# Bayesian analysis for a generalised Dirichlet process prior

**Nils Lid Hjort**

**University of Oslo**

*November 2000*

Abstract. A family of random probabilities is defined and studied. This family contains the Dirichlet process as a special case, corresponding to an inner point in the appropriate parameter space. The extension makes it possible to have random means with larger or smaller skewnesses as compared to skewnesses under the Dirichlet prior, and also in other ways amounts to additional modelling flexibility. The usefulness of such random probabilities for use in nonparametric Bayesian statistics is discussed. The posterior distribution is complicated, but inference can nevertheless be carried out via simulation, and some exact formulae are derived for the case of random means. The class of nonparametric priors provides an instructive example where the speed with which the posterior forgets its prior with increasing data sample size depends on special aspects of the prior, which is a different situation from that of parametric inference.

Key words: *consistency, Dirichlet process, jump sizes, nonparametric Bayes, random means, speed of memory loss, stochastic equation*

## 1. Introduction and summary

The Dirichlet process, introduced in Ferguson (1973, 1974), continues to be a cornerstone of nonparametric Bayesian statistics, where it may be used as a prior for an unknown probability distribution for data. Various generalisations have been proposed and investigated in the literature, making the Dirichlet a special favourite case of Pólya trees, of Beta processes, of neutral to the right and of tailfree processes and of Dirichlet mixtures; see Walker, Damien, Laud and Smith (1998) and Hjort (2001) for recent overviews. The purpose of this article is to provide yet another generalisation of the Dirichlet process and to study some of its properties.

Write $P \sim \mathrm{Dir}(b, P_0)$ to signify that $P$ is a Dirichlet process with parameters $(b, P_0)$ on some sample space $\Omega$, where $b$ is positive and $P_0$ a probability distribution; in particular, for each set $A$, the random probability $P(A)$ has a Beta distribution with parameters $(bP_0(A), b\{1-P_0(A)\})$. For a function $g$ of interest, consider the random mean $\theta = \mathrm{E}\{g(X)\,|\,P\} = \int g\,\mathrm{d}P$. Its moments are most conveniently given in terms of a variable $Y = g(\xi)$ with $\xi$ being drawn from $P_0$. Then

$$\mathrm{E}\theta = \theta_0, \quad \mathrm{Var}\,\theta = \frac{1}{1+b}\mathrm{E}_0(Y-\theta_0)^2 \quad \text{and} \quad \mathrm{E}(\theta-\theta_0)^3 = \frac{2}{(1+b)(2+b)}\mathrm{E}_0(Y-\theta_0)^3, \tag{1.1}$$

where $\theta_0 = \mathrm{E}_0 Y$; here '$\mathrm{E}_0$' signifies expectations with respect to the base distribution $Q_0 = P_0 g^{-1}$ for $Y$. The two first results are in Ferguson (1973) while the third may be proved using similar arguments; in Section 3 below we actually give a method for finding the full moment sequence. For $g$ the indicator function of a set $A$, (1.1) specialises to

$$\mathrm{E}P(A) = p_0, \quad \mathrm{Var}\, P(A) = \frac{p_0(1-p_0)}{1+b} \quad \text{and} \quad \mathrm{E}\{P(A)-p_0\}^3 = \frac{2p_0(1-p_0)(1-2p_0)}{(1+b)(2+b)},$$

where $p_0 = P_0(A)$, agreeing by necessity with moment calculations from the Beta distribution $(bp_0, b - bp_0)$. These equations make clear that there is a good amount of modelling flexibility with the Dirichlet prior, as one may centre the random $P$ at any chosen prior mean distribution $P_0$ and attune $b$ to a desired level of variability. One is then stuck with the consequences, however, regarding all further aspects of the prior, such as the implied skewnesses of random probabilities and of random means.

The generalised class of nonparametric priors to be worked with below makes it possible to adjust to further aspects of prior knowledge, for example regarding the skewnesses of random means. Let $B_1, B_2, \ldots$ be independent from a suitable distribution on $(0,1)$, and define random probabilities $\gamma_1 = B_1$, $\gamma_2 = \bar{B}_1 B_2$, $\gamma_3 = \bar{B}_1 \bar{B}_2 B_3$ and so on, where $\bar{B}_i = 1 - B_i$. Here $1 - \sum_{j=1}^n \gamma_i = \bar{B}_1 \cdots \bar{B}_n$, making it easy to show that the $\gamma_j$s sum to 1 with probability 1. This allows us the possibility of defining a random probability

$$P = \sum_{j=1}^{\infty} \gamma_j \delta(\xi_j), \quad \text{with independently drawn } \xi_j\text{s from } P_0. \tag{1.2}$$

Here $\delta(\xi)$ indicates the measure giving full mass 1 to location $\xi$. In other words, $P(A)$ for a set $A$ can be represented as a random sum of random probabilities $\sum_{\xi_j \in A} \gamma_j$. As demonstrated in Sethuraman and Tiwari (1982), the Dirichlet process can also be represented in such a form, corresponding to the particular case where the distribution for $B_j$ is chosen as the Beta$(1, b)$; see also Sethuraman (1994).

The prior process given in (1.2) is described by two distributions, the prior mean $P_0$ and the distribution $H$ on $(0,1)$ governing the $B_j$s. Below some properties of this general prior, say $\mathrm{GD}(H, P_0)$, are investigated. An attractive class of priors emerges by allowing $B_j$ the Beta$(a,b)$ distribution. We write $P \sim \mathrm{GD}(a, b, P_0)$ to indicate this particular extension of the Dirichlet, which corresponds to having $a = 1$. It is important to note that the Dirichlet becomes an 'inner point' in the enlarged class, in contrast to what is the case for some other proposals, where the Dirichlet is at 'a corner' of the underlying parameter space; see Hjort and Ongaro (2001) for some constructions of that type.

In Section 2 we demonstrate that the prior (1.2) has large support in the space of probability measures on the sample space, indicating that these priors are genuinely nonparametric. Section 3 deals with Markov chain Monte Carlo simulation methods for $P$ and for random means thereof, and uses stochastic identities to derive formulae for the full moment sequence of such random means. This is used in Section 4 to show that the added flexibility afforded by the $\mathrm{GD}(a, b, P_0)$ class, and a fortiori the more general $\mathrm{GD}(H, P_0)$

class, indeed allows skewnesses a larger range than that dictated by the Dirichlet. In Section 5 the posterior mean estimator for $P$ is exhibited, leading also to explicit expressions for Bayes estimators of mean parameters. These are convex combinations of prior means and sample averages. This is supplemented in Section 6 with posterior variance formulae, which are used in Section 7 to show that not only the posterior mean estimator but also the posterior distribution as such becomes consistent, in the sense of being able to recapture any true distribution underlying the data, as the sample size increases. Interestingly we find that the speed with which this convergence takes place depends on aspects of the prior process; in particular, this speed is sometimes faster and sometimes slower than the rate $O(n^{-1})$ found for all models with a finite number of parameters. The posterior process itself is somewhat complicated. It is exhibited in Section 8. Then in Section 9 some results for distributions of random means are given before we offer a list of concluding remarks in Section 10.

## 2. Large supports

In parametric Bayesian statistics a prior is in effect placed on some set of densities, say on $\mathcal{M}_0 = \{f_\alpha(\cdot)\colon \alpha \in R\}$, where $R$ is a finite-dimensional parameter set indexing the $f_\alpha$ densities on the sample space $\Omega$. But this set is a thin one in the space $\mathcal{M}$ of all distributions on $\Omega$, topologically speaking; natural neighbourhoods around given distributions are given zero prior probability. The situation is different for genuinely nonparametric priors, as demonstrated for example by Ferguson (1973) for the Dirichlet process prior. We show here that the generalised priors (1.2) continue to have such large supports in the space $\mathcal{M}$. The support is the set of probability distributions $Q$ such that every neighbourhood around it has positive probability.

In this section we take the distribution $H$ of $B_j$s to have full support $[0,1]$. The key observation is that if only $P_0(A) > 0$, then the possible outcomes of $P(A) = \sum_{\xi_j \in A} \gamma_j$ fill out all of $(0,1)$. More generally, if $A_1, \ldots, A_k$ are disjoint sets with positive $P_0$ probability and union probability less than 1, then the distribution of $(P(A_1), \ldots, P(A_k))$ has a density being positive on the $k$-simplex of $(p_1, \ldots, p_k)$ with positive components and sum less than 1.

2.1. Support under strong convergence. Strong or set-wise convergence $Q_n \to Q$, for probability measures on $\Omega$, means that $Q_n(A) \to Q(A)$ for all sets $A$. A basis for neighbourhoods around a given $Q$ is the class of

$$U = U(Q; A_1, \ldots, A_m, \varepsilon_1, \ldots, \varepsilon_m) = \{Q'\colon |Q'(A_j) - Q(A_j)| < \varepsilon_j \text{ for } j = 1, \ldots, m\}, \quad (2.1)$$

where $m$ is any integer, the $A_j$s are measurable subsets, and the $\varepsilon_j$s positive. Here

$$\mathrm{supp}\{\mathrm{GD}(H, P_0)\} = \{Q\colon Q << P_0\}, \quad (2.2)$$

the set of measures absolutely continuous with respect to $P_0$. To see this, let $Q << P_0$. It suffices to show that $U$ has positive probability when the $A_j$s form a measurable partition.

If $P_0(A_j) = 0$ then $Q(A_j) = 0$ and $P(A_j) = 0$ too. Hence $U$ has positive probability if $U'$ has, where $U' = \{Q' \colon |Q'(A_j) - Q(A_j)| < \varepsilon_j \text{ for } j = i_1, \ldots, i_k\}$, where these are the indexes for which $P_0(A_j) > 0$. But it follows from the comment made above that $\mathrm{GD}(H, P_0)$ gives positive probability to this event. Hence $Q$ is in the support. If on the other hand $Q$ is such that $Q(A) > 0$ but $P_0(A) = 0$ for some $A$. Then $P(A) = 0$ a.s., and $\{P \colon |P(A) - Q(A)| < \frac{1}{2}Q(A)\}$ does not have positive probability.

