up earlier work by Feigin and Tweedie (1989).

An equivalent stochastic equation for θ, which will be employed several times in the following, can be obtained by multiplying both sides of equation (3.2) by an independent Gamma random variable with shape parameter $a+1$. By exploiting independence properties of the Beta distribution one finds

$$\theta G_{a+1} =_d G_1 Y + G_a \theta, \tag{3.3}$$

where θ is independent of G_{a+1} on the left hand side and G_1, G_a, θ and Y are independent on the right hand side, with G_s denoting a Gamma random variable with parameter $(s, 1)$.

There is some literature on similarly-structured stochastic equations, but in contexts very different from the present. See Gjessing and Paulsen (1997), Dufresne (1998) and references therein for an integro-differential equation approach and for a list of similar equations with exact solutions.

As a simple example, consider the case $Y \sim \text{Beta}(c, 1-c)$, with $0<c<1$. Then (3.3) becomes $\theta G_{a+1} =_d G_c + G_a \theta$, where all the random variables on the left and the right hand side are independent. As a direct consequence of Theorem 2 in Dufresne (1998), this equation admits the simple solution $\theta \sim \text{Beta}(a+1/2, a+1/2)$ when $c=1/2$; see Remark (7) in Section 2.

The stochastic equation (3.2) can be shown to establish, for any given a, a one-to-one correspondence between the law of a random Dirichlet mean and the distribution Q_0: for any choice of Q_0, for which the mean $\mathrm{E}\log(1+|Y|)$ is finite, the corresponding random mean θ is uniquely determined by (3.2) and, conversely, any random variable distributed as a Dirichlet mean uniquely determines through (3.2) the corresponding distribution $Q_0 = P_0 g^{-1}$. Furthermore the stochastic equation (3.3) yields a characterisation, via generalised Stieltjes transforms, of Q_0 in terms of θ.

PROPOSITION 3. *Let $P \sim \text{Dir}(aP_0)$ and $\theta = \int g\,\mathrm{d}P$, where g is a real measurable function. For any fixed $a>0$, there is a one-to-one correspondence between the class $\mathcal{Q} = \{Q_0 = P_0 g^{-1} \colon \int \log(1+|y|)\,\mathrm{d}Q_0(y) < \infty\}$ and the class $\mathcal{L} = \{\mathcal{L}(\theta) \colon \theta \text{ exists finite a.s.}\}$, where $\mathcal{L}(\theta)$ denotes the law of θ. Such a correspondence is given by the distributional equation* (3.2) *and by the transforms* (2.5) *and*

$$\mathrm{E}\left(\frac{1}{1-itY}\right) = \mathrm{E}\left(\frac{1}{1-it\theta}\right)^{a+1} \Big/ \mathrm{E}\left(\frac{1}{1-it\theta}\right)^{a} \quad \textit{for } t \in \mathcal{R}, \tag{3.4}$$

where $Y = g(\xi) \sim Q_0$.

Proof. If we fix Q_0, then equation (3.2) uniquely defines the law of θ by Lemma 3.3 in Sethuraman (1994), whenever θ exists and is finite, i.e. whenever $Q_0 \in \mathcal{Q}$. Conversely, let us fix the law of θ. Suppose that both $Y \sim Q_0 \in$

$\mathcal{Q}$ and $Y' \sim Q'_0 \in \mathcal{Q}$ satisfy equation (3.2) and therefore equation (3.3). It follows that

$$G_1 Y + G_a\theta =_d G_1 Y' + G_a\theta. \tag{3.5}$$

By computing the characteristic functions of both sides of equation (3.5) and noticing that $\mathrm{E}\exp(itG_a\theta) \neq 0$, by (2.5), one can show that equation (3.5) implies $G_1 Y =_d G_1 Y'$, which in turn implies $Y =_d Y'$.

Relation (3.4) can be derived by computing the characteristic function of the left and the right hand side of equation (3.3). □

The existence of a one-to-one correspondence is proved in Lijoi and Regazzini (2001) via a completely different route, making use of complex theory arguments. Relation (3.4) appears to be new, as well as the fact that the stochastic equation (3.2) determines a bijection between random Dirichlet means and the corresponding distribution Q_0. See also Cifarelli and Regazzini (1993, Section 7).

Another application of the stochastic equation (3.2) gives a description of θ in terms of its moments. Expressions for moments and sufficient conditions for their existence are obtained in Regazzini (1998) and Epifani (1999) in terms of complete Bell exponential polynomials. The proofs of such elaborated expressions are rather technical; they are based on the theory of special functions (Regazzini, 1998) and on approximating the Dirichlet process by Bernshteĭn polynomials (Epifani, 1999). See also Cifarelli and Melilli (2000, Remark 2.1).

Here we shall derive, as a straightforward consequence of (3.2), a simple recursive formula for direct moments $\mathrm{E}\theta^p$ and centralised moments $\mathrm{E}(\theta - \theta_0)^p$. As a further consequence of the stochastic equations (3.2) and (3.3), we shall also give new necessary and sufficient conditions for existence of such moments.

PROPOSITION 4. *Let* $P \sim \mathrm{Dir}(aP_0)$ *and* $\theta = \int g\,\mathrm{d}P$ *where* g *is a measurable function such that* $\mathrm{E}\log(1+|g(\xi)|)$ *is finite. Then, for any positive integer* p, *the following three conditions are equivalent*: (1) $\mathrm{E}|\theta|^p < \infty$; (2) $\mathrm{E}|g(\xi)|^p < \infty$; *and* (3) $\mathrm{E}(\int |g|\,\mathrm{d}P)^p < \infty$. *Furthermore, under any of the above conditions, we have*

$$\mathrm{E}(\theta - x)^p = a(p-1)! \sum_{j=0}^{p-1} \frac{1}{j!(a+j)^{[p-j]}} \mathrm{E}(Y - x)^{p-j}\,\mathrm{E}(\theta - x)^j, \tag{3.6}$$

where x *is an arbitrary real number and* $y^{[p]} = y(y+1)\cdots(y+p-1)$.

Proof. Let us first prove expression (3.6) assuming that $\mathrm{E}|\theta|^p<\infty$ and $\mathrm{E}|Y|^p<\infty$, where $Y=g(\xi)$. From (3.2),

$$(\theta-x)^p =_d \sum_{j=0}^{p}\binom{p}{j}\{B(Y-x)\}^{p-j}(1-B)^j(\theta-x)^j$$

for any x, implying

$$\mathrm{E}(\theta-x)^p=\frac{1}{1-\mathrm{E}(1-B)^p}\sum_{j=0}^{p-1}\binom{p}{j}\mathrm{E}B^{p-j}(1-B)^j\,\mathrm{E}(Y-x)^{p-j}\,\mathrm{E}(\theta-x)^j.$$

Formula (3.6) then follows by using the formula $\mathrm{E}B^{p-j}(1-B)^j=a(p-j)!\,\Gamma(a+j)/\Gamma(1+a+p)$ for the Beta$(1,a)$ variable B.

