DP Mixture inference derivations: an example

ziyisong

May 2020

We are familiar with finite Gaussian mixture model in Bayesian world. Routinely construct a generative model augmented with assignment variables and finite k different Gaussian distributions, give priors and observations, compute posterior distributions for all model parameters, directly estimate density, and evaluate predictive distributions.

Now we discuss these issues using data distributions derived as normal mixtures in the framework of Dirichlet processes. Besides dealing with above issues, as a natural by-product, we develop approaches to inference about the number of components and modes in a population distribution.

The important content and framework of this report was introduced by Escobar and West (1995). But detailed derivations and steps are skipped in their paper. Our report here provides you a whole story with every single derivation.

Contents

1 Common distribution and inference in Bayes

1.1 Common distribution

We first present several common distributions before showing the usual Bayesian inference formulas.

Gamma Distribution

shape-rate parameterization:

$$Gamma(x|\text{shape} = a, \text{rate} = b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb} \tag{1}$$

for $a, b, x > 0$

$$\begin{aligned}
\text{mean} &= \frac{a}{b} \\
\text{mode} &= \frac{a-1}{b} \text{ for } a \geq 1 \\
\text{var} &= \frac{a}{b^2}
\end{aligned}$$

Inverse Gamma Distribution

Let $X \sim Gamma(\text{shape} = a, \text{rate} = b)$ and $Y = \frac{1}{X}$, then $Y \sim IG(\text{shape} = a, \text{scale} = b)$, where inverse Gamma distribution is

$$IG(x|\text{shape} = a, \text{scale} = b) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} e^{-\frac{b}{x}} \tag{2}$$

for $a, b, x > 0$

$$\begin{aligned}
\text{mean} &= \frac{b}{a-1} \text{ for } a > 1 \\
\text{mode} &= \frac{b}{a+1} \\
\text{var} &= \frac{b^2}{(a-1)^2(a-2)} \text{ for } a > 2
\end{aligned}$$

Student-t distribution
pdf:

$$t_\nu(x|\mu,\sigma^2) = c \left[1 + \frac{1}{\nu}(\frac{x-\mu}{\sigma})^2\right]^{-(\frac{\nu+1}{2})}$$
$$c = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\nu\pi}\sigma} \tag{3}$$

where c is the normalization constant, and ν is the degree of freedom.

$$\text{mean} = \mu$$
$$\text{mode} = \mu$$
$$var = \frac{\nu\sigma^2}{\nu-2} \text{ for } \nu > 2$$

Student-t distribution is like an infinite sum of Gaussians, where each Gaussian has a different precision:

$$\int N(x|\mu,\tau^{-1})\ Gamma(\tau|a,\text{rate}=b)d\tau = t_{2a}\left(x|\mu,\frac{b}{a}\right)$$

1.2 Common Bayesian Inference

Let $D = (x_1,\ldots,x_n)$ be the data. Assume data follows a Gaussian distribution, we estimate its parameters using Bayesian inference and conjugate priors.

1.2.1 Normal-inverse-Gamma (NIG) prior

likelihood

$$p(D|\mu,\sigma^2) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \tag{4}$$

prior

$$\begin{aligned} p(\mu,\sigma^2) &= NIG(m_0,V_0,a_0,b_0) \\ &= N(\mu|m_0,\sigma^2 V_0)\ IG(\sigma^2|a_0,b_0) \end{aligned} \tag{5}$$

posterior

posterior is also NIG

$$\begin{aligned} p(\mu,\sigma^2|D) &= NIG(m_n,V_n,a_n,b_n) \\ V_n^{-1} &= V_0^{-1}+n \\ \frac{m_n}{V_n} &= V_0^{-1}m_0+n\overline{X} \\ a_n &= a_0+n/2 \\ b_n &= b_0+\frac{1}{2}\left[m_0^2V_0^{-1}+\sum_i x_i^2-m_n^2V_n^{-1}\right] \end{aligned} \tag{6}$$

marginal likelihood

$$\begin{aligned} p(D) &= \frac{\Gamma(a_n)}{\Gamma(a_0)}\sqrt{\frac{V_n}{V_0}}\frac{(2b_0)^{a_0}}{(2b_n)^{a_n}}\frac{1}{\pi^{n/2}} \\ &= \frac{|V_n|^{1/2}}{|V_0|^{1/2}}\frac{b_0^{a_0}}{b_n^{a_n}}\frac{\Gamma(a_n)}{\Gamma(a_0)}\frac{1}{\pi^{n/2}2^{n/2}} \end{aligned} \tag{7}$$

posterior predictive

$$p(x|D) = t_{2a_n}\left(m_n,\frac{b_n(1+V_n)}{a_n}\right) \tag{8}$$

1.2.2 Normal prior

Let us consider Bayesian estimation of the mean, variances are assumed to be known but different.

