5

Dirichlet Process Mixtures

The discreteness of the Dirichlet process makes it useless as a prior for estimating a density. This can be remedied by convolving it with a kernel. Although it is rarely possible to characterize the resulting posterior distribution analytically, a variety of efficient and elegant algorithms allow us to approximate it numerically. These algorithms exploit the special properties of the Dirichlet process derived in the preceding chapter. The present chapter starts with a general introduction to Dirichlet process mixtures and next discusses a number of computational strategies.

5.1 Dirichlet Process Mixtures

For each θ in a parameter set Θ let $x \mapsto \psi(x, \theta)$ be a probability density function, measurable in its two arguments. Densities $p_F(x) = \int \psi(x, \theta)\, dF(\theta)$, with F equipped with a Dirichlet process prior, are known as *Dirichlet process 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 process mixture," but the name Dirichlet process mixture even for this case seems more convenient.

In this section we discuss a method of posterior computation for these mixtures. To include also other problems such as regression, spectral density estimation and so on, we allow the observations to be non-i.i.d. For $x \mapsto \psi_i(x; \theta, \varphi)$ probability density functions (relative to a given σ-finite dominating measure ν), consider

$$X_i \overset{\text{ind}}{\sim} p_{i,F,\varphi}(x) = \int \psi_i(x; \theta, \varphi)\, dF(\theta), \qquad i = 1, \ldots, n. \tag{5.1}$$

We equip F and φ with independent priors $F \sim \mathrm{DP}(\alpha)$ and $\varphi \sim \pi$. (Independence of the observations is not crucial; later we also consider mixture models for estimating the transition density of a Markov chain.) The resulting model can be equivalently written in terms of n latent variables $\theta_1, \ldots, \theta_n$ as

$$X_i \,|\, \theta_i, \varphi, F \overset{\text{ind}}{\sim} \psi_i(\cdot; \theta_i, \varphi), \qquad \theta_i \,|\, F, \varphi \overset{\text{iid}}{\sim} F, \qquad F \sim \mathrm{DP}(\alpha), \qquad \varphi \sim \pi. \tag{5.2}$$

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 $X_1, \ldots, X_n$ are ancillary with respect to (F, φ), whence the conditional

distribution of F given $\theta_1, \dots, \theta_n, X_1, \dots, X_n$ is free of the observations. In particular, for any measurable function ψ,

$$\mathrm{E}\Big(\int \psi \, dF \,|\, \varphi, \theta_1, \dots, \theta_n, X_1, \dots, X_n\Big) = \frac{1}{|\alpha| + n}\Big[\int \psi \, d\alpha + \sum_{j=1}^{n} \psi(\theta_j)\Big]. \tag{5.3}$$

The advantage of this representation is that the infinite-dimensional parameter F has been eliminated. To compute the posterior expectation, it now suffices to average out the right-hand side with respect to the posterior distribution of $(\theta_1, \dots, \theta_n)$, and that of φ. Proposition 5.2 and Theorem 5.3 below give two methods for the first, an analytic one and one based on simulation, where the second one is the more practical one.

Example 5.1 (Density estimation) The choice $\psi(\theta) = \psi_i(x, \theta, \varphi)$ in (5.3) gives the Bayes estimate of the density $p_{i,F,\varphi}(x)$. This consists of a part attributable to the prior and a part due to the observations. We might ignore the prior part and view the conditional expectation

$$\frac{1}{n} \sum_{j=1}^{n} \int \psi_i(x, \theta_j, \varphi) \, d\Pi_n(\theta_j \,|\, X_1, \dots, X_n)$$

of the second part as a partial Bayesian estimator, which has the influence of the prior guess reduced.

Let $\mathscr{S}_n$ stand for the class of all partitions of the set $\{1, \dots, n\}$, and let W be the probability measure on $\mathscr{S}_n$ satisfying

$$W(\mathcal{S}) \propto \prod_{j=1}^{\#\mathcal{S}} (\#S_j - 1)! \int \prod_{l \in S_j} \psi_l(X_l; \theta) \, d\alpha(\theta), \qquad \mathcal{S} = \{S_1, \dots, S_k\}.$$

Proposition 5.2 *In the model* (5.2)*, for any nonnegative or integrable function* ψ*,*

$$\mathrm{E}\Big(\int \psi \, dF \,|\, \varphi, X_1, \dots, X_n\Big)$$
$$= \frac{1}{|\alpha| + n}\Bigg[\int \psi \, d\alpha + \sum_{\mathcal{S} \in \mathscr{S}_n} W(\mathcal{S}) \sum_{j=1}^{\#\mathcal{S}} \#S_j \frac{\int \psi(\theta) \prod_{l \in S_j} \psi_l(X_l; \theta, \varphi) \, d\alpha(\theta)}{\int \prod_{l \in S_j} \psi_l(X_l; \theta, \varphi) \, d\alpha(\theta)}\Bigg].$$

Proof It suffices to show that the conditional expectation of the second term on the right of (5.3) given $X_1, \dots, X_n$, agrees with the second term of the formula. As φ is fixed throughout, we drop it from the notation.

Because $\theta_1, \dots, \theta_n$ are a random sample from a DP(α)-distribution, their (marginal) prior distribution Π is described by Proposition 4.7. By Bayes's rule the posterior density of $(\theta_1, \dots, \theta_n)$ is proportional to $\prod_{i=1}^{n} \psi_i(X_i; \theta_i) d\Pi(\theta_1, \dots, \theta_n)$. Hence, for any measurable functions g_i,

$$\mathrm{E}\Big(\prod_{i=1}^{n} g_i(\theta_i)\,|\,X_1,\dots,X_n\Big) = \frac{\int\cdots\int \prod_{i=1}^n g_i(\theta_i)\prod_{i=1}^n \psi_i(X_i;\theta_i)d\Pi(\theta_1,\dots,\theta_n)}{\int\cdots\int \prod_{i=1}^n \psi_i(X_i;\theta_i)d\Pi(\theta_1,\dots,\theta_n)}$$
$$= \frac{\sum_{\mathcal{S}\in\mathscr{S}_n}\prod_{j=1}^{\#\mathcal{S}}(\#S_j-1)!\int\prod_{l\in S_j}(g_l(\theta_l)\psi_l(X_l;\theta))\,d\alpha(\theta)}{\sum_{\mathcal{S}\in\mathscr{S}_n}\prod_{j=1}^{\#\mathcal{S}}(\#S_j-1)!\int\prod_{l\in S_j}\psi_l(X_l;\theta)\,d\alpha(\theta)},$$

by Proposition 4.7. For $g_i = \psi$ and $g_j = 1$ for $j \neq i$, we can rewrite this in the form

$$\mathrm{E}(\psi(\theta_i)\,|\,X_1,\dots,X_n) = \sum_{\mathcal{S}\in\mathscr{S}_n} W(\mathcal{S})\frac{\int\prod_{l\in S(i)}\psi(\theta)\psi_l(X_l;\theta)\,d\alpha(\theta)}{\int\prod_{l\in S(i)}\psi_l(X_l;\theta)\,d\alpha(\theta)},$$

where $S(i)$ is the partitioning set in $\mathcal{S} = \{S_1,\dots,S_k\}$ that contains i. When summing this over $i = 1,\dots,n$, we obtain $\#\mathcal{S}$ different values, with multiplicities $\#S_1,\dots,\#S_k$. □

Unfortunately, the analytic formula of Proposition 5.2 is of limited practical importance because it involves a sum with a large number of terms (cf. Problem 4.17). In practice, we use a simulation technique: we draw repeatedly from the posterior distribution of $(\theta_1,\dots,\theta_n,\varphi)$, evaluate the right-hand side of (5.3) for each realization, and average out over many realizations.

The next theorem explains a Gibbs sampling scheme to simulate from the posterior distribution of $(\theta_1,\dots,\theta_n)$, based on a weighted generalized Pólya urn scheme. Inclusion of a possible parameter φ and other hyperparameters is tackled in the next section.

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

Theorem 5.3 (Gibbs sampler) *In the model* (5.2) *the conditional posterior distribution of* θ_i *is given by*

$$\theta_i\,|\,\theta_{-i},\varphi,X_1,\dots,X_n \sim \sum_{j\neq i} q_{i,j}\delta_{\theta_j} + q_{i,0}G_{b,i}, \tag{5.4}$$

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

$$q_{i,j} \propto \begin{cases}\psi_i(X_i;\theta_j,\varphi), & j\neq i,\ j\geq 1,\\ \int\psi_i(X_i;\theta,\varphi)\,d\alpha(\theta), & j=0,\end{cases} \tag{5.5}$$

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

$$dG_{b,i}(\theta\,|\,\varphi,X_i) \propto \psi_i(X_i;\theta,\varphi)\,d\alpha(\theta). \tag{5.6}$$

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

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

Because (θ_i,X_i) is conditionally independent of (θ_{-i},X_{-i}) given F, the inner conditional expectation is equal to $\mathrm{E}(\mathbb{1}_A(X_i)\mathbb{1}_B(\theta_i)\,|\,F) = \int\int\mathbb{1}_A(x)\mathbb{1}_B(\theta)\psi_i(x;\theta)\,d\mu(x)\,dF(\theta)$. Upon inserting this expression, in which F is the only variable, in the display we see that 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 Proposition 4.3 the preceding display is equal to

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

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

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

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

Remark 5.4 The posterior distribution of F is actually a mixture of Dirichlet processes: for Π^* described in (5.4),

$$F \,|\, X_1, \ldots, X_n, \varphi, \alpha \sim \mathrm{MDP}\Big(\alpha + \sum_{i=1}^n \delta_{\theta_i}, (\theta_1, \ldots, \theta_n) \sim \Pi^*\Big).$$

5.2 MCMC Methods

In this section we present five algorithms to simulate from the posterior distribution in the Dirichlet process mixture model:

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

where φ, M and ξ are independently generated hyperparameters. The basic algorithm uses the Gibbs sampling scheme of Theorem 5.3 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 4.5, and/or additional Gibbs steps. The prior densities of the hyperparameters are denoted by a generic π.