2.2. Support under weak convergence. Assume now that the sample space has a metric and study the topology determined by weak convergence, where $Q_n \to Q$ means convergence in distribution. A basis for neighbourhoods under this topology is the class of (2.1) type sets, but with the restriction that the $A_j$ sets are $Q$-continuous, that is, $Q(\partial A_j) = 0$, where $\partial A_j$ is the boundary set of $A_j$. Here

$$\mathrm{supp}\{\mathrm{GD}(H, P_0)\} = \{Q \colon \mathrm{supp}(Q) \subset \mathrm{supp}(P_0)\}. \tag{2.3}$$

Let $Q$ have a support contained in the support of $P_0$. In general $P_0(A) = 0$ does not imply $Q(A) = 0$, but this is seen to hold when the set $A$ is $Q$-continuous. Hence the arguments used to prove (2.2) can be used with small modifications to prove (2.3).

## 3. Stochastic equations, MCMC and random means

In this section a fruitful stochastic equation is exhibited which characterises the $\mathrm{GD}(H, P_0)$ prior process. This is used to give a Markov chain Monte Carlo method for simulating realisations of the processes and to derive results about random means. If one only needs simulated realisations for one or more random means a simpler Monte Carlo Markov chain suffices.

3.1. Stochastic equations and MCMC simulation. Let $P$ have the prior given in (1.2), with a general distribution $H$ for the $B_j$s. Then

$$\begin{aligned} P &= B_1\delta(\xi_1) + \bar{B}_1B_2\delta(\xi_2) + \bar{B}_1\bar{B}_2B_3\delta(\xi_3) + \cdots \\ &= B_1\delta(\xi_1) + \bar{B}_1\{B_2\delta(\xi_2) + \bar{B}_2B_3\delta(\xi_3) + \bar{B}_2\bar{B}_3B_4\delta(\xi_4) + \cdots\} \\ &= B_1\delta(\xi_1) + \bar{B}_1P', \end{aligned}$$

where $P'$ is constructed in the very same manner $P$. Letting '$=_d$' mean equality in distribution there is accordingly a stochastic equation

$$P =_d B\delta(\xi) + \bar{B}P, \tag{3.1}$$

where on the right hand side $B$, $\xi$, $P$ are independent, with $B \sim H$ and $\xi \sim P_0$. One may show that this identity fully characterises the distribution of $P$. Applying (3.1) to a random mean functional $\theta = \int g \,\mathrm{d}P$ one finds that this variable, which may also be expressed as $\sum_{j=1}^{\infty} \gamma_j g(\xi_j)$, satisfies the stochastic equation

$$\theta =_d BY + \bar{B}\theta, \quad \text{where } Y = g(\xi) \sim P_0 g^{-1} \text{ and } B \sim H. \tag{3.2}$$

A Markov chain $P_1, P_2, \ldots$ may be constructed in the space of probability measures on the sample space via

$$P_n = B_n \delta(\xi_n) + \bar{B}_n P_{n-1},$$

where $(B_n, \xi_n)$ are independent copies of $(B, \xi)$. With arguments parallelling those in Feigin and Tweedie (1989) the equilibrium distribution for the chain may be shown to be exactly that of our prior process (1.2). For a random mean functional, the Markov chain scheme becomes $\theta_n = B_n Y_n + \bar{B}_n \theta_{n-1}$, with the distribution of $\theta$ under (1.2) as its equilibrium. This is akin to similar simulation strategies for means of the Dirichlet process, worked with in Feigin and Tweedie (1989), Guglielmi and Tweedie (2000) and Guglielmi, Holmes and Walker (2001). See also Paulsen and Hove (1999) for precise results about speed of convergence and quality of approximation to the real distribution with the empirical one observed from simulations.

Note that when interest lies in one or more random means the simpler simulation scheme suffices, as there is no need for the full process $P$. We also point out that the moment-correcting methods used in Hjort and Ongaro (2000) apply here too and amount to ways of easily improving the simulation-based approximations of Paulsen and Hove (1999), Guglielmi and Tweedie (2000) and Guglielmi, Holmes and Walker (2001). The key is that the full moment sequence may be uncovered, as we demonstrate next.

3.2. Finding the moments. A recursive method of finding all moments for such a $\theta$, in terms of moments for the null distribution $Q_0$ for $Y = g(\xi)$, emerges by writing

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

which implies

$$\mathrm{E}(\theta - x)^p = \frac{1}{1 - \mathrm{E}\bar{B}^p} \sum_{j=0}^{p-1} \mathrm{E}B^{p-j} \bar{B}^j \, \mathrm{E}_0 (Y - x)^{p-j} \, \mathrm{E}(\theta - x)^j. \tag{3.3}$$

This is valid for all $p \geq 1$ for which $\mathrm{E}|Y|^p$ is finite, and for all $x$. One finds in particular $\mathrm{E}\theta = \theta_0 = \mathrm{E}_0 Y$ and

$$\begin{aligned}
\mathrm{Var}\,\theta &= \frac{\mathrm{E}B^2}{1 - \mathrm{E}\bar{B}^2} \sigma_0^2, \\
\mathrm{E}(\theta - \theta_0)^3 &= \frac{\mathrm{E}B^3}{1 - \mathrm{E}\bar{B}^3} \mathrm{E}_0 (Y - \theta_0)^3, \\
\mathrm{E}(\theta - \theta_0)^4 &= \frac{1}{1 - \mathrm{E}\bar{B}^4} \Big\{ \mathrm{E}B^4 \, \mathrm{E}_0 (Y - \theta_0)^4 + 6 \, \mathrm{E}B^2 \bar{B}^2 \frac{\mathrm{E}B^2}{1 - \mathrm{E}\bar{B}^2} \sigma_0^4 \Big\},
\end{aligned} \tag{3.4}$$

in terms of $\sigma_0^2 = \mathrm{E}_0 (Y - \theta_0)^2$. Further formulae for centralised moments follow from (3.3), expressed in terms of

$$M_{i,j} = \mathrm{E}B^i \bar{B}^j = \int_0^1 s^i (1 - s)^j \, \mathrm{d}H(s). \tag{3.5}$$

With $g$ the indicator of a set $A$, the $\theta$ becomes the random probability $P(A)$, for which we therefore have found $\mathrm{E}P(A) = P_0(A) = p_0$ and

$$\mathrm{Var}\, P(A) = \frac{M_{2,0}}{1-M_{0,2}} p_0(1-p_0) \quad \text{and} \quad \mathrm{E}\{P(A)-p_0\}^3 = \frac{M_{3,0}}{1-M_{0,3}} p_0(1-p_0)(1-2p_0).$$

## 4. Skewness factors and added flexibility

In this section the increased flexibility of the nonparametric prior class is discussed in relation to the skewness of random means.

4.1. Skewnesses under the $\mathrm{GD}(a,b,P_0)$ prior. When $H$ is the $\mathrm{Beta}(a,b)$,

$$M_{i,j} = \mathrm{E}B^i\bar{B}^j = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+i)\Gamma(b+j)}{\Gamma(a+b+i+j)} = \frac{a^{[i]}b^{[j]}}{(a+b)^{[i+j]}}$$

for integers $i,j$, where $x^{[i]} = x(x+1)\cdots(x+i-1) = \Gamma(x+i)/\Gamma(x)$. This may be used to derive moment formulae under the $\mathrm{GD}(a,b,P_0)$ prior. From (3.4) one finds

$$\mathrm{Var}\,\theta = \frac{M_{2,0}}{1-M_{0,2}}\sigma_0^2 = \frac{a+1}{a+2b+1}\sigma_0^2 = \frac{\sigma_0^2}{1+b^*}, \quad \text{with } b^* = \frac{2b}{1+a}, \tag{4.1}$$

and similarly $\mathrm{Var}\, P(A) = P_0(A)\{1-P_0(A)\}/(1+b^*)$. Furthermore,

$$\begin{aligned}\mathrm{E}(\theta-\theta_0)^3 &= \frac{a(a+1)(a+2)/\{(a+b)(a+b+1)(a+b+2)\}}{1-b(b+1)(b+2)/\{(a+b)(a+b+1)(a+b+2)\}}\mathrm{E}_0(Y-\theta_0)^3 \\ &= \frac{(a+1)(a+2)}{a^2+3a(b+1)+3b^2+6b+2}\mathrm{E}_0(Y-\theta_0)^3.\end{aligned}$$

The Dirichlet case is $a=1$ for which the skewness factor is $2/\{(b+1)(b+2)\}$; cf. (1.1).

Assume a $\mathrm{Dir}(b_0,P_0)$ has been chosen, and consider using a more general $\mathrm{GD}(a,b,P_0)$ instead; using the same base measure $P_0$ automatically ensures that the expected values of any random mean are being matched for the two priors. We may also precisely match all pairs of variances of random means through putting $2b/(1+a) = b_0$, compare (4.1) and (1.1). This amounts to $a = 2b/b_0 - 1 = 2x-1$ as a function of $x = b/b_0$; notice that $x > \frac{1}{2}$, or $b > \frac{1}{2}b_0$, is required. We may then study the skewness of $\theta$ under the Dirichlet versus its value under the $\mathrm{GD}(a,b,P_0)$. The ratio of skewnesses becomes

$$\begin{aligned}\rho(x) &= \frac{(a+1)(a+2)}{a^2-1+3a(b+1)+3(b+1)^2}\Big/\frac{2}{(b_0+1)(b_0+2)} \\ &= \frac{2x(2x+1)}{(2x-1)^2-1+3(2x-1)(b_0x+1)+3(b_0x+1)^2}\frac{(b_0+1)(b_0+2)}{2}.\end{aligned}$$

This is a decreasing function, starting for $b = \frac{1}{2}b_0$ with ratio value $\rho_{\max}$ and ending for large $b$ with ratio value $\rho_{\min}$, say, where

$$\rho_{\max} = \frac{(b_0+1)(b_0+2)}{2+3b_0+(3/4)b_0^2} \quad \text{and} \quad \rho_{\min} = \frac{2(b_0+1)(b_0+2)}{4+6b_0+3b_0^2}.$$

This interval contains the value 1 as an inner point, corresponding to $b = b_0$ and $a = 1$, the Dirichlet case, and otherwise portrays the added flexibility through the additional $a$ parameter. With $b < b_0$ and accompanying $a = 2b/b_0 - 1$, the $\mathrm{GD}(a, b, P_0)$ prior leads to skewnesses bigger in absolute size for all random means than with the correspondingly matched Dirichlet prior; similarly, with $b > b_0$ the skewnesses are reduced in absolute size. The interval of skewness ratios stretches from $2/3$ to $4/3$ when $b_0$ becomes large.