Let us consider now the equivalence of the three conditions. Clearly, (3) implies (1). Let us show that (1) implies (2). As a consequence of stochastic equation (3.3), condition (1) implies that $\mathrm{E}|\theta G_{a+1}|^p$ and therefore $\mathrm{E}|G_1Y+G_a\theta|^p$ are finite. By independence of G_1Y and $G_a\theta$, this implies (see, for example, Lemma 3 of Section V.6 in Feller, 1966) that $\mathrm{E}|G_1Y|^p<\infty$, which in turn implies $\mathrm{E}|Y|^p<\infty$.

We conclude the proof by showing that (3) is a consequence of (2). We prove it by induction. It is easy to check, using the Sethuraman–Tiwari representation, that the results hold for $n=1$. Suppose now that it is true for an arbitrary n, that is, suppose that $\mathrm{E}|Y|^n<\infty$ implies $\mathrm{E}(\theta^+)^n<\infty$, where $\theta^+=\int|g|\,\mathrm{d}P$, and let us prove it for $n+1$. If $\mathrm{E}|Y|^{n+1}<\infty$ then $\mathrm{E}|Y|^n<\infty$ which then implies, by hypothesis, $\mathrm{E}(\theta^+)^n<\infty$. Consider now, for a positive integer k, the function $t_k(x)=|x|I\{|x|\leqslant k\}+kI\{|x|>k\}$. Clearly, $0\leqslant t_k(x)\leqslant k$ and $t_k(x)\uparrow|x|$. Moreover, let $\theta_k=\int t_k(g)\,\mathrm{d}P$ and notice that, by monotone convergence, $\theta_k\uparrow\theta^+$ a.s.. As θ_k and $Y_k=t_k(Y)$ are bounded random variables, they admit moments of any order and we can therefore apply the recursive formula (3.6) obtaining

$$\mathrm{E}(\theta_k)^{n+1}=a\,n!\sum_{j=0}^{n}\frac{1}{j!(a+j)^{[n+1-j]}}\mathrm{E}(Y_k)^{n+1-j}\,\mathrm{E}(\theta_k)^j.$$

By the theorem on monotone convergence one then finds

$$\mathrm{E}(\theta^+)^{n+1}=a\,n!\sum_{j=0}^{n}\frac{1}{j!(a+j)^{[n+1-j]}}\mathrm{E}(|Y|)^{n+1-j}\,\mathrm{E}(\theta^+)^j,$$

the expression on the right being finite, as $\mathrm{E}|Y|^{n+1}$ and $\mathrm{E}(\theta^+)^n$ are finite. □

Formula (3.6) gives an easily implementable recursive computational scheme for finding even higher order centralised moments for θ. These may e.g. be used to obtain numerical approximations to its density; for one such idea, see Section 6.3. For Proposition 4 it is also worth noticing the equivalence of condition (1) with the apparently stronger condition (3); this equivalence is by no means true for general random probability measures.

4. Results for Symmetric, Unimodal, and Stable Distributions

In this section we shall apply the two representations above, when P_0 belongs to suitable general classes of interest, to infer aspects of the distribution of the consequent random Dirichlet mean. Some of these results might also be helpful in situations where the statistician needs to elicit the prior measure P_0.

4.1. SYMMETRIC AND UNIMODAL DISTRIBUTIONS

As a direct consequence of Proposition 3 and of the stochastic equation (3.2) we may obtain a simple proof of the following result which characterises the class of symmetric random means. A different proof, based on a contour integral expression of the characteristic function of θ derived through multiple hypergeometric functions, can be found in Lijoi and Regazzini (2001).

PROPOSITION 5. *Suppose θ exists and is finite. Then θ is symmetric if and only if Q_0 is symmetric.*

Proof. Suppose that the law of θ is symmetric, i.e. $\theta =_d -\theta$. One has $\theta =_d BY + (1 - B)\theta$ and $-\theta =_d B(-Y) + (1 - B)(-\theta)$ and by the symmetry of θ follows $\theta =_d B(-Y) + (1 - B)\theta$ which then implies, by Proposition 3, that $Y =_d (-Y)$. An analogous argument shows the converse. □

A partially analogous result can be obtained using representation (2.2) if one considers symmetric and unimodal distribution. We shall say that a distribution function F is unimodal with vertex x_0 if there exists a real number x_0 such that $F(x)$ is convex for $x < x_0$ and concave for $x > x_0$. This means that F is absolutely continuous, except possibly at x_0, and that its density is monotone in the intervals $\{x < x_0\}$ and $\{x > x_0\}$.

PROPOSITION 6. *Suppose that the distribution of $Y = g(\xi)$ is symmetric and unimodal with vertex c and such that* $\mathrm{E}\log(1 + |Y|) < \infty$. *Then the distribution of $\theta = \int g\,\mathrm{d}P$ is symmetric and unimodal with the same vertex.*

Proof. We can assume without loss of generality that $c = 0$, as $\int (g - c)\,\mathrm{d}P = \theta - c$. Consider the approximation $\theta_m = \sum_{j=1}^m \beta_j Y_j$ to θ, as in (2.2), and let us first prove that θ_m is symmetric and unimodal with vertex 0 (denoted by S-U). Clearly, $\beta_j Y_j \mid \beta_j$ is S-U. The same is true for $\sum_{j=1}^m \beta_j Y_j \mid (\beta_1, \ldots, \beta_m)$, as a sum of independent S-U random variables is S-U; see, for example, Theorem 4.5.5 in Lukacs (1970). But this holds also unconditionally, i.e. for θ_m, since it is easy to check that if the conditional distribution of $T \mid V$ is a.s. S-U then T is also S-U. The limit θ is symmetric by Proposition 5, or, alternatively, by noticing that since θ_m and $-\theta_m$ converge, respectively, to θ and $-\theta$ and $\theta_m =_d -\theta_m$, then necessarily $\theta =_d -\theta$. Finally, θ is also unimodal, being a limit of unimodal random variables; for a proof of this see Theorem 4.5.4 in Lukacs (1970). □

The converse to Proposition 6 is not true: a counterexample is given by the case $Y \sim \text{Beta}(1/2, 1/2)$ and $\theta \sim \text{Beta}(a+1/2, a+1/2)$ considered in Section 2's Remark (7).

4.2. STABLE DISTRIBUTIONS

Proposition 7 below gives a characterisation for the family of random means obtained starting from stable distributions. It essentially says that if $Y = g(\xi) \sim Q_0$ has a stable law, then $\theta = \int g\,\mathrm{d}P$ is a scale mixture of Y. Several characterisations of the mixing distribution are also given.

Recall that a general nondegenerate stable law $\text{Stab}(\mu, \sigma, p, \gamma)$ has characteristic function given by

$$\phi(t, \mu, \sigma, p, \gamma) = \exp\{i\mu t - |\sigma t|^p (1 + i\gamma K(p, t))\}, \tag{4.1}$$

where

$$K(p, t) = (\operatorname{sign} t)\tan(\tfrac{1}{2}p\pi) \quad \text{if } p \neq 1,\ 0 < p \leqslant 2,$$

while being equal to $(2/\pi)(\operatorname{sign} t)\log|t|$ if $p = 1$, and $\sigma > 0$, $\mu \in \mathcal{R}$, and $-1 \leqslant \gamma \leqslant 1$. We shall denote by Φ the class of all stable laws $\text{Stab}(\mu, \sigma, p, \gamma)$ and by Φ^- the subclass of Φ obtained by excluding the case $\{p = 1, \gamma \neq 0\}$.

