Collapsed Gibbs sampler

Stephen Coleman

February 10, 2020

1 Gibbs sampling

Consider Gibbs sampling of some vector of variables $\theta = (\theta_1, \ldots, \theta_p)$. Gibbs sampling works by iterating over each variable to be predicted, updating it based upon the current values of all the other variables and then repeating this a large number of times. Let $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_n^{(j)})$ be the predicted values of θ in the j^{th} iteration of Gibbs sampling. Our update for $\theta_i^{(j)}$ is conditioned on all the current values for the other variables - this means that the first $(i-1)$ variables have already been updated j times, but the remaining $p-i$ variables are still based upon the $(j-1)^{th}$ iteration, i.e. our update probability is of the form:

$$p\left(\theta_i^{(j)} | \theta_1^{(j)}, \ldots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \ldots, \theta_p^{(j-1)}\right)$$

Now, consider Gibbs sampling for a mixture of K components for data $x = (x_1, \ldots, x_n)$, allocation variables $z = (z_1, \ldots, z_n)$, component parameters $\theta = (\theta_1, \ldots, \theta_K)$, and component weights $\pi = (\pi_1, \ldots, \pi_K)$. Let x_{-i} indicate the vector $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ and similarly for z_{-i}. Let our model be that described in figure 1.

We are interested in the sampling of the z variables. First recall that:

$$\begin{aligned}
p(A, B|C) &= p(B|A, C)p(A|C) && (1)\\
p(A|B, C) &= \frac{p(A, B|C)}{P(B|C)} && (2)\\
&= \frac{p(B|A, C)p(A|C)}{P(B|C)} && (3)\\
p(A|C) &= \int_B p(A|B', C)p(B'|C)dB' && (4)
\end{aligned}$$

Now consider the sampling of z_i. As this can only hold a relatively small number of values we can consider the probability for each possible k. From our hierarchical model in figure 1 and equations 3 and 4:

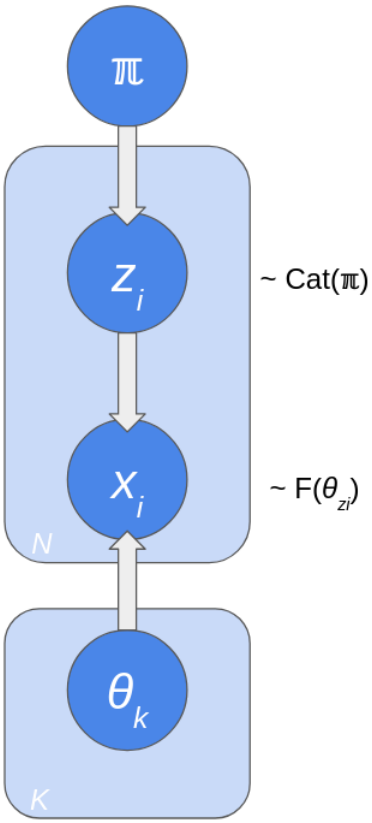

Figure 1: Hierarchical model for mixture model. Further hyperparameters can be included, but for our purposes of describing a collapsed Gibbs sampler this is sufficient.

$$
\begin{aligned}
p(z_i = k|x, z_{-i}, \pi) &= \frac{p(z_i = k|\pi, x_{-i}, z_{-i}, \pi)p(x_i|z, x_{-i}, \pi)}{p(x_i|x_{-i}, z_{-i}, \pi)} && (5)\\
&\propto p(z_i = k|\pi_k) \int_\theta p(x_i|\theta, z, x_{-i}, \pi)p(\theta|z, x_{-i}, \pi)d\theta && (6)\\
&= \pi_k \int_\theta p(x_i|\theta)p(\theta|z, x_{-i})d\theta && (7)
\end{aligned}
$$

Note that $p(x_i|x_{-i}, z_{-i}, \pi)$ in the denominator is independent of z_i and thus the same for all values of k.

The integral in equation 7 is the posterior predictive distribution for x_i given the other observations, x_{-i}. Thus, one may think of this as how well each component predicts x_i.

An alternative way of describing this involves the ratio of marginal likelihoods. As we are component specific (given $z_i = k$), I drop the z and π from my conditional and assume we are referring only to the x_j for which $z_j = k$.

$$p(x_i|z, x_{-i}, \pi) = \frac{p(x|z)}{p(x_{-i}|z)} \tag{8}$$

$$= \frac{\int_\theta p(x|\theta)p(\theta)d\theta}{\int_\theta p(x_{-i}|\theta)p(\theta)d\theta} \tag{9}$$

Therefore we can write the posterior predictive distribution as this ratio of marginal likelihoods:

$$p(z_i = k|x, z_{-i}, \pi) \propto \pi_k \frac{p(x)}{p(x_{-i})} \tag{10}$$

Thus we can create a K-vector of probabilities for the allocation of x_i to each component by finding the ratio of marginal likelihoods for each component including and excluding x_i, and multiplying these by the associated component weight, π_k. One can normalise these by dividing by the sum of the members of this vector due to the independence of the normalising constant from z_i.

2 Gaussian mixture models

In this section we derive the marginal likelihood for a component of the Gaussian mixture model assuming that the mean μ and the precision λ are unknown. Before we can continue we state the associated probability density functions of the Normal and Gamma distributions:

$$\mathcal{N}(x|\mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right) \tag{11}$$

$$Ga(x|\alpha, \text{rate} = \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \tag{12}$$

2.1 Likelihood

The model likelihood for n observations is:

$$p(x|\mu, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \tag{13}$$

Considering specifically the sum within the exponent here in equation 13, and letting $\bar{x}$ be the sample mean:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2 && (14)\\
&= \sum_{i=1}^{n}\left[(x_i-\bar{x})^2+(\mu-\bar{x})^2+2(x_i\bar{x}-\bar{x}^2-x_i\mu+\bar{x}\mu)\right] && (15)\\
&= n(\mu-\bar{x})^2+\sum_{i=1}^{n}(x_i-\bar{x})^2 && (16)
\end{aligned}$$

Substituting this back into equation 13, we have:

$$p(x|\mu,\lambda) = \left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}}\exp\left(-\frac{\lambda}{2}\left[n(\mu-\bar{x})^2+\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\right) \quad (17)$$

2.1.1 Prior

The conjugate prior for this model is the *Normal-Gamma* distribution. This has the probability density function:

$$\begin{aligned}
NG(\mu,\lambda|\mu_0,\kappa_0,\alpha_0,\beta_0) &\coloneqq \mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})Ga(\lambda|\alpha_0,\beta_0) && (18)\\
&= \sqrt{\frac{\kappa_0\lambda}{2\pi}}\exp\left(-\frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2\right) && (19)\\
&\quad\times\ \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\lambda^{\alpha_0-1}\exp(-\beta_0\lambda) && (20)\\
&= \sqrt{\frac{\kappa_0}{2\pi}}\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} && (21)\\
&\quad\times\ \lambda^{\alpha_0-\frac{1}{2}}\exp\left(-\frac{\lambda}{2}\left[\kappa_0(\mu-\mu_0)^2+2\beta_0\right]\right) && (22)
\end{aligned}$$

Here the normalising constant is:

$$Z_0^{-1} = \sqrt{\frac{\kappa_0