4

Dirichlet process

The Dirichlet process is the "normal distribution of Bayesian nonparametrics". It is the default prior on spaces of probability measures, and a building block for priors on other structures.

4.1 Finite-dimensonal Dirichlet distribution

A random vector $X = (X_1, \dots, X_k)$ with values in the k-dimensional unit simplex $\mathbb{S}_k := \{(s_1, \dots, s_k) : s_j \geq 0, \sum_{j=1}^k s_j = 1\}$ is said to possess a *Dirichlet distribution* with parameters $k \in \mathbb{N}$ and $\alpha_1, \dots, \alpha_k > 0$ if it has density proportional to $x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1}$ with respect to the Lebesgue measure on $\mathbb{S}_k$.

The unit simplex $\mathbb{S}_k$ is a subset of a $(k-1)$-dimensional affine space, and so "its Lebesgue measure" is to be understood to be $(k-1)$-dimensional Lebesgue measure appropriately mapped to $\mathbb{S}_k$. The norming constant of the Dirichlet density depends on the precise construction. Alternatively, the vector X may be described through the vector $(X_1, \dots, X_{k-1})$ of its first $k-1$ coordinates, the last coordinate being fixed by the relationship $X_k = 1 - \sum_{i=1}^{k-1} X_i$. This vector has density proportional to $x_1^{\alpha_1 - 1} \cdots x_{k-1}^{\alpha_{k-1} - 1} (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1}$ with respect to the usual $(k-1)$-dimensional Lebesgue measure restricted to the set $\mathbb{D}_k = \{(x_1, \dots, x_{k-1}) : \min_i x_i \geq 0, \sum_{i=1}^{k-1} x_i \leq 1\}$. The inverse of the normalizing constant is given by the *Dirichlet form*

$$\int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-x_1-\cdots-x_{k-2}} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_{k-1}^{\alpha_{k-1} - 1} \times (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1} \, dx_{k-1} \cdots dx_2 \, dx_1. \tag{4.1}$$

The Dirichlet distribution takes its name from this integral, which was can be evaluated to $\Gamma(\alpha_1) \cdots \Gamma(\alpha_k) / \Gamma(\alpha_1 + \cdots + \alpha_k)$ by successive integrations and scalings to beta integrals.

Definition 4.1 (Dirichlet distribution) The *Dirichlet distribution* $\mathrm{Dir}(k; \alpha)$ with parameters $k \in \mathbb{N} - \{1\}$ and $\alpha = (\alpha_1, \dots, \alpha_k) > 0$ is the distribution of a vector $(X_1, \dots, X_k)$ such that $\sum_{i=1}^k X_i = 1$ and such that $(X_1, \dots, X_{k-1})$ has density

$$\frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \, x_1^{\alpha_1} x_2^{\alpha_2 - 1} \cdots x_{k-1}^{\alpha_{k-1} - 1} (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1}, \quad x \in \mathbb{D}_k. \tag{4.2}$$

The Dirichlet distribution with parameters k and $\alpha \geq 0$, where $\alpha_i = 0$ for $i \in I \subsetneq \{1, \dots, k\}$, is the distribution of the vector $(X_1, \dots, X_k)$ such that $X_i = 0$ for $i \in I$

and such that $(X_i: i \notin I)$ possesses a lower-dimensional Dirichlet distribution, given by a density of the form (4.2).

For $k = 2$ the vector (X_1, X_2) is completely described by a single coordinate, where $X_1 \sim \mathrm{Be}(\alpha_1, \alpha_2)$ and $X_2 = 1 - X_1 \sim \mathrm{Be}(\alpha_2, \alpha_1)$. Thus the Dirichlet distribution is a multivariate generalization of the Beta distribution. The $\mathrm{Dir}(k; 1, \ldots, 1)$-distribution is the uniform distribution on $\mathbb{S}_k$.

Throughout the section we write $|\alpha|$ for $\sum_{i=1}^k \alpha_i$.

Proposition 4.2 (Gamma representation) *If $Y_i \overset{ind}{\sim} \mathrm{Ga}(\alpha_i, 1)$, then $(Y_1/Y, \ldots, Y_k/Y) \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$, and is independent of and $Y := \sum_{i=1}^k Y_i$.*

Proof We may assume that all α_i are positive. The Jacobian of the inverse of the transformation $(y_1, \ldots, y_k) \mapsto (y_1/y, \ldots, y_{k-1}/y, y) =: (x_1, \ldots, x_{k-1}, y)$ is given by $y^{k-1}(1 - x_1 - \cdots - x_{k-1})$. The density of the $\mathrm{Ga}(\alpha_i, 1)$-distribution is proportional to $e^{-y_i} y_i^{\alpha_i - 1}$. Therefore the joint density of $(Y_1/Y, \ldots, Y_{k-1}/Y, Y)$ is, proportional to,

$$e^{-y} y^{|\alpha|-1} x_1^{\alpha_1 - 1} \cdots x_{k-1}^{\alpha_{k-1}-1} (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1}.$$

This factorizes into a Dirichlet density of dimension $k - 1$ and the $\mathrm{Ga}(|\alpha|, 1)$-density of Y. □

Proposition 4.3 (Aggregation) *If $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$ and $Z_j = \sum_{i \in I_j} X_i$ for a given partition $I_1, \ldots, I_m$ of $\{1, \ldots, k\}$, then*

(i) *$(Z_1, \ldots, Z_m) \sim \mathrm{Dir}(m; \beta_1, \ldots, \beta_m)$, where $\beta_j = \sum_{i \in I_j} \alpha_i$, for $j = 1, \ldots, m$.*
(ii) *$(X_i/Z_j: i \in I_j) \overset{ind}{\sim} \mathrm{Dir}(\#I_j; \alpha_i, i \in I_j)$, for $j = 1, \ldots, m$.*
(iii) *$(Z_1, \ldots, Z_m)$ and $(X_i/Z_j: i \in I_j, j = 1, \ldots, m)$ are independent.*

Conversely, if X is a random vector such that (i)–(iii) hold, for a given partition $I_1, \ldots, I_m$ and $Z_j = \sum_{i \in I_j} X_i$, then $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$.

Proof In terms of the Gamma representation $X_i = Y_i/Y$ of Proposition 4.2 we have

$$Z_j = \frac{\sum_{i \in I_j} Y_i}{Y}, \qquad \text{and} \qquad \frac{X_i}{Z_j} = \frac{Y_i}{\sum_{i \in I_j} Y_i}.$$

Because $W_j := \sum_{i \in I_j} Y_i \overset{\text{ind}}{\sim} \mathrm{Ga}(\beta_j, 1)$ for $j = 1, \ldots, m$, and $\sum_j W_j = Y$, the Dirichlet distributions in (i) and (ii) are immediate from Proposition 4.2. The independence in (ii) is immediate from the independence of the groups $(Y_i: i \in I_j)$, for $j = 1, \ldots, m$. By Proposition 4.2 W_j is independent of $(Y_i/W_j: iin I_j)$, for every j, whence by the independence of the groups the variables $W_j, (Y_i/W_j: i \in I_j)$, for $j = 1, \ldots, m$, are jointly independent. Then (iii) follows, because $(X_i/Z_j: i \in I_j, j = 1, \ldots, m)$ is a function of $(Y_i/W_j: i \in I_j, j = 1, \ldots, m)$ and $(Z_1, \ldots, Z_m)$ is a function of $(W_j: j = 1, \ldots, m)$.