likelihood

Let $D = (x_1, \ldots, x_n)$ be the data. The likelihood is

$$p\left(D|\mu, \sigma^2\right) = \prod_{i=1}^{n} p\left(x_i|\mu, \sigma_i^2\right) = \left(2\pi\sigma_i^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma_i^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

prior

The natural conjugate prior for mean has the form

$$p(\mu) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \propto \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right)$$

posterior

The posterior still follows a Gaussian distribution with

$$\sigma_n^2 = \frac{1}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2} + \frac{1}{\sigma_0^2}}$$

$$\frac{\mu_n}{\sigma_n^2} = \sum_{i=1}^{n}\frac{x_i}{\sigma_i^2} + \frac{\mu_0}{\sigma_0^2}$$

2 Bayesian density estimation and inference in DP Mixture

The basic normal mixture model is described as follows. Suppose that data $Y_1, \ldots, Y_n$ are conditionally independent and normally distributed, $(Y_i|\pi_i) \sim N(\mu_i, V_i)$, with means μ_i and variances V_i determining the parameters $\pi_i = (\mu_i, V_i), i = 1, \ldots, n$. Suppose further that the π_i come from some prior distribution on $\Re \times \Re^+$. Having observed data $D_n = \{y_i, \ldots, y_n\}$, with y_i the observed value of Y_i, the distribution of a future case is a mixture of normals; the relevant density function $Y_{n+1} \sim N(\mu_{n+1}, V_{n+1})$ mixed with respect to the posterior predictive distribution for $(\pi_{n+1}|D_n)$. If the common prior distribution for the π_i is uncertain and modeled as a Dirichlet process, then the data come from a Dirichlet mixture of normals.

$$\begin{aligned} G &\sim Dir(\alpha G_0) \\ \pi_i|G &\sim G \\ Y_i &\sim N(y_i|\pi_i) \end{aligned} \tag{9}$$

Suppose that the normal means and variances π_i come from some prior distribution $G(\cdot)$ on $\Re \times \Re^+$. If $G(\cdot)$ is uncertain and modeled as a Dirichlet process, then the data come from a Dirichlet mixture of normals. In particular, we suppose that $G \sim Dir(\alpha G_0)$, a Dirichlet process defined by α, a positive scalar, and $G_0(\cdot)$, a specified bivariate distribution function over $\Re \times \Re^+$. $G_0(\cdot)$ is the prior expectation of $G(\cdot)$. Write $\pi = \{\pi_1, \ldots, \pi_n\}$.

The setup is described as follows. The model will have a prior after we give bivariate distribution function $G_0(\cdot)$ a prior distribution. Provided observations $\{y_1, \ldots, y_n\}$, we can compute unconditional predictive distribution $P(Y_{n+1}|\pi)$, and the Bayesian prediction, or density estimation, problem is solved by summarizing the conditional predictive distribution with respect to the posterior $(\pi|D_n)$.

A key feature of the model structure, and of its analysis, relates to the discreteness of $G(\cdot)$ under the Dirichlet process assumption. Let $\pi^{(i)}$ be π without π_i, $\pi^{(i)} = \{\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n\}$.

Then the conditional prior for $(\pi_i|\pi^{(i)})$ is

$$(\pi_i|\pi^{(i)}) \sim \alpha a_{n-1} G_0(\pi_i) + a_{n-1} \sum_{j=1, j\neq i}^{n} \delta_{\pi_j}(\pi_i) \tag{10}$$

where $a_r = 1/(\alpha + r)$ for positive integers r. Similarly, the distribution of $(\pi_{n+1}|\pi)$ is given by

$$(\pi_{n+1}|\pi) \sim \alpha a_n G_0(\pi_{n+1}) + a_n \sum_{i=1}^{n} \delta_{\pi_i}(\pi_{n+1}) \tag{11}$$

. Thus, given π, a sample of size n from $G(\cdot)$, the next case π_{n+1} represents a new, distinct value with probability αa_n and is otherwise drawn uniformly from among the first n values. These first n values with positive probability will reduce to some $k < n$ distinct values. Write the k distinct values among the n elements of π as $\pi_j^* = (\mu_j^*, V_j^*),\ j = 1, \ldots, k$; there are n_j occurrences of π_j^* with $n_1 + \cdots + n_k = n$. Thus

$$(\pi_{n+1}|\pi) \sim \alpha a_n G_0(\pi_{n+1}) + a_n \sum_{j=1}^{k} n_j \delta_{\pi_j^*}(\pi_{n+1}). \tag{12}$$

Antoniak (1974) gave the prior for k induced by this Dirihclet process model. The prior distribution for k depends critically on α, stochastically increasing with α.