4.2. More flexible kurtosis. A similar exercise is to compute $\mathrm{E}(\theta - \theta_0)^4$, first under the Dirichlet prior $(b_0, P_0)$, and compare it with the value obtained under the $\mathrm{GD}(a, b, P_0)$ process, having fixed $2b/(1 + a) = b_0$ to have the same mean and same variance. This gives a suitable kurtosis ratio curve $\kappa(x)/\kappa(1)$ to study, where $\kappa(x)$ is $\mathrm{E}(\theta-\theta_0)^4$ computed with $a = 2x - 1$ and $b = b_0 x$, for $x > \frac{1}{2}$. This ratio curve depends on $b_0$ and the underlying population kurtosis $\mathrm{E}_0(Y - \theta_0)/\sigma_0^4 - 3$. Carrying out this exercise one finds that the kurtosis is larger than under the Dirichlet for $a < 1$ and smaller than under the Dirichlet for $a > 1$. The ratio interval spans for each $b_0$ a reasonable interval containing 1.

## 5. Marginal distributions and posterior means

Conditionally on the random $P$, let $X_1, \ldots, X_n$ be independently sampled from $P$ in the sample space $\Omega$. In this section we consider the marginal distribution of data and go on to a formula for the predictive distribution, that is, the posterior mean of $P$.

5.1. Marginal distributions. The simultaneous distribution of $P$ and the random sample is given by

$$\Pr\{P \in C, X_1 \in A_1, \ldots, X_n \in A_n\} = \mathrm{E}\, I\{P \in C\} P(A_1) \cdots P(A_n), \tag{5.1}$$

required to hold for measurable subsets $C$ of the space of probability measures on the space and for all measurable subsets $A_i$; see e.g. Ferguson (1973). Here $I$ denotes an indicator function. In particular,

$$\Pr\{X_1 \in A_1, \ldots, X_n \in A_n\} = \mathrm{E}\, P(A_1) \cdots P(A_n).$$

For $n = 1$ one finds

$$\Pr\{X_i \in A\} = \mathrm{E}P(A) = P_0(A),$$

adding significance to the interpretation of $P_0$ as the marginal distribution of a single observation. For $n = 2$,

$$\begin{aligned}
\mathrm{E}P(A)P(B) &= \mathrm{E} \sum_{j,k} \gamma_j \gamma_k I\{\xi_j \in A, \xi_k \in B\} \\
&= \sum_j \mathrm{E}\gamma_j^2\, P_0(A \cap B) + 2 \sum_{j<k} \mathrm{E}\gamma_j \gamma_k\, P_0(A) P_0(B) \\
&= (1 - a_2) P_0(A \cap B) + a_2 P_0(A) P_0(B),
\end{aligned}$$

where $a_2 = \Pr\{X_2 \neq X_1\} = 1 - M_{2,0}/(1 - M_{0,2}) = 2M_{1,1}/(1 - M_{0,2})$.

The following identity will be useful.

LEMMA. *Let $P$ come from the* $\mathrm{GD}(H,P_0)$ *prior (1.2) and let $A_1,\ldots,A_n$ be disjoint sets. Then*

$$\mathrm{E}P(A_1)\cdots P(A_n)=n!\prod_{j=1}^{n-1}\frac{M_{1,j}}{1-M_{0,j+1}}P_0(A_1)\cdots P_0(A_n) \tag{5.2}$$

*for all $n\ge 2$, in terms of the product moments $M_{i,j}$ of (3.5).*

PROOF. The (1.2) definition leads to the formula

$$\mathrm{E}P(A_1)\cdots P(A_n)=P_0(A_1)\cdots P_0(A_n)\,n!\sum_{i_1<\cdots<i_n}\mathrm{E}\gamma_{i_1}\cdots\gamma_{i_n},$$

which indeed may be worked further by careful cataloguing of $B_k$ and $\bar B_k$ factors entering the product of $\gamma_j$s; one finds for example that $\mathrm{E}P(A_1)P(A_2)=a_2\,P_0(A_1)P_0(A_2)$ with the $a_2$ exhibited above. It is easier, however, to use the simultaneous stochastic equations

$$P(A_i)=_d BY_i+\bar BP(A_i)\quad\text{for } i=1,\ldots,n,$$

where $Y_i=I\{\xi\in A_i\}$ for a $\xi$ drawn from $P_0$, independently of $B\sim H$. That these equations hold simultaneously follows from (3.1). All products of two or more $Y_j$s vanish since the $A_j$s are disjoint. This simplifies the structure of

$$\prod_{i=1}^n P(A_i)=_d\prod_{i=1}^n\{BY_i+\bar BP(A_i)\}=\sum_{i=1}^n BY_i\bar B^{n-1}\theta_{(i)}+\bar B^n\prod_{i=1}^n P(A_i),$$

writing $\theta_{(i)}$ for the product of those $n-1$ terms $P(A_j)$ for which $j\ne i$. Hence

$$\mathrm{E}\prod_{i=1}^n P(A_i)=\frac{1}{1-M_{0,n}}\sum_{i=1}^n M_{1,n-1}P_0(A_i)\,\mathrm{E}\theta_{(i)}.$$

This may now be used to demonstrate (5.2) by induction on $n$, noting that the formula was seen to hold for $n=2$ above. ■

5.2. THE POSTERIOR MEAN. For the next development, define

$$w_n=(n+1)\frac{M_{1,n}}{1-M_{0,n+1}}=(n+1)\frac{\mathrm{E}B\bar B^n}{1-\mathrm{E}\bar B^{n+1}}=(n+1)\delta_n. \tag{5.3}$$

We take the sample space to be a metric space, for example a subset of any Euclidean space, where we condition on the information in a data point $X_i=x_i$ via conditioning on the information $X_i\in S(x_i,\varepsilon)$, say, an $\varepsilon$-neighbourhood around $x_i$, and then letting $\varepsilon\to 0$. For an observed sample, consider the predictive distribution $\widehat P$ given by $\widehat P(A)=\mathrm{E}\{P(A)\,|\,\mathrm{data}\}$; this is also the Bayes estimator of $P$ under squared error loss.

PROPOSITION. *Let $P$ follow the prior process $\mathrm{GD}(H, P_0)$ with an atom-free prior mean measure $P_0$, and assume data points $X_1 = x_1, \ldots, X_n = x_n$ have been observed, and that these are distinct. Then the Bayes estimator of $P$ can be represented as*

$$\widehat{P} = w_n P_0 + (1 - w_n)\frac{1}{n}\sum_{i=1}^{n} \delta(x_i) = w_n P_0 + (1 - w_n) P_n, \tag{5.4}$$

*a convex combination of prior mean and the empirical distribution.*

PROOF. From (5.1) one may show that

$$\mathrm{E}\{\psi(P) \,|\, X_1 \in A_1, \ldots, X_n \in A_n\} = \frac{\mathrm{E}\psi(P)P(A_1)\cdots P(A_n)}{\mathrm{E}P(A_1)\cdots P(A_n)}$$

for all bounded measurable functions $\psi$, provided the $A_i$s have positive $P_0$ measure. In particular, therefore,

$$\mathrm{E}\{P(A) \,|\, X_1 \in A_1, \ldots, X_n \in A_n\} = \frac{\mathrm{E}P(A)P(A_1)\cdots P(A_n)}{\mathrm{E}P(A_1)\cdots P(A_n)}.$$

Take first a set $A$ not meeting the data, which means that it is outside the union of data windows $A_i = S(X_i, \varepsilon)$ for small enough $\varepsilon$. Then the above quotient, by the use of the lemma, reduces to

$$\mathrm{E}\{P(A) \,|\, \mathrm{info}_\varepsilon\} = (n+1)\{M_{1,n}/(1 - M_{0,n+1})\}\, P_0(A),$$

where $\mathrm{info}_\varepsilon$ signifies the information $X_i \in S(x_i, \varepsilon)$ for $i = 1, \ldots, n$. Since the answer is independent of $\varepsilon$, the probability measure $\widehat{P}$ must be equal to $w_n P_0$ on $\Omega - \{x_1, \ldots, x_n\}$, that is, outside the data values.

Being a probability measure it must distribute its remaining mass $1 - w_n$ on the $n$ data values $x_1, \ldots, x_n$. With these being distinct there must be full symmetry, and $\widehat{P}$ must assign value $(1 - w_n)/n$ to each of these. This proves assertion (5.4). ∎

5.3. NONPARAMETRIC BAYES ESTIMATION OF MEANS. Consider Bayesian estimation of a random mean $\theta = \int g \,\mathrm{d}P$. Under squared error loss and with the $\mathrm{GD}(H, P_0)$ prior, for which the prior guess is $\theta_0 = \int g \,\mathrm{d}P_0$, the estimator is

$$\widehat{\theta} = \mathrm{E}(\theta \,|\, \mathrm{data}) = \int g \,\mathrm{d}\widehat{P} = w_n\theta_0 + (1 - w_n)\bar{g}_n \quad \text{with } \bar{g}_n = n^{-1}\sum_{i=1}^{n} g(x_i). \tag{5.5}$$

This follows from (5.4), again under the assumption on there being no ties in data. With a little more formality, this concerns

$$\mathrm{E}(\theta \,|\, \mathrm{data}) = \int_{\mathcal{M}} \theta(P)\, \mathcal{P}(\mathrm{d}P \,|\, \mathrm{data}),$$

where $\mathcal{P}(\cdot\,|\,\text{data})$ is the posterior distribution on the space $\mathcal{M}$ of probability measures on the sample space, and an ingredient is existence and measurability of $\theta = \theta(P)$. A more careful argument, therefore, starts with $g$ equal to a simple function, a linear combination of indicator functions. For such a $g$ the result follows directly from (5.4). Then pass to the limit via monotone convergence to make formula (5.5) valid for all $g$ for which $\int |g|\,\mathrm{d}P_0$ is finite. Measurability comes from it being a limit of linear combinations of $P(A)$ variables, and existence is guaranteed under the minimal condition $\int \log(1+|g|)\,\mathrm{d}P_0$, see Hjort and Ongaro (2000).