The converse also follows from the Gamma representation. □

Proposition 4.4 (Moments) *If $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$, then $X_i \sim \mathrm{Be}(\alpha_i, |\alpha| - \alpha_i)$. In*

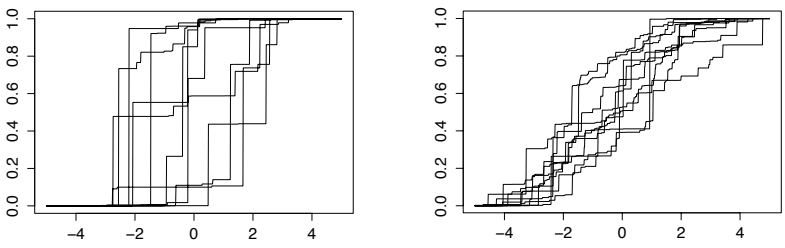

Figure 4.1 Cumulative distribution functions of 10 draws from the Dirichlet process with base measures $N(0, 2)$ (left) and $10N(0, 2)$ (right). (Computations based on Sethuraman representation truncated to 1000 terms.)

particular, $\mathrm{E}(X_i) = \alpha_i/|\alpha|$ *and* $\mathrm{var}(X_i) = \alpha_i(|\alpha| - \alpha_i)/(|\alpha|^2(|\alpha| + 1))$. *Furthermore,* $\mathrm{cov}(X_i, X_j) = -\alpha_i\alpha_j/(|\alpha|^2(|\alpha| + 1))$ *and, with* $r = r_1 + \cdots + r_k$,

$$\mathrm{E}(X_1^{r_1} \cdots X_k^{r_k}) = \frac{\Gamma(\alpha_1 + r_1) \cdots \Gamma(\alpha_k + r_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \times \frac{\Gamma(|\alpha|)}{\Gamma(|\alpha| + r)}. \tag{4.3}$$

In particular, if $r_1, \ldots, r_k \in \mathbb{N}$, *then the expression in* (4.3) *is equal to* $\alpha_1^{[r_1]} \cdots \alpha_k^{[r_k]}/|\alpha|^{[r]}$, *where* $x^{[m]} = x(x + 1) \cdots (x + m - 1)$, $m \in \mathbb{N}$, *stands for the ascending factorial.*

Proof The first assertion follows from Proposition 4.3 by taking $m = 2$, $I_i = \{i\}$, $I_2 = I - \{i\}$, for $I = \{1, \ldots, k\}$. Next the expressions for expectation and variance follow by the properties of the beta distribution.

For the second assertion, we take $m = 2$, $I_1 = \{i, j\}$ and $I_2 = I - I_1$ in Proposition 4.3 to see that $X_i + X_j \sim \mathrm{Be}(\alpha_i + \alpha_j, |\alpha| - \alpha_i - \alpha_j)$. This gives $\mathrm{var}(X_i + X_j) = (\alpha_i + \alpha_j)(|\alpha| - \alpha_i - \alpha_j)/\big(|\alpha|^2(|\alpha| + 1)\big)$, and allows to obtain the expression for the covariance from the identity $2\,\mathrm{cov}(X_i, X_j) = \mathrm{var}(X_i + X_j) - \mathrm{var}(X_i) - \mathrm{var}(X_j)$.

For the derivation of (4.3), observe that the mixed moment is the ratio of two Dirichlet forms (4.1) with parameters $(\alpha_1 + r_1, \ldots, \alpha_k + r_k)$ and $(\alpha_1, \ldots, \alpha_k)$. □

4.2 Dirichlet process

Definition 4.5 (Dirichlet process) A random measure P on $(\mathfrak{X}, \mathscr{X})$ is said to possess a *Dirichlet process* distribution $\mathrm{DP}(\alpha)$ with *base measure* α, if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,

$$\big(P(A_1), \ldots, P(A_k)\big) \sim \mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big). \tag{4.4}$$

In this definition α is a given finite positive Borel measure on $(\mathfrak{X}, \mathscr{X})$. We write $|\alpha| = \alpha(\mathfrak{X})$ for its total mass and $\bar{\alpha} = \alpha/|\alpha|$ for the probability measure obtained by normalizing α, respectively, and use the notations $P \sim \mathrm{DP}(\alpha)$ and $P \sim \mathrm{DP}\big(|\alpha|, \bar{\alpha}\big)$ interchangeably to say that P has a Dirichlet process distribution with base measure α.

Existence of the Dirichlet process is not obvious, but proved below.

Definition 4.5 specifies the joint distribution of the vector $\big(P(A_1), \ldots, P(A_k)\big)$, for any

measurable partition $\{A_1, \ldots, A_k\}$ of the sample space. In particular, it specifies the distribution of $P(A)$, for every measurable set A, and hence the *mean measure* $A \mapsto \mathrm{E}P(A)$. By Proposition 4.4,

$$\mathrm{E}(P(A)) = \bar{\alpha}(A).$$

Thus the mean measure is the normalized base measure $\bar{\alpha}$, which is a valid Borel measure by assumption. Therefore Theorem 3.5 implies existence of the Dirichlet process $\mathrm{DP}(\alpha)$ provided the specification of distributions can be consistently extended to any vector of the type $\big(P(A_1), \ldots, P(A_k)\big)$, for arbitrary measurable sets and not just partitions, in such a way that it gives a finitely-additive measure.

An arbitrary collection $A_1, \ldots, A_k$ of measurable sets defines a collection of 2^k atoms of the form $A_1^* \cap A_2^* \cap \cdots \cap A_k^*$, where A^* stands for A or A^c. These atoms $\{B_j \colon j = 1, \ldots, 2^k\}$ (some of which may be empty) form a partition of the sample space, and hence the joint distribution of $\big(P(B_j) \colon j = 1, \ldots, 2^k\big)$ is defined by Definition 4.5. Every A_i can be written as a union of atoms, and $P(A_i)$ can be defined accordingly as the sum of the corresponding $P(B_j)$'s. This defines the distribution of the vector $\big(P(A_1), \ldots, P(A_k)\big)$.

To prove the existence of a stochastic process $\big(P(A) \colon A \in \mathscr{X}\big)$ that possesses these marginal distributions, it suffices to verify that this collection of marginal distributions is consistent in the sense of Kolmogorov's extension theorem. Consider the distribution of the vector $\big(P(A_1), \ldots, P(A_{k-1})\big)$. This has been defined using the coarser partitioning in the 2^{k-1} sets of the form $A_1^* \cap A_2^* \cap \cdots \cap A_{k-1}^*$. Every set in this coarser partition is a union of two sets in the finer partition used previously to define the distribution of $\big(P(A_1), \ldots, P(A_k)\big)$. Therefore, consistency pertains if the distributions specified by Definition 4.5 for two partitions, where one is finer than the other, are consistent.