To proceed, we need to specify the prior mean $G_0(\cdot)$ of $G(\cdot)$. A convention form is the Normal-inverse-Gamma conjugate to the normal sampling model. With respect to predicting Y_{n+1}, it is clear that $P(Y_{n+1}|\pi, D_n) \equiv P(Y_{n+1}|\pi)$.

$$\begin{aligned}
\underbrace{P(Y_{n+1}|\pi)}_{\text{marginal likelihood}} &= \int P(Y_{n+1}, \pi_{n+1}|\pi) d\pi_{n+1} \\
&= \int \underbrace{P(Y_{n+1}|\pi_{n+1})}_{likelihood} \underbrace{P(\pi_{n+1}|\pi)}_{prior} d\pi_{n+1} \\
&= \int P(Y_{n+1}|\pi_{n+1}) dP(\pi_{n+1}|\pi)
\end{aligned}$$

Given that

$$(\pi_{n+1}|\pi) \sim \alpha a_n G_0(\pi_{n+1}) + a_n \sum_{i=1}^{n} \delta_{\pi_i}(\pi_{n+1})$$

For the first part:

- **prior**: Normal-inverse-Gamma $G_0(\pi_{n+1}) \sim N(\mu_j; m, \tau V_j) \times IG(V_j; s/2, S/2)$; for the moment, assume that the prior parameters $s, S, m,$ and τ are specified.
- **likelihood**: Normal $Y_j \sim N(\mu_j, V_j)$

By Section 1.2.1, we compute the first part of the marginal likelihood $P(Y_{n+1}|\pi)$ as follows:

$$\begin{aligned}
&= \frac{(\frac{\tau}{1+\tau})^{1/2}}{\tau^{1/2}} \frac{\Gamma(\frac{s+1}{2})}{\Gamma(\frac{s}{2})} \frac{1}{\pi^{\frac{1}{2}}2^{\frac{1}{2}}} \frac{(S/2)^{s/2}}{\left(\frac{S}{2}+\frac{1}{2}\left[\frac{m^2}{\tau}+y^2-(\frac{m}{\tau}+y)^2\frac{\tau}{1+\tau}\right]\right)^{\frac{s+1}{2}}} \\
&= \frac{\Gamma(\frac{s+1}{2})}{\Gamma(\frac{s}{2})} \frac{1}{\sqrt{\pi}\sqrt{(1+\tau)S}} \frac{(\frac{S}{2})^{\frac{s+1}{2}}}{\left[\frac{S}{2}+\frac{1}{2}\frac{(y-m)^2}{1+\tau}\right]^{\frac{s+1}{2}}} \\
&= \frac{\Gamma(\frac{s+1}{2})}{\Gamma(\frac{s}{2})} \frac{1}{\sqrt{\pi}\sqrt{(1+\tau)S}} \left[1+\frac{(y-m)^2}{(1+\tau)S}\right]^{-\frac{s+1}{2}} \\
&= \frac{\Gamma(\frac{s+1}{2})}{\Gamma(\frac{s}{2})} \frac{1}{\sqrt{s\pi}\sqrt{\frac{(1+\tau)S}{s}}} \left[1+\frac{1}{s}\frac{(y-m)^2}{\frac{(1+\tau)S}{s}}\right]^{-\frac{s+1}{2}} \\
&= \frac{\Gamma(\frac{s+1}{2})}{\Gamma(\frac{s}{2})} \frac{1}{\sqrt{s\pi}M^{1/2}} \left[1+\frac{1}{s}\frac{(y-m)^2}{M}\right]^{-\frac{s+1}{2}} \\
&= T_s\left(y_{n+1}|m, M\right)
\end{aligned}$$

where $M = \frac{(1+\tau)S}{s}$.