For the following proposition, let $T = (\sum_{j=1}^{\infty} \gamma_j^p)^{1/p}$, for the given positive p, in terms of the Sethuraman–Tiwari representation (3.1). It is also the limit in distribution of $T_m = (\sum_{j=1}^{m} \beta_j^p)^{1/p}$, in terms of our representation (2.1) (see Hjort and Ongaro, 2005). One may show that the distribution of T is uniquely determined by the stochastic equation

$$T^p =_d B^p + (1-B)^p T^p, \tag{4.2}$$

where $B \sim \text{Beta}(1, a)$ is independent of T. Another characterisation of the distribution of T^p via Laplace transforms is given in Hjort and Ongaro (2005), where the asymptotic behaviour of T as a tends to infinity is also established.

PROPOSITION 7. *Let* $Q_0 \in \Phi^-$. *This is equivalent to assuming that if* $Y, Y_1, \ldots, Y_n$ *are i.i.d. random variables from* Q_0, *then there exist* $0 < p \leqslant 2$ *and* $\mu \in \mathcal{R}$ *such that one of the following two equivalent conditions hold*:

(a) $\sum_{i=1}^{n} Y_i \sim n^{1/p} Y + \mu(n - n^{1/p})$ *for* $n \geqslant 1$;
(b) $a_1(Y_1 - \mu) + a_2(Y_2 - \mu) \sim (a_1^p + a_2^p)^{1/p}(Y - \mu)$ *for* $a_1, a_2 \geqslant 0$.

Then $\theta = \int g \, \mathrm{d}P$, *where* $P \sim \mathrm{Dir}(aP_0)$, *has distribution given by*

$$\theta =_d \mu + (Y - \mu)\, T,$$

where $Y \sim Q_0$ *is independent of* T.

Proof. It is immediate to see that if $Q_0 \in \Phi^-$ then (b) is satisfied. Furthermore (b) implies (a). The fact that (a) implies $Q_0 \in \Phi^-$ follows from general theory of stable laws (see e.g. Hoffmann-Jorgensen, 1994). Let us next work with the characteristic function of $\theta_m = \sum_{j=1}^{m} \beta_j Y_j$. By conditioning on the β_js, we have

$$\mathrm{E}\exp(it\theta_m) = \mathrm{E}\,\phi(t, \mu, T_m\sigma, p, \gamma),$$

which implies $\theta_m - \mu =_d (Y - \mu)\, T_m$, where $Y \sim P_0$ is independent of $T_m = (\sum_{j=1}^{m} \beta_j^p)^{1/p}$. The required result follows from convergence of θ_m to θ and of T_m to T. □

Special examples of the above result are the Cauchy and the normal distributions corresponding, respectively, to the cases $p = 1$ (and $\gamma = 0$) and $p = 2$. In particular, for the Cauchy distribution one has $T_m = T = 1$, which gives the known result θ is also a Cauchy, first discovered by Yamato (1984).

A further representation of θ can be given for a mixture of Cauchy distributions: if Y is a scaled mixture of Cauchys, then so is θ, and the mixing distribution for θ can be written as a random mean from a Dirichlet process with parameter proportional to the distribution mixing the Cauchy.

PROPOSITION 8. *Let* $X \sim R$ *be a nonnegative random variable and let* $Q_0 = P_0 g^{-1}$ *be an* R*-mixture of scaled Cauchys, i.e.* Q_0 *is the law of* XZ, *where* $Z \sim$ Cauchy *is independent of* X. *Then* $\gamma = \int g \, \mathrm{d}P$, *with* $P \sim \mathrm{Dir}(aP_0)$, *is distributed as* $\gamma =_d Z\theta$, *where* Z *is independent of the nonnegative random variable* θ. *Furthermore* $\theta =_d \int x \, \mathrm{d}P$, *where* $P \sim \mathrm{Dir}(aR)$, *and is therefore uniquely determined by the stochastic equation*

$$\theta =_d BX + (1 - B)\theta, \quad \textit{where } B \sim \mathrm{Beta}(1, a)$$

with X, B *and* θ *independent.*

Proof. Consider $\gamma_m = \sum_{j=1}^{m} \beta_j \varepsilon_j$, where ε_j is from the Cauchy mixture Q_0, that is, $\varepsilon_j \,|\, \sigma_j$ is Cauchy (σ_j) where σ_j is drawn from R. Hence $\mathrm{E}\exp(iu\varepsilon_j) = \mathrm{E}\exp(-|u|\sigma_j)$ and

$$\mathrm{E}[\exp(iu\beta_j\varepsilon_j) \,|\, \beta] = \mathrm{E}[\exp(-|u|\beta_j\sigma_j) \,|\, \beta].$$

It follows that $\mathrm{E}\exp(iu\gamma_m) = \mathrm{E}\exp(-|u|\theta_m)$ where θ_m is our usual approximation to θ. This implies that $\mathrm{E}\exp(-|u|\theta_m)$ converges to $\mathrm{E}\exp(-|u|\theta)$, which can be proven to be the characteristic function of $Z\theta$.

□

5. Mixtures of Dirichlet Means and Means of Dirichlet Mixtures

This section first gives characterisations and representations of random Dirichlet means $\theta = \int g\,\mathrm{d}P$ when the base measure P_0 is a finite mixture. This leads in particular to an exact representation of the posterior distribution of θ, to new exact simulation recipes, and to a new proof of the Sethuraman–Tiwari representation of a random mean. Then we go on to reach results for random means from mixtures of Dirichlet process priors.

5.1. RANDOM MEANS WHEN P_0 IS A MIXTURE

Our first result is as follows.

PROPOSITION 9. *Suppose* $P \sim \mathrm{Dir}(aP_0)$, *where* P_0 *has a mixture representation of the form* $\sum_{i=1}^k p_i P_{0,i}$. *If* g *is a measurable function such that* $\mathrm{E}\log\{1+|g(\xi)|\}$ *is finite, then* $\theta = \int g\,\mathrm{d}P$ *admits the decomposition*

$$\theta =_d \sum_{i=1}^{k} D_i\theta_i \tag{5.1}$$

in which $D = (D_1, \dots, D_k)$ *is Dirichlet* $(ap_1, \dots, ap_k)$ *and independent of* $\theta_1, \dots, \theta_k$. *These are independent among themselves and* $\theta_i = \int g\,\mathrm{d}P_i$, *where* $P_i \sim \mathrm{Dir}(ap_i P_{0,i})$.

Proof. We first mention that a constructive proof may be given based on the approximation θ_m to θ. The following proof is somewhat simpler, checking directly that the given decomposition of θ has the correct Stieltjes transform, i.e. that it satisfies relation (2.5). Notice also that existence of the θ_is follows from the finiteness assumption on $\mathrm{E}\log\{1+|g(\xi)|\}$.