Let $\{A_1, \ldots, A_k\}$ be a measurable partition and let $\{A_{i1}, A_{i2}\}$ be a further measurable partition of A_i, for $i = 1, \ldots, k$. Then Definition 4.5 specifies that

$$\begin{aligned}\big(P(A_{11}), P(A_{12}), P(A_{21}), \ldots, P(A_{k1}), P(A_{k2})\big) \\ \sim \mathrm{Dir}\big(2k; \alpha(A_{11}), \alpha(A_{12}), \alpha(A_{21}), \ldots, \alpha(A_{k1}), \alpha(A_{k2})\big).\end{aligned}$$

In view of the group additivity of finite dimensional Dirichlet distributions given by Proposition 4.3, this implies

$$\Big(\sum_{j=1}^{2} P(A_{1j}), \ldots, \sum_{j=1}^{2} P(A_{kj})\Big) \sim \mathrm{Dir}\Big(k; \sum_{j=1}^{2} \alpha(A_{1j}), \ldots, \sum_{j=1}^{2} \alpha(A_{kj})\Big).$$

Consistency follows as the right side is $\mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big)$, since α is a measure.

That $P(\emptyset) = 0$ and $P(\mathfrak{X}) = 1$ almost surely follow from the fact that $\{\emptyset, \mathfrak{X}\}$ is an eligible partition in Definition 4.5, whence $\big(P(\emptyset), P(\mathfrak{X}),\big) \sim \mathrm{Dir}(2; 0, |\alpha|)$ by (4.4). That $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ almost surely for every disjoint pair of measurable sets A_1, A_2, follows similarly from consideration of the distributions of the vectors $\big(P(A_1), P(A_2), P(A_1^c \cap A_2^c)\big)$ and $\big(P(A_1 \cup A_2), P(A_1^c \cap A_2^c)\big)$, whose three and two components both add up to 1.

We have proved the existence of the Dirichlet process distribution $\mathrm{DP}(\alpha)$ for every Polish sample space and every base measure α.

4.3 The Sethuraman representation

The *Sethuraman representation* of the Dirichlet process is a random discrete measure of the type discussed in Section 3.2, with stick-breaking weights, as in Section 3.2.1, based on the Beta-distribution. The random support points are generated from the normalized base measure.

The representation gives an easy method to simulate a Dirichlet process, at least approximately. It also proves the remarkable fact that realizations from the Dirichlet measure are discrete measures, with probability one.

In view of the results of Section 3.2, we also infer that the Dirichlet process is fully supported relative to the weak topology.

Theorem 4.6 (Sethuraman) *If* $\theta_1, \theta_2, \ldots \overset{iid}{\sim} \bar{\alpha}$ *and* $Y_1, Y_2, \ldots \overset{iid}{\sim} \mathrm{Be}(1, M)$ *are independent random variables and* $V_j = Y_j \prod_{l=1}^{j-1}(1 - Y_l)$*, then* $\sum_{j=1}^{\infty} V_j \delta_{\theta_j} \sim \mathrm{DP}(M\bar{\alpha})$.

Proof Because $\mathrm{E}\big(\prod_{l=1}^{j}(1 - Y_l)\big) = (M/(M+1))^j \to 0$, the stick-breaking weights V_j form a probability vector a.s. (c.f. Lemma 3.4), so that P is a probability measure a.s..

For $j \geq 2$ define $V_j' = Y_j \prod_{l=2}^{j-1}(1 - Y_l)$ and $\theta_j' = \theta_{j+1}$. Then $V_j = (1 - Y_1)V_{j-1}'$ for every $j \geq 1$ and hence

$$P := V_1\delta_{\theta_1} + \sum_{j=2}^{\infty} V_j\delta_{\theta_j} = Y_1\delta_{\theta_1} + (1 - Y_1)\sum_{j=1}^{\infty} V_j'\delta_{\theta_j'}.$$

The random measure $P' := \sum_{j=1}^{\infty} V_j'\delta_{\theta_j'}$ has exactly the same structure as P, and hence possesses the same distribution. Furthermore, it is independent of (Y_1, θ_1).

We conclude that P satisfies the distributional equation (4.5) given below, and the theorem follows from Lemma 4.7. □

The distributional equation for the Dirichlet process used in the preceding proof is of independent interest. For independent random variables $Y \sim \mathrm{Be}\big(1, |\alpha|\big)$ and $\theta \sim \bar{\alpha}$, consider the equation

$$P =_d Y\delta_\theta + (1 - Y)P. \tag{4.5}$$

We say that a random measure P that is independent of (Y, θ) is a solution to equation (4.5) if for every measurable partition $\{A_1, \ldots, A_k\}$ of the sample space the random vectors obtained by evaluating the random measures on its left and right sides are equal in distribution in $\mathbb{R}^k$.

Lemma 4.7 *For given independent* $\theta \sim \bar{\alpha}$ *and* $Y \sim \mathrm{Be}\big(1, |\alpha|\big)$*, the Dirichlet process* $\mathrm{DP}(\alpha)$ *is the unique solution of the distributional equation* (4.5).

Proof For a given measurable partition $\{A_1, \ldots, A_k\}$, the equation requires that $Q := \big(P(A_1), \ldots, P(A_k)\big)$ has the same distribution as the vector $YN + (1 - Y)Q$, for $N \sim \mathrm{MN}(1; \bar{\alpha}(A_1), \ldots, \bar{\alpha}(A_k))$ and (Y, N) independent of Q.

We first show that the solution is unique in distribution. Let (Y_n, N_n) be a sequence of i.i.d. copies of (Y, N), and for two solutions Q and Q' that are independent of this sequence and suitably defined on the same probability space, set $Q_0 = Q$, $Q_0' = Q'$, and recursively define $Q_n = Y_nN_n + (1 - Y_n)Q_{n-1}$, $Q_n' = Y_nN_n + (1 - Y_n)Q_{n-1}'$, for $n \in \mathbb{N}$. Then every

Q_n is distributed as Q and every Q'_n is distributed as Q', because each of them satisfies the distributional equation. Also

$$\|Q_n - Q'_n\| = |1 - Y_n| \, \|Q_{n-1} - Q'_{n-1}\| = \prod_{i=1}^{n} |1 - Y_i| \, \|Q - Q'\| \to 0$$

with probability 1, since the Y_i are i.i.d. and are in $(0, 1)$ with probability one. This forces the distributions of Q and Q' to agree.