Algorithm 1 Generate samples by sequentially executing steps (i)–(iv) below:

(i) Given the observations and φ, M and ξ, update each θ_i sequentially using (5.4) inside a loop $i = 1, \ldots, n$.
(ii) Update $\varphi \sim p(\varphi \,|\, \theta_1, \ldots, \theta_n, X_1, \ldots, X_n) \propto \pi(\varphi) \prod_{i=1}^n \psi_i(X_i;\theta_i,\varphi)$.
(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 Pólya scheme (4.13).
(iv) Update M and next the auxiliary variable η using (4.30), for K_n the number of distinct values in $\{\theta_1, \ldots, \theta_n\}$.

Because sampling from the generalized Pólya urn will rarely yield a new value (cf. Theorem 4.8), the θ-values will involve many ties and can be stored and updated more efficiently. Let $\{\mu_1, \ldots, \mu_k\}$ be the distinct values in $\{\theta_1, \ldots, \theta_n\}$, let $N_1, \ldots, N_k$ be their

multiplicities, and let $(s_1, \ldots, s_n)$ be pointers such that $s_i = s$ if $\theta_i = \mu_s$. Then $(\theta_1, \ldots, \theta_n)$ is in one-to-one correspondence with $(k, \mu_1, \ldots, \mu_k, s_1, \ldots, s_n)$. The pointers $(s_1, \ldots, s_n)$ determine the partition of $\{1, \ldots, n\}$ corresponding to the ties in $\theta_1, \ldots, \theta_n$. In the Pólya urn scheme for generating $\theta_1, \ldots, \theta_n$ the values μ_s of the ties are drawn independently from the center measure $\bar{\alpha}$. Given the pointers $s_1, \ldots, s_n$ they completely determine $\theta_1, \ldots, \theta_n$. Given the pointers $s_1, \ldots, s_n$, the observations $X_1, \ldots, X_n$ can be viewed as next generated as k independent samples $(X_i : s_i = s)$ of sizes $N_s = \#(i : s_i = s)$ from the densities $(x_i : s_i = s) \mapsto \prod_{i:s_i=s} \psi_i(x_i; \mu_s, \varphi)$. The conditional density of the tie values given the observations and still given the pointers follows by Bayes's rule. This leads to the following algorithm, which alternates between updating the pointers $(s_1, \ldots, s_n)$ and the values $\mu_1, \ldots, \mu_k$ of the ties.

The parameters φ, ξ and M are treated as in Algorithm 1 and therefore are dropped from the notation.

Algorithm 2 Perform the two steps:

(i) Update the full configuration vector $(s_1, \ldots, s_n)$ in a Gibbs loop according to $\mathrm{P}(s_i = s \,|\, s_{-i}, \text{rest}) = q^*_{i,s}$, for, with $N_{-i,s} = N_s - \mathbb{1}\{s = s_i\}$,

$$q^*_{i,s} \propto \begin{cases} N_{-i,s}\psi_i(X_i; \mu_s), & \text{if } s = s_j,\ j \neq i, \\ \int \psi_i(X_i; \theta)\, d\alpha(\theta), & \text{if } s = s_{i,\text{new}}. \end{cases}$$

Here $s_{i,\text{new}}$ is a value that does not appear in the current vector s_{-i}. If a new value $s = s_{i,\text{new}}$ is selected in the ith step, generate a corresponding μ_s from $G_{b,i}$ given in (5.6). Update k as the number of different elements in $s_1, \ldots, s_n$.

(ii) Update the full vector $(\mu_1, \ldots, \mu_k)$ by drawing the coordinates independently from the densities proportional to $\mu_s \mapsto \prod_{i:s_i=s} \psi_i(X_i; \mu_s)\, d\alpha(\mu_s)$.

In some applications, such as clustering problems, the values $\{\mu_1, \ldots, \mu_k\}$ may not be of interest. The following algorithm "integrates them out" and simulates only values of the configuration vector $(s_1, \ldots, s_n)$. A disadvantage of this algorithm is that it cannot update hyperparameters, but these could be estimated using an empirical Bayes approach.

Algorithm 3 Update the configuration vector $(s_1, \ldots, s_n)$ by

$$\mathrm{P}(s_i = s \,|\, s_{-i}, \text{rest}) \propto \frac{\int N_{-i,s}\psi_i(X_i; \theta) \prod_{j:s_j=s} \psi_j(X_j; \theta)\, d\alpha(\theta)}{\int \prod_{j\neq i:s_j=s} \psi_j(X_j; \theta)\, d\alpha(\theta)},$$

where an empty product means 1.

All three algorithms are feasible, even for large n, but only if the integrals in (5.6) and (5.5) are analytically computable and drawing from the baseline posterior distribution $G_{b,i}$ is straightforward. This virtually restricts to the case where the base measure α is conjugate to the kernels ψ_i. In many applications conjugacy arises naturally and is not too restrictive, especially if α is modeled flexibly using some hyperparameters. Some common cases are discussed in Section 5.5.

In the case where a conjugate base measure does not exist none of the three algorithms is suitable. Several algorithms have been developed for this situation. We present one such

method, the *no gaps algorithm*. The basis of reasoning in this algorithm is that the configuration indices $s_1, \ldots, s_n$ ought to form a consecutive sequence without a gap, and this constraint should be respected in all moves of the MCMC. With the same notations as before, the algorithm has the following steps:

Algorithm 4 ("No gaps" algorithm) For k_{-i} the number of distinct elements in the set $\{s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n\}$:

(a) For $i = 1, \ldots, n$, repeat the following steps.
   (i) If $s_i \neq s_j$ for all $j \neq i$, leave s_i unchanged with probability $k_{-i}/(k_{-i}+1)$ and with probability $1/(k_{-i}+1)$ assign it to the new label $k_{-i}+1$. Proceed to step (iii).
   (ii) If $s_i = s_j$ for some $j \neq i$, draw a value for $\mu_{k_{-i}+1} \sim G$ if this is not available in the previous iteration.
   (iii) Label s_i to be $s \in \{1, \ldots, k_{-i}+1\}$ according to the probabilities

$$\mathrm{P}(s_i = s \,|\, s_{-i}, \text{rest}) \propto \begin{cases} N_{-i,s}\psi_i(X_i; \mu_s), & 1 \leq s \leq k_{-i}, \\ \frac{M}{k_{-i}+1}\psi_i(X_i; \mu_s), & s = k_{-i}+1. \end{cases}$$

(b) For all $s \in \{1, \ldots, k\}$, draw a new value $(\mu_s \,|\, X_i : s_i = s)$ by a one-step Metropolis-Hastings algorithm, or some other method of sampling for which the target distribution is invariant.

The "no gaps" algorithm deliberately uses a unidentifiable model to improve mixing. We outline why this algorithm works in the no-data case. The basic reasoning is the same in general.

The probability of obtaining a fixed configuration $(s_1, \ldots, s_n)$ is given by (4.20) as

$$P(s_1, \ldots, s_n) = \frac{M^k \Gamma(M) \prod_{j=1}^{k} \Gamma(N_j)}{\Gamma(M+n)}.$$

If gaps are to be avoided, there are $k!$ labeling possible. Thus

$$P(s_1, \ldots, s_n, \text{labels}) = \frac{M^k \Gamma(M) \prod_{j=1}^{k} \Gamma(N_j)}{k!\,\Gamma(M+n)}.$$

An observation cannot move if it would empty a cluster, unless it is in cluster k. The probabilities for index i to move to the jth cluster are proportional to

$$\begin{aligned} &\frac{\Gamma(M) M^{k_{-i}} N_{-i,j} \prod_{l=1}^{k_{-i}} \Gamma(N_{-i,l})}{\Gamma(M+n)(k_{-i})!}, && j = 1, \ldots, k_{-i}, \\ &\frac{\Gamma(M) M^{k_{-i}+1} \prod_{l=1}^{k_{-i}} \Gamma(N_{-i,l})}{\Gamma(M+n)(k_{-i}+1)!}, && j = k_{-i}+1. \end{aligned}$$

This is equivalent to saying that the probabilities are proportional to $N_{-i,j}$ for $j = 1, \ldots, k_{-i}$ and $M/(k_{-i}+1)$ for $j = k_{-i}+1$.