As a special case, when an unknown distribution function $F$ of one-dimensional data is to be estimated, the Bayes estimator takes the form $\widehat{F}(t) = w_n F_0(t) + (1-w_n)F_n(t)$, where $F_0$ is the distribution function of $P_0$ and $F_n$ is the empirical distribution function.

Remark. The Dirichlet case corresponds to a $\mathrm{Beta}(1,b)$ distribution for the $B_j$s, and a little algebra on (5.3) shows that for this case $w_n = b/(b+n)$. This is a well-known formula for the weight a posterior Dirichlet distribution still attaches to its prior, also lending strength to the 'prior sample size' interpretation of the $b$ parameter. More nuances are at play for the general $\mathrm{GD}(H, P_0)$ case, however, as shown in section 7. ■

## 6. Posterior variance

The aim of the following efforts is to supplement the posterior mean result above with an explicit formula for the posterior variance of $P$, and more generally for the posterior variance of a $\int g\,\mathrm{d}P$ parameter. This makes construction of credibility intervals possible, and is used to assess full posterior consistency in the next section.

6.1. Posterior variance of $P(A)$. To do the posterior mean calculation, formula (5.2) sufficed. To calculate posterior variances requires a little list of further formulae. Let $A_1,\ldots,A_n$ be disjoint sets, and let $\theta_i = P(A_i)$ with prior mean $P_0(A_i) = p_i$. We show later that the various means-of-products take the following form:

$$\begin{aligned}
\mathrm{E}\theta_1\theta_2\cdots\theta_n &= a_n p_1\cdots p_n,\\
\mathrm{E}\theta_1^2\theta_2\cdots\theta_n &= b_n p_1\cdots p_n + c_n p_1^2 p_2\cdots p_n,\\
\mathrm{E}\theta_1^3\theta_2\theta_3\cdots\theta_n &= d_n p_1p_2p_3\cdots p_n + e_n p_1^2p_2p_3\cdots p_n + f_n p_1^3p_2p_3\cdots p_n,\\
\mathrm{E}\theta_1^2\theta_2^2\cdots\theta_n &= g_n p_1p_2\cdots p_n + h_n p_1p_2(p_1+p_2)\cdots p_n + i_n p_1^2p_2^2\cdots p_n.
\end{aligned} \tag{6.1}$$

Here $a_n,\ldots,i_n$ are sequences of constants, to be returned to below.

Take a prior mean distribution $P_0$ free of atoms, and consider a set $A$ not meeting the data, which we again take to be $n$ distinct values $x_1,\ldots,x_n$. Then, with notation as in Section 5.2 and with $p_0 = P_0(A)$,

$$\mathrm{E}\{P(A)\,|\,\mathrm{info}_\varepsilon\} = \frac{\mathrm{E}P(A)P(A_1)\cdots P(A_n)}{\mathrm{E}P(A_1)\cdots P(A_n)} = \frac{a_{n+1}p_0p_1\cdots p_n}{a_np_1\cdots p_n} = \frac{a_{n+1}}{a_n}p_0,$$

while

$$\mathrm{E}\{P(A_k)\,|\,\mathrm{info}_\varepsilon\}=\frac{\mathrm{E}P(A_k)P(A_1)\cdots P(A_n)}{\mathrm{E}P(A_1)\cdots P(A_n)}=\frac{b_n+c_np_k}{a_n}=\frac{b_n}{a_n}+O(\varepsilon),$$

showing that $\mathrm{E}[P\{x_k\}\,|\,\mathrm{data}]=b_n/a_n$. Ingredients required for second moment calculations include

$$\mathrm{E}\{P(A)^2\,|\,\mathrm{data}\}=\frac{b_{n+1}p_0p_1\cdots p_n+c_{n+1}p_0^2p_1\cdots p_n}{a_np_1\cdots p_n}=\frac{b_{n+1}}{a_n}p_0+\frac{c_{n+1}}{a_n}p_0^2,$$

$$\mathrm{E}\{P(A_k)^2\,|\,\mathrm{info}_\varepsilon\}=\frac{\mathrm{E}P(A_k)^3\prod_{i\neq k}P(A_i)}{\mathrm{E}\prod_{i=1}^nP(A_i)}=\frac{d_n+e_np_k+f_np_k^2}{a_n}=\frac{d_n}{a_n}+O(\varepsilon),$$

$$\mathrm{E}\{P(A)P(A_k)\,|\,\mathrm{info}_\varepsilon\}=\frac{b_{n+1}p_0+c_{n+1}p_0p_k}{a_n}=\frac{b_{n+1}}{a_n}p_0+O(\varepsilon).$$

These and similar efforts entail

$$\mathrm{E}[P\{x_k\}^2\,|\,\mathrm{data}]=d_n/a_n\quad\text{and}\quad\mathrm{E}[P\{x_k\}P\{x_l\}\,|\,\mathrm{data}]=g_n/a_n\text{ for }k\neq l,$$

while $\mathrm{E}[P\{x_k\}P(A)\,|\,\mathrm{data}]=(b_{n+1}/a_n)p_0$.

Let now $A$ be any set, containing say $j$ of the data values, and split it into $A\cap\mathrm{data}=\{x_{i_1},\ldots,x_{i_j}\}$ and $A_0=A-\mathrm{data}$. Then $P(A)=P\{x_{i_1},\ldots,x_{i_j}\}+P(A_0)$ and, with $p_0=P_0(A)$,

$$\mathrm{E}\{P(A)\,|\,\mathrm{data}\}=(a_{n+1}/a_n)p_0+jb_n/a_n=w_np_0+(1-w_n)(j/n),\tag{6.2}$$

agreeing of course with (5.4). Next, collecting together the various contributions to $P(A)^2$,

$$\mathrm{E}\{P(A)^2\,|\,\mathrm{data}\}=j\frac{d_n}{a_n}+j(j-1)\frac{g_n}{a_n}+\frac{b_{n+1}}{a_n}p_0+\frac{c_{n+1}}{a_n}p_0^2+2j\frac{b_{n+1}}{a_n}p_0.\tag{6.3}$$

We also record a formula for the cross-moment for two disjoint sets $A$ and $B$, catching respectively $j$ and $k$ data points:

$$\mathrm{E}\{P(A)P(B)\,|\,\mathrm{data}\}=jk\frac{g_n}{a_n}+j\frac{b_{n+1}}{a_n}P_0(B)+k\frac{b_{n+1}}{a_n}P_0(A)+\frac{a_{n+2}}{a_n}P_0(A)P_0(B).$$

6.2. Posterior variance of a random mean. We have found formulae for posterior variance of a $P(A)=\int I_A\,\mathrm{d}P$. More generally we need the posterior variance of a random mean $\theta=\int g\,\mathrm{d}P$, for which the posterior mean is given in (5.5). Start with a simple $g=\sum_{j=1}^m y_jI_{A_j}$ with disjoint sets $A_j$, so that $\theta=\sum_{j=1}^m y_jP(A_j)$. With a little work,

$$\begin{aligned}\mathrm{E}(\theta^2\,|\,\mathrm{data})=\sum_{j=1}^m y_j^2\Big\{&\frac{nd_n}{a_n}P_n(A_j)+\frac{n^2g_n}{a_n}P_n(A_j)^2-\frac{ng_n}{a_n}P_n(A_j)+\frac{c_{n+1}}{a_n}P_0(A_j)^2\\&+\frac{b_{n+1}}{a_n}P_0(A_j)+\frac{2nb_{n+1}}{a_n}P_0(A_j)P_n(A_j)\Big\}\\+\sum_{j\neq k}y_jy_k\Big\{&\frac{n^2g_n}{a_n}P_n(A_j)P_n(A_k)+\frac{nb_{n+1}}{a_n}P_n(A_j)P_0(A_k)\\&+\frac{nb_{n+1}}{a_n}P_0(A_j)P_n(A_k)+\frac{a_{n+2}}{a_n}P_0(A_j)P_0(A_k)\Big\},\end{aligned}$$

which in terms of $\theta_n = \int g\,\mathrm{d}P_n$ and $\theta_0 = \int g\,\mathrm{d}P_0$ simplifies to

$$\frac{n^2 g_n}{a_n}\theta_n^2 + 2\frac{n b_{n+1}}{a_n}\theta_n\theta_0 + \frac{a_{n+2}}{a_n}\theta_0^2 + \frac{b_{n+1}}{a_n}\int g^2\,\mathrm{d}P_0 + \frac{n d_n - n g_n}{a_n}\int g^2\,\mathrm{d}P_n. \tag{6.4}$$

Used here is the fact that $c_n = a_{n+1}$, proved below.

That this gives a formula $\mathrm{E}(\theta^2\,|\,\mathrm{data}) - \{\mathrm{E}(\theta\,|\,\mathrm{data})\}^2$ for the posterior variance also for the case of any random $\int g\,\mathrm{d}P$, provided only that $\int g^2\,\mathrm{d}P_0$ is finite, follows by passing to the limit via simple functions and multiple uses of the monotone convergence theorem.