To prove that the Dirichlet process is a solution let $W_0, W_1, \ldots, W_k \overset{\text{ind}}{\sim} \text{Ga}(\alpha_i, 1)$, $i = 0, 1, \ldots, k$, where $\alpha_0 = 1$. Then by Proposition 4.3 the vector (W_0, W), for $W = \sum_{i=1}^{k} W_i$, is independent of the vector $Q := (W_1/W, \ldots, W_k/W) \sim \text{Dir}(k, \alpha_1, \ldots, \alpha_k)$. Furthermore, $Y := W_0/(W_0+W) \sim \text{Be}(1, |\alpha|)$ and $(Y, (1-Y)Q) \sim \text{Dir}(k+1; 1, \alpha_1, \ldots, \alpha_k)$. Thus for any $i = 1, \ldots, k$, merging the 0th cell with the ith, we obtain from Proposition 4.3 that, with e_i the ith unit vector,

$$Ye_i + (1 - Y)Q \sim \text{Dir}(k; \alpha + e_i), \quad i = 1, \ldots, k. \tag{4.6}$$

This gives the conditional distribution of the vector $YN + (1 - Y)Q$ given $N = e_i$. It follows that $YN + (1 - Y)Q$ given N possesses a $\text{Dir}(k; \alpha + N)$-distribution, just as p given N in Proposition 4.8. Because also the marginal distributions of N in the two cases are the same, so must be the marginal distributions of $YN + (1 - Y)N$ and p, where the latter is $p \sim \text{Dir}(k; \alpha)$. □

Proposition 4.8 (Conjugacy) *If $p \sim \text{Dir}(k; \alpha)$ and $N|\, p \sim \text{MN}(n, k; p)$, then $p|\, N \sim \text{Dir}(k; \alpha + N)$.*

Proof If some coordinate α_i of α is zero, then the corresponding coordinate p_i of p is zero with probability one, and hence so is the coordinate N_i of N. After removing these coordinates we can work with densities. The product of the Dirichlet density and the multinomial likelihood is proportional to

$$p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1} \times p_1^{N_1} \cdots p_k^{N_k} = p_1^{\alpha_1 + N_1 - 1} \cdots p_k^{\alpha_k + N_k - 1}.$$

This is proportional to the density of the $\text{Dir}(k; \alpha_1 + N_1, \ldots, \alpha_k + N_k)$-distribution. □

4.3.1 Self-similarity

For a measure P and measurable set B, let $P_{|B}$ stand for the restriction measure $P_{|B}(A) = P(A \cap B)$, and P_B for the conditional measure $P_B(A) = P(A|\, B)$, for B with $P(B) > 0$.

Theorem 4.9 (Self-similarity) *If $P \sim \text{DP}(\alpha)$, then $P_B \sim \text{DP}(\alpha_{|B})$, and the variable and processes $P(B)$, $\big(P_B(A)\colon A \in \mathscr{X}\big)$ and $\big(P_{B^c}(A)\colon A \in \mathscr{X}\big)$ are mutually independent, for any $B \in \mathscr{X}$ such that $\alpha(B) > 0$.*

Proof Because $P(B) \sim \text{Be}\big(\alpha(B), \alpha(B^c)\big)$ the condition that $\alpha(B) > 0$ implies that $P(B) > 0$ a.s., so that the conditional probabilities given B are well defined.

For given partitions $A_1, \ldots, A_r$ of B and $C_1, \ldots, C_s$ of B^c, the vector

$$X := \big(P(A_1), \ldots, P(A_r), P(C_1), \ldots, P(C_s)\big)$$

possesses a Dirichlet distribution $\mathrm{Dir}\big(r+s; \alpha(A_1), \ldots, \alpha(A_r), \alpha(C_1), \ldots, \alpha(C_s)\big)$. By Proposition 4.3 the four variables or vectors

$$Z_1 := \sum_{i=1}^{r} X_i, \quad Z_2 := \sum_{i=r+1}^{r+s} X_i, \quad (X_1/Z_1, \ldots, X_r/Z_1), \quad (X_{r+1}/Z_2, \ldots, Z_{r+s}/Z_2)$$

are mutually independent, and the latter two vectors have Dirichlet distributions with the restrictions of the original parameters. These are precisely the variables $P(B)$, $P(B^c)$, and vectors with coordinates $P_B(A_i)$ and $P_{B^c}(C_i)$. □

Theorem 4.9 shows that the Dirichlet process "localized" by conditioning to a set B is again a Dirichlet process, with base measure the restriction of the original base measure. Furthermore, processes at disjoint localities are independent of each other, and also independent of the "macro level" variable $P(B)$. Within any given locality, mass is further distributed according to a Dirichlet process, independent of what happens to the "outside world". This property may be expressed by saying that locally a Dirichlet process is like itself; in other words it is *self similar*.

Exercises

4.1 Show that if $P \sim \mathrm{DP}(\alpha)$ and $\psi\colon \mathfrak{X} \to \mathfrak{Y}$ is a measurable mapping, then $P \circ \psi^{-1} \sim \mathrm{DP}(\beta)$, for $\beta = \alpha \circ \psi^{-1}$.

4.2 Show that if $P \sim \mathrm{DP}(\alpha)$, then $\mathrm{E} \int \psi\, dP = \int \psi\, d\bar{\alpha}$, and $\mathrm{var} \int \psi\, dP = \int (\psi - \int \psi\, d\bar{\alpha})^2\, d\bar{\alpha}/(1+|\alpha|)$, for any measurable function ψ for which the integrals make sense (e.g. bounded). [Hint: proof this first for $\psi = 1_A$.]

4.3 Let $0 = T_0 < T_1 < T_2 < \cdots$ be the events of a standard Poisson process and let $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} \bar{\alpha}$ and independent of $(T_1, T_2, \ldots)$. Show that

$$P = \sum_{k=1}^{\infty} (e^{-T_{k-1}} - e^{-T_k}) \delta_{\theta_k}$$

follows a Dirichlet process $\mathrm{DP}(\bar{\alpha})$. How can we change the prior precision to $M \neq 1$?

4.4 Let $F \sim \mathrm{DP}(MG)$ be a Dirichlet process on $\mathfrak{X} = \mathbb{R}$, for a constant $M > 0$ and probability distribution G, identified by its cumulative distribution function $x \mapsto G(x)$. So F can be viewed as a random cumulative distribution function. Define its median as any value m_F such that $F(m_F-) \leq 1/2 \leq F(m_F)$. Show that

$$\Pr\big(m_F \leq x\big) = \int_{1/2}^{1} \beta\big(u, MG(x), M(1-G(x))\big)\, du,$$

where $\beta(\cdot, \alpha, \beta)$ is the density of the Beta-distribution.

4.5 Simulate and plot the cumulative distribution function of the Dirichlet processes $F \sim \mathrm{DP}(\Phi)$, $F \sim \mathrm{DP}(0.1\Phi)$, and $F \sim \mathrm{DP}(10\Phi)$. Do the same with the Cauchy base measure. [Suggestion use Sethuraman's presentation. Cut the series at an appropriate point.]

6

Dirichlet process (2)

Consider observations $X_1, X_2, \ldots, X_n$ sampled independently from a distribution P that was drawn from a Dirichlet prior distribution, i.e.

$$P \sim \mathrm{DP}(\alpha), \qquad X_1, X_2, \ldots \,|\, P \overset{\text{iid}}{\sim} P.$$

By an abuse of language, which we shall follow, such observations are often termed a *sample from the Dirichlet process*.

6.1 Posterior distribution

One of the most remarkable properties of the Dirichlet process prior is that the posterior distribution is again Dirichlet.