The permutation step immediately before each transition says that a cluster i of size 1 is kept the same with probability $k_{-i}/(k_{-i}+1)$ and moved with probability $1/(k_{-i}+1)$. Given a move, the destination is j with probability proportional to $N_{-i,j}$, for $j = 1, \ldots, k_{-i}$, and is $k_{-i}+1$ (that is, a new cluster is opened) with probability proportional to $M/(k_{-i}+1)$.

The next algorithm is based on the stick-breaking representation $F = \sum_{j=1}^{\infty} W_j \delta_{\theta_j}$ of the mixing distribution and "slicing." The stick-breaking weights $W_j = V_j \prod_{l<j}(1 - V_l)$ in this representation are in one-to-one correspondence with the relative cuts V_j of the sticks, whence we can use the V_j and W_j interchangeably in the description of the algorithm. In terms of the stick-breaking representation the mixed density (5.1) (with the parameter φ omitted) becomes

$$p_{i,F}(x) = \sum_j W_j \psi_i(x; \theta_j).$$

We shall write this presently as $p_i(x \mid W, \theta)$, for $W = (W_1, W_2, \ldots)$ and $\theta = (\theta_1, \theta_2, \ldots)$ the vectors of all latent variables. We use similar notation to collect all coordinates of other variables.

It is helpful to rewrite the sampling scheme (5.2) in terms of the W_j and θ_j, and also to augment it with latent variables $s_1, \ldots, s_n$ that denote the components (or θ_j) from which $X_1, \ldots, X_n$ are drawn. (These variables have another meaning than in Algorithm 2; also the present θ_j have another interpretation than the variables θ_i in (5.2).) This gives the equivalent description

$$X_i \mid \theta, s, W \overset{\text{ind}}{\sim} \psi_i(\cdot; \theta_{s_i}), \quad s_i \mid V, \theta \overset{\text{iid}}{\sim} W, \quad V_j \overset{\text{iid}}{\sim} \text{Be}(1, M), \quad \theta_j \overset{\text{iid}}{\sim} \bar{\alpha}. \tag{5.7}$$

It follows that the conditional likelihood of $X_1, \ldots, X_n, s_1, \ldots, s_n$ takes the form

$$p(X_1, \ldots, X_n, s_1, \ldots, s_n \mid \theta, W) = \prod_{i=1}^{n} \psi_i(X_i; \theta_{s_i}) W_{s_i}.$$

The next step is to augment the scheme with further latent variables $U_1, \ldots, U_n$, the *slicing variables*, such that the joint (conditional) density is given by

$$p(X_1, \ldots, X_n, s_1, \ldots, s_n, U_1, \ldots, U_n \mid \theta, W) = \prod_{i=1}^{n} \psi_i(X_i; \theta_{s_i}) \mathbb{1}\{U_i < W_{s_i}\}.$$

Given a uniform dominating measure this indeed integrates out relative to $U_1, \ldots, U_n$ to the preceding display, and hence provides a valid augmentation.

The next algorithm is a Gibbs sampler for the posterior density obtained by multiplying the preceding display by prior densities of θ and W (or equivalently V, which is a product of the beta densities $M(1 - v_j)^{M-1}$).

Algorithm 5 (Slicing algorithm) Sequentially update the vectors U, θ, V, s using the conditional densities in the given order:

(i) $p(U_i \mid U_{-i}, \text{rest}) \propto \mathbb{1}\{U_i < W_{s_i}\}, \qquad i = 1, \ldots, n;$
(ii) $p(\theta_j \mid \theta_{-j}, \text{rest}) \propto \prod_{i: s_i = j} \psi_i(X_i; \theta_j)\, d\alpha(\theta_j), \qquad j = 1, 2 \ldots;$
(iii) $p(V_j \mid W_{-j}, \text{rest}) \propto M(1 - V_j)^{M-1} \prod_{i=1}^{n} \mathbb{1}\{U_i < W_{s_i}\}, \qquad j = 1, 2 \ldots;$
(iv) $p(s_i \mid s_{-i}, \text{rest}) \propto \psi_i(X_i; \theta_{s_i}) \mathbb{1}\{U_i < W_{s_i}\}, \qquad i = 1, \ldots, n.$

The products in (ii) or (iii) can be vacuous, in which case they are interpreted as 1. As in the other algorithms described above, conjugacy makes step (ii) easier, but is not essential.

Steps (ii) and (iii) are written as updates of the infinite sequences θ and W, but in fact only their coordinates θ_{s_i} and W_{s_i} appear in the algorithm, and $\max(s_1, \ldots, s_n)$ will typically remain small, so that in practice truncation to a small set is feasible. Furthermore, the conditional distributions in steps (ii) and (iii) are identical to the prior for $j > \max(s_1, \ldots, s_n)$; no update is needed for such θ_j. The set $\{s_1, \ldots, s_n\}$ varies over the cycles of the Gibbs sampler. Because the conditional density of s_i vanishes at k if $W_k \le U_i$, all s_i will be smaller than N if $\max_{k>N} W_k \le \min_i U_i$, which is implied by $\sum_{j=1}^{N} W_j \ge 1 - \min_i U_i$. In view of Proposition 4.20 the minimal N is distributed as one plus a Poisson random variable with parameter $M \log_-(\min_i U_i)$.

The density for updating V_j (or W_j) in (iii) is a beta-density restricted to the set $\cap_i \{U_i < W_{s_i}\}$. Since $W_j = \prod_{l<j}(1 - V_l)V_j$, this set can by some algebra be rewritten in the form $a_j < V_j < b_j$, for

$$a_j = \frac{\max_{i:s_i=j} U_i}{\prod_{l<j}(1 - V_l)}, \qquad b_j = \min_{i:s_i>j}\Big(1 - \frac{U_i}{V_{s_i}\prod_{l<s_i, l\neq j}(1 - V_l)}\Big).$$

Sampling from this truncated distribution can be simply implemented through the quantile method, as the cumulative distribution function of the $\mathrm{Be}(1, M)$-distribution has the explicit form $v \mapsto (1 - v)^M$. In fact, the algorithm can be easily extended to more general stick-breaking distributions as long as the truncated cumulative distribution function has a form suitable for inversion.

An easy way of computing with Dirichlet mixtures is replacing $F \sim \mathrm{DP}(\alpha)$ by the process F_N following the Dirichlet-multinomial distribution described by Theorem 4.19. This reduces the parameter of the problem from infinite-dimensional to finite-dimensional, allowing a treatment by common MCMC methods, such as the Metropolis-Hastings algorithm. However, the right choice of N is unclear. Thumb rules such as $N = cn$ for small n and $N = \sqrt{n}$ for large n have been proposed, but a full theoretical justification is unavailable. (However, see Problem 4.39.)

5.3 Variational Algorithm

A *variational algorithm* is an iterative, deterministic method to determine an approximation to the posterior density of a given predefined form. Given a flexible but tractable class $\mathcal{Q}$ of probability densities on the parameter space Θ it seeks to find $q \in \mathcal{Q}$ that minimizes the Kullback-Leibler divergence

$$K(q; \pi(\cdot\,|\,X)) = -\int q(\theta) \log\Big(\frac{p(X\,|\,\theta)\pi(\theta)}{q(\theta)}\Big)\, d\nu(\theta) + \log p(X) \tag{5.8}$$

between q and the posterior density $\theta \mapsto \pi(\theta\,|\,X) = p(X\,|\,\theta)\pi(\theta)/p(X)$. Typically $\mathcal{Q}$ is taken to be a high-dimensional parametric family, and maximization is carried out by sequentially optimizing over a single parameter, keeping the other parameters fixed, iterating "until convergence." The method is similar in spirit to the EM algorithm in that it involves computing and maximizing integrals, and shares the coordinatewise iterations with Gibbs sampling.

Because the second term on the right of (5.8), the log-marginal density of X, does not depend on q, the minimization can be carried out by maximizing the integral. If the posterior density is contained in $\mathcal{Q}$, then the left side will be minimized to zero, and hence the maximum value of the integral will equal $\log p(X)$. Therefore the value of the maximization problem is an estimate of $\log p(X)$. This by-product of the algorithm is useful, e.g. for model selection.

Variational methods were originally developed for exponential families with incomplete data and conjugate priors. In that situation the integral of the log-density reduces to an expression involving the mean of the variable, whence the approach is commonly known as a *mean-field variational method*. Here we discuss an implementation of the algorithm for approximating the posterior density in Dirichlet process mixture models.