By conditioning on $\sum_{i=1}^k D_i\theta_i$ we have

$$\mathrm{E}\exp\left\{it\left(G_a\sum_{i=1}^{k} D_i\theta_i\right)\right\} = \mathrm{E}\left(\frac{1}{1-it\sum_{i=1}^k D_i\theta_i}\right)^a.$$

Well known independence properties of the Dirichlet distribution yield the distributional equation $G_a\left(\sum_{i=1}^k D_i\theta_i\right) =_d \sum_{i=1}^k G_{ap_i}\theta_i$, where on the left hand side G_a is independent of the sum and all the random variables variables on the right hand side are independent. Consequently, the left hand side of the above equation is equal to

$$\mathrm{E}\exp\left(it\sum_{i=1}^{k}G_{ap_i}\theta_i\right)=\exp\left[-a\sum_{i=1}^{k}p_i\int\log\{1-itg(x)\}\,\mathrm{d}P_{0,i}(x)\right],$$

completing the proof. □

Proposition 9 has a number of interesting consequences.

(1) Consider the simple case where $P_{0,i}=P_0$ and $p_i=1/k$. This gives a representation of θ as a symmetric Dirichlet mixture of i.i.d. copies of a random mean from a Dirichlet process with total mass a/k. This allows one to extend knowledge about the distribution of θ from a $\mathrm{Dir}(cP_0)$ to the distribution of θ from $\mathrm{Dir}(kcP_0)$, for any integer k.

For example, when a is an integer one has $\theta=\sum_{i=1}^{a}D_i\theta_i$ where D is from the flat Dirichlet$(1,\dots,1)$ and the θ_is are from a Dirichlet process with total mass 1. As the density of such random means is known, see (2.7), this permits exact simulation from θ. Simulation schemes of this type are related to the one described after expression (2.8).

As a special case, let P_0 correspond to the density of $\xi=\exp(C)/\{1+\exp(C)\}$, where C is a Cauchy distributed random variable, and let $a=1$. Then θ has a uniform distribution on $(0,1)$ (Diaconis and Kemperman, 1996). Thus when $P\sim\mathrm{Dir}(aP_0)$, θ can be represented as $\sum_{i=1}^{a}D_i\theta_i$, where $\theta_1,\dots,\theta_a$ are i.i.d. and uniform, and $(D_1,\dots,D_a)$ is a flat Dirichlet. With some efforts one finds the density for $a=2$, for example;

$$d_2(t)=-2[(1-t)\log(1-t)+t\log t)]\quad\text{for } t\in(0,1).$$

(2) Next note that if we take the limit for $k\to\infty$ in (5.1) with $P_{0,i}=P_0$ and $p_i=1/k$, we actually obtain the Sethuraman–Tiwari sum representation $\theta=\sum_{j=1}^{\infty}\gamma_j g(\xi_j)$. That this is true rests on two arguments. The first is that $P\sim\mathrm{Dir}(\varepsilon P_0)$ with a small ε is close to a point mass at a random point $\xi\sim P_0$, so that $\theta_i=\int g\,\mathrm{d}P_i$ in (5.1) is close to $g(\xi_i)$, when $p_i=1/k$ is small. The second is that the symmetric Dirichlet distribution with parameters $(a/k,\dots,a/k)$ converges in distribution to the collection of random weights in the Sethuraman–Tiwari representation (3.1), after a re-ordering of the weights according to their sizes; this is related to material presented in Billingsley (1999, Section 4) and is more formally proved in Ongaro (2005).

(3) Consider partitioning the sample space into separate regions $A_1,\dots,A_k$ such that $P_0(A_i)>0$ and let $P_{0,i}(A)=P_0(A\cap A_i)/P_0(A_i)$ be the normalised restriction of P_0 to A_i. We thus obtain a decomposition of θ in terms of independent random means from Dirichlet processes defined on different regions. If the $A_1,\dots,A_k$ division is into many small cells, centred at say $x_1,\dots,x_k$, and g does not vary much over any of these, then $\theta\approx\sum_{i=1}^{k}D_i g(x_i)$, where $(D_1,\dots,D_k)$ is a Dirichlet vector with parameters $(aP_0(A_1),\dots,aP_0(A_k))$.

For another example, by taking $\Omega^+ = \{x : g(x) \geqslant 0\}$ and Ω^- its complement, one has

$$\theta =_d B\theta^+ - (1-B)\theta^-,$$

where $B = \mathrm{Beta}(aP_0(\Omega^+), a(1 - P_0(\Omega^+))$, $\theta^+ = \int_{\Omega^+} g^+ \,\mathrm{d}P^+$ and $\theta^- = \int_{\Omega^-} g^- \,\mathrm{d}P^-$. Here g^+ and g^- denote the positive and negative part of g, so that $g = g^+ - g^-$, and P^+ and P^- are Dirichlet processes with parameters $aP_0/P_0(\Omega^+)$ and $aP_0/P_0(\Omega^-)$ on Ω^+ and Ω^-. This allows studying properties of a general random mean from properties of random means of positive functions, extending in particular the known examples of distributions of random means; in fact the list of examples provided in Cifarelli and Melilli (2000) of distributions for random means all deal only with positive functions.

(4) Consider next the general case where the $P_{0,i}$ are neither identical nor defined on different regions. Let us first take the simple case where $P_0 = pP_0' + (1-p)\delta(x)$, with x a fixed point belonging to the sample space. Then θ admits the simple representation

$$\theta =_d g(x) + B\{\theta' - g(x)\},$$

where $B \sim \mathrm{Beta}(ap, a(1-p))$ is independent of $\theta' = \int g \,\mathrm{d}P'$, with $P' \sim \mathrm{Dir}(apP_0')$. For example, if $P_0' = \mathrm{Beta}(1/2, 1/2)$, $g(x) = 0$ and $p = 1 - 1/(2a)$, with $a > 1/2$, we obtain $\theta =_d \mathrm{Beta}(a - 1/2, a + 1/2)$, with an analogous result holding for $g(x) = 1$.

More substantially, the above mixture model can be used to construct nonparametric priors centred at a multi-modal distribution. For example, consider the bimodal distribution $P_0 = pP_{0,1} + (1-p)P_{0,2}$, where $P_{0,i} = \mathrm{Cauchy}(\mu_i, \sigma^2)$. Then a direct application of Propositions 7 and 9 gives

$$\theta =_d B\mu_1 + (1-B)\mu_2 + Y,$$

where $Y =_d \mathrm{Cauchy}(0, \sigma^2)$ is independent of B. As a second example, consider a mixture $P_0 = \sum_{i=1}^k p_i P_{0,i}$ of normals, with $P_{0,i} = \mathrm{N}(\mu_i, \sigma^2)$. Then, again appealing to Propositions 7 and 9, one finds the representation

$$\theta =_d \sum_{i=1}^k D_i(\mu_i + T_i \sigma Z_i) =_d \sum_{i=1}^k D_i \mu_i + \sigma Z \left(\sum_{i=1}^k D_i^2 T_i^2 \right)^{1/2}$$

in terms of independent standard normals $Z_1, \ldots, Z_k, Z$. Here $(D_1, \ldots, D_k)$ is a Dirichlet $(ap_1, \ldots, ap_k)$, independent of $T_1, \ldots, T_k$, also independent among themselves, with $T_i \sim W(ap_i)$, say; here $W(b)$ is the distribution of $T(b) = (\sum_{j=1}^\infty \gamma_j^2)^{1/2}$, where the random weights γ_j are as in (3.1), in terms of a sequence of $\mathrm{Beta}(1, b)$ variables.