Theorem 6.1 (Conjugacy) *The posterior distribution of P given an i.i.d. sample $X_1, \ldots, X_n$ from a $\mathrm{DP}(\alpha)$ process is $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{X_i})$.*

Proof Because the Dirichlet process is tail free for any sequence of partitions by Theorem 4.9, and a given measurable partition $\{A_1, \ldots, A_k\}$ of $\mathfrak{X}$ can be viewed as part of a sequence of successive binary partitions, the posterior distribution of the random vector $\big(P(A_1), \ldots, P(A_k)\big)$ given $X_1, \ldots, X_n$ is the same as the posterior distribution of this vector given the vector $N = (N_1, \ldots, N_k)$ of cell counts, defined by $N_j = \#(1 \le i \le n\colon X_i \in A_j)$. Given P the vector N possesses a multinomial distribution with parameter $\big(P(A_1), \ldots, P(A_k)\big)$, which has a $\mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big)$ prior distribution. The posterior distribution can be obtained using Bayes' rule applied to these finite-dimensional vectors, as in Proposition 4.8. □

Theorem 6.1 can be remembered as the updating rule $\alpha \mapsto \alpha + \sum_{i=1}^n \delta_{X_i}$ for the base measure of the Dirichlet distribution. In terms of the parameterization $\alpha \leftrightarrow \big(M = |\alpha|, \bar\alpha\big)$ of the base measure, this rule takes the form

$$M \mapsto M + n \quad \text{and} \quad \bar\alpha \mapsto \frac{M}{M+n}\bar\alpha + \frac{n}{M+n}\mathbb{P}_n, \tag{6.1}$$

where $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ is the *empirical distribution* of $X_1, \ldots, X_n$. Because the mean measure of a Dirichlet process is the normalized base measure, we see that

$$\mathrm{E}\big(P(A)|\, X_1, \ldots, X_n\big) = \frac{|\alpha|}{|\alpha|+n}\bar\alpha(A) + \frac{n}{|\alpha|+n}\mathbb{P}_n(A). \tag{6.2}$$

Thus the posterior mean (the "Bayes estimator" of P) is a convex combination of the prior

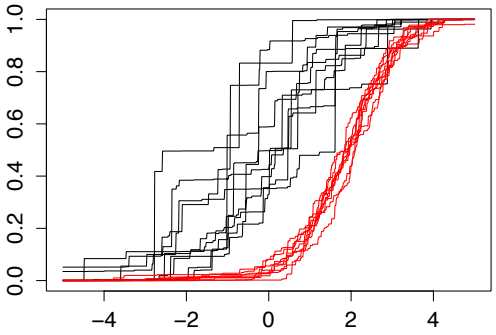

Figure 6.1 Cumulative distribution functions of 10 draws (black) from the Dirichlet process with base measure $5N(0, 2)$, and of 10 draws (red) from the realization of the posterior distribution based on a sample of size 100 from a $N(2, 1)$ distribution.

mean $\bar{\alpha}$ and the empirical distribution, with weights $M/(M + n)$ and $n/(M + n)$, respectively. For a given sample it is close to the prior mean if M is large, and close to the empirical distribution (which is based only on the data) if M is small. Thus M determines the extent to which the prior controls the posterior mean — a Dirichlet process prior with precision M contributes information equivalent to a sample of size M (although M is not restricted to integer values). This invites to view M as the *prior sample size*, or the "number of pre-experiment samples". In this interpretation the sum $M + n$ is the "posterior sample size".

For a fixed prior (i.e. fixed M) the posterior mean (6.2) behaves asymptotically as $n \to \infty$ like the empirical distribution $\mathbb{P}_n$ to the order $O(n^{-1})$, a.s.. Thus it possesses the same asymptotic properties as the empirical distribution. In particular, if $X_1, X_2, \ldots$ are sampled from a "true distribution"P_0, then the posterior mean will tend a.s. to P_0.

In addition the full posterior distribution will contract to its mean, whenever the posterior sample size tends to infinity. Indeed, by combining Theorem 6.1 and the formula for the variance of a Dirichlet variable, we see, for $\tilde{\mathbb{P}}_n$ the posterior mean (6.2),

$$\operatorname{var}\bigl(P(A)\bigm| X_1, \ldots, X_n\bigr) = \frac{\tilde{\mathbb{P}}_n(A)\tilde{\mathbb{P}}_n(A^c)}{1 + M + n} \leq \frac{1}{4(1 + M + n)}. \tag{6.3}$$

Consequently, if the data are sampled from a true distribution P_0, then the posterior distribution of P converges weakly to the measure degenerate at P_0. Formally, we can can state this as follows.

Corollary 6.2 *For any set A the posterior distribution of $P(A)$ given a random sample $X_1, \ldots, X_n$ of size n from a Dirichlet process tends in distribution to $\delta_{P_0(A)}$ as $n \to \infty$ for a.e. sequence $X_1, X_2, \ldots$ generated independently from a given distribution P_0.*

6.2 Predictive distribution

The joint distribution of a sequence $X_1, X_2, \ldots$ generated from a Dirichlet process, has a complicated structure, but can be conveniently described by its sequence of *predictive distributions*: the laws of X_1, $X_2 | X_1$, $X_3 | X_1, X_2$, etc.

Because $\Pr(X_1 \in A) = \mathrm{E}\Pr(X_1 \in A | P) = \mathrm{E}P(A) = \bar{\alpha}(A)$, the marginal distribution of X_1 is $\bar{\alpha}$.

Because $X_2 | (P, X_1) \sim P$ and $P | X_1 \sim \mathrm{DP}(\alpha + \delta_{X_1})$, we can apply the same reasoning again, but now conditionally given X_1, to see that $X_2 | X_1$ follows the normalization of $\alpha + \delta_{X_1}$. This is a mixture of α and δ_{X_1} with weights $|\alpha|/(|\alpha| + 1)$ and $1/(|\alpha| + 1)$, respectively.

Repeating this argument, using that $P | X_1, \ldots, X_{i-1} \sim \mathrm{DP}(\alpha + \sum_{j=1}^{i-1} \delta_{X_j})$, we find that

$$X_i | X_1, \ldots, X_{i-1} \sim \begin{cases} \delta_{X_1}, & \text{with probability } \frac{1}{|\alpha|+i-1}, \\ \vdots & \vdots \\ \delta_{X_{i-1}}, & \text{with probability } \frac{1}{|\alpha|+i-1}, \\ \bar{\alpha}, & \text{with probability } \frac{|\alpha|}{|\alpha|+i-1}. \end{cases} \tag{6.4}$$

Being a mixture of a product of identical distributions, the joint distribution of $X_1, X_2, \ldots$ is exchangeable, so re-labeling does not affect the structure of (6.4).