We consider the model (5.7), with the difference that presently we truncate the vectors θ and W to vectors of length N. This corresponds to using a truncated stick-breaking approximation $F_N = \sum_{j=1}^N W_j \delta_{\theta_j}$ for the mixing distribution F (which follows a Dirichlet process in the "ideal" model). The weights in this approximation are for $j = 1, \ldots, N-1$, given as usual by $W_j = V_j \prod_{l=1}^{j-1}(1 - V_l)$, for $V_j \overset{\text{iid}}{\sim} \text{Be}(1, M)$, but presently we set $V_N = 1$ to ensure that $\sum_{j=1}^N W_j = 1$. (This corresponds to taking $\theta_0 = \theta_{N_\epsilon}$ in the ϵ-Dirichlet process (4.23).) The posterior density of the set of latent variables now satisfies

$$p(s_1, \ldots, s_n, \theta_1, \ldots, \theta_N, V_1 \ldots, V_{N-1} | X_1, \ldots, X_n) \propto \prod_{i=1}^n \psi_i(X_i; \theta_{s_i}) \prod_{i=1}^n W_{s_i} \prod_{j=1}^N g(\theta_j) \prod_{j=1}^{N-1} \text{be}(V_j; 1, M).$$

We approximate this by a density q under which all the latent variables s_i, θ_j, V_j are independent, every s_i has a completely unknown distribution on $\{1, \ldots, N\}$ given by a probability vector $(\pi_{i,1} \ldots, \pi_{i,N})$ that is allowed to be different for different coordinates, $\theta_j \sim g_{\xi_j}$ for some parameterized family of densities g_ξ, and $V_j \overset{\text{ind}}{\sim} \text{Be}(a_j, b_j)$ for parameters a_j, b_j. Thus

$$q(s_1, \ldots, s_n, \theta_1, \ldots, \theta_N, V_1, \ldots, V_{N-1}) = \prod_{i=1}^n \pi_{i,s_i} \prod_{j=1}^N g_{\xi_j}(\theta_j) \prod_{j=1}^{N-1} \text{be}(V_j; a_j, b_j).$$

Next we determine the parameters $\pi_{i,j}, \xi_j, a_j, b_j$ by the variational method.

The integral in (5.8) is the expectation under q of

$$\sum_{i=1}^n \Big(\log \psi_i(X_i; \theta_{s_i}) + \log \frac{W_{s_i}}{\pi_{i,s_i}} \Big) + \sum_{j=1}^N \log \frac{g}{g_{\xi_j}}(\theta_j) + \sum_{j=1}^{N-1} \log \frac{\text{be}(V_j; 1, M)}{\text{be}(V_j; a_j, b_j)}.$$

After substituting $W_j = V_j \prod_{l<j}(1 - V_l)$ and the expressions for the beta densities, we can evaluate this expectation to

$$\sum_{i=1}^{n}\sum_{j=1}^{N}\pi_{i,j}\Big[\mathrm{E}_q(\log\psi_i(X_i;\theta_j))+\mathrm{E}_q\log V_j+\sum_{l<j}\mathrm{E}_q\log(1-V_l)-\log\pi_{i,j}\Big]$$
$$-\sum_{j=1}^{N}K(g_{\xi_j};g)$$
$$+\sum_{j=1}^{N-1}\Big[\log\frac{M\Gamma(a_j)\Gamma(b_j)}{\Gamma(a_j+b_j)}-(a_j-1)\mathrm{E}_q\log V_j+(M-b_j)\mathrm{E}_q\log(1-V_j)\Big].$$

A beta variable $V\sim\mathrm{Be}(a,b)$ has logarithmic moment $\mathrm{E}(\log V)=\Psi(a)-\Psi(a+b)$ and hence $\mathrm{E}(\log(1-V))=\Psi(b)-\Psi(a+b)$, for the *digamma function* $\Psi(x)=\frac{d}{dx}\log\Gamma(x)$. This allows to replace $\mathrm{E}_q\log V_j$ and $\mathrm{E}_q\log(1-V_j)$ by concrete expressions involving (a_j,b_j). The resulting expression must be maximized with respect to the variational parameters $\xi_j,\pi_{i,j},a_j,b_j$, for $i=1,\dots,n$ and $j=1,\dots,N$. A typical method is "coordinatewise ascent": the expression is repeatedly maximized with respect to one parameter at a time, with the other parameters fixed. The updating formulas are known as *variational updates*.

Example 5.5 We illustrate the method for normal mixtures, with conjugate choices for the densities constituting q. (We fix the standard deviation σ of the mixing distribution; extensions to varying σ are straightforward.) Specifically, we let $\psi_i(\cdot;\theta)$ be the density of the $\mathrm{Nor}(\theta,\sigma^2)$-distribution, the center measure G of the Dirichlet prior a $\mathrm{Nor}(\mu,\tau^2)$-distribution, and replace g_ξ by the density of a $\mathrm{Nor}(\xi,\eta^2)$-distribution (and ξ by (ξ,η)). The integral in (5.8) then becomes

$$-\sum_{i=1}^{n}\sum_{j=1}^{N}\pi_{i,j}\Big[\log(\sigma\sqrt{2\pi})+\frac{(X_i-\xi_j)^2+\eta_j^2}{2\sigma^2}\Big]-\sum_{i=1}^{n}\sum_{j=1}^{N}\pi_{i,j}\log\pi_{i,j}$$
$$+\sum_{i=1}^{n}\sum_{j=1}^{N-1}\pi_{i,j}(\Psi(a_j)-\Psi(a_j+b_j))+\sum_{i=1}^{n}\sum_{j=1}^{N}\pi_{i,j}\sum_{l<j}(\Psi(b_l)-\Psi(a_l+b_l))$$
$$-\sum_{j=1}^{N}\Big[\log\frac{\tau}{\eta_j}-\frac{\tau^2-\eta_j^2}{2\tau^2}+\frac{(\mu-\xi_j)^2}{2\tau^2}\Big]+\sum_{j=1}^{N-1}\log\frac{M\Gamma(a_j)\Gamma(b_j)}{\Gamma(a_j+b_j)}$$
$$-\sum_{j=1}^{N-1}\Big[(a_j-1)(\Psi(a_j)-\Psi(a_j+b_j))-(M-b_j)(\Psi(b_j)-\Psi(a_j+b_j))\Big].$$

Maximizing this with respect to one parameter at a time leads to the following variational updates, which must be executed cyclically until convergence occurs:

(i) $\xi_j\to\Big(\frac{\sum_{i=1}^n\pi_{i,j}X_i}{\sigma^2}+\frac{\mu}{\tau^2}\Big)\Big(\frac{\sum_{i=1}^n\pi_{i,j}}{\sigma^2}+\frac{1}{\tau^2}\Big)^{-1}$;

(ii) $\eta_j^2\to\Big(\frac{\sum_{i=1}^n\pi_{i,j}}{\sigma^2}+\frac{1}{\tau^2}\Big)^{-1}$;

(iii) update $\pi_{i,j}$ for $j = 1, \ldots, N-1$, to $Kc_{i,j}/(1+K\sum c_{i,l})$, and $\pi_{i,N}$ to $1/(1+K\sum c_{i,l})$, where $K = \exp\big[-((X_i-\xi_N)^2+\eta_N^2)/(2\sigma^2)\big]$ and

$$c_{i,j} = \exp\Big(-\frac{(X_i-\xi_j)^2+\eta_j^2}{2\sigma^2} - \sum_{k=1}^{j}\Psi(a_k+b_k) + \Psi(a_j) + \sum_{k=1}^{j-1}\Psi(b_k)\Big).$$

(iv) a_j is updated to the solution a of the equation

$$\Psi'(a)\Big(a-1+\sum_{i=1}^{n}\pi_{i,j}\Big) - \Psi'(a+b_j)\Big(\sum_{i=1}^{n}\sum_{l=j}^{N}\pi_{i,l}+M-b_j-a\Big) = 0.$$

(v) b_j is updated to the solution b of of the equation

$$\Psi'(b)\Big(\sum_{i=1}^{n}\sum_{l=j}^{N}\pi_{i,l}+M-b\Big) - \Psi'(a_j+b)\Big(\sum_{i=1}^{n}\sum_{l=j}^{N}\pi_{i,l}+M+1-a_j-b\Big) = 0.$$

The updates of a_j and b_j are easy to implement, since the *trigamma function* $\Psi'(x) = \frac{d^2}{dx^2}\log\Gamma(x)$ is available in the library of most mathematical software packages.

5.4 Predictive Recursion Deconvolution Algorithm

Consider the mixture model $p_F(x) = \int \psi(x;\theta)\,dF(\theta)$ as in (5.1), but with a kernel $\psi(\cdot;\theta)$ that does not depend on i or an additional nuisance parameter. Given $F \sim \mathrm{DP}(MG)$ and a single observation X_1 from the mixture density p_F, the posterior mean of the mixing distribution satisfies, by Proposition 5.2 with $n = 1$ (or (5.3) in combination with Theorem 5.3 for the conditional distribution of θ_1 given X_1),

$$\mathrm{E}(F(A)|\,X_1) = \frac{M}{M+1}G(A) + \frac{1}{M+1}\frac{\int_A \psi(X_1;\theta)\,dG(\theta)}{\int \psi(X_1;\theta)\,dG(\theta)}.$$

Given a center measure G with a density g this suggests as estimator for a density f of the mixing distribution F:

$$\hat f_1(\theta) = \frac{M}{M+1}g(\theta) + \frac{1}{M+1}\frac{\psi(X_1;\theta)g(\theta)}{\int \psi(X_1;t)\,dG(t)}.$$

This is a convex combination of the prior density g and the "baseline posterior density," which treats g as the prior for θ and the kernel $\psi(X_1;\theta)$ as the likelihood. This formula can be forwarded into a sequential updating algorithm, which incorporates the observations one by one. The initial step starts from a true prior density $g = \hat f_0$, but each following step uses the output $\hat f_{i-1}$ of the previous step as the prior. Generalizing also the weights of the convex combination to given numbers $w_1, w_2, \ldots \in (0,1)$, we are led to the recursive algorithm, often called *Newton's algorithm*:

$$\hat f_i(\theta) = (1-w_i)\hat f_{i-1}(\theta) + w_i\frac{\psi(X_i;\theta)\,\hat f_{i-1}(\theta)}{\int \psi(X_i;t)\,d\hat F_{i-1}(t)}. \tag{5.9}$$

The corresponding estimates of the mixture densities are given by $\hat{p}_i(x) := \int \psi(x;\theta)\, d\hat{F}_i(\theta)$.