6.3. Formulae for the constants. It remains to give formulae for the $a_n, \ldots, i_n$ sequences of (6.1). We have already found that $a_n = n!\,\delta_{n-1}\cdots\delta_1$, in terms of $\delta_j = M_{1,j}/(1-M_{0,j+1})$. For $(b_n, c_n)$, write $\theta_1^2 =_d B^2 Y_1 + 2B\bar{B}Y_1 + \bar{B}^2\theta_1^2$ and $\theta_j =_d BY_j + \bar{B}\theta_j$ for $j = 2,\ldots,n$, as in the arguments used to prove (5.2) above. Writing out the product $\theta_1^2\theta_2\cdots\theta_n$ in a distributional identity and discarding all terms involving two or more $Y_j$s gives an expression for its mean, which after simplification delivers

$$\begin{aligned} b_n &= \frac{1}{1-M_{0,n+1}}\{M_{2,n-1}a_{n-1} + (n-1)M_{1,n}b_{n-1}\},\\ c_n &= \frac{1}{1-M_{0,n+1}}\{2M_{1,n}a_n + (n-1)M_{1,n}c_{n-1}\}. \end{aligned}$$

Finding $\mathrm{E}\theta_1^2$ explicitly gives start values $b_1 = M_{2,0}/(1-M_{0,2})$ and $c_1 = 2M_{1,1}/(1-M_{0,2})$ for these recursive relations. Some investigations lead to $b_n = (n-1)!\,\delta_{n-1}\cdots\delta_1(1-w_n)$ and to $c_n = (n+1)!\,\delta_n\cdots\delta_1 = a_{n+1}$. Working similarly with $\theta_1^3\theta_2\cdots\theta_n$ gives

$$\begin{aligned} d_n &= \frac{1}{1-M_{0,n+2}}\{M_{3,n-1}a_{n-1} + (n-1)M_{1,n+1}d_{n-1}\},\\ e_n &= \frac{1}{1-M_{0,n+2}}\{3M_{2,n}a_n + 3M_{1,n+1}b_n + (n-1)M_{1,n+1}e_{n-1}\},\\ f_n &= \frac{1}{1-M_{0,n+2}}\{3M_{1,n+1}c_n + (n-1)M_{1,n+1}f_{n-1}\}, \end{aligned}$$

with start values

$$d_1 = M_{3,0}/(1-M_{0,3}), \quad e_1 = 3(M_{2,1} + M_{1,2}b_1)/(1-M_{0,3}), \quad f_1 = 3M_{1,2}c_1/(1-M_{0,3})$$

determined from $\mathrm{E}\theta_1^3$. Finally, studying $\theta_1^2\theta_2^2\theta_3\cdots\theta_n$ leads to

$$\begin{aligned} g_n &= \frac{1}{1-M_{0,n+2}}\{2M_{2,n}b_{n-1} + (n-2)M_{1,n+1}g_{n-1}\},\\ h_n &= \frac{1}{1-M_{0,n+2}}\{M_{2,n}c_{n-1} + 2M_{1,n+1}b_n + (n-2)M_{1,n+1}h_{n-1}\},\\ i_n &= \frac{1}{1-M_{0,n+2}}\{4M_{1,n+1}c_n + (n-2)M_{1,n+1}i_{n-1}\} \end{aligned}$$

for $n \geq 2$, where

$$g_2 = 2M_{2,2}b_1/(1-M_{0,4}), \quad h_2 = (M_{2,2}c_1 + 2M_{1,3}b_2)/(1-M_{0,4}), \quad i_2 = 4M_{1,3}c_2/(1-M_{0,4}).$$

One may easily compute the $d_n, \ldots, i_n$ constants via these recursive schemes.

To learn more about these sequences, observe that formula (6.3) implies

$$1 = nd_n/a_n + n(n-1)g_n/a_n + (2n+1)b_{n+1}/a_n + c_{n+1}/a_n, \tag{6.5}$$

simply by letting $A$ be the full sample space. Other helpful formulae for the constants involved in (6.1) emerge as follows. Let $A_1, \ldots, A_n$ form a measurable partition and write $\theta_i = P(A_i)$. Then equating $\mathrm{E}(\sum_{i=1}^n \theta_i)\theta_1 \cdots \theta_n$ with $\mathrm{E}\theta_1 \cdots \theta_n$ leads to $nb_n + c_n = a_n$, which with $nb_n = a_n(1-w_n)$ gives $c_n = a_{n+1}$ (again). Similarly, equating $\mathrm{E}(\sum_{i=1}^n \theta_i)^2\theta_1 \cdots \theta_n$ with $\mathrm{E}\theta_1 \cdots \theta_n$ gives

$$1 = \frac{nd_n}{a_n} + \frac{e_n}{a_n} + \frac{f_n}{a_n}\sum_{i=1}^n p_i^2 + \frac{n(n-1)g_n}{a_n} + \frac{(2n-2)h_n}{a_n} + \frac{i_n}{a_n}\sum_{i \neq j} p_i p_j. \tag{6.6}$$

Since this is an identity valid for all $p_i$s summing to 1, and since $\sum_{i\neq j} p_i p_j = 1 - \sum_{i=1}^n p_i^2$, one must have $f_n = i_n$ for all $n \geq 2$. Helped by this, one may show by induction, using the recursive relations, that

$$f_n = i_n = c_{n+1} = a_{n+2} = (n+2)!\,\delta_{n+1} \cdots \delta_1 \quad \text{for } n \geq 2.$$

Combining (6.5) with (6.6) it is also clear that $(2n+1)b_{n+1}/a_n + c_{n+1}/a_n = e_n/a_n + f_n/a_n + 2(n-1)h_n/a_n$.

Let us work out what happens to the iteratively defined sequence $d_n$. It is helpful to write

$$d_n = (n-1)!\,y_{n-1} + (n-1)\delta_{n+1}d_{n-1} \quad \text{for } n \geq 1,$$

with $(n-1)!\,y_{n-1} = M_{3,n-1}(1-M_{0,n+2})^{-1}a_{n-1}$ for $n \geq 1$. Some minutes of investigation yield $d_n = (n-1)!\sum_{j=0}^{n-1} y_j\delta_{j+3}\cdots\delta_{n+1}$. Going back to $y_n$, one sees that $y_j = \varepsilon_j\delta_j\cdots\delta_1$, where $\varepsilon_j = M_{3,j}/(1-M_{0,j+2})$. Hence $d_n = (n-1)!\,\delta_{n+1}\cdots\delta_1\sum_{j=0}^{n-1}\varepsilon_j/(\delta_{j+1}\delta_{j+2})$, which with $a_n = n!\,\delta_{n-1}\cdots\delta_1$ leads to

$$\frac{nd_n}{a_n} = \delta_{n+1}\delta_n\sum_{j=1}^n \frac{\varepsilon_{j-1}}{\delta_j\delta_{j+1}}. \tag{6.7}$$

We may similarly work out an expression for the $g_n$ sequence. Write

$$g_n = (n-2)!\,z_{n-1} + (n-2)\delta_{n+2}g_{n-1} \quad \text{for } n \geq 2,$$

where $(n-2)!\,z_{n-1} = 2M_{2,n}(1-M_{0,2+n})^{-1}b_{n-1}$; in particular, $g_2 = z_1$. This gives expressions for $g_3, g_4, \ldots$, and the general pattern is discovered to be

$$g_n = (n-2)!\,(z_{n-1} + \delta_{n+1}z_{n-2} + \cdots + \delta_{n+1}\cdots\delta_4 z_1) = (n-2)!\sum_{j=1}^{n-1}\delta_{n+1}\cdots\delta_{j+3}z_j.$$

Going back to $z_n$, which may be expressed as $\eta_j\delta_{j-1}\cdots\delta_1$ for $\eta_j = 2M_{2,j+1}(1-w_j)/(1-M_{0,3+j})$, one finds $g_n = (n-2)!\,\delta_{n+1}\cdots\delta_1\sum_{j=1}^{n-1}\eta_j/(\delta_j\delta_{j+1}\delta_{j+2})$. In conjunction with $a_n = n!\,\delta_{n-1}\cdots\delta_1$ this implies

$$\frac{n(n-1)g_n}{a_n} = \delta_{n+1}\delta_n\sum_{j=1}^{n-1}\frac{\eta_j}{\delta_j\delta_{j+1}\delta_{j+2}}.$$

It will be seen in the next section that of the parts summing to 1 in (6.5) and (6.6), the $n(n-1)g_n/a_n$ is the dominant one.

## 7. Consistency, and how quickly do we forget?

Assume data $X_1, \ldots, X_n$ in reality follow some underlying distribution $P_{\rm true}$. It is well known that the empirical distribution $P_n$ converges to $P_{\rm true}$ with probability 1, even uniformly over all subsets, as the data volume increases. A question of importance is whether the Bayes estimator $\widehat{P}$ matches this feat, and, more generally, whether the posterior distribution converges to the measure concentrated in $P_{\rm true}$.

For parametric models it is known that Bayes inference agrees for large samples with that based on maximum likelihood. A more informative statement is that for Bayes and likelihood estimators $\widehat\theta_{B,n}$ and $\widehat\theta_{L,n}$ based on the $n$ first data points, it holds that $n^{1/2}(\widehat\theta_{B,n} - \widehat\theta_{L,n}) \rightarrow_p 0$, even when the parametric model used to generate these likelihoods and posteriors is incorrect, under very mild regularity assumptions; see Hjort and Pollard (1993). It is furthermore the case that the posterior 'forgets its prior' at a speed linear with $n$, in the sense that aspects of the posterior traceable to the prior has weight exactly or approximately equal to $b/(b+n)$ for a suitable $b$, which then can be interpreted as 'prior sample size'. The very same behaviour is observed for the Dirichlet process prior, as shown in Ferguson (1973, 1974). We shall see that the situation can be quite different for other members of the $\mathrm{GD}(H, P_0)$ class.

7.1. Consistency of the posterior mean. In what follows take $P_{\rm true}$ to be free of atoms on its sample space, making all realisations $X_1, X_2, \ldots$ a.s. distinct. From (5.4) it is clear that $\widehat{P}$ also goes to $P_{\rm true}$ almost surely provided only that $w_n \rightarrow 0$. Under this key condition $\widehat{P}$ and the nonparametric frequentist estimator $P_n$ agree asymptotically. It turns out that indeed $w_n \rightarrow 0$, but with a speed depending upon aspects of the distribution $H$ of the $B_j$s.