The recipe (6.4) is called the *generalized Polya urn scheme*, and can be viewed as a continuous analog of the familiar Polya urn scheme. Consider balls which can carry a continuum $\mathfrak{X}$ of "colors". Initially the "number of balls" is $M = |\alpha|$, which may be any positive number, and the colors are distributed according to $\bar{\alpha}$. We draw a ball from the collection, observe its color X_1, and return it to the urn along with an additional ball of the same color. The total number of balls is now $M + 1$, and the colors are distributed according to $(M\bar{\alpha} + \delta_{X_1})/(M + 1)$. We draw a ball from this updated urn, observe its color X_2, and return it to the urn along with an additional ball of the same color. The probability of picking up the ball that was added after the first draw is $1/(M + 1)$, in which case $X_2 = X_1$; otherwise, with probability $M/(M + 1)$, we make a fresh draw from the original urn. This process continues indefinitely, leading to the conditional distributions in (6.4).

6.3 Number of distinct values

It is clear from the preceding description that a realization of $(X_1, \ldots, X_n)$ will have ties (equal values) with positive probability. For instance, with probability at least

$$\frac{1}{M+1}\frac{2}{M+2}\cdots\frac{n-1}{M+n-1}$$

all X_i will even be identical. For simplicity assume that the base measure α is non-atomic, so that the ith value X_i in the Polya scheme (6.4) is different from the previous $X_1, \ldots, X_{i-1}$ if it is drawn from $\bar{\alpha}$. The vector $(X_1, \ldots, X_n)$ then induces a random partition $\{\mathcal{P}_1, \ldots, \mathcal{P}_{K_n}\}$ of the set of indices $\{1, 2, \ldots, n\}$, corresponding to the ties, and given this partition the K_n distinct values are an i.i.d. sample from $\bar{\alpha}$.

The number of distinct values is remarkably small.

For $i \in \mathbb{N}$ define $D_i = 1$ if the ith observation X_i is a "new value", i.e. if $X_i \notin \{X_1, \ldots, X_{i-1}\}$, and set $D_i = 0$ otherwise. Then $K_n = \sum_{i=1}^n D_i$ is the number of distinct values among the first n observations.

Proposition 6.3 *If the base measure α is nonatomic and of strength $|\alpha| = M$, then the variables $D_1, D_2, \ldots$ are independent Bernoulli variables with success probabilities $\Pr(D_i = 1) = M/(M + i - 1)$. Consequently, for fixed M, as $n \to \infty$,*

(i) $\mathrm{E}(K_n) \asymp M \log n \asymp \mathrm{var}(K_n)$.
(ii) $K_n / \log n \to M$, *a.s.*
(iii) $(K_n - \mathrm{E}K_n)/\operatorname{sd}(K_n) \to_d \mathrm{Nor}(0, 1)$.

Proof The first assertion follows, because given $X_1, \ldots, X_{i-1}$ the variable X_i is "new" if and only if it is drawn from $\bar{\alpha}$, which happens with probability $M/(M + i - 1)$. Then assertion (i) can be derived from the exact formulas

$$\mathrm{E}(K_n) = \sum_{i=1}^n \frac{M}{M + i - 1}, \qquad \mathrm{var}(K_n) = \sum_{i=1}^n \frac{M(i-1)}{(M + i - 1)^2}.$$

Furthermore, assertion (ii) follows from Kolmogorov's strong law of large numbers for independent variables, since

$$\sum_{i=1}^\infty \frac{\mathrm{var}(D_i)}{(\log i)^2} = \sum_{i=1}^\infty \frac{M(i-1)}{(M + i - 1)^2 (\log i)^2} < \infty.$$

Next (iii) is a consequence of the Lindeberg central limit theorem. □

Thus the number of distinct values in a (large) sample from a distribution taken from a fixed Dirichlet prior is logarithmic in the sample size. Furthermore, the fluctuations of this number around its mean are of the order $\sqrt{\log n}$.

The following proposition gives the distribution of the partition $\{\mathcal{P}_1, \ldots, \mathcal{P}_{K_n}\}$ induced by $(X_1, \ldots, X_n)$. (This can be more formally defined as the equivalence classes under the relation $i \equiv j$ iff $X_i = X_j$.)

Proposition 6.4 *A random sample $X_1, \ldots, X_n$ from a Dirichlet process with nonatomic base measure of strength $|\alpha| = M$ induces a given partition of $\{1, 2, \ldots, n\}$ into k sets of sizes $n_1, \ldots, n_k$ with probability equal to*

$$\frac{M^k \Gamma(M) \prod_{j=1}^k \Gamma(n_j)}{\Gamma(M + n)}. \tag{6.5}$$

Proof By exchangeability the probability depends on the sizes of the partitioning sets only. The probability that the partitioning set of size n_1 consists of the first n_1 variables, the one of size n_2 of the next n_2 variables, etc. can be obtained by multiplying the appropriate conditional probabilities for the consecutive draws in the Polya urn scheme in their natural order of occurrence. For $r_j = \sum_{l=1}^j n_l$, it is given by

$$\begin{aligned} &\frac{M}{M} \frac{1}{M+1} \frac{2}{M+2} \cdots \frac{n_1 - 1}{M + n_1 - 1} \frac{M}{M + n_1} \frac{1}{M + n_1 + 1} \times \cdots \\ &\qquad \cdots \times \frac{M}{M + r_{k-1}} \frac{1}{M + r_{k-1} + 1} \cdots \frac{n_k - 1}{M + r_{k-1} + n_k - 1}. \end{aligned}$$

This can be rewritten as in the proposition. □

6.4 Mixtures of Dirichlet processes

Application of the Dirichlet prior requires a choice of a base measure α. It is often reasonable to choose the center measure $\bar{\alpha}$ from a specific family such as the normal family, but then the parameters of the family must still be specified. It is natural to give these a further prior. Similarly, one may put a prior on the precision parameter $|\alpha|$.

For a base measure α_ξ that depends on a parameter ξ the Bayesian model then consists of the hierarchy

$$X_1, \ldots, X_n | P, \xi \overset{\text{iid}}{\sim} P, \qquad P | \xi \sim \mathrm{DP}(\alpha_\xi), \qquad \xi \sim \pi. \tag{6.6}$$

We denote the induced (marginal) prior on P by $\mathrm{MDP}(\alpha_\xi, \xi \sim \pi)$. Many properties of this *mixture Dirichlet prior* follow immediately from those of a Dirichlet process. For instance, any P following an MDP is almost surely discrete. However, unlike a Dirichlet process, an MDP is not tail free.

Given ξ we can use the posterior updating rule for the ordinary Dirichlet process, and obtain that

$$P | \xi, X_1, \ldots, X_n \sim \mathrm{DP}(\alpha_\xi + n\mathbb{P}_n).$$

To obtain the posterior distribution of P given $X_1, \ldots, X_n$, we need to mix this over ξ relative to its posterior distribution given $X_1, \ldots, X_n$. By Bayes's theorem the latter has density proportional to

$$\xi \mapsto \pi(\xi)\, p(X_1, \ldots, X_n | \xi). \tag{6.7}$$

Here the marginal density of $X_1, \ldots, X_n$ given ξ (the second factor) is described by the generalized Polya urn scheme (6.4) with α_ξ instead of α. In general, this has a somewhat complicated structure due to the ties between the observations. However, for a posterior calculation we condition on the observed data $X_1, \ldots, X_n$, and know the partition that they generate. Given this information the density takes a simple form. For instance, if the observations are distinct (which happens with probability one if the observations actually follow a continuous distribution), then the Polya urn scheme must have simply generated a random sample from the normalized base measure $\bar{\alpha}_\xi$, in which case the preceding display becomes

$$\pi(\xi) \prod_{i=1}^n d\alpha_\xi(X_i) \prod_{i=1}^n \frac{1}{|\alpha_\xi| + i - 1},$$

for $d\alpha_{xi}$ a density of α_ξ. Further calculations depend on the specific family and its parameterization.