Attractive features of this algorithm are simplicity, making for fast computations, and the built-in absolute continuity of the estimate of the mixing distribution. An unappealing aspect is that the final estimate $\hat{f}_n$ depends on the ordering of the observations. In particular, it is not a Bayes estimator,[1] even if the initial motivation came from the posterior mean of a Dirichlet process mixture prior. The dependence on the ordering can be removed (or alleviated) by averaging over all (or many random) permutations of the observations. We can hope that the resulting estimator will be a reasonable approximation to a proper Bayesian estimator.

On the theoretical side, the estimator's properties are difficult to study, since it is neither a Bayes estimator nor an optimizer of a criterion function, nor does it have a closed form. The next theorem shows consistency; its proof uses martingale convergence techniques.

The weights are assumed to satisfy $\sum_{i=1}^{\infty} w_i = \infty$ and $\sum_{i=1}^{\infty} w_i^2 < \infty$. The first can be motivated by the fact that the weight $\prod_{i=1}^{n}(1-w_i)$ of the initial estimate $\hat{f}_0$ in $\hat{f}_n$ converges to zero if and only if $\sum_{i=1}^{\infty} w_i = \infty$. The second ensures that the weights decay to zero fast enough that the final estimate is not too much affected by the later terms, but information provided by additional observations accumulates properly. The theorem makes the following technical assumption:

$$\sup_{\theta_1,\theta_2,\theta_3\in\Theta} \int \frac{\psi^2(x;\theta_1)}{\psi^2(x;\theta_2)}\, \psi(x;\theta_3)\, d\nu(x) < \infty. \tag{5.10}$$

For consistency of $\hat{f}_n$ (in a weak sense) the mixing distribution is assumed identifiable. Most common kernels, including $\mathrm{Nor}(\theta,\sigma^2)$, σ fixed, $\mathrm{Ga}(\alpha,\theta)$, α fixed and $\mathrm{Poi}(\theta)$, satisfy this condition.

Let f_0 denote the density of the true mixing measure F_0 and let $p_0(x) = \int \psi(x;\theta)\, dF_0(\theta)$ be the true density of the observations.

Theorem 5.6 *If $\sum_{i=1}^{\infty} w_i = \infty$ and $\sum_{i=1}^{\infty} w_i^2 < \infty$, $K(p_0; p_G) < \infty$, and (5.10) holds, then $K(p_0; \hat{p}_n) \to 0$ a.s. If, moreover, the map $F \mapsto p_F$ is one-to-one, the kernel $\psi(x;\theta)$ is bounded and continuous in θ for every $x \in \mathfrak{X}$, and for every $\epsilon > 0$ and compact subset $\mathfrak{X}_0 \subset \mathfrak{X}$, there exists a compact subset $\Theta_0 \subset \Theta$ such that $\int_{\mathfrak{X}_0} \psi(x;\theta)\, d\nu(x) < \epsilon$ for all $\theta \notin \Theta_0$, then $\hat{F}_n \rightsquigarrow F_0$ a.s.*

Proof Because convergence in Kullback-Leibler divergence is stronger than $\mathbb{L}_1$- and weak convergence, the last assertion of the theorem follows from the first and Lemma K.11.

For the proof of the first assertion we first show that $K(p_0; \hat{p}_n)$ converges a.s. to a finite limit. We can write

$$\log \frac{\hat{p}_n}{\hat{p}_0}(x) = \sum_{i=1}^{n} \log \frac{\hat{p}_i}{\hat{p}_{i-1}}(x) = \sum_{i=1}^{n} \log\,(1 + w_i Z_i(x)),$$

[1] The only exception is $n = 1$, when $\hat{f}_1$ coincides with the Bayes estimator for the Dirichlet mixture prior where $M = (w_1^{-1} - 1)$ and G has density $\hat{f}_0$.

for the stochastic processes

$$Z_i(x) = \int \frac{\psi(x;\theta)\psi(X_i;\theta)}{\hat{p}_{i-1}(x)\hat{p}_{i-1}(X_i)}\, d\hat{F}_{i-1}(\theta) - 1.$$

For R defined by the relation $\log(1+x) = x - x^2 R(x)$, this gives that

$$K(p_0;\hat{p}_0) - K(p_0;\hat{p}_n) = \sum_{i=1}^n \int w_i Z_i\, dP_0 - \sum_{i=1}^n \int w_i^2 Z_i^2 R(w_i Z_i)\, dP_0. \tag{5.11}$$

By Taylor's theorem there exists a constant $\xi \in [0,1]$ such that $R(x) = \frac{1}{2}/(1+\xi x)^2$. Therefore $R(x) \le \frac{1}{2}/(1-w)^2$ for $x \ge -w$, whence the second term on the right is bounded above by $\frac{1}{2}\sum_{i=1}^n w_i^2/(1-w_i)^2 \int Z_i^2\, dP_0$. For $\mathscr{F}_i = \sigma(X_1,\ldots,X_i)$,

$$\begin{aligned}
\mathrm{E}\Big(\int Z_i\, dP_0 |\, \mathscr{F}_{i-1}\Big) &= \int \Big(\int \frac{\psi(x;\theta)}{\hat{p}_{i-1}(x)}\, dP_0(x)\Big)^2 d\hat{F}_{i-1}(\theta) - 1 \\
&\ge \Big(\iint \frac{\psi(x;\theta)}{\hat{p}_{i-1}(x)}\, dP_0(x)\, d\hat{F}_{i-1}(\theta)\Big)^2 - 1 = \Big(\int dP_0(x)\Big)^2 - 1 = 0.
\end{aligned}$$

Thus $S_n := \sum_{i=1}^n w_i \int Z_i\, dP_0$ is a submartingale. By the inequality $(a+b)^2 \le 2a^2 + 2b^2$ and the Cauchy-Schwarz inequality $Z_i^2(x) \le 2k_{i-1}(x)k_{i-1}(X_i) + 2$, for the function defined by $k_{i-1}(x) = \int \psi^2(x;\theta)/\hat{p}_{i-1}^2(x)\, d\hat{F}_{i-1}(\theta)$. Consequently,

$$\begin{aligned}
\tfrac{1}{2}\mathrm{E}\Big(\int Z_i^2\, dP_0 |\, \mathscr{F}_{i-1}\Big) &\le \Big(\iint \frac{\psi^2(x;\theta)}{\hat{p}_{i-1}^2(x)}\, d\hat{F}_{i-1}(\theta)\, dP_0(x)\Big)^2 + 1 \\
&\le \Big(\iiiint \frac{\psi^2(x;\theta_1)\psi(x;\theta_3)}{\psi^2(x;\theta_2)}\, d\mu(x)\, d\hat{F}_{i-1}(\theta_1) d\hat{F}_{i-1}(\theta_2)\, dF_0(\theta_3)\Big)^2 + 1.
\end{aligned}$$

In the last step we apply Jensen's inequality to $1/\hat{p}_{i-1}^2(x)$, using the convexity of $x \mapsto x^{-2}$ on $(0,\infty)$. The right is bounded by 1 plus the square of the constant given in assumption (5.10). Because $R(x) \ge 0$ it follows that $Q := \frac{1}{2}\sum_{i=1}^\infty w_i^2/(1-w_i)^2 \int Z_i^2\, dP_0$ is nonnegative and integrable. Because $S_n \le K(p_0;\hat{p}_0) - K(p_0;\hat{p}_n) + Q \le K(p_0;\hat{p}_0) + Q$ by (5.11), where $\hat{p}_0 = p_G$ is fixed, it follows that S_n is a submartingale that is bounded above by an integrable random variable. By a martingale convergence theorem it converges a.s. (and in $\mathbb{L}_1$) to a finite limit.

If the limit were strictly positive, then the sequence $\sum_{i=1}^n w_i K(p_0;\hat{p}_{i-1})$ would increase to infinity, as the series $\sum_i w_i$ diverges by assumption. Thus to prove that the limit is zero, it suffices to show that the sequence $\sum_{i=1}^n w_i K(p_0;\hat{p}_{i-1})$ is almost surely bounded. By the inequality $\log x \le x - 1$, for $x > 0$, this sequence is bounded above by $\sum_{i=1}^n w_i \int (p_0/\hat{p}_{i-1} - 1)\, dP_0$. We show that the latter sequence tends to a limit by a similar argument as the preceding, but now at the level of f rather that p_F.

