Hyperparameter estimation
in Dirichlet process mixture models

By MIKE WEST

Institute of Statistics and Decision Sciences
Duke University, Durham NC 27706, USA.

SUMMARY
In Bayesian density estimation and prediction using Dirichlet process mix-
tures of standard, exponential family distributions, the precision or total
mass parameter of the mixing Dirichlet process is a critical hyperparam-
eter that strongly influences resulting inferences about numbers of mixture
components. This note shows how, with respect to a flexible class of prior
distributions for this parameter, the posterior may be represented in a sim-
ple conditional form that is easily simulated. As a result, inference about
this key quantity may be developed in tandem with the existing, routine
Gibbs sampling algorithms for fitting such mixture models. The concept
of data augmentation is important, as ever, in developing this extension of
the existing algorithm. A final section notes an simple asymptotic approx-
imation to the posterior.

Some key words: Data augmentation; Dirichlet process; Gibbs sampling; Mixture
models

This research was partially financed by the National Science Foundation
under grant DMS-9024793
1. INTRODUCTION

Escobar and West (1991) develop Gibbs sampling methods for the computations involved in Bayesian density estimation and problems of deconvolution using mixtures of Dirichlet processes. We adopt the normal mixture model of this reference for example here – note that the assumed normal data distributions may be replaced by any exponential family form, if required. In analysis of random samples from these nonparametric Bayesian models, data \( y_1, \ldots, y_n \) are assumed conditionally independently normally distributed, \( (y_j|\mu_j, v_j) \sim N(\mu_j, v_j), (j = 1, \ldots, n) \), with the bivariate quantities \( (\mu_j, v_j) \) drawn independently from an uncertain prior distribution \( G(\mu, v) \). The Dirichlet mixture model supposes that \( G(\mu, v) \) is a bivariate Dirichlet process, \( G \sim D(\alpha G_0) \) with mean function \( G_0(\mu, v) = E(G(\mu, v)) \), for any point \( (\mu, v) \), and precision, or total mass, parameter \( \alpha > 0 \) (Ferguson, 1973; Antoniak, 1974). Following Escobar and West, we take \( G_0 \) to be the conjugate normal/inverse gamma distribution under which \( v^{-1} \sim G(s/2, S/2) \) and \( (\mu|v) \sim N(m, \tau v) \). A key feature of the Dirichlet is its discreteness, which in our context implies that the pairs \( (\mu_j, v_j), (j = 1, \ldots, n) \), concentrate on a set of some \( k \leq n \) distinct pairs \( (\theta_i, w_i), (i = 1, \ldots, k) \). As a result, supposing these pairs together with \( G_0 \) and \( \alpha \) to be known, the conditional predictive distributions in such models are mixtures of normals,

\[
p(y|\theta) = \alpha a_n N(\mu_0, v_0) + a_n \sum_{i=1}^{k} N(\theta_i, w_i)
\]

where \( \theta = \{k, (\theta_i, w_i, i = 0, \ldots, k), G_0\} \) and \( a_n = 1/(\alpha + n) \). Observing a sample \( D = (y_1, \ldots, y_n) \) leads to a posterior distribution \( P(\theta|D) \) and predictive inference is based on the Bayesian density estimate \( p(y|D) = \int p(y|\theta)dP(\theta|D) \). Even if \( n \) is extremely small, posteriors \( P(\theta|D) \) are extraordinarily complicated, and simulation is the only method available. If \( P(\theta|D) \) can be sampled, the integral defining \( p(y|D) \) can be approximated by a summation over samples of \( \theta \) values for efficient Monte Carlo inference (Gelfand and Smith 1991). Escobar and West (1991) shown how this may be easily achieved, and how the basic simulation algorithm extends to incorporate learning about the hyperparameters \( m, \tau \) and, less critically, \( S \) of \( G_0 \). Illustrations appear in that paper, and some further applications appear in West and Cao (1992) (the latter in the case of constant variances \( v_i = v \) for each \( i \).)

Central to this class of models is the precision parameter \( \alpha \) of the underlying Dirichlet process. This directly determines the prior distribution for \( k \), the number of additional normal components in the mixture (1), and is thus a critical smoothing parameter for the model and previously has been assumed specified by the investigator. At each stage of the simulation analysis, a specific value of \( k \) is simulated
from the posterior for \( k \) (together with sampled values of the means and variances of the normal components) which also depends critically on this hyperparameter \( \alpha \).

We now show how, based on a specific but flexible family of prior distributions for \( \alpha \), the parameter vector \( \theta \) may be augmented to allow for simulation of the full joint posterior now including \( \alpha \).