Lemma. *For any distribution $H$ for $B$, $w_n$ of (5.3) goes to zero with growing $n$.*

PROOF. It suffices to show

$$
\begin{aligned}
\mathrm{E}\bar{B}^{n+1} &= \int_0^1 (1-s)^{n+1}\,\mathrm{d}H(s) \to 0, \\
(n+1)\,\mathrm{E}B\bar{B}^n &= \int_0^1 (n+1)s(1-s)^n\,\mathrm{d}H(s) \to 0.
\end{aligned}
$$

The first follows quickly by dominated convergence, as does actually also the second. The point is that the integrand $(n+1)s(1-s)^n$ goes pointwise to zero, and has a maximum value bounded in $n$. Inspection shows that the maximum occurs for $s_0 = 1/(n+1)$ and that the resulting maximum value converges to $e^{-1}$. Hence there is uniform integrability and the claim follows. ■

Consider next the $\mathrm{GD}(a,b,P_0)$ case, for which (3.5) and (5.3) yield

$$
\begin{aligned}
w_n &= (n+1)\frac{ab^{[n]}}{(a+b)^{[n+1]}}\Big/\Big\{1-\frac{b^{[n+1]}}{(a+b)^{[n+1]}}\Big\} \\
&= (n+1)\frac{a\Gamma(b+n)}{\Gamma(b)}\Big/\Big\{\frac{\Gamma(a+b+n+1)}{\Gamma(a+b)}-\frac{\Gamma(b+n+1)}{\Gamma(b)}\Big\} \\
&= \frac{n+1}{n+b}\frac{a}{\Gamma(b)}\Big/\Big\{\frac{\Gamma(a+b+n+1)}{\Gamma(b+n+1)\Gamma(a+b)}-\frac{1}{\Gamma(b)}\Big\}.
\end{aligned}
\tag{7.1}
$$

This answer generalises the well-known formula $w_n = b/(n+b)$ valid for the posterior mass outside data points for the Dirichlet process. Formula (7.1) gives the precise weight the Bayes estimator attaches to outside-of-data information, that is, as caused by the prior. The speed with which $w_n \to 0$ is different from the traditional $O(n^{-1})$, when $a \neq 1$, as we shall see.

Since the denominator of (5.3) goes to 1 it suffices for large $n$ to study $u_n = (n+1)\,\mathrm{E}B\bar{B}^n$ and the speed with which this sequence tends to zero. For the $\mathrm{GD}(a,b,P_0)$ case,

$$
u_n = (n+1)a\frac{\Gamma(a+b)}{\Gamma(b)}\frac{\Gamma(b+n)}{\Gamma(a+b+n+1)},
$$

and we may use the Stirling approximation, for example in the form of

$$
\log\Gamma(x) = (x-\tfrac{1}{2})\log x - x + \tfrac{1}{2}\log(2\pi) + 1/(12x) + O(1/x^2) \quad \text{for large } x,
$$

to assess its size. Some algebra efforts reveal $\log u_n = -a\log n + \log\{a\Gamma(a+b)/\Gamma(b)\} - 2(a+1) + O(n^{-1})$, which means

$$
u_n = n^{-a}\{a\Gamma(a+b)/\Gamma(b)\}\,\exp\{-2(a+1)\}\{1+O(n^{-1})\} \quad \text{when } n \text{ grows.}
$$

Hence, only for the Dirichlet case $a=1$ does the posterior process forget its origin with speed $O(n^{-1})$, which is the traditional speed with which memory loss sets in for Bayesian parametric statistics. For $a > 1$ the prior is forgotten more quickly and for $a < 1$ more slowly than the traditional rate.

These calculations also lead to

$$n^{1/2}\{\widehat{P}(A) - P_n(A)\} \to_p 0 \quad \text{provided } a > \tfrac{1}{2}.$$

Under this condition, inferential statements made by the Bayesian, such as credibility intervals, will agree asymptotically with those of the frequentist using the empirical distribution. For smaller values of $a$, however, the speed with which the posterior is able to forget where it came from is really too slow; the predictive distribution is consistent, but converges slowly, and credibility intervals will not match frequentist confidence intervals, even for large $n$.

7.2. Consistency of the posterior distribution. We wish to find out whether the posterior distribution as such is consistent, in the sense that for any small neighbourhood around $P_{\mathrm{true}}$, the posterior probability of such a set converges to 1 as $n$ grows. This is a stronger statement than merely knowing that the posterior mean is a consistent estimator of $P_{\mathrm{true}}$.

Proposition. *Assume $X_1, X_2, \ldots$ are independent from some atom-free $P_{\mathrm{true}}$, and consider $\theta = \int g \,\mathrm{d}P$ for an arbitrary $g$ for which $g^* = \int g \,\mathrm{d}P_{\mathrm{true}}$ is finite. Then, for almost all sample paths, $\theta \,|\, \mathrm{data} \to_p g^*$.*

Proof. We know that the empirical mean $\theta_n = \int g \,\mathrm{d}P_n$ goes a.s. to $g^*$, and as above it is clear that $\mathrm{E}(\theta \,|\, \mathrm{data}) \to g^*$ a.s. in that $w_n \to 0$. It will suffice to show that $\mathrm{E}(\theta^2 \,|\, \mathrm{data}) \to (g^*)^2$ a.s.; this implies that the posterior variance goes to zero, and there is convergence in probability by the usual Chebyshov inequality argument.

To this end we work with expression (6.4), and aim to demonstrate that $n^2 g_n/a_n \to 1$ while the other terms go to zero. This causes $\mathrm{E}(\theta^2 \,|\, \mathrm{data})$ to go to $(g^*)^2$ for exactly those sample paths for which $\theta_n \to g^*$. From established formulae for $a_n, b_n, c_n$ we see that the third and fourth terms of the right hand side of (6.5) go to zero; this also secures that the terms $f_n/a_n$, $e_n/a_n$ and $2(n-1)h_n/a_n$ of (6.6) go to zero. It will therefore be enough to show that also the first term there goes to zero. For this we use formula (6.7). Note that $\delta_j \ge \delta_{j+1}$, and one finds $\varepsilon_j \le \delta_j$. A constant $K$ can be found such that $\delta_{j-1}/\delta_j \le K$ for all $j$. This implies

$$\frac{nd_n}{a_n} \le \delta_{n+1}\delta_n \sum_{j=1}^{n} \frac{\delta_{j-1}}{\delta_j \delta_{j+1}} \le K\delta_n \sum_{j=1}^{n} \frac{\delta_{n+1}}{\delta_{j+1}} \le Kn\delta_n,$$

which goes to zero since $w_n$ does. ■

Inspection of the details in these calculations show that the speed with which the variance goes to zero is $O(w_n)$. As we have seen, this corresponds to the traditional $O(1/n)$ variance rate for the Dirichlet process, whereas the speed may be both slower and faster for the more general prior process.

## 8. Bayesian inference and the posterior process

Let $P \sim \mathrm{GD}(H, P_0)$ and assume data $x_1, \ldots, x_n$ have been observed. This section looks into aspects of the posterior process, which turns out to be quite complicated. Only in the Dirichlet case, where $H$ is the $\mathrm{Beta}(1, b)$, does the posterior seem to have an easy structure. Bayesian inference can nevertheless be carried out via stochastic simulation.

8.1. One data point. We may take the view that $P$ of (1.2) is described in terms of $(B, \xi)$, where $B$ is the sequence of $B_j$s from $H$, leading in their turn to probability weights $\gamma_j = \bar{B}_1 \ldots \bar{B}_{j-1} B_j$, and where $\xi$ is the sequence of $\xi_j$s from $P_0$. Let in addition $J$ be a random variable in $\{1, 2, 3, \ldots\}$ which conditionally on $(B, \xi)$ has distribution given by these $\gamma_j$s, and define $X = \xi_J$. Then $X$ given $P$ has distribution $P$. The task is to pass from this simultaneous representation of $(P, X)$ to the conditional process $P$ given $X = x$.

When $X = \xi_J = x$ and $J = j$, one has $\xi_j = x$, without further knowledge about the other $\xi_k$s. Furthermore, the fact that this happened with probability $\gamma_j$ upgrades the information about the distributions $B_1, \ldots, B_j$, but does not affect the prior information about $B_k$ for $k > j$. Using arguments partly parallelling those in in Sethuraman (1994, Section 4), one finds that

$$P \mid \{X = x, J = j\} \sim P_{x,j} = \sum_{k=1}^{\infty} \gamma'_k \delta(\xi'_k), \tag{8.1}$$

where on the right hand side the $\{\gamma'_k\}$ sequence is formed from a $\{B'_k\}$ sequence independent of the $\xi'_k$, which are independently drawn from $P_0$ with the exception of $\xi'_j$, which is equal to the fixed $x$. Now $B'_k \sim H'_k$ for $k = 1, 2, \ldots$, where these $H'_k$s are not equal anymore; $\mathrm{d}H'_k(s) \propto (1 - s)\,\mathrm{d}H(s)$ for $k \leq j - 1$, $\mathrm{d}H'_k(s) \propto s\,\mathrm{d}H(s)$ for $k = j$, while $H'_k = H$ for $k \geq j + 1$. Thus there is a mixture representation of the posterior as

$$P \mid \{X = x\} \sim \sum_{j=1}^{\infty} q(j \mid x) P_{x,j}, \quad \text{or} \quad \Pr\{P \in C \mid x\} = \sum_{j=1}^{\infty} q(j \mid x) \Pr\{P_{x,j} \in C\},$$

valid for measurable subsets $C$ of the space of probability measures on the sample space (the Borel subsets under the topology of set-wise convergence). It remains only to identify

$$q(j \mid x) = \Pr\{J = j \mid X = x\} = \mathrm{E}\,\gamma_j = M_{0,1}^{j-1} M_{1,0} \quad \text{for } j = 1, 2, \ldots. \tag{8.2}$$

This is since the information $X = x$ from a single data point does not change the marginal distribution $J$ has from the $(P, J)$ model. Notice that in (8.1) there is dependence on $x$ in the $\xi'_k$, without overburdening the notation to indicate this.

8.2. The posterior in the general case. Conditionally on $(B, \xi)$, the two sequences determining $P$, let $J_1, \ldots, J_n$ be independent integer variables with distribution given by the $\gamma_j$s, and define $X_1 = \xi_{J_1}, \ldots, X_n = \xi_{J_n}$. Then, given $P$, these really form an independent $n$-sample from $P$. This provides a simultaneous representation of $(P, X_1, \ldots, X_n)$.