Writing $\log(\hat{f}_n/\hat{f}_0) = \sum_{i=1}^n \log(\hat{f}_i/\hat{f}_{i-1}) = \sum_{i=1}^n \log(1+w_i Y_i)$, we obtain in analogy with (5.11),

$$K(f_0;\hat{f}_0) - K(f_0;\hat{f}_n) = \sum_{i=1}^n \int w_i Y_i\, dF_0 - \sum_{i=1}^n \int w_i^2 Y_i^2 R(w_i Y_i)\, dF_0,$$

for the stochastic processes $Y_i(\theta) = \psi(X_i;\theta)/\hat{p}_{i-1}(X_i) - 1$. The second term on the right is bounded above by $Q' := \frac{1}{2}\sum_{i=1}^{\infty} w_i^2/(1-w_i)^2 \int Y_i^2\, dF_0$, which can be seen to be an integrable random variable by similar arguments as before. Since $\int Y_i\, dF_0 = (p_0/\hat{p}_{i-1})(X_i) - 1$, the increments of the first term satisfy

$$\mathrm{E}\Big(\int Y_i\, dF_0 |\, \mathscr{F}_{i-1}\Big) = \int \Big(\frac{p_0}{\hat{p}_{i-1}} - 1\Big)\, dP_0 \geq \int \log \frac{p_0}{\hat{p}_{i-1}}\, dP_0 \geq 0.$$

Thus $S_n' := \sum_{i=1}^n w_i \int Y_i\, dF_0$ is a submartingale. It is bounded above by $K(f_0; g) + Q'$ and hence converges a.s. to a finite limit by a martingale convergence theorem.

The increments of the martingale $M_n' := S_n' - \sum_{i=1}^n \mathrm{E}(\int w_i Y_i\, dF_0 |\, \mathscr{F}_{i-1})$ have conditional second moments bounded by $\mathrm{E}(\int w_i^2 Y_i^2\, dF_0 |\, \mathscr{F}_{i-1})$, whose sum converges in $\mathbb{L}_1$. Therefore, the martingale M_n' converges a.s., and hence the compensator $\sum_{i=1}^n \mathrm{E}(\int w_i Y_i\, dF_0 |\, \mathscr{F}_{i-1})$ converges, which is what we wanted to prove. □

5.5 Examples of Kernels

Example 5.7 (Density estimation on $\mathbb{R}$ using normal mixtures) For estimating a density function supported on the whole of $\mathbb{R}$ the density $x \mapsto \phi_\sigma(x-\mu)$ of the normal $\mathrm{Nor}(\mu, \sigma^2)$-distribution is a natural kernel. One may consider either a location-only mixture or a location-scale mixture.

For a location-only mixture, we equip $\theta = \mu$ with a mixing distribution F and consider $\varphi = \sigma$ an additional parameter of the kernel. The conjugate center measure G for the Dirichlet process prior on F is a normal distribution $\mathrm{Nor}(m, \eta^2)$. Given σ, the updating rule given by the generalized Pólya urn scheme of Theorem 5.3 becomes

$$q_{i,j} \propto \begin{cases} \sigma^{-1} e^{-(X_i-\mu_j)^2/(2\sigma^2)}, & j \neq i,\ j \geq 1, \\ M(\sigma^2+\eta^2)^{-1/2} e^{-(x-m)^2/(2(\sigma^2+\eta^2))}, & j = 0, \end{cases}$$

with the baseline posterior described by

$$dG_b(\mu |\, \sigma, X_i) \sim \mathrm{Nor}\Big(\frac{\sigma^{-2}X_i + \eta^{-2}m}{\sigma^{-2}+\eta^{-2}}, \frac{1}{\sigma^{-2}+\eta^{-2}}\Big).$$

The bandwidth parameter σ may be kept constant (but dependent on the sample size), or, more objectively, be equipped with a prior. The inverse-gamma prior $\sigma^{-2} \sim \mathrm{Ga}(s, \beta)$ is conjugate for the model given $(X_1, \ldots, X_n, \mu_1, \ldots, \mu_n)$, which is essentially a parametric model $X_i - \mu_i \overset{\mathrm{iid}}{\sim} \mathrm{Nor}(0, \sigma^2)$. By elementary calculations,

$$\sigma^{-2} |\, X_1, \ldots, X_n, \mu_1, \ldots, \mu_n \sim \mathrm{Ga}\Big(s + n/2, \beta + \tfrac{1}{2}\sum_{i=1}^n (X_i - \mu_i)^2\Big).$$

For location-scale mixtures, we equip the parameter $\theta = (\mu, \sigma)$ with a mixing distribution, which in turn is modeled by a Dirichlet process prior. The normal-inverse-gamma distribution (that is, $\sigma^{-2} \sim \mathrm{Ga}(s, \beta)$ and $\mu |\, \sigma \sim \mathrm{Nor}(m, \sigma^2/a)$) is conjugate in this case. A simple computation shows that

$$q_{i,j} \propto \begin{cases} \dfrac{M\sqrt{a}\,\Gamma(s+1/2)\beta^s}{\sqrt{1+a}\,\Gamma(s)\{\beta + a(X_i-m)^2/(2(1+a))\}^{s+1/2}}, & j=0, \\ \sigma_j^{-1} e^{-(X_i-\mu_j)^2/(2\sigma_j^2)}, & j \neq i,\ j \geq 1, \end{cases}$$

and the baseline posterior is described by

$$\mu|\,\sigma, X_i \sim \mathrm{Nor}((X_i + am)/(1+a), \sigma^2/(1+a)),$$
$$\sigma^{-2}|\,X_i \sim \mathrm{Ga}(s+1/2, \beta + a(X_i-m)^2/(2(1+a))).$$

Example 5.8 (Uniform scale mixtures) Any nonincreasing probability density p on $[0,\infty)$ can be represented as a mixture of the form (see Williamson 1956)

$$p(x) = \int_0^\infty \theta^{-1} \mathbb{1}\{0 \leq x \leq \theta\}\, dF(\theta).$$

This lets us put a prior on the class of *decreasing densities* by putting a Dirichlet process prior on F, with center measure G supported on $(0,\infty)$. In this case, the posterior updating formulas of Theorem 5.3 hold with

$$q_{i,j} \propto \begin{cases} \theta_j^{-1}\mathbb{1}\{X_i < \theta_j\}, & j \neq 0, \\ M\int_{X_i}^\infty \theta^{-1}\, dG(\theta), & j=0, \end{cases}$$
$$dG_b(\theta|\,X_i) \propto \theta^{-1}\mathbb{1}\{\theta > X_i\}\, dG(\theta).$$

Symmetrization of decreasing densities yield *symmetric, unimodal densities*. Such densities are representable as mixtures $p(x) = \int_0^\infty (2\theta)^{-1}\mathbb{1}\{-\theta \leq x \leq \theta\}\, dF(\theta)$ and can be equipped with a prior by, as before, putting a Dirichlet process prior on F. It is left as an exercise to derive computational formulas for the posterior.

Possibly *asymmetric unimodal densities* can be written as two-parameter mixtures of the form

$$p(x) = \int (\theta_2+\theta_1)^{-1}\mathbb{1}\{-\theta_1 \leq x \leq \theta_2\}\, dF(\theta_1,\theta_2).$$

We may put a Dirichlet process prior on F, with center measure G supported on $(0,\infty)\times(0,\infty)$, or alternatively model F a priori as $F = F_1 \times F_2$, for independent $F_1 \sim \mathrm{DP}(M_1G_1)$ and $F_2 \sim \mathrm{DP}(M_2G_2)$. In the second case it is convenient to index by a pair of integers (i,j), and write the updating formulas as $q_{i_1j_1:i_2j_2} = q^{(1)}_{i_1,j_1} q^{(2)}_{i_2,j_2}$, where

$$q^{(1)}_{i_1j_1} \propto \begin{cases} \theta_{1j_1}^{-1}\mathbb{1}\{X_i > -\theta_{1j_1}\}, & j_1 \neq 0, \\ M_1\int_{X_i^-}^\infty \theta_1^{-1}\mathbb{1}\{X_i > -\theta_1\}\, dG_1(\theta_1), & j_1 = 0, \end{cases}$$
$$q^{(2)}_{i_2j_2} \propto \begin{cases} \theta_{2j_2}^{-1}\mathbb{1}\{X_i < \theta_{2j_2}\}, & j_2 \neq 0, \\ M_2\int_{X_i^+}^\infty \theta_2^{-1}\mathbb{1}\{X_i < \theta_2\}\, dG_2(\theta_2), & j_2 = 0, \end{cases}$$

and

$$dG_{1b}(\theta_1|\,X_i) \propto \int_{\theta_2 \in (X_i^+,\infty)} (\theta_2+\theta_1)^{-1}\, dG_2(\theta_2)\mathbb{1}\{\theta_1 > -X_i\}\, dG_1(\theta_1),$$
$$dG_{2b}(\theta_2|\,X_i) \propto \int_{\theta_1 \in (X_i^-,\infty)} (\theta_2+\theta_1)^{-1}\, dG_1(\theta_1)\mathbb{1}\{\theta_2 < X_i\}\, dG_2(\theta_2).$$

Example 5.9 (Mixtures on the half line) Densities on the half line occur in applications in survival analysis and reliability. Also, the error distribution in quantile regression, which has zero as a fixed quantile, may be modeled by a combination of two densities on the half line.