Typically the precision parameter M and center measure G in $\alpha = MG$ will be modelled as independent under the prior. The posterior calculation then factorizes in these two parameters. To see this, consider the following scheme to generate the parameters and observations:

(i) Generate M from its prior.

(ii) Given M generate a random partition $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_{K_n}\}$ according to the distribution given in Proposition 6.4.
(iii) Generate G from its prior, independently of $(M, \mathcal{P})$.
(iv) Given $(\mathcal{P}, G)$ generate a random sample of size K_n from G, independently of M, and set X_i with $i \in \mathcal{P}_j$ equal to the jth value in this sample.

By the description of the Polya urn scheme this indeed gives a sample $X_1, \ldots, X_n$ from the mixture of Dirichlet processes $\mathrm{MDP}(MG, M \sim \pi, G \sim \pi)$. We may now formally write the density of $(M, \mathcal{P}, G, X_1, \ldots, X_n)$ in the form, with π abusively denoting prior densities for both M and G and p conditional densities of observed quantities,

$$\pi(M)\, p(\mathcal{P}|\, M)\, \pi(G)\, p(X_1, \ldots, X_n|\, G, \mathcal{P}).$$

Since this factorizes in terms involving M and G, these parameters are also independent under the posterior distribution, and the computation of their posterior distributions can be separated.

The term involving M depends on the data through K_n only (the latter variable is *sufficient* for M). Indeed, by Proposition 6.4 it is proportional to,

$$M \mapsto \pi(M)\frac{M^{K_n}\Gamma(M)}{\Gamma(M+n)} \propto \pi(M) M^{K_n} \int_0^1 \eta^{M-1}(1-\eta)^{n-1}\, d\eta.$$

Rather than by (numerically) integrating this expression, the posterior density is typically computed by simulation. xsSuppose that $M \sim \mathrm{Ga}(a, b)$ a priori, and consider a fictitious random vector (M, η) with $0 \le \eta \le 1$ and joint (Lebesgue) density proportional to

$$\pi(M) M^{K_n} \eta^{M-1}(1-\eta)^{n-1} \propto M^{a+K_n-1} e^{-M(b-\log\eta)} \eta^{-1}(1-\eta)^{n-1}.$$

Then by the preceding display the marginal density of M is equal to its posterior density (given K_n, which is fixed for the calculation). Thus simulating from the distribution of (M, η) and dropping η simulates M from its posterior distribution. The conditional distributions are given by

$$M|\, \eta, K_n \sim \mathrm{Ga}(a + K_n, b - \log\eta), \qquad \eta|\, M, K_n \sim \mathrm{Be}(M, n). \tag{6.8}$$

We can use these in a *Gibbs sampling scheme*: given an arbitrary starting value η_0 we generate a sequence $M_1, \eta_1, M_2, \eta_2, M_3, \ldots$, by repeatedly generating M from its conditional distribution given (η, K_n) and η from its conditional distribution given (M, K_n), each time setting the conditioning variable (η or M) equal to its last value. After an initial *burn-in* the values $M_k, M_{k+1}, \ldots$ will be approximately from the posterior distribution of M given K_n.

6.5 Dirichlet process mixtures

Because the Dirichlet process is discrete, it is a useless prior when we wish to estimate a density. This can be remedied by convolving it with a kernel. For each θ in a parameter set Θ let $x \mapsto \psi(x, \theta)$ be a probability density function, measurable in its two arguments. For a measure F on Θ define a *mixture density* by

$$p_F(x) = \int \psi(x, \theta)\, dF(\theta).$$

By equipping F with a prior, we obtain a prior on densities. Densities p_F with F following a Dirichlet process prior are known as *Dirichlet mixtures*. If the kernel also depends on an additional parameter $\varphi \in \Phi$, giving mixtures $p_{F,\varphi}(x) = \int \psi(x,\theta,\varphi)\, dF(\theta)$, it is more appropriate to call the result a "mixture of Dirichlet mixture", but the nomenclature Dirichlet mixture even for this case seems more convenient.

In this section we discuss methods of posterior computation for these mixtures. For $x \mapsto \psi(x;\theta,\varphi)$ a probability density function (relative to a given σ-finite dominating measure ν), consider

$$X_i \overset{\text{iid}}{\sim} p_{F,\varphi}(x) = \int \psi(x;\theta,\varphi)\, dF(\theta), \qquad i = 1,\ldots,n, \tag{6.9}$$

We equip F and φ with independent priors $F \sim \mathrm{DP}(\alpha)$ and $\varphi \sim \pi$. The resulting model can be equivalently written in terms of n *latent variables* $\theta_1,\ldots,\theta_n$ as

$$X_i|\, \theta_i,\varphi,F \overset{\text{ind}}{\sim} \psi(\cdot;\theta_i,\varphi), \qquad \theta_i|\, F,\varphi \overset{\text{iid}}{\sim} F, \qquad F \sim \mathrm{DP}(\alpha), \qquad \varphi \sim \pi. \tag{6.10}$$

The posterior distribution of any object of interest can be described in terms of the posterior distribution of (F,φ) given $X_1,\ldots,X_n$. The latent variables $\theta_1,\ldots,\theta_n$ help to make the description simpler, since $F|\, \theta_1,\ldots,\theta_n \sim \mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{\theta_i})$, and given $\theta_1,\ldots,\theta_n$, the observations $X_1,\ldots,X_n$ and F are independent. Hence the conditional distribution of F given $\theta_1,\ldots,\theta_n,X_1,\ldots,X_n$ is free of the observations. In particular, for any measurable function ψ, in view of Exercise 4.1,

$$\mathrm{E}\Big(\int \psi\, dF|\, \varphi,\theta_1,\ldots,\theta_n,X_1,\ldots,X_n\Big) = \frac{1}{|\alpha|+n}\Big[\int \psi\, d\alpha + \sum_{j=1}^n \psi(\theta_j)\Big]. \tag{6.11}$$

The advantage of this representation is that the infinite-dimensional parameter F has been eliminated. To compute the posterior expectation it now suffies to average out the right hand side of (6.11) with respect to the posterior distribution of $(\theta_1,\ldots,\theta_n)$, and that of φ.