Suppose for representational simplicity that the data points $x_1, \ldots, x_n$ are distinct. One may generalise the first result above to

$$P \,|\, \{X_1 = x_1, \ldots, X_n = x_n, J_1 = j_1, \ldots, J_n = j_n\} \sim P_{\mathrm{data}, j_1, \ldots, j_n} = \sum_{k=1}^{\infty} \gamma'_k \delta(\xi'_k),$$

where the $\{\gamma'_k\}$ is formed from a sequence of independent variables $\{B'_k\}$ and independently of the $\{\xi'_k\}$; these are such that $\xi'_{j_1}, \ldots, \xi'_{j_n}$ are fixed at values $x_1, \ldots, x_n$, respectively, while the remaining $\xi'_k$s are independent from $P_0$. The upgraded distributions $H'_k$ for $B'_k$ are given by

$$\mathrm{d}H'_k(s) = \mathrm{const.}\, (1-s)^{Y(k)-\Delta N(k)} s^{\Delta N(k)} \,\mathrm{d}H(s),$$

in which $Y(k) = \sum_{i=1}^n I\{j_i \geq k\}$ and $\Delta N(k) = \sum_{i=1}^n I\{j_i = k\}$. Hence

$$P \,|\, \mathrm{data} \sim \sum_{j_1, \ldots, j_n \ \mathrm{distinct}} q(j_1, \ldots, j_n \,|\, \mathrm{data}) P_{\mathrm{data}, j_1, \ldots, j_n}.$$

It remains to give the posterior distribution of indexes. Say that $G$ has a geometric distribution with parameter $M$ if $\Pr\{G = g\} = (1-M)M^g$ for $g = 1, 2, \ldots$.

Proposition. *Let there be $n$ distinct data points, and order the random indexes $J_1, \ldots, J_n$ as $J_{(1)} < \cdots < J_{(n)}$. Then*

$$(J_{(1)}, \ldots, J_{(n)}) =_d (G_1, G_1 + G_2, \ldots, G_1 + \cdots + G_n),$$

*where $G_1, \ldots, G_n$ are independent and geometric with parameters $M_{0,n}, \ldots, M_{0,1}$, respectively.*

Proof. Knowledge of data values $\xi_{j_i} = x_i$ does not change the distribution of the labels as long as these are distinct. For the ordered labels one therefore finds the distribution

$$\begin{aligned} \bar{q}(j_1, \ldots, j_n) &= n!\, \mathrm{E}\, \gamma_{j_1} \cdots \gamma_{j_n} / \Pr(D_n) \\ &= \frac{n!}{\Pr(D_n)} \mathrm{E} \prod_{k=1}^{\infty} \bar{B}_k^{Y(k)-\Delta N(k)} B_k^{\Delta N(k)} = \frac{n!}{\Pr(D_n)} \prod_{k=1}^{\infty} M_{\Delta N(k), Y(k)-\Delta N(k)} \end{aligned}$$

for $j_1 < \cdots < j_n$, where $D_n$ is the event that data points are distinct. The product may be expressed as

$$M_{0,n}^{j_1-1} M_{1,n-1} M_{0,n-1}^{j_2-j_1-1} M_{1,n-2} \cdots M_{0,2}^{j_{n-1}-j_{n-2}-1} M_{1,1} M_{0,1}^{j_n-j_{n-1}-1} M_{1,0},$$

while it is shown in Section 10.2 that $\Pr(D_n)$ is equal to the $a_n$ of formula (5.2). Combining these facts one is left with

$$(1-M_{0,n}) M_{0,n}^{j_1-1} (1-M_{0,n-1}) M_{0,n-1}^{j_2-j_2-1} \cdots (1-M_{0,2}) M_{0,2}^{j_{n-1}-j_{n-2}-1} (1-M_{0,1}) M_{0,1}^{j_n-j_{n-1}-1},$$

which is seen to be equivalent to the claim. ■

Remark. The description above is valid for the general $\mathrm{GD}(H, P_0)$ case, and can even be generalised further to the case of different distributions $H_1, H_2, \ldots$ for $B_1, B_2, \ldots$ in the prior. Note that for the particular $\mathrm{GD}(a, b, P_0)$ family, in which the Dirichlet is the $a = 1$ case, at least the $H$ to $H'_k$ updating is easy, in that $H'_k \sim \mathrm{Beta}(a + \Delta N(k), b + Y(k) - \Delta N(k))$. For $k$ larger than the largest $j_i$ the $H'_k$ is the same as the original $H$.

For the Dirichlet case the posterior can of course be described in a much simpler way than the scheme above. One may deduce from (8.1) and (8.2) that $P \mid x$ is simply another Dirichlet with total measure $bP_0 + \delta(x)$, via various identities for Beta distributions; see Sethuraman (1994, Section 4).

## 9. Distribution of random means

Recently there has been much attention given to studying aspects of the distributions of random Dirichlet means; see Diaconis and Kemperman (1996), Regazzini, Guglielmo and di Nunno (2000) and Hjort and Ongaro (2000) for discussion and references. Here we look at the more general version of this problem, where $P$ is a generalised Dirichlet process.

9.1. General transform identities. That equation (3.2) characterises the distribution of $\theta$ uniquely can be seen as in a parallel situation in Hjort and Ongaro (2000); see also Lemma 3.3 in Sethuraman (1994). Exhibiting this distribution is however a difficult task and can rarely be done in closed form. The list of explicit solutions to this problem for the Dirichlet case is very short, so a fortiori one cannot expect explicit answers for the more general $\mathrm{GD}(H, P_0)$ case. We point out, however, that equation (3.2) translates into an identity for characteristic or moment generating functions and which can be worked with to extract information about the $\theta$ distribution. Let $L(u) = \mathrm{E}\exp(iu\theta)$ and $L_0(u) = \mathrm{E}_0 \exp(iuY)$. Via conditioning on $(B, Y)$ and then integrating over $Y$ one finds from (3.2) that

$$L(u) = \int_0^1 L_0(us) L(u(1-s))\,\mathrm{d}H(s). \tag{9.1}$$

In principle $L$ is determined from knowledge of $L_0$. Similarly a convolution-type identity can be put up for the density $f$ of $\theta$ in terms of the density $f_0$ for $Y$ under $P_0$.

An exception admitting a straight answer is when $Y$ is Cauchy. One then sees that the Cauchy distribution for $\theta$ fits the stochastic equation (3.2), and is hence the answer; $\theta$ is Cauchy when $Y$ is. This is valid for any distribution $H$ for the $B_j$s, as can also be seen via (9.1), and therefore generalises a classic result for the Dirichlet process.

9.2. Results for normal and stable laws. Another situation of interest where some progress can be made is the case of a normal base measure. Let $W = \sum_{j=1}^{\infty} \gamma_j^2$ in (1.2); this is a well-defined variable on $(0, 1)$ with a distribution determined via its stochastic equation

$$W =_d B^2 + (1 - B)^2 W, \quad \text{where } B \sim H \text{ in } (0, 1). \tag{9.2}$$

This follows from (1.2) in the same way as (3.1) was derived. If now $P_0$ is standard normal, $\theta = \sum_{j=1}^\infty \gamma_j Y_j$ is for given weights a normal $(0, W)$. This shows that $\theta$ is a scale-mixture of normals, with density of the form $\int_0^1 \sigma^{-1}\phi(\sigma^{-1}t)p(\sigma)\,\mathrm{d}\sigma$, the $p$ density in question being the density of $W^{1/2}$. This density cannot be written down in closed form, but may be arbitrarily well approximated via its moment sequence, which may be found in a simple recursive manner; see Hjort and Ongaro (2000) for illustrations for the special Dirichlet process case.

These arguments also work for general stable laws. For $\alpha \in (0, 2]$ and $c$ positive, say that $Y$ is stable $(\alpha, c)$ if its characteristic function is $\mathrm{E}\exp(iuY) = \exp(-c^\alpha|u|^\alpha)$; notice that $Y/c$ then is stable $(\alpha, 1)$. Now take $P \sim \mathrm{GD}(H, P_0)$ where $P_0$ is stable $(\alpha, 1)$, and consider $\theta = \int x\,\mathrm{d}P(x)$. This random mean can be expressed as $\sum_{j=1}^\infty \gamma_j Y_j$ where $Y_j \sim P_0$. Let $W = (\sum_{j=1}^\infty \gamma_j^\alpha)^{1/\alpha}$. Then $\theta$ given $\{\gamma_j\}$ is a stable $(\alpha, W)$. It follows that $\theta$ is a scale mixture of such stable laws; its density is $\int_0^1 w^{-1}g_\alpha(w^{-1}t)\,p_\alpha(w)\,\mathrm{d}w$, where $p_\alpha$ is the density of $W$ and $p_\alpha$ the density of a stable $(\alpha, 1)$ variable.

## 10. Concluding remarks

In these final remarks a couple of further uses of the generalised Dirichlet process are identified, and possibilities for further research are noted.

10.1. Bayesian robustness. If a statistician uses the Dirichlet $(b_0, P_0)$ process as a prior, or as an element in a more complicated prior, one may supplement such analysis with that using the $\mathrm{GD}(a, b, P_0)$ prior, preferably with the proviso $2b/(1+a) = b_0$, as indicated in Section 4. Answers derived under the Dirichlet should then be compared to those obtained with the more general prior, say corresponding to values of $a$ inside $(\frac{1}{2}, \frac{3}{2})$. Small variation in results indicates Bayesian robustness.

10.2. Marginal distribution when data are distinct. Let $P \sim \mathrm{GD}(H, P_0)$ with consequent observations $X_1, X_2, \ldots$. Consider $D_n$, the event that the $n$ first observations are distinct. From the definition (1.2),

$$\Pr(D_n) = n!\,\Pr\{X_1 < \cdots < X_n\} = n! \sum_{i_1 < \cdots < i_n} \mathrm{E}\,\gamma_{i_1}\cdots\gamma_{i_n}.$$

But from the proof of the lemma of Section 5.1 it is clear that $\Pr(D_n) = a_n = n!\,\delta_{n-1}\cdots\delta_1$, in the notation of Section 5.3. It also follows that for disjoint sets $A_1, \ldots, A_n$,

$$\Pr\{X_1 \in A_1, \ldots, X_n \in A_n \,|\, D_n\} = P_0(A_1)\cdots P_0(A_n),$$

that is, conditional on data points being distinct, the observations form an i.i.d. sequence from $P_0$. This generalises a result for the Dirichlet process due to Korwar and Hollander (1973).