The second part of the marginal likelihood $P(Y_{n+1}|\pi)$ is:

$$\begin{aligned}
\int P(Y_j|\pi_j)\delta_{\pi_i}(\pi_j)d\pi_j &= P(Y_i|\pi_i) \\
&= N(\mu_i, V_i)
\end{aligned}$$

These imply

$$(Y_{n+1}|\pi) \sim \alpha a_n T_s(m, M) + a_n \sum_{i+1}^{n} N(\mu_i, V_i). \qquad (13)$$

Equivalently, we have

$$(Y_{n+1}|\pi) \sim \alpha a_n T_s(m, M) + a_n \sum_{j=1}^{k} n_j N(\mu_j^*, V_j^*) \tag{14}$$

The Bayesian prediction, or density estimation, problem is solved by summarizing the unconditional predictive distribution

$$P(Y_{n+1}|D_n) = \int P(Y_{n+1}|\pi) dP(\pi|D_n) \tag{15}$$

The integral is difficult because $\pi = \{\pi_1, \ldots, \pi_n\}$. Direct evaluation is extremely computationally involved for even rather small sample size n, due to the inherent complexity of the posterior $P(\pi|D_n)$. Use Monte Carlo to sample $(\pi|D_n)$ and approximate $P(Y_{n+1}|D_n)$.

Recall that for each i, $\pi^{(i)} = \{\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n\}$. For each i, the conditional posterior for

$$\begin{aligned}\left(\pi_i|\pi_{(i)}, D_n\right) &= \left(\pi_i|\pi^{(i)}, y_i\right) \\ &\sim \frac{\left(\pi_i|\pi^{(i)}\right) \times \left(y_i|\pi_i, \pi^{(i)}\right)}{\left(y_i|\pi^{(i)}\right)} \\ &= \frac{\left(\pi_i|\pi^{(i)}\right) \times (y_i|\pi_i)}{(y_i)}\end{aligned}$$

Recall that

$$(\pi_i|\pi^{(i)}) \sim \alpha a_{n-1} G_0(\pi_i) + a_{n-1} \sum_{j=1, j\neq i}^{n} \delta_{\pi_j}(\pi_i).$$

Thus

$$\left(\pi_i|\pi_{(i)}, D_n\right) \propto \left(\pi_i|\pi^{(i)}\right) \times (y_i|\pi_i).$$

For the first part:

- **prior**: Normal-inverse-Gamma $\pi_i|\pi^{(i)} \sim N(\mu_i; m, \tau V_i) \times IG(V_i; s/2, S/2)$; for the moment, assume that the prior parameters $s, S, m,$ and τ are specified.
- **likelihood**: Normal $Y_i|\pi_i \sim N(\mu_i, V_i)$

According to Section 1.2.1, posterior distributions for parameters μ_i, V_i are

$$\begin{aligned} V_i^{-1} &\sim Gamma\left(\frac{s+1}{2},\ \frac{S}{2} + \frac{(y_i-m)^2}{2(1+\tau)}\right) \\ \mu_i|V_i &\sim N\left(\left(\frac{m}{\tau}+y_i\right)\frac{\tau}{1+\tau},\ \frac{\tau}{1+\tau}V_i\right) \end{aligned}$$

$$G_i(\pi_i) = \pi_i|y_i = (\mu_i, V_i)|y_i \sim IG\left(V_i|\frac{s+1}{2}, \frac{S_i}{2}\right) \times N\left(\mu_i|x_i, XV_i\right)$$

where $G_i(\pi_i)$ is the posterior and

$$\begin{aligned} S_i &= S + \frac{(y_i-m)^2}{1+\tau} \\ X &= \frac{\tau}{1+\tau} \\ x_i &= \frac{m+y_i}{1+\tau} = \left(\frac{m}{\tau}+y_i\right)\frac{\tau}{1+\tau} \end{aligned}$$

. Similar to we have derived before, the marginal likelihood, which is also the normalization constant is

$$T_s\left(y_i|m, M\right) = \frac{\Gamma(\frac{s+1}{2})}{\Gamma(\frac{s}{2})}\frac{1}{\sqrt{s\pi}M^{1/2}}\left[1+\frac{1}{s}\frac{(y_i-m)^2}{M}\right]^{-\frac{s+1}{2}}$$

where $M = \frac{(1+\tau)S}{s}$.

For the second part:

- **prior** $\delta_{\pi_j}(\pi_i)$
- **likelihood** $N(y_i|\pi_i) = N(y_i|\mu_i, V_i)$
- **marginal likelihood** $\int N(y_i|\pi_i)\delta_{\pi_j}(\pi_i)d\pi_i = N(y_i|\pi_j)$
- **posterior** $\frac{N(y_i|\pi_i)\delta_{\pi_j}(\pi_i)}{N(y_i|\pi_j)} = \frac{N(y_i|\pi_j)\delta_{\pi_j}(\pi_i)}{N(y_i|\pi_j)} = \delta_{\pi_j}(\pi_i)$

so normalization constant for the second part is the marginal likelihood $N(y_i|\pi_j) = N(y_i|\mu_j, V_j)$.