(5) As a final and statistically important illustration of the use of Proposition 9, consider the posterior distribution of θ, given a random sample of

observations $X_1, \dots, X_n$. This is a random mean from a Dirichlet process with strength parameter $a+n$ and probability measure given by the mixture $w_n P_0 + (1-w_n)P_n$, where $P_n = n^{-1}\sum_{i=1}^n \delta(X_i)$ is the empirical distribution and $w_n = a/(a+n)$. The random mean from a Dirichlet process with parameter nP_n is $\theta_{P_n} = \sum_{i=1}^n D_i g(X_i)$ where the D_is are drawn from a symmetric Dirichlet $(1, \dots, 1)$; an explicit (though still cumbersome) expression for the density $d_n(\theta)$ of such a random mean can be found, see Cifarelli and Melilli (2000, Example 3.1). By applying the above proposition it follows that the posterior random mean $\theta^{(n)}$ can be decomposed as

$$\theta^{(n)} =_d B\theta + (1-B)\theta_{P_n}, \tag{5.2}$$

where $B \sim \mathrm{Beta}(a, n)$, θ and θ_{P_n} are independent. This is a useful description of $\theta^{(n)}$, in terms of available information for the prior random mean θ, and gives also an exact and easy to implement simulation recipe. The exact posterior density may be written

$$d(\theta \,|\, \mathrm{data}) = \int\int d_n\left(\frac{\theta - bt}{1-b}\right)\beta(b; a, n+a)g_0(t)\,\mathrm{d}b\,\mathrm{d}t,$$

where $\beta(b; a, n+a)$ is the Beta density and $g_0(t)$ is the prior density for θ.

5.2. MIXTURES OF DIRICHLET PRIORS

Our theory can also be fruitfully applied to reach results for and make inference in nonparametric hierarchical models. In particular mixtures of Dirichlet process priors (MDPs) have received considerable attention in the literature (see, for example, West et al., 1994; Escobar and West, 1998; MacEachern, 1998). We give two types of illustrations.

Assume first that σ has a prior $\pi(\sigma)$ and that P for given σ is a Dirichlet process with parameter $a\mathrm{N}(0, \sigma^2)$, and consider the random mean $\theta = \int x\,\mathrm{d}P$ for this particular MDP. Using Proposition 7 we see that $\theta \,|\, \sigma$ may be represented as $\sigma T Z$, where Z is standard normal and T is the limit of $T_m = (\sum_{j=1}^m \beta_j^2)^{1/2}$. The marginal distribution of θ is accordingly that of UZ, another random scaling of the standard normal, where $U = \sigma T$, with $\sigma \sim \pi(\cdot)$, independent of T.

A similar MDP example is when μ has a prior $\pi_0(\mu)$ and $P \,|\, \mu$ is a Dirichlet $a\mathrm{N}(\mu, \sigma_1^2)$, for some fixed σ_1. Then examination of earlier arguments shows that $\theta \,|\, \mu$ can be represented as $\mu + T\,\mathrm{N}(0, \sigma_1^2)$. The unconditional distribution of θ, which in typical MDP applications would mean the inferred prior distribution of the mean parameter, is then obtained by integrating out μ w.r.t. π_0. If the prior for μ is a normal $(0, \sigma_0^2)$, for example, then $(\theta \,|\, T) \sim \mathrm{N}(0, \sigma_0^2 + T^2\sigma_1^2)$. In particular, if μ in this construction is given a flat (improper) prior, then $\theta = \int x\,\mathrm{d}P$ has a flat prior too.

Our second type of illustration corresponds to the more typical use of MDPs in Bayesian analysis, where the model framework is as follows. At level 1 there is a Dirichlet process $P \sim \mathrm{Dir}(aP_0)$. Conditionally on P, at level 2, parameters $\alpha_1, \ldots, \alpha_n$ are drawn i.i.d. from P. These are not observed. At level 3 we have the observed data $= (X_1, \ldots, X_n)$, which conditionally on P and the α_is are independent, with X_i having distribution say $F(x_i \,|\, \alpha_i)$. There are more general versions of this framework, perhaps with the parameters α_i being vectors, and perhaps with a further prior on parameters inside P_0, cf. references mentioned above.

For such hierarchical models, the posterior distribution of P is a mixture of Dirichlet processes (Antoniak, 1974). More specifically, it can be represented by mixing the posterior Dirichlet distribution of P given the sample $\mathcal{A} = \{\alpha_1, \ldots, \alpha_n\}$ with the distribution $H_n(\cdot \,|\, \mathrm{data})$ of $\mathcal{A} \,|\, \mathrm{data}$. In symbols,

$$P^{(n)} = (P \,|\, \mathrm{data}) \sim \int \mathrm{Dir}\left(aP_0 + \sum_{i=1}^{n} \delta(\alpha_i)\right) \mathrm{d}H_n(\alpha_1, \ldots, \alpha_n \,|\, \mathrm{data}).$$

If the conditional distribution $F(\cdot \,|\, \alpha_i)$ of $X_i \,|\, \alpha_i$ admits a density $f(\cdot \,|\, \alpha_i)$ then $H_n(\mathcal{A} \,|\, \mathrm{data})$ has density proportional to $\prod_{i=1}^{n} f(X_i \,|\, \alpha_i)$ with respect to the marginal distribution of $\mathcal{A}$. Efficient algorithms for simulating from $H_n(\mathcal{A} \,|\, \mathrm{data})$ have been developed, cf. Escobar (1994), MacEachern (1994, 1998), Escobar and West (1995), and Ishwaran and James (2001).

Consider now the posterior random mean $\theta^{(n)} = \int g \,\mathrm{d}P^{(n)}$. Then, as $(\theta^{(n)} \,|\, \mathcal{A}) =_d (\theta \,|\, \mathcal{A})$, the discussion leading to (5.2) gives the representation $B\theta + (1-B)\sum_{i=1}^{n} D_i g(\alpha_i)$ for $\theta^{(n)}$ given the latent $\mathcal{A}$ parameters, again with $B \sim \mathrm{Beta}(a, n)$ independent of a flat Dirichlet$(1, \ldots, 1)$ for $(D_1, \ldots, D_n)$. If we then average with respect to the distribution of $\mathcal{A} \,|\, \mathrm{data}$, we obtain the representation

$$\theta^{(n)} =_d B\theta + (1-B)\sum_{i=1}^{n} D_i g(Z_i), \tag{5.3}$$

where $(Z_1, \ldots, Z_n) \sim H_n(\mathcal{A} \,|\, \mathrm{data})$ is independent of all the other random quantities. It follows that also in this more general hierarchical model we are able to perform exact simulation for the posterior mean $\theta^{(n)}$ whenever simulation procedures are available for the prior random mean θ.

From (5.3) we also find the Bayes estimate under quadratic loss, namely

$$\widehat{\theta} = \mathrm{E}(\theta \,|\, \mathrm{data}) = w_n\theta_0 + (1-w_n)\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\{g(\alpha_i) \,|\, \mathrm{data}\}$$

with θ_0 the prior mean of θ. Explicit formulae may be put up for the conditional mean of $g(\alpha_i)$ here, for some typical models, but these would be cumbersome and computationally demanding. They would typically involve sums over all possible clusters patterns formed by the $\alpha_1, \ldots, \alpha_n$ sample. It is therefore easiest to compute $\mathrm{E}\{g(\alpha_i) \mid \text{data}\}$, along with other quantities of interest, by simulation of $\mathcal{A}$ vectors conditionally on the data. The posterior variance may also be found along such lines. Such schemes lead to more accurate inference than when using Markov chain simulation methods for the full $(P, \mathcal{A})$.