10.3. A semiparametric prior giving density estimates. Assume that $Y_i = \theta + \varepsilon_i$ for $i = 1, \ldots, n$ with $\varepsilon_1, \ldots, \varepsilon_n$ being independent from a $P$ centred at zero. For this signal plus noise model a sensible prior could be to give $\theta$ a prior $\pi(\theta)\,\mathrm{d}\theta$ and $P$ an independent $\mathrm{GD}(H, P_0)$ process prior, where $P_0$ has a density $p_0$ centred at zero. Then calculations similar to but more general than those of Section 5 show that $\theta$ given observations $y_1, \ldots, y_n$ has posterior density $\pi(\theta \,|\, \mathrm{data}) = c\,\pi(\theta)\prod_{i=1}^n p_0(y_i - \theta)$, which is also the posterior computed under the simple parametric model where $P = P_0$. It is assumed here that the $y_i$s are distinct. Since knowing data and $\theta$ amounts to knowing the $\varepsilon_i$s, results of Sections 5 and 6 apply, giving

$$\mathrm{E}\{P(A) \,|\, \mathrm{data}, \theta\} = w_n P_0(A) + (1 - w_n) n^{-1} \sum_{i=1}^n I\{y_i - \theta \in A\}.$$

But this gives

$$\widehat{P}(A) = \mathrm{E}\{P(A) \,|\, \mathrm{data}\} = w_n P_0(A) + (1 - w_n) n^{-1} \sum_{i=1}^n \Pr\{\theta \in y_i - A \,|\, \mathrm{data}\},$$

which is found to be an integral of a smooth density estimate,

$$\widehat{p}(t) = w_n p_0(t) + (1 - w_n) n^{-1} \sum_{i=1}^n \pi(y_i - t \,|\, \mathrm{data}).$$

This is a mixture of the prior guess density and a kernel type density estimator, with bandwidth approximately proportional to $n^{-1/2}$. The construction here can be generalised to include scale parameters and covariates.

10.4. Prior process with different $B_i$ distributions. As the complicated posterior indicates, it may be useful to allow different distributions $H_1, H_2, \ldots$ for the $B_1, B_2, \ldots$ in (1.2). A condition guaranteeing a.s. convergence of $\bar{B}_1 \cdots \bar{B}_n$ to zero is needed. Tsilevich (1997) has actually worked with a particular construction of this type, but in a different probabilistic framework, and she does not discuss applications or implications for Bayesian statistics. For the general prior process indexed by $H_1, H_2, \ldots$ the posterior of $P$ given a set of data becomes of the same type, with updated $H_1', H_2', \ldots$, following the lines of Section 8. Accordingly, at least in a technical sense of the term, we have constructed a large conjugate class of nonparametric priors.

10.5. Ties in data. Formulae for posterior mean and variance were derived above for the case of data points $x_1, \ldots, x_n$ being distinct, as they would be if stemming from an underlying atom-free distribution. When the $X_i$s really come from a $P$ chosen by the generalised Dirichlet process there will be multiple ties with positive probability, however. A more complete description should therefore include generalised versions of say (5.4) and (6.4) for multiplicities among the data points. This is possible but requires cumbersome extensions of arguments and recursive schemes developed in Section 6.3. To illustrate, and to compare issues of data weighting with the distinct case and with the Dirichlet case, we indicate here results for the case of $x_1 = x_2$ distinct from $n-2$ distinct values $x_3, \ldots, x_n$. Let $\mathrm{info}_\varepsilon$ indicate the information $X_1, X_2 \in S(x_1, \varepsilon)$ and $X_i \in S(x_i, \varepsilon)$ for $i = 3, \ldots, n$, and write $\theta_i = P(S(x_i, \varepsilon))$. With arguments and notation as in Section 6.1 one first finds that

$$\mathrm{E}\{P(A)\,|\,\mathrm{info}_\varepsilon\} = \frac{\mathrm{E}P(A)\theta_1^2\theta_3\cdots\theta_n}{\mathrm{E}\theta_1^2\theta_3\cdots\theta_n} = \frac{b_nP_0(A) + c_nP_0(A)p_1}{b_{n-1} + c_{n-1}p_1} = \frac{b_n}{b_{n-1}}P_0(A) + O(\varepsilon)$$

for sets $A$ not meeting the data, which means that $\widehat{P} = \mathrm{E}\{P(\cdot)\,|\,\mathrm{data}\}$ is the same as $(b_n/b_{n-1})P_0$ outside the data set. Furthermore,

$$\mathrm{E}\{P(A_1)\,|\,\mathrm{info}_\varepsilon\} = \frac{\mathrm{E}\theta_1^3\theta_3\cdots\theta_n}{\mathrm{E}\theta_1^2\theta_3\cdots\theta_n} = \frac{d_{n-1} + e_{n-1}p_1 + f_{n-1}p_1^2}{b_{n-1} + c_{n-1}p_1} = \frac{d_{n-1}}{b_{n-1}} + O(\varepsilon),$$

$$\mathrm{E}\{P(A_3)\,|\,\mathrm{info}_\varepsilon\} = \frac{\mathrm{E}\theta_1^2\theta_3^2\cdots\theta_n}{\mathrm{E}\theta_1^2\theta_3\cdots\theta_n} = \frac{g_{n-1} + h_{n-1}(p_1+p_3) + i_{n-1}p_1p_3}{b_{n-1} + c_{n-1}p_1} = \frac{g_{n-1}}{b_{n-1}} + O(\varepsilon).$$

Accordingly, $\mathrm{E}[P\{x_1\}\,|\,\mathrm{data}] = d_{n-1}/b_{n-1}$ while $\mathrm{E}[P\{x_i\}\,|\,\mathrm{data}] = g_{n-1}/b_{n-1}$ for the $n-2$ other data points.

For the $\mathrm{GD}(a, b, P_0)$ process one learns that for $a < 1$, there is slightly less weight $b_n/b_{n-1}$ to the outside-the-data set with the $x_1 = x_2$ tie than without such a tie; the situation is reversed for $a > 1$. The expected weight $d_{n-1}/b_{n-1}$ given to the double data point $x_1$ can similarly be compared with $2b_n/a_n$, the expected weight given to $\{x_1, x_2\}$ when these are distinct. Here sometimes the first is bigger than the second and sometimes the other way around, for a given $a \neq 1$. Quite generally, these probability weights given to the outside-the-data set and to the individual data points are independent of ties if and only if the process is a Dirichlet, that is, the $H$ distribution is a $\mathrm{Beta}(1, b)$ for some $b$.

10.6. Other nonparametric priors. One sees from the results of Section 8 that the posterior distribution of an arbitrary random mean $\theta = \int g\,\mathrm{d}P$ has the structure

$$\theta\,|\,\mathrm{data} \sim D_0T + \sum_{i=1}^n D_ig(x_i), \tag{10.1}$$

where $D_0, D_1, \ldots, D_n$ are random weights summing to 1 and $T$ is a variable with mean $\theta_0 = \int g\,\mathrm{d}P_0$. The distribution of $(D_1, \ldots, D_n)$ is symmetric when the data values are distinct, securing equal weight to each data point. The distribution is more complicated with ties in the data, as indicated above. It is interesting that several different unrelated nonparametric priors lead to the structure (10.1), among them two constructions of Hjort and Ongaro (2001), and, of course, the Dirichlet. For each such prior the predictive distribution takes the form $w_nF_0 + (1-w_n)\widetilde{F}_n$, say, where $w_n = \mathrm{E}D_0$ and $\widetilde{F}_n$ is a 'modified empirical distribution function', being equal to the empirical $F_n$ when data are distinct and otherwise awarding somewhat modified weights to data points with different multiplicities.

A characterisation theorem of Lo (1991) implies that only when the prior is a Dirichlet do these weights become proportional to the multiplicities, that is, only then is $\widetilde{F}_n$ equal to $F_n$ for all data configurations.

## References


Diaconis, P. and Kemperman, J. (1996). Some new tools for Dirichlet priors. In *Bayesian Statistics 5*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), 97–106. Oxford University Press, Oxford.

Feigin, P.B. and Tweedie, R.L. (1989). Linear functionals and Markov chains associated with Dirichlet processes. *Mathematics Proceedings of the Cambridge Philosophical Society* **105**, 579–585.

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

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615–629.

Guglielmi, A., Holmes, C.C. and Walker, S.G. (2001). Perfect simulation involving functionals of a Dirichlet process. To appear.

Guglielmi, A. and Tweedie, R.L. (2000). MCMC estimation of the law of the mean of a Dirichlet process. Technical report TR 00.15, CNR–IAMI, Milano.

Hjort, N.L. (2001). Topics in nonparametric Bayesian statistics. In *Highly Structured Stochastic Systems*, to be published.

Hjort, N.L. and Ongaro, A. (2000). On the distribution of random Dirichlet means. Statistical Research Report, University of Oslo.

Hjort, N.L. and Ongaro, A. (2001). Two generalisations of the Dirichlet process. In progress.

Hjort, N.L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. Statistical Research Report, University of Oslo.

Korwar, R.M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Annals of Probability* **1**, 705–711.

Lo, A.Y. (1991). A characterization of the Dirichlet process. *Statistics and Probability Letters* **12**, 185–187.

Paulsen, J. and Hove, A. (1999). Markov chain Monte Carlo simulation of the distribution of some perpetuities. *Advances of Applied Probability* **31**, 112–134.

Regazzini, E., Guglielmi, A. and di Nunno, G. (2000). Theory and numerical analysis for exact distributions of functionals of a Dirichlet process. Research report, Università di Pavia.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Sethuraman, J. and Tiwari, R. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Proceedings of the Third Purdue Symposium* *on Statistical Decision Theory and Related Topics*, S.S. Gupta and J. Berger (eds.), 305–315. Academic Press, New York.

Цилевич, Н.В. (1997). Распределение среднего значения для некоторых случайных мер. Записки научных семинаров ПОМИ, Petersburg Department of Mathematical Institute, University of Sankt-Peterburg, 268–279.

Walker, S.G., Damien, P., Laud, P.W. and Smith, A.F.M. (1998). Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society* **B 61**, 485–527.