Mixtures of exponential distributions are reasonable models for a decreasing, convex density on the positive half line. More generally, mixtures of gamma densities may be considered. Other possible choices of kernels include the inverse-gamma, Weibull and log-normal. Multi-dimensional parameters may be mixed using a joint mixing distribution or separately modeled by one-dimensional mixing distributions. Weibull and lognormal mixtures are dense in the space of densities on the positive half line relative to the $\mathbb{L}_1$-distance, provided that the shape parameter can assume arbitrarily large (for the Weibull) or small (for the lognormal) values. This follows because these two kernels form location-scale families in the log-scale, so that (2.4) applies.

We discuss computation for Weibull mixtures, with the prior on the mixing distribution equal to the Dirichlet-multinomial. The latter was seen to approximate the Dirichlet process in Theorem 4.19. Other examples can be treated similarly.

The Dirichlet-multinomial distribution is particularly convenient for application using BUGS. It suffices to describe the model and the prior, since the appropriate algorithm for drawing from conditional distributions is obtained by the program itself. Let $\text{Wei}(\alpha, \lambda)$ stand for the Weibull distribution with shape parameter α and rate parameter λ; that is, the density of X is given by $\alpha\lambda^\alpha x^{\alpha-1}e^{-\lambda^\alpha x^\alpha}$, $x > 0$. Let g stand for the density of the center measure. Then the model and the prior can be described through the following hierarchical scheme:

(i) $X_i \overset{\text{ind}}{\sim} \text{Wei}(\alpha, \lambda_{s_i})$,
(ii) $(s_1, \ldots, s_N)$ i.i.d. with $\text{P}(s_i = j) = w_j$, $j = 1, \ldots, N$,
(iii) $(w_1, \ldots, w_N) \sim \text{Dir}(N; M/N, \ldots, M/N)$,
(iv) $\lambda_i \overset{\text{iid}}{\sim} g$,
(v) $\alpha \sim \pi$.

Example 5.10 (Mixtures of beta and Bernstein polynomials) Beta distributions form a flexible two-parameter family of distributions on the unit interval, and generate a rich class of mixtures. Particular shapes of mixtures may be promoted by restricting the values of the beta parameters. For instance, a (vertically reflected) J-shaped density may be modeled by mixtures of $\text{Be}(a, b)$, $a < 1 \leq b$. This type of density is used to model the density of p-values under the alternative in multiple hypothesis testing or to model service time distributions.

We may consider a two-dimensional mixing distribution F on the parameters (a, b) of the beta distributions or mix a and b separately using two univariate mixing distributions, F_1 and F_2. In both cases we can employ Dirichlet process priors, where in the second case it is natural to model F_1 and F_2 as a priori independent. To obtain the reflected J-shape, it suffices to restrict the support of the base measure of F within $(0, 1) \times [1, \infty)$, or that of F_1 and F_2 within $(0, 1)$ and $[1, \infty)$, respectively.

In fact, mixtures of beta densities with *integer* parameters are sufficient to approximate any probability density. These are closely related to *Bernstein polynomials*, as discussed in Section E.1. The kth Bernstein polynomial associated to a given cumulative distribution function F on $(0, 1]$ is

$$B(x; k, F) := \sum_{j=0}^{k} F\Big(\frac{j}{k}\Big)\binom{k}{j} x^j (1-x)^{k-j}.$$

This is in fact itself a cumulative distribution function on $[0, 1]$, which approximates F as $k \to \infty$. The corresponding density is

$$b(x; k, F) := \sum_{j=1}^{k} F\Big(\frac{j-1}{k}, \frac{j}{k}\Big] \mathrm{be}(x; j, k-j+1). \tag{5.12}$$

If F has a continuous density f, then $b(\cdot; k, F)$ approximates f uniformly as $k \to \infty$. The approximation rate is k^{-1} if f has bounded second derivative (see Proposition E.3), which is lower than the rate k^{-2} obtainable with other k-dimensional approximation schemes (such as general polynomials, splines or Gaussian mixtures). This slower approximation property will lead to slower contraction of a posterior distribution. However, the constructive nature of the approximation and preservation of nonnegativity and monotonicity are desirable properties of Bernstein polynomials.

The Bernstein density (5.12) is a mixture of beta densities. Modeling F as a Dirichlet process $\mathrm{DP}(MG)$ and independently, the degree k by a prior ρ on $\mathbb{N}$ yields the *Bernstein-Dirichlet process*. The beta densities $\mathrm{be}(x; j, k-j+1)$ may be thought of as smooth analogs of the uniform kernels $k\mathbb{1}\{(j-1)/k < x \le j/k\}$, and thus the Bernstein-Dirichlet process can be considered a smoothing of the Dirichlet process. The tuning parameter k works like the reciprocal of a bandwidth parameter.

For computing the posterior distribution based on a random sample $X_1, \ldots, X_n$ from the Bernstein-Dirichlet process, it is convenient to introduce auxiliary latent variables $Y_1, \ldots, Y_n$ that control the label of the beta component from where $X_1, \ldots, X_n$ are sampled: if $(j-1)/k < Y_i \le j/k$, then $X_i \sim \mathrm{Be}(\cdot; j, k-j+1)$. The prior then consists of the hierarchy

(i) $k \sim \rho$;
(ii) $F \sim \mathrm{DP}(MG)$;
(iii) $Y_1, \ldots, Y_n \,|\, k, F \overset{\mathrm{iid}}{\sim} F$;
(iv) $X_1, \ldots, X_n \,|\, k, F, Y_1, \ldots, Y_n \overset{\mathrm{ind}}{\sim} p(\cdot \,|\, k, Y_i)$, where

$$p(x \,|\, k, Y_i) = \sum_{j=1}^{k} \mathrm{be}(x; j, k-j+1)\mathbb{1}\{(j-1)/k < Y_i \le j/k\}.$$

The posterior distribution can be computed by Gibbs sampling, as follows, with $z(y, k) := j$ if $(j-1)/k < y \le j/k$:

(i) $k \,|\, X_1, \ldots, X_n, Y_1, \ldots, Y_n \sim \rho(\cdot \,|\, \mathrm{X}, \mathrm{Y})$, for

$$\rho(k \,|\, \mathrm{X}, \mathrm{Y}) \propto \rho(k) \prod_{i=1}^{n} \mathrm{be}(X_i; z(Y_i, k), k - z(Y_i, k) + 1);$$

(ii) $Y_i \,|\, k, X_1, \ldots, X_n, Y_{-i} \sim \sum_{j \neq i} q_{i,j}\delta_{Y_j} + q_{i,0}G_{b,i}$, for

$$q_{i,j} \propto \begin{cases} Mb(X_i; k, G), & j = 0, \\ \mathrm{be}(X_i; z(Y_j, k), k - z(Y_j, k) + 1), & j \neq i, 1 \leq j \leq k, \end{cases}$$
$$dG_{b,i}(y \,|\, k, Y_i) \propto g(y)\mathrm{be}(X_i; z(y, k), k - z(y, k) + 1).$$

Example 5.11 (Random histograms) Random histograms were briefly discussed in Section 2.3.2 as a method for prior construction via binning. Here we consider the particular situation where the corresponding probability distribution is given a Dirichlet process prior. The sample space, an interval in the real line, is first partitioned into intervals $\{I_j(h)\colon j \in J\}$ of (approximately) length h, which is chosen from a prior. Next, a probability density is formed by distributing mass 1 first to the intervals according to the Dirichlet process and next uniformly within every interval.

The resulting prior can be cast in the form of a Dirichlet mixture, with the kernel

$$\psi(x, \theta, h) = \frac{1}{h} \sum_{j \in J} \mathbb{1}\{x \in I_j(h)\}\mathbb{1}\{\theta \in I_j(h)\}.$$

The corresponding mixture with mixing distribution F is

$$p_{h,F}(x) = \frac{1}{h} \sum_{j \in J} F(I_j(h))\mathbb{1}\{x \in I_j(h)\}.$$

We now equip h with a prior π and, independently, F with a DP(α) prior.

Analytic computation of the posterior mean of $p_{h,F}(x)$, given a random sample $X_1, \ldots, X_n$ from this density, is possible, assisted by a single one-dimensional numerical integration, without using simulation. As a first step, given the bandwidth h,

$$\mathrm{E}(p_{h,F}(x) \,|\, h, X_1, \ldots, X_n) = \frac{1}{h} \sum_{j \in J} \frac{\alpha(I_j(h)) + N_j(h)}{|\alpha| + n} \mathbb{1}\{x \in I_j(h)\}, \tag{5.13}$$

for $N_j(h) = \#\{i\colon X_i \in I_j(h)\} = n\mathbb{P}_n(I_j(h))$ the number of observations falling in the jth partitioning set. It remains to integrate out the right side with respect to the posterior distribution of h.