2. COMPUTING \( p(\alpha|k) \)

2.1 Introduction

We begin by recalling results from Antoniak (1974). There it is shown that the prior distribution of \( k \) in (1) may be written as

\[
P(k|\alpha, n) = c_n(k)n!\alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \quad (k = 1, 2, \ldots, n),
\]

and \( c_n(k) = P(k|\alpha = 1, n) \), not involving \( \alpha \). If required, the factors \( c_n(k) \) are easily computed using recurrence formulae for Stirling numbers (further details available on request from the author). This is important, for example, in considering the implications for priors over \( k \) of specific choices of priors for \( \alpha \) (and vice-versa) in the initial prior elicitation process.

Assume a continuous prior density \( p(\alpha) \) (which may depend on the sample size \( n \)), hence an implied prior \( P(k|n) = \int P(k|\alpha, n)p(\alpha)d\alpha \). Now suppose we have sampled values of the parameter \( k \) and all the other model parameters. From our model, the data \( D \) are initially conditionally independent of \( \alpha \) given all the other parameters, and we deduce

\[
p(\alpha|k, \pi, D) \propto p(\alpha|k) \propto p(\alpha)P(k|\alpha),
\]

with likelihood function given in (2) (the sample size \( n \) should appear in conditioning, of course, but is omitted for clarity of notation.) Thus the Gibbs sampling analysis can be extended: for given \( \alpha \), we sample parameters \( \pi \), and hence \( k \), as usual from the conditional posterior \( p(\pi|\alpha, D) \). Then, at each iteration, we can include \( \alpha \) in the analysis by sampling from the conditional posterior (2) based on the previously sampled value of \( k \) – no other information is needed. Sampling from (2) may involve using a rejection, or other, method depending on the form of the prior \( p(\alpha) \). Alternatively, we may discretise the range of \( \alpha \) so that (2) provides a discrete approximation to the posteriors – the so-called ‘griddy Gibbs’ approach. More attractively, sampling from the exact, continuous posterior (2) is possible in the Gibbs iterations when the prior \( p(\alpha) \) comes from the class of mixtures of gamma distributions. We develop the results initially for a single gamma prior.
2.2 Gamma prior for $\alpha$

Suppose $\alpha \sim G(a, b)$, a gamma prior with shape $a > 0$ and scale $b > 0$ (which we may extend to include a ‘reference’ prior (uniform for $\log(\alpha)$) by letting $a \to 0$ and $b \to 0$.) In this case, (2) may be expressed as a mixture of two gamma posteriors, and the conditional distribution of the mixing parameter given $\alpha$ and $k$ (and, of course, $n$) is a simple beta. See this as follows. For $\alpha > 0$, the gamma functions in (1) can be written as

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} = \frac{\beta(\alpha + 1, n)}{\alpha \Gamma(n)},$$

where $\beta(., .)$ is the usual beta function. Then, in (2), and for any $k = 1, 2, \ldots, n$,

$$p(\alpha | k) \propto p(\alpha) \alpha^{k-1} (\alpha + n) \beta(\alpha + 1, n)$$

$$\propto p(\alpha) \alpha^{k-1} (\alpha + n) \int_0^1 x^\alpha (1-x)^{n-1} dx,$$

using the definition of the beta function. This implies that $p(\alpha | k)$ is the marginal distribution from a joint for $\alpha$ and a continuous quantity $x$ ($0 < x < 1$) such that

$$p(\alpha, x | k) \propto p(\alpha) \alpha^{k-1} (\alpha + n) x^\alpha (1-x)^{n-1}, \quad (0 < \alpha, 0 < x < 1).$$

Hence we have conditional posteriors $p(\alpha | x, k)$ and $p(x | \alpha, k)$ determined as follows. Firstly, under the $G(a, b)$ prior for $\alpha$,

$$p(\alpha | x, k) \propto \alpha^{a+k-2} (\alpha + n) e^{-\alpha (b - \log(x))}$$

$$\propto \alpha^{a+k-1} e^{-\alpha (b - \log(x))} + n \alpha^{a+k-2} e^{-\alpha (b - \log(x))}$$

for $\alpha > 0$, which reduces easily to a mixture of two gamma densities, viz

$$(\alpha | x, k) \sim \pi_x G(a + k, b - \log(x)) + (1 - \pi_x) G(a + k - 1, b - \log(x))$$

(4)

with weights $\pi_x$ defined by

$$\frac{\pi_x}{1 - \pi_x} = \frac{(a + k - 1)}{n(b - \log(x))}$$

(note that these distributions are well defined for all gamma priors, all $x$ in the unit interval and all $k > 1$). Secondly,

$$p(x | \alpha, k) \propto x^\alpha (1-x)^{n-1}, \quad (0 < x < 1)$$

(5)

so that $x | \alpha, k \sim B(\alpha + 1, n)$, a beta distribution with mean $(\alpha + 1)/(\alpha + n + 1)$.
It should now be clear how $\alpha$ can be sampled at each stage of the simulation — at each Gibbs iteration, the currently sampled values of $k$ and $\alpha$ allow us to draw a new value of $\alpha$ by (i) first sampling an $x$ value from the simple beta distribution (5), conditional on $\alpha$ and $k$ fixed at their most recent values; then (ii) sampling the new $\alpha$ value from the mixture of gammas in (4) based on the same $k$ and the $x$ value just generated in (i).