Therefore the conditional posterior for $(\pi_i|\pi^{(i)}, D_n)$ is the mixture

$$(\pi_i|\pi^{(i)}, D_n) \sim \text{const } T_s(y_i|m, M)G_i(\pi_i) + \sum_{j=1, j\neq i}^{n} \text{const } N(y_i|\mu_j, V_j)\delta_{\pi_j}(\pi_i)$$

equivalently,

$$\left(\pi_i|\pi^{(i)}, D_n\right) \sim q_0 G_i\left(\pi_i\right) + \sum_{j=1, j\neq i}^{n} q_j \delta_{\pi_j}\left(\pi_i\right) \qquad (16)$$

where

(a) $G_i\left(\pi_i\right)$ is the bivariate Normal-inverse-Gamma distribution whose components are $V_i^{-1} \sim G\left((1+s)/2, S_i/2\right)$ with $S_i = S + \left(y_i - m\right)^2/(1+\tau)$, and $(\mu_i|V_i) \sim N\left(x_i, XV_i\right)$ with $X = \tau/(1+\tau)$ and $x_i = \left(m\ +\tau y_i\right)/(1+\tau)$;

(b) the weights q_j are defined as

$$q_0 \propto \alpha \mathbf{c}(s) \left[1 + \left(y_i - m\right)^2/(sM)\right]^{-(1+s)/2} /M^{1/2}$$

and

$$q_j \propto \exp\left\{-\left(y_i - \mu_j\right)^2/\left(2V_j\right)\right\}\left(2V_j\right)^{-1/2} \quad j = 1, \ldots, n; j \neq i$$

subject to $q_0+\cdots+q_{i-1}+q_{i+1}+\cdots+q_n=1$ with $M=(1+\tau)S/s$ and $\mathbf{c}(s)=\Gamma((1+s)/2)\Gamma(s/2)^{-1}s^{-1/2}$.

Here $G_i(\cdot)$ is just the posterior distribution of $(\pi_i|y_i)$ under a prior $G_0(\cdot)$, and the weight q_0 is proportional to α times the marginal density of Y_i evaluated at the datum y_i using $G_0(\cdot)$ as the prior for π_i. In our model, therefore, q_0 is proportional to α times the density function of $T_s(m, M)$ evaluated at y_i. The weight q_j is proportional to the likelihood of data y_i being a sample from the normal distribution $(Y_i|\pi_j)$ or just the density function of $N(\mu_j, V_j)$ at the point y_i.

Algorithm 2.1. 1. Choose a starting value of π; reasonable initial values are samples from the individual conditional posteriors $G_i(\cdot)$ in

$$\left(\pi_i|\pi^{(i)}, D_n\right) \sim q_0 G_i\left(\pi_i\right) + \sum_{j=1, j\neq i}^{n} q_j \delta_{\pi_j}\left(\pi_i\right)$$

2. Sample elements of π sequentially by drawing from the distribution of $(\pi_1|\pi^{(1)}, D_n)$, then $(\pi_2|\pi^{(2)}, D_n)$, and so on up to $(\pi_n)|\pi^{(n)}, D_n)$, with the relevant elements of the most recently sampled $\pi_{(i)}$ values inserted in the conditioning vectors at each step.

3. Return to Step 2 and proceed iteratively until convergence.

The sampling process is computationally very straightforward. Note that in implementation, the required computations are reduced through the fact that each of the mixtures

$$\left(\pi_i|\pi^{(i)}, D_n\right) \sim q_0 G_i\left(\pi_i\right) + \sum_{j=1, j\neq i}^{n} q_j \delta_{\pi_j}\left(\pi_i\right)$$

will reduce to typically fewer than the apparent n components, due to the clustering of the elements of $\pi^{(i)}$. Using the earlier superscript $*$ to denote distinct values, suppose that the conditioning quantities $\pi^{(i)}$ in

$$\left(\pi_i|\pi^{(i)}, D_n\right) \sim q_0 G_i\left(\pi_i\right) + \sum_{j=1, j\neq i}^{n} q_j \delta_{\pi_j}\left(\pi_i\right)$$

concentrate on $k_i \leq n-1$ distinct values $\pi_j^* = (\mu_j^*, V_j^*)$, with some n_j taking this common value. Then

$$\left(\pi_i|\pi^{(i)}, D_n\right) \sim q_0 G_i\left(\pi_i\right) + \sum_{j=1, j\neq i}^{n} q_j \delta_{\pi_j}\left(\pi_i\right)$$

reduces to $(\pi_i|\pi_{(i)}, D_n) \sim q_0 G_i(\pi_i) + \sum_{j=1}^{k_i} q_j^* \delta_{\pi_j^*}(\pi_i)$, where the weights now include the n_j, viz., $q_j \propto n_j exp\left\{-(y_i - \mu_j^*)^2/(2V_j^*)\right\}(2V_j^*)^{-1/2}$. The sampling process results in an approximate draw from $p(\pi|D_n)$.