Because the prior enters the likelihood only through the probabilities $F(I_j(h))$ of the partitioning sets $I_j(h)$, the posterior distribution is only dependent on the vector of cell counts $(N_j(h)\colon j \in J)$, by Theorem 3.14. The likelihood given (h, F) is proportional to $\prod_j F(I_j(h))^{N_j(h)}$, and hence the likelihood given only h is proportional to the expectation of this expression with respect to F, that is

$$\mathrm{E}\Big(\prod_{j \in J} F(I_j(h))^{N_j(h)}\Big) = \prod_{j \in J} (\alpha(I_j(h)))^{[N_j(h)]}/|\alpha|^{[n]},$$

by Corollary G.4; here $a^{[n]} = a(a + 1) \cdots (a + n - 1)$ is the ascending factorial. (Note that the product is a finite product, since $N_j(h) = 0$ except for finitely many j.) Thus, by Bayes's theorem, the posterior density satisfies

$$\pi(h|\, X_1, \ldots, X_n) \propto \pi(h) \prod_{j \in J} \alpha(I_j(h))^{[N_j(h)]}. \tag{5.14}$$

Thus the posterior distribution of $p_{h,F}$ is conjugate, where the updating rule is given by $\alpha \mapsto \alpha + \sum_{i=1}^n \delta_{X_i}$ and $\pi \mapsto \pi(\cdot|\, X_1, \ldots, X_n)$.

Finally, the posterior mean $\mathrm{E}(p_{h,F}(x)|\, X_1, \ldots, X_n)$ can be computed as the posterior expectation of (5.13), and is given by

$$\frac{\int h^{-1} \sum_{j \in J} \frac{\alpha(I_j(h)) + N_j(h)}{|\alpha| + n} \mathbb{1}\{x \in I_j(h)\} \pi(h) \prod_{k \in J} \alpha(I_k(h))^{[N_k(h)]}\, dh}{\int \pi(h) \prod_{k \in J} \alpha(I_k(h))^{[N_k(h)]}\, dh}. \tag{5.15}$$

Example 5.12 (Hierarchical Dirichlet process) This example is qualitatively different from the previous examples in that a nonparametric mixture of nonparametric distributions is considered. Suppose

$$X_{i,1}, \ldots, X_{i,n_j} |\, F_i, \bar{\alpha} \overset{\text{iid}}{\sim} F_i, \quad F_1, \ldots, F_k |\, \bar{\alpha} \overset{\text{iid}}{\sim} \mathrm{DP}(M\bar{\alpha}), \quad \bar{\alpha} \sim \mathrm{DP}(\alpha_0).$$

Thus, k distributions are independently generated from a Dirichlet process whose base measure is itself randomly drawn from a source Dirichlet process, and a sample of observations is drawn from every of these k distributions. The discreteness of the Dirichlet process forces $\bar{\alpha}$ to be discrete, and hence the F_i will share support points. Such "Dirichlet mixtures of Dirichlet processes" are useful for hierarchical clustering. For further discussion, see Section 14.1.1.

Example 5.13 (Feller priors) The *Feller random sampling scheme* introduced in Section 2.3.4 approximates a given density f on an interval $I \subset \mathbb{R}$ by a mixture

$$a(x; k, F) = \int h_k(x; z)\, dF(z),$$

with the kernel h_k derived from the density $g_k(\cdot; x)$ of a Feller random sampling scheme $Z_{k,x}$. Specifically, if $Z_{i,k,x} = k^{-1} \sum_{i=1}^k Y_{i,x}$ with the $Y_{i,x}$ i.i.d. variables following an exponential family with parameters so that $\mathrm{E}Y_{i,x} = x$, the relationship between h_k and g_k can be made explicit and the scheme shown to be consistent in the sense that $a(\cdot; k, F)$ converges pointwise to f, at least if f is bounded and continuous.

We can form a prior on densities by putting a Dirichlet process on the mixing distribution F in (2.5). In view of Proposition 2.6 and Scheffe's theorem, the resulting Feller prior will have full total variation support if the Dirichlet process prior for F has full weak support.

The Feller prior for the $\mathrm{Nor}(\theta, 1)$ family is a Dirichlet process mixture of normal location densities with $\sigma = k^{-1}$, while a Bernoulli sampling scheme gives the Bernstein polynomial prior. Another interesting choice is the Poisson family, which leads to a discrete gamma Dirichlet mixture $a(x; k, F) = \sum_{j=1}^\infty F((j-1)/k, j/k]\mathrm{ga}(x; j, k)$. This is thus seen to

have every continuous density on the half line in its $\mathbb{L}_1$-support. Finally, the gamma family leads to a mixture of inverse-gamma distributions $\int (kz)^k x^{-(k+1)} e^{-kz/x}\, dF(z)/\Gamma(k)$.

5.6 Historical Notes

Bayesian density estimation was first considered by Ghorai and Rubin (1982). Dirichlet process mixtures were introduced by Ferguson (1983) and Lo (1984). Computational formulas were discussed by Kuo (1986) and Lo (1984). The Gibbs sampler based on the weighted generalized Pólya urn scheme was contributed by Escobar (1994), Escobar and West (1995) and MacEachern (1994), and refined by MacEachern and Müller (1998), Dey et al. (1998) and Neal (2000). These papers also discuss extensive applications of Dirichlet process mixtures of normals. The slicing algorithm is due to Walker (2007). The variational algorithm has its roots in the statistical physics literature. Beal and Ghahramani (2003) developed the ideas in the context of parametric exponential families. Blei and Jordan (2006) adapted the ideas to suit posterior computations in Dirichlet process mixture models. The predictive recursion deconvolution algorithm appeared in the papers Newton et al. (1998) and Newton and Zhang (1999). An initial proof of convergence appeared in Newton (2002), and was corrected by Ghosh and Tokdar (2006) and Tokdar et al. (2009). Brunner and Lo (1989) and Brunner (1992) considered uniform scale mixtures. Bernstein polynomials priors were studied by Petrone (1999b,a). Gasparini (1996) studied random histograms. Feller approximation priors were introduced by Petrone and Veronese (2010).

Problems

5.1 Generalize (5.14) and (5.15) when α is allowed to depend on h.

5.2 Obtain variational updates in a Dirichlet process mixture of normal model when σ^2 is unknown and is given an inverse-gamma prior distribution.

5.3 (Ghosal et al. 2008) Let a density p be a mixture of $\text{Be}(a, 1)$, $0 < a \leq 1$. Show that $H(y) := \int_0^{e^{-y}} p(x)\, dx$ is a completely monotone function of y on $[0, \infty)$, that is, $(-1)^k H^{(k)}(y) \geq 0$ for all $k \in \mathbb{N}$, $y \geq 0$. Conversely, if $H(y)$ is a completely monotone function of y on $[0, \infty)$, then p is a mixture of $\text{Be}(a, 1)$, $0 < a < \infty$.

Similarly, if p is a mixture of $\text{Be}(1, b)$, $b \geq 1$, then $H(y) := \int_{e^{-y}}^{\infty} p(x)\, dx$ is a completely monotone function of y on $[0, \infty)$. Conversely, if $H(y)$ is a completely monotone function of y, then p is a mixture of $\text{Be}(1, b)$, $0 < b < \infty$.

5.4 (Petrone 1999b) If for all k, $\rho(k) > 0$ and w_k has full support on $\mathbb{S}_k$, then show that every distribution on (0,1] is in the weak support of the Bernstein polynomial prior, and every continuous distribution is in the support for the topology induced by the Kolmogorov-Smirnov distance d_{KS}.

5.5 Verify that $b(x; k, F)$ is the density of $B(x; k, F)$ in Example 5.10.

5.6 (Petrone 1999a) Consider a Bernstein polynomial prior $p(x) = b(x; k, F)$, where $F \sim \text{DP}(MG)$ and $k \sim \rho$. Let $\xi_{jk} = G((j-1)/k, j/k])$ and let $p^*(x)$ be the Bayesian density estimate given observations $X_1, \ldots, X_n$. Put $N_{jk} = \#\{i\colon X_i \in ((j-1)/k, j/k]\}$. Show that

(a) $p^*(x) \to \sum_{k=1}^{\infty} b(x; k, G)\rho_\infty(k|\, X_1, \ldots, X_n)$ as $M \to \infty$, for the probability mass function

$$\rho_\infty(k|\, X_1, \ldots, X_n) \propto \rho(k) \sum_{(z_{1k},\ldots,z_{kk})\in\{1,\ldots,k\}^k} \prod_{j=1}^{k} \xi_{jk}^{N_{jk}} \prod_{i=1}^{n} \mathrm{be}(X_i; z_{ik}, k - z_{ik} + 1);$$

(b) $p^*(x) \to \sum_{k=1}^{\infty} b(x; k, w_k)\rho_\infty(k|\, X_1, \ldots, X_n)$ as $M \to 0$, where the jth component of $w_k \in \mathbb{S}_k$ is proportional to $\xi_{jk} \prod_{i=1}^{n} \mathrm{be}(X_i; j, k - j + 1)$ and

$$\rho_\infty(k|\, X_1, \ldots, X_n) \propto \rho(k) \sum_{j=1}^{k} \xi_{jk} \prod_{i=1}^{n} \mathrm{be}(X_i; j, k - j + 1).$$