In applications of the MDP apparatus statisticians sometimes use flat noninformative priors on some parameters, which as we have seen above may lead to an implied improper prior for θ. It then follows from (5.3) that also the posterior is improper, with probability one, for all sample sizes, rendering Bayes analysis meaningless. This is not easily detected by the McMC algorithms.

6. Bernshteĭn–von Mises Theorems and Other Developments

Here we briefly outline some further themes that can be worked with using the machinery we have developed in this paper. These include extensions to the multidimensional case, nonparametric Bernshteĭn–von Mises theorems, and a general approximation algorithm for the densities of random means.

6.1. THE MULTIDIMENSIONAL CASE

Diaconis and Kemperman (1996) mentioned specifically that there seems to be no results in the literature on the joint distribution of several random means. This would be needed for determining the distribution of a random variance, for example. The methods developed in this paper lend themselves nicely also to the multidimensional framework. The following can be proved along the lines of results arrived at above.

Let $g_1, \ldots, g_k$ be nonnegative functions on the sample space Ω with finite integrals $\int \log(1+g_i)\,\mathrm{d}P_0$. Then the joint distribution of the vector of k Dirichlet means $\theta_i = \int g_i\,\mathrm{d}P$ is determined by the following transform, which we may term 'the simultaneous Stieltjes transform of order a', as follows:

$$\mathrm{E}\exp\Big\{-a\log\Big(1+\sum_{i=1}^{k} u_i\theta_i\Big)\Big\} = \exp\Big[-a\int\log\Big\{1+\sum_{i=1}^{k} u_i g_i(\xi)\Big\}\mathrm{d}P_0(\xi)\Big].$$

Similar results have appeared in Lijoi and Regazzini (2001), Kerov and Tsilevich (2001), and Regazzini et al. (2002). We point out that the two quantities appearing here are identical to the joint Laplace transform $\mathrm{E}\exp(-\sum_{i=1}^{k} u_i R_i)$, where $R_i = G_a\theta_i$ with G_a a Gamma variable with parameters $(a, 1)$, independent of the θ_is, cf. (2.4). We may also view these

R_is as integrals of functions w.r.t. a Gamma process. The distribution of $(\theta_1, \dots, \theta_k)$ is determined by the above transform, and can also be characterised via the simultaneous Laplace transform of $(R_1, \dots, R_k, G_a)$, which becomes

$$\mathrm{E}\exp\Big(-\sum_{i=1}^{k} u_i R_i - vG_a\Big) = \exp\Big[-a\int \log\Big\{1+\sum_{i=1}^{k} u_i g_i(\xi)+v\Big\}\,\mathrm{d}P_0(\xi)\Big].$$

One may also prove a vector version of the distributional equation (3.2), cf. Hjort (2000).

6.2. NONPARAMETRIC BERNSHTEĬN–VON MISES THEOREMS

For classical parametric inference there are various results commonly referred to as Bernshteĭn–von Mises theorems about connections and comparisons with maximum likelihood and Bayes estimators (see e.g. LeCam and Yang, 1990, Ch. 7). To explain the nature of these, consider a parametric model indexed by θ and with maximum likelihood estimator $\widehat{\theta}_{\mathrm{ml}}$, and for which a classic result is that $\sqrt{n}(\widehat{\theta}_{\mathrm{ml}}-\theta)$ tends to $\mathrm{N}_p(0, J(\theta)^{-1})$, where J is the information matrix and p the dimension of the parameter vector. The accompanying Bernshteĭn–von Mises theorem is that if $\widehat{\theta}_B$ is the Bayes estimator, for any given prior, then the posterior distribution of $\sqrt{n}(\theta-\widehat{\theta}_B)$ also tends to the $\mathrm{N}(0, J(\theta)^{-1})$, under mild regularity conditions. This also leads to the approximation

$$\theta \,|\, \mathrm{data} \approx_d \mathrm{N}_p(\widehat{\theta}_{\mathrm{ml}}, J(\widehat{\theta}_{\mathrm{ml}})^{-1}/n) \quad \text{with probability 1,} \tag{6.1}$$

which indicates that Bayesian schemes for large n will lead to inference equivalent to that of maximum likelihood.

Such results can not be taken for granted in Bayesian nonparametrics, as there are many examples to the contrary; see comments in e.g. Hjort (2003). It is comforting, then, to see that results similar to (6.1) will be guaranteed for most parameters $\theta=\theta(P)$, when the P is given the nonparametric Dirichlet process prior. To demonstrate this, we start with a multivariate version of Proposition 1 of Section 2. We observe that the vectors studied in Section 6.1 are jointly infinitely divisible, due to the form of the joint Laplace transforms given there. Let θ_0 be the vector of means and $\Sigma(P_0)$ the covariance matrix of $g_1(\xi), \dots, g_k(\xi)$, where $\xi \sim P_0$. The result is then that as a goes to infinity,

$$a^{1/2}(\theta-\theta_0) \to_d \mathrm{N}_k(0, \Sigma(P_0)). \tag{6.2}$$

This is close to implying limiting normality of the posterior distribution of smooth functions of θ, as the prior is kept fixed and the sample size n

grows, the key being that the posterior distribution of a Dirichlet process is still a Dirichlet process with parameter $(a+n)\widehat{P}_n$, where the centre distribution is

$$\widehat{P}_n = w_n P_0 + (1-w_n)P_n, \quad \text{where } w_n = a/(a+n)$$

and P_n is the empirical distribution of the data points x_i. Result (6.2) can not be applied immediately, since $\widehat{P}_n$ changes with n. Assume that the data points in reality have emerged from a distribution $P_{\rm true}$. Then the event that P_n (and hence $\widehat{P}_n$) converge to this $P_{\rm true}$ is 1, by the Glivenko–Cantelli theorem. Modifications of previous arguments now show that

$$(n+a)^{1/2}\{\theta(P)-\widehat{\theta}_n\}\,|\,\text{data} \to_d \mathrm{N}_k(0, \Sigma(P_{\rm true})) \quad \text{a.s.}, \tag{6.3}$$

where $\widehat{\theta}_n = \theta(\widehat{P}_n) = w_n\theta_0 + (1-w_n)\bar{g}_n$, writing $\bar{g}_n = n^{-1}\sum_{i=1}^n g(x_i)$, and $\Sigma(Q)$ is defined as the variance matrix of the $g_j(\xi)$ when $\xi \sim Q$.