For this DP Mixture model, the prior in our setup is the Normal-inverse-Gamma $G_0(\pi_i) \sim N(\mu_i; m, \tau V_i) \times IG(V_i; s/2, S/2)$; so far, we assume that the prior hyperparameters $s, S, m,$ and τ are specified. The initial prior variance τ plays a critical role in determining the extent of smoothing in the analysis. For a given k distinct values among the elements of π, a larger value of τ leads to increased dispersion among the k group means μ_j^*, which, for fixed V_j^*, leads to a greater chance of multimodality in the resulting predictive distribution.

To learn about the prior hyperparameters m and/or τ, suppose independent priors of the form $m \sim N(a, A)$ and $\tau^{-1} \sim Gamma\left(w/2, W/2\right)$, for some specified hyperparameters $a, A, w,$ and W.

(a) Given τ and π, m is conditionally independent of D_n.

$$\underbrace{p(m|\pi,\tau)}_{posterior} \propto \underbrace{p(m)}_{prior} \times \underbrace{\prod_{j=1}^{k} \pi_j^*|m,\tau}_{likelihood}$$

where

$$\begin{aligned}
p(m) &= \frac{1}{\sqrt{2\pi}\sqrt{A}} exp\left[-\frac{1}{2A}(m-a)^2\right] \\
\prod_{j=1}^{k} \pi_j^*|m,\tau &= \prod_{j=1}^{k} N(\mu_j^*|m,\tau V_j^*) \\
&= \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi}\sqrt{\tau V_j^*}} exp\left[-\frac{1}{2\tau_j^*(\mu_j^*-m)^2}\right]
\end{aligned}$$

According to Section 1.2.2,

$$\begin{aligned}
\sigma_n^2 &= \frac{1}{\frac{1}{A}+\sum_{j=1}^{k}\frac{1}{\tau V_j^*}} \\
&= \frac{A\tau}{\tau + A\sum_{j=1}^{k}\frac{1}{V_j^*}} \\
&= \frac{A\tau\overline{V}}{\tau\overline{V}+A},
\end{aligned}$$

where $\overline{V}^{-1} = \sum_{j=1}^{k} \frac{1}{V_j^*}$ and $\overline{V} = \frac{1}{\sum_{j=1}^{k} \frac{1}{V_j^*}}$.

$$\begin{aligned}
\frac{\mu_n}{\sigma_n^2} &= \sum_{j=1}^{k} \frac{\mu_j^*}{\tau V_j^*} + \frac{a}{A} \\
\mu_n &= \frac{A\tau\overline{V}}{\tau\overline{V} + A} \left(\frac{1}{\tau} \sum_{j=1}^{k} \frac{\mu_j^*}{V_j^*} + \frac{a}{A} \right) \\
&= a\frac{\tau\overline{V}}{A + \tau\overline{\overline{V}}} + \frac{A}{A + \tau\overline{\overline{V}}}\overline{V} \sum_{j=1}^{k} \frac{\mu_j^*}{V_j^*} \\
&= (1-x)a + x\overline{V} \sum_{j=1}^{k} \frac{\mu_j^*}{V_j^*}
\end{aligned}$$

where $x = \frac{A}{A+\tau\overline{\overline{V}}}$.