Example 6.5 (Density estimation) The choice $\psi(\theta) = \psi(x,\theta,\varphi)$ in (6.11) gives the density $\int \psi(x,\theta,\varphi)\, dF(\theta) = p_{F,\varphi}(x)$. Thus the posterior mean density satisfies

$$\mathrm{E}\big(p_{F,\varphi}(x)|\, \varphi,X_1,\ldots,X_n\big) = \frac{1}{|\alpha|+n}\Big[\int \psi(x;\theta,\varphi)\, d\alpha(\theta) + \mathrm{E}\Big(\sum_{j=1}^n \psi(x;\theta_j,\varphi)|\, X_1,\ldots,X_n\Big)\Big].$$

This consists of a part attributable to the prior and a part due to observations. In practice the latter is computed by simulating many samples $(\theta_1\ldots,\theta_n)$ from its posterior distribution.

Analytical formulas for the posterior distribution corresponding to a Dirichlet mixture are possible, but too unwieldy for practical implementation. Computation is typically done by simulation. The next theorem explains a *Gibbs sampling scheme* to simulate from the posterior distribution of $(\theta_1,\ldots,\theta_n)$, based on a weighted generalized Polya urn scheme. Inclusion of a possible parameter φ and other hyperparameters is tackled in the next section.

A *Gibbs sampler* in general is a method for simulating from the joint distribution of a number of variables. It simply updates the variables one-by-one by simulating a new variable from its conditional distribution given the other variables. By repeating this indefinitely a sequence of vectors is created that after an initial "burn-in period" can be viewed as sampled

from the target distribution. More precisely, the sequence of vectors forms a Markov chain with the target distribution as its stationary distribution.

We use the subscript $-i$ to denote every index $j \neq i$, and $\theta_{-i} = (\theta_j \colon j \neq i)$.

Theorem 6.6 (Gibbs sampler) *The conditional posterior distribution of θ_i is given by:*

$$\theta_i | \, \theta_{-i}, \varphi, X_1, \ldots, X_n \sim \sum_{j \neq i} q_{i,j} \delta_{\theta_j} + q_{i,0} G_{b,i}, \tag{6.12}$$

where $(q_{i,j} \colon j \in \{0, 1, \ldots, n\} - \{i\})$ is the probability vector satisfying

$$q_{i,j} \propto \begin{cases} \psi(X_i; \theta_j, \varphi), & j \neq i, j \geq 1, \\ \int \psi(X_i; \theta, \varphi) \, d\alpha(\theta), & j = 0, \end{cases} \tag{6.13}$$

and $G_{b,i}$ is the "baseline posterior measure" given by

$$dG_{b,i}(\theta | \, \varphi, X_i) \propto \psi(X_i; \theta, \varphi) \, d\alpha(\theta). \tag{6.14}$$

Proof Since the parameter φ is fixed throughout, we suppress it from the notation. For measurable sets A and B,

$$\mathrm{E}\big(\mathbb{1}_A(X_i) \mathbb{1}_B(\theta_i) | \, \theta_{-i}, X_{-i}\big) = \mathrm{E}\Big(\mathrm{E}\big(\mathbb{1}_A(X_i) \mathbb{1}_B(\theta_i) | \, F, \theta_{-i}, X_{-i}\big) | \, \theta_{-i}, X_{-i}\Big).$$

Because (θ_i, X_i) is conditionally independent of (θ_{-i}, X_{-i}) given F, the inner conditional expectation is equal to $\mathrm{E}\big(\mathbb{1}_A(X_i) \mathbb{1}_B(\theta_i) | \, F\big) = \int\int \mathbb{1}_A(x) \mathbb{1}_B(\theta) \psi(x; \theta) \, d\mu(x) \, dF(\theta)$. In the outer layer of conditioning the variables X_{-i} are superfluous, by the conditional independence of F and X_{-i} given θ_{-i}. Therefore, by Exercise 4.1 the preceding display is equal to

$$\frac{1}{|\alpha| + n} \int\int \mathbb{1}_A(x) \mathbb{1}_B(\theta) \psi(x; \theta) \, d\mu(x) \, d\Big(\alpha + \sum_{j \neq i} \delta_{\theta_j}\Big)(\theta).$$

This determines the joint conditional distribution of (X_i, θ_i) given (θ_{-i}, X_{-i}). By Bayes's rule (applied to this joint law conditionally given (θ_{-i}, X_{-i})) we infer that

$$\Pr\big(\theta_i \in B | \, X_i, \theta_{-i}, X_{-i}\big) = \frac{\int_B \psi(X_i; \theta) \, d(\alpha + \sum_{j \neq i} \delta_{\theta_j})(\theta)}{\int \psi(X_i; \theta) \, d(\alpha + \sum_{j \neq i} \delta_{\theta_j})(\theta)}.$$

This in turn is equivalent to the assertion of the theorem. □

6.5.1 MCMC method

In this section we present an algorithm to simulate from the posterior distribution in the MDP model:

$$X_i | \, \theta_i, \varphi, M, \xi, F \overset{\text{ind}}{\sim} \psi(\cdot; \theta_i, \varphi), \quad \theta_i | \, F, \varphi, M, \xi \overset{\text{iid}}{\sim} F, \quad F | \, M, \xi \sim \mathrm{DP}(M, G_\xi),$$

where φ, M and ξ are independently generated hyperparameters. The basic algorithm uses the Gibbs sampling scheme of Theorem 6.6 to generate $\theta_1, \ldots, \theta_n$ given $X_1, \ldots, X_n$ in combination with the Gibbs sampler for the posterior distribution of M given in Section 6.4,

and/or additional Gibbs steps. The prior densities of the hyperparameters are denoted by a generic π.

Algorithm Generate samples by sequentially executing steps (i)–(iv) below:

(i) Given the observations and φ, M and ξ, update each θ_i sequentially using (6.12) inside a loop $i = 1, \ldots, n$.
(ii) Update $\varphi \sim p(\varphi|\, \theta_1, \ldots, \theta_n, X_1, \ldots, X_n) \propto \pi(\varphi) \prod_{i=1}^n \psi(X_i; \theta_i, \varphi)$.
(iii) Update $\xi \sim p(\xi|\, \theta_1, \ldots, \theta_n) \propto \pi(\xi) p(\theta_1, \ldots, \theta_n|\, \xi)$, where the marginal distribution of $(\theta_1, \ldots, \theta_n)$ is as in the Polya scheme (6.4).
(iv) Update M and next the auxiliary variable η using (6.8), for K_n the number of distinct values in $\{\theta_1, \ldots, \theta_n\}$.

Exercises

6.1 Let ψ be a given bounded measurable function. Show that if $P \sim \mathrm{DP}(\alpha)$ and $X_1, \ldots, X_n|\, P \overset{\text{iid}}{\sim} P$, then the posterior distribution of $\int \psi\, dP$ given $X_1, \ldots, X_n$ tends in distribution to a Dirac measure at $\int \psi\, dP_0$ for a.e. sequence $X_1, X_2, \ldots$ generated iid from P_0.

6.2 In the model (6.6) assume that the total mass $|\alpha_\xi|$ is bounded uniformly in ξ. Show that the posterior distribution of $P(A)$ is consistent.

6.3 Simulate and plot the cumulative distribution functions of realizations of some posterior Dirichlet processes. First use several fixed prior strengths. Second put a Gamma prior on the prior strength.