Result (6.3) may be formally proved by first demonstrating a generalisation of (6.2), valid for Dirichlet process centre parameters that may change with a as a increases. Let $\theta_a = \int g\,\mathrm{d}P_a$, where $P_a \sim \mathrm{Dir}(a, P_{0,a})$, and assume that $P_{0,a} \to P_0$ as $a \to \infty$, in the specific sense that $\int g\,\mathrm{d}P_{0,a} \to \int g\,\mathrm{d}P_0$ and $\Sigma(P_{0,a}) \to \Sigma(P_0)$. Then

$$a^{1/2}\left(\theta_a - \int g\,\mathrm{d}P_{0,a}\right) \to_d \mathrm{N}_k(0, \Sigma(P_0)) \quad \text{as } a \to \infty.$$

From (6.3) we also find the following relevant approximation, which is even more in line with the Bernshteĭn–von Mises result (6.1):

$$(n+a)^{1/2}\Sigma(\widehat{P}_n)^{-1/2}\{\theta(P)-\widehat{\theta}_n\}\,|\,\text{data} \to_d \mathrm{N}_k(0, I) \quad \text{a.s.}$$

For a simple illustration, let $k=1$ and $g(x)=x$. Then

$$(n+a)^{1/2}\left(\int x\,\mathrm{d}P - \widehat{\theta}_n\right)\Big/\widehat{\sigma}_n \,\Big|\, \text{data} \to_d \mathrm{N}(0,1) \quad \text{a.s.},$$

where

$$\begin{aligned}\widehat{\sigma}_n^2 = \sigma^2(\widehat{P}_n) &= \int \{x-\widehat{\theta}_n\}^2\,\mathrm{d}\widehat{P}_n \\ &= w_n\{\sigma_0^2 + (\widehat{\theta}_n-\theta_0)^2\} + (1-w_n)\{s_n^2 + (\widehat{\theta}_n - \bar{x}_n)^2\}\end{aligned}$$

in terms of sample mean $\bar{x}_n$ and sample standard deviation s_n. Here $\widehat{\sigma}_n$ is first-order equivalent to s_n for large n, but the above provides a more accurate approximation. Similar results are reached via the delta method for smooth functions of mean parameters. Consider for example the random standard deviation $\sigma(P)$. Then we find

$$\sigma(P)\,|\,\text{data} \approx_d \mathrm{N}(\widehat{\sigma}_n, v_n^2/(n+a)) \quad \text{a.s.}$$

with $v_n^2 = \widehat{\sigma}_n^2(1/2 + 1/4\widehat{\kappa}_n)$, and where $\widehat{\kappa}_n = \int\{(x - \widehat{\theta}_n)/\widehat{\sigma}_n\}^4 \,\mathrm{d}\widehat{P}_n - 3$ is the nonparametric Bayes estimator of the kurtosis. A less refined approximation, also in the Bernshteĭn–von Mises spirit, is that $\sigma(P)\,|\,\text{data}$ is approximately a normal $(s_n, s_n^2(1/2+1/4\kappa_n/n))$, with κ_n being the sample kurtosis.

The above results have been given under the a fixed and n growing scenario. They are also valid when a is allowed to grow with n, slowly enough for $a/\sqrt{n}$ to go to zero. In the more extreme case of $a = cn$, a case reflecting a more persistent belief in the prior, the results above need to be modified with $P_\infty = cP_0 + (1-c)P_{\text{true}}$ replacing P_{true}. The results presented here may also be derived via general empirical process theory, see Hjort (1991), but have been demonstrated here in a simpler fashion, using the theory and infinite divisibility representations of Section 6.1.

6.3. APPROXIMATING THE DENSITY OF A RANDOM DIRICHLET MEAN

To explain the following method, assume for simplicity of presentation that P_0 is confined to the unit interval. Our task is to approximate the density f of $\theta = \int_0^1 x \,\mathrm{d}P$. To this end, consider

$$f_m(t) = f_0(t)c_m(b)^{-1}\exp(b_1 t + b_2 t^2 + \cdots + b_m t^m) \quad \text{on } [0, 1],$$

where f_0 is a suitably chosen start approximation (which could be $f_0(t) = 1$) and $c_m(b) = \int_0^1 f_0(t)\exp(b_1 t + \cdots + b_m t^m)\,\mathrm{d}t$ the necessary integration constant. The idea is to select the coefficients $b_1, \ldots, b_m$ so as to give optimal approximation quality of f_m to the real f. We do this by minimising the Kullback–Leibler distance $\int f \log(f/f_m)\,\mathrm{d}t$, which is seen to be the same as maximising the function $A_m(b) = \sum_{j=1}^m b_j\xi_j - \log c_m(b)$ with respect to $b = (b_1, \ldots, b_m)$, where $\xi_j = \int t^j f(t)\,\mathrm{d}t$ is the real jth moment of θ. These moments can easily be found numerically via (3.6), say for j up to $m = 100$. The maximisation task is also rather easily done using optimisation algorithms available in software packages like `Splus`, helped here by the fact that $A_m(b)$ is concave in b.

There is also a closely related sister method that uses $b_1(t-x) + b_2(t-x)^2 + \cdots$ instead of $b_1 t + b_2 t^2 + \cdots$ above, where x is any fixed value chosen for the convenience of the situation, like the point of symmetry in a case where P_0 is symmetric. The algorithm then needs the $\xi_j(x) = \mathrm{E}(\theta - x)^j$ quantities that are found via Proposition 4.

We have tried out this method in some situations and found it to work very well, even with $f_0 = 1$. The approximation quality improves, for smaller m, when a better start approximation f_0 is used, for example a nonparametric estimator of the real density based on simulations from f. We lack however precise results about the quality of the resulting approximation. Results reached in e.g. Barron and Sheu (1991) have some relevance, but are typically derived under too restrictive conditions.

Acknowledgements

Our article has benefitted from perceptive and constructive comments from Alessandra Guglielmi and two careful reviewers, who also pointed out some literature that we had missed in an earlier version of our manuscript.

References

1. Antoniak, C. E.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *Ann. Stat.* **2** (1974), 1152–1174.
2. Barron, A. R. and Sheu, C.: Approximations of density functions by sequences of exponential families, *Ann. Stat.* **19** (1991), 1347–1369.
3. Billingsley, P.: *Convergence of Probability Measures* 2nd ed. Wiley, New York, 1999.
4. Cifarelli, D. M. and Regazzini, E.: Distribution functions of means of a Dirichlet process, *Ann. Stat.* **18** (1990), 429–442; corrigendum, ibid. **22** (1994), 1633–1634.
5. Cifarelli, D. M. and Regazzini, E.: Some remarks on the distribution functions of means of a Dirichlet process, Quaderno IAMI 93.4, CNR-IAMI, Milano, 1993.
6. Cifarelli, D. M. and Melilli, E.: Some new results for Dirichlet priors, *Ann. Stat.* **28** (2000), 1390–1413.
7. Diaconis, P. and Kemperman, J.: Some new tools for Dirichlet priors, In: J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds) *Bayesian Statistics 5*, Oxford University Press, Oxford, 1996, pp. 97–106.
8. Dufresne, D.: Algebraic properties of Beta and Gamma distributions, and applications, *Adv. Appl. Math.* **20** (1998), 285–299.
9. Epifani, I.: Some results on the moments of vector-valued functionals of the Ferguson-Dirichlet process, Quaderno 380/P, Dipartimento di Matematica, Politecnico di Milano, 1999.
10. Escobar, M. D.: Estimating normal means with a Dirichlet process prior, *J. Am. Stat. Ass.* **89** (1994), 268–277.
11. Escobar, M. D. and West, M.: Bayesian density estimation and inference using mixtures, *J. Am. Stat. Ass.* **90** (1995), 577–588.
12. Escobar, M. D. and West, M.: Computing nonparametric hierarchical models, In: D. Day, P. Müller and D. Sinha (eds) *Practical Nonparametric and Semiparametric Bayesian Statistics*, Lecture Notes in Statistics vol. 133, Springer-Verlag, New York, 1998, pp. 1–22.
13. Feigin, P. B. and Tweedie, R. L.: Linear functionals and Markov chains associated with Dirichlet processes, *Math. Proc. Cambridge Phil. Soc.* **105** (1989), 579–585.
14. Feller, W.: *An Introduction to Probability Theory and Its Applications*, Vol. II, Wiley, New York, 1966.
15. Ferguson, T. S.: A Bayesian analysis of some nonparametric problems, *Ann. Stat.* **1** (1973), 209–230.
16. Ferguson, T. S.: Prior distributions on spaces of probability measures, *Ann. Stat.* **2** (1974), 615–629.