(b) Given m and π, τ is conditionally independent of D_n.

$$p(\tau|m,\pi) \propto p(\tau) \times \prod_{j=1}^{k} \pi_j^*|\tau, m$$

where

$$\tau^{-1} \sim Gamma(w/2, W/2)$$

$$\prod_{j=1}^{k} \pi_j^*|\tau, m \sim \prod_{j=1}^{k} N(\mu_j^*|m, \tau V_j^*) = \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi}\sqrt{\tau V_j^*}} exp\left[-\frac{1}{2\tau V_j^*} \left(\mu_j^* - m)^2\right) \right]$$

thus, posterior

$$p(\tau|m,\pi) \propto (\tau^{-1})^{\frac{w}{2}-1} e^{-(\tau^{-1})\frac{W}{2}} \tau^{-\frac{k}{2}} exp\left[-\frac{1}{2\tau} \underbrace{\sum_{j=1}^{k} \frac{(\mu_j^* - m)^2}{V_j^*}}_{K} \right]$$

$$\propto \left(\tau^{-1}\right)^{\frac{w+k}{2}-1} e^{-(\tau-1)\frac{W+K}{2}}$$

so $\tau^{-1} \sim Gamma\left(\frac{w+k}{2}, \frac{W+K}{2}\right)$, where $K = \sum_{j=1}^{k} \frac{\left(\mu_j^* - m\right)^2}{V_j^*}$.

Incorporating m and/or τ into the iterative resampling scheme provides for sampling from the complete joint posterior of $(\pi, m, \tau|D_n)$.

Algorithm 2.2. 1. Generate an initial π conditional on a preliminary chosen value of m and τ. Reasonable initial values π are samples from the individual conditional posteriors $G_i(\pi_i)$ in

$$\left(\pi_i|\pi^{(i)}, D_n\right) \sim q_0 G_i\left(\pi_i\right) + \sum_{j=1, j\neq i}^{n} q_j \delta_{\pi_j}\left(\pi_i\right)$$

where

$$V_i^{-1} \sim Gamma\left(\frac{s+1}{2}, \ \frac{S}{2} + \frac{(y_i - m)^2}{2(1+\tau)}\right)$$

$$\mu_i|V_i \sim N\left(\left(\frac{m}{\tau} + y_i\right)\frac{\tau}{1+\tau}, \ \frac{\tau}{1+\tau}V_i\right)$$

$$G_i(\pi_i) = \pi_i|y_i = (\mu_i, V_i)|y_i \sim IG\left(V_i|\frac{s+1}{2}, \frac{S_i}{2}\right) \times N\left(\mu_i|x_i, XV_i\right)$$

2. Sample m and τ in some order using the relevant distribu-

tions as just described.

$$m|\pi, \tau \sim N(\mu_n, \sigma_n^2) \text{ with } \sigma_n^2 = \frac{A\tau\overline{V}}{\tau\overline{V} + A} \text{ and } \mu_n = (1-x)a + x\overline{V}\sum_{j=1}^{k}\frac{\mu_j^*}{V_j^*}$$

$$\tau^{-1}|\pi, m \sim Gamma\left(\frac{w+k}{2}, \frac{W+K}{2}\right), \text{ where } K = \sum_{j=1}^{k}\frac{\left(\mu_j^* - m\right)^2}{V_j^*}$$

3. Using the most recently sampled values of m and τ, sample elements of π sequentially by drawing from the distribution of $(\pi_1|\pi^{(1)}, D_n)$, then $(\pi_2|\pi^{(2)}, D_n)$, and so on up to $(\pi_n|\pi^{(n)}, D_n)$, with the relevant elements of the most recently sampled $\pi^{(i)}$ values inserted in the conditioning vectors at each step

4. Return to Step 2, and proceed iteratively until convergence.

From specified initial values, we first iterate the sampling procedure to "burn-in" the process to (approximate) convergence. Following burn-in, successively generated values of $\pi, m,$ and τ are ssumed to be drawn from the posterior; denote these values by $(\pi(r), m(r), \tau(r))$, for $r = 1, \ldots, N$, where N is the specified simulation sample size required. Approximate predictive inference now follows through the Monte Carlo approximation to

$$\begin{aligned} P(Y_{n+1}|D_n) &= \int P(Y_{n+1}|\pi)dP(\pi|D_n) \\ &= E_{\pi|D_n}\left[P(Y_{n+1}|\pi)\right] \end{aligned}$$

given by

$$P(Y_{n+1}|D_n) \approx N^{-1}\sum_{r=1}^{N} P(Y_{n+1}|\pi(r), m(r), \tau(r)), \qquad (17)$$

with the summands given by the mixtures in

$$(Y_{n+1}|\pi) \sim \alpha a_n T_s(m, M) + a_n \sum_{j=1}^{k} n_j N(\mu_j^*, V_j^*),$$

and the notation now explicitly recognizes the dependence on the sampled values of m and τ. Additional information available includes the sampled values of k, $\{k(r), r = 1, \ldots, N\}$, which directly provide a histogram approximation to $p(k|D_n)$, of interest in assessing the number of components. The posteriors for m and/or τ may also be approximated by mixture of their conditional posteriors noted earlier. For m, this leads to the mixture of normals $p(m|D_n) \approx N^{-1}\Sigma p(m|\tau(r), \pi(r))$; for τ, to the mixture of inverse gammas $p(\tau|D_n) \approx N^{-1}\Sigma p(\tau|m(r), \pi(r))$, the sums being over $r = 1, \ldots, N$ in each case.