On completion of the simulation, we will have a series of sampled values of $k$, $\alpha$, $x$, and all the other parameters. Suppose that the Monte Carlo sample size is $N$, and denote the sampled values $k_s$, $x_s$, etc, for $s = 1, \ldots, N$. Only the sampled values $k_s$ and $x_s$ are needed in estimating the posterior $p(\alpha|D)$ via the usual Monte Carlo average of conditional posteriors, viz

$$p(\alpha|D) \approx N^{-1} \sum_{s=1}^{N} p(\alpha|x_s, k_s),$$

where the summands are simply the conditional gamma mixtures in equation (4).

2.3 Mixture of gamma priors for $\alpha$

Now consider cases in which the prior for $\alpha$ is a mixture of gammas, $\alpha \sim \sum_{i=1}^{h} c_i G(a_i, b_i)$ based on specified shapes $a_i$, scales $b_i$ and mixture weights $c_i$ that sum to unity. The above analysis generalises easily – the conditional posterior for $(x|\alpha, k)$ remains as in (5), but the two-component mixture (4) for $(\alpha|x, k)$ extends to the $2h$-component mixture

$$(\alpha|x, k) \sim \sum_{i=1}^{m} \pi_{i,x} G(a_i + k, b_i - \log(x)) + (1 - \pi_{i,x}) G(a_i + k - 1, b_i - \log(x))\right)$$

where, for $i = 1, \ldots, h$,

$$\frac{\pi_{i,x}}{1 - \pi_{i,x}} = \frac{(a_i + k - 1)}{n(b_i - \log(x))}$$

and

$$c_{i,x} \propto c_i \frac{\Gamma(a_i + k - 1)}{(b_i - \log(x))^{a_i + k - 1}} \{n + \frac{(a_i + k - 1)}{(b_i - \log(x))}\},$$

subject, of course, to unit sum. This extended mixture is trivially sampled.

3. AN ASYMPTOTIC RESULT

A simple asymptotic approximation to the posterior for $\alpha$ can be derived by developing an asymptotic approximation to the sampling density (1) directly. Take equation (1),

$$P(k|\alpha, n) = c_n(k)n!\alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \quad (k = 1, 2, \ldots, n),$$

subject, of course, to unit sum. This extended mixture is trivially sampled.

3. AN ASYMPTOTIC RESULT

A simple asymptotic approximation to the posterior for $\alpha$ can be derived by developing an asymptotic approximation to the sampling density (1) directly. Take equation (1),

$$P(k|\alpha, n) = c_n(k)n!\alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \quad (k = 1, 2, \ldots, n),$$
with $c_n(k) \propto |S_n^{(k)}|$ (using results in Antoniak 1974), where $S_n^{(k)}$ ($k = 1, \ldots, n$) are Stirling numbers of the first kind (Abramowitz & Stegun section 24.1.3, p824). As $n \to \infty$ and with $k = o(\log(n))$, Abramowitz & Stegun gives the asymptotic approximation

$$|S_n^{(k)}| \approx \frac{(n-1)!}{(k-1)!} (\gamma + \log(n))^{k-1}$$

where $\gamma$ is Euler’s constant. It trivially follows that, under these conditions, (1) reduces to

$$P(k|\alpha, n) \propto \alpha^k (\gamma + \log(n))^{k}/(k-1)!, \quad k = o(\log(n)),$$

and so the asymptotic Poisson distribution $k = 1 + X$ with $X$ Poisson with mean $\alpha(\gamma + \log(n))$. In our framework, $k$ is typically much smaller than $n$ when $n$ is large so the approximation should be useful; the Poisson distribution for numbers of groups is very nice indeed.

As a corollary, we have the following approximation to the posterior for $\alpha$ given $k$ (and $n$) discussed above. Taking the gamma prior $\alpha \sim G(a, b)$, the Poisson approximation here implies directly an approximate gamma posterior

$$p(\alpha|k, n) \approx G(a + k - 1, b + \gamma + \log(n))$$

for large $n$ and with $k = o(\log(n))$. Just how good these approximations are remains to be explored. The approximate posterior may be used to simulate values of $\alpha$ in the analysis, though the earlier exact mixture representation is only slightly more involved computationally.

REFERENCES