17. Ferguson, T. S.: Bayesian density estimation by mixtures of normal distributions, In: H. Rizvi, J. S. Rustagi and D. O. Siegmund (eds) *Recent Advances in Statistics*, Chernoff Festschrift Volume, Academic Press, New York, 1983, pp. 287–302.
18. Gjessing, H. and Paulsen, J.: Present value distributions with applications to ruin theory and stochastic equations, *Stochastic Processes Appl.* **71** (1997), 123–144.
19. Guglielmi, A.: A simple procedure calculating the generalized Stieltjes transform of the mean of a Dirichlet process, *Stat. Prob. Lett.* **38** (1998), 299–303.
20. Guglielmi, A., Holmes, C. C. and Walker, S. G.: Perfect simulation involving functionals of a Dirichlet process, *J. Comput. Graphical Stat.* **29** (2002), 306–310.
21. Guglielmi, A. and Tweedie, R. L.: Markov chain Monte Carlo estimation of the law of the mean of a Dirichlet process, *Bernoulli* **7** (2001), 593–604.
22. Hannum, R., Hollander, M. and Langberg, N.: Distributional results for random functionals of a Dirichlet process, *Ann. Prob.* **9** (1981), 665–670.
23. Henrici, P.: *Applied and Computational Complex Analysis*, Vol. 3, Wiley, New York, 1993.
24. Hjort, N. L.: Nonparametric Bayes estimators based on Beta processes in models for life history data, *Ann. Stat.* **18** (1990), 1259–1294.
25. Hjort, N. L.: *Bayesian and empirical Bayesian bootstrapping, Statistical Research Report*, Department of Mathematics, University of Oslo, 1991.
26. Hjort, N. L.: Bayesian approaches to non- and semiparametric density estimation [with discussion], In: J. Berger, J. Bernardo, A. P. Dawid, A. F. M. Smith (eds) *Bayesian Statistics 5*, Oxford University Press, Oxford, 1996, pp. 223–253.
27. Hjort, N. L.: *Bayesian analysis for a generalised Dirichlet process prior, Statistical Research Report*, Department of Mathematics, University of Oslo, 2000.
28. Hjort, N. L.: Topics in nonparametric Bayesian statistics [with discussion], In: P. J. Green, N. L. Hjort and S. Richardson (eds) *Highly Structured Stochastic Systems*, Oxford University Press, Oxford, 2003, pp. 455–478.
29. Hjort, N. L. and Ongaro, A.: On the distribution of random Dirichlet jumps, to appear, 2005.
30. Hoffmann-Jorgensen, J.: *Probability With a View Toward Statistics*, Chapman and Hall, New York, 1994.
31. Ishwaran, H. and James, L. F.: Gibbs sampling methods for stick-breaking priors, *J. Am. Stat. Ass.* **96** (2001), 161–173.
32. Ishwaran, H. and Zarepour, M.: Exact and approximate sum-representations for the Dirichlet process, *Can. J. of Stat.* **30** (2002), 1–16.
33. Kallenberg, O.: *Random Measures*, Akademie-Verlag, Berlin and Academic Press, London, 1986.
34. Kerov, S.: Interlacing measures, *Am. Mathe. Soc. Translations* **181** (1998), 35–83.
35. Kerov, S. and Tsilevich, N.: The Markov–Krein correspondence in several dimensions (in Russian), Transactions from the scientific seminars of POMI **283** (2001), 98–122.
36. LeCam, L. and Yang, G. L.: *Asymptotics in Statistics*, Springer-Verlag, New York, 1990.
37. Lijoi, A. and Regazzini, E.: Means of a Dirichlet process and multiple hypergeometric functions, Quaderno IAMI 01.09, CNR-IAMI, Milano, 2001.
38. Lo, A. Y.: On a class of Bayesian nonparametric estimates: I. Density estimates, *Ann. Stat.* **12** (1984), 351–357.
39. Lukacs, E.: *Characteristic Functions* 2nd ed. Griffin, London, 1970.
40. MacEachern, S. N.: Estimating normal means with a conjugate style Dirichlet process prior, *Comm. Stat. – Simulation Comput.* **23** (1994), 727–741.

41. MacEachern, S. N.: Computational methods for mixture of Dirichlet process models, In: D. Day, P. Müller and D. Sinha (eds) *Practical Nonparametric and Semiparametric Bayesian Statistics*, Lecture Notes in Statistics Vol 133, Springer-Verlag, New York, 1998, pp. 23–43.
42. Ongaro, A.: Size-biased sampling and discrete nonparametric Bayesian inference, *J. Stat. Plan. Inference* **128** (2005), 123–148.
43. Regazzini, E.: An example of the interplay between statistics and special functions, in *Tricomi's ideas and contemporary applied mathematics, Atti dei Convegni Lincei* **147**, Accademia Nazionale dei Lincei e Accademia delle Scienze di Torino, 1998, pp. 303–320.
44. Regazzini, E., Guglielmi, A. and Di Nunno, G.: Theory and numerical analysis for exact distributions of functionals of a Dirichlet process, *Ann. Stat.* **30** (2002), 1376–1411.
45. Sethuraman, J.: A constructive definition of Dirichlet priors, *Statistica Sinica* **4** (1994), 639–650.
46. Sethuraman, J. and Tiwari, R.: Convergence of Dirichlet measures and the interpretation of their parameter, In: S. S. Gupta and J. Berger (eds) *Proceedings of the Third Purdue Symposium on Statistical Decision Theory and Related Topics*, Academic Press, New York, 1982, pp. 305–315.
47. Vershik, A. M., Yor, M. and Tsilevich, N. V.: O tozhdestve Markova–Kreina i kvaziinvariantnosti gamma-protsessa [On the Markov–Krein identity and quasi-invariant Gamma processes (in Russian)], Transactions from the scientific seminars of POMI **283** (2001), 21–36.
48. Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M.: Bayesian nonparametric inference for random distributions and related functions [with discussion], *J. R. Stat. Soc.* **B 61** (1999), 485–527.
49. Widder, D. V.: *An Introduction to Transform Theory*, Academic Press, New York, 1971.
50. Yamato, H.: Characteristic functions of means of distributions chosen from a Dirichlet process, *Ann. Prob.* **12** (1984), 262–267.