The precision parameter α of the underlying Dirichlet process is a critical smoothing parameter for the model. Learning about α from the data may be addressed with a view to incorporating α into the Gibbs sampling analysis. Using results of Antoniak (1974), the prior distribution of k can be written as

$$P(k|\alpha, n) = c_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \; k = 1, 2, \ldots, n \tag{18}$$

and $c_n(k) = P(k|\alpha = 1, n)$, not involving α. If required, the factors $c_n(k)$ are easily computed using recurrence formulae for Stirling numbers.

Now suppose that we have sampled values of the parameters π_i. By sampling the parameters π_i, we have in fact sampled a value for k, the number of distinct components, and have also sampled a specific configuration of the data D_n into k groups. From our model, the data are initially conditionally independent of α when k, π, and the configuration are known, and the parameters π are also conditionally independent of α when k and the configuration are known. So

$$p(\alpha|k, \pi, D_n) = p(\alpha|k) \propto p(\alpha)P(k|\alpha),$$

with likelihood function given in

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

Since $\Gamma(\alpha+n)\Gamma(n) = \Gamma(\alpha+n)B(\alpha, n)$ and $B(\alpha+1, n) = B(\alpha, n)\frac{\alpha}{\alpha+n}$, then

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

So

$$\begin{aligned} p(\alpha|k) &\propto p(\alpha)P(k|\alpha) \\ &\propto p(\alpha)\alpha^k \frac{B(\alpha+1, n)}{\Gamma(n)}\frac{(\alpha+n)}{\alpha} \\ &\propto p(\alpha)\alpha^{k-1}(\alpha+n)B(\alpha+1, n) \\ &\propto p(\alpha)\alpha^{k-1}(\alpha+n)\int_0^1 x^\alpha(1-x)^{n-1}dx \end{aligned}$$

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

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

Suppose prior $\alpha \sim Gamma(a,b)$, viz., $p(\alpha) = \frac{1}{b^a\Gamma(a)}\alpha^{a-1}e^{-\alpha b}$. Firstly, under the $Gamma(a,b)$ prior for α,

$$\begin{aligned} p(\alpha|x,k) &\propto p(\alpha,x|k) \\ &\propto \alpha^{a-1}e^{-\alpha b}\alpha^{k-1}(\alpha+n)x^{\alpha}(1-x)^{n-1} \\ &\propto \alpha^{a+k-2}(\alpha+n)e^{-\alpha(b-log(x))} \\ &\propto \alpha^{a+k-1}e^{-\alpha(b-log(x))} + n\alpha^{a+k-2}e^{-\alpha(b-log(x))}, \end{aligned}$$

which reduces easily to a mixture of two Gamma densities, viz

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

with weights π_x defined by

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

because

$$\frac{\Gamma(\alpha+k)}{(b-\mathrm{log}x)^{a+k}}\frac{(b-\mathrm{log}x)^{a+k-1}}{\Gamma(\alpha+k-1)} = \frac{a+k-1}{b-\mathrm{log}x}$$

Secondly,

$$\begin{aligned} p(x|\alpha,k) &= \frac{p(\alpha,x|k)}{p(\alpha|k)} \\ &\propto \frac{p(\alpha)\alpha^{k-1}(\alpha+n)x^{\alpha}(1-x)^{n-1}}{p(\alpha)\alpha^{k-1}(\alpha+n)\int_0^1 x^{\alpha}(1-x)^{n-1}dx} \\ &\propto x^{\alpha}(1-x)^{n-1} \end{aligned}$$

so that $(x|\alpha,k) \sim Beta(\alpha+1,n)$, a beta distribution with mean $(\alpha+1)/(\alpha+n+1)$.

It is now clear how α can be sampled at each stage of the simulation. At each Gibbs iteration, the currently sampled values of k and α allow us to draw a new value of α by (a.) first sampling an x value from the simple beta distribution

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

conditional on α and k fixed at their most recent values;

(b.) then sampling the new α value from the mixture of gammas in

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

based on the same k and the x value just generated in (a).

On completion of the simulation, $p(\alpha|D_n)$ will be estimated by the usual Monte Carlo average of conditional forms

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

viz.

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

where x_s are the sampled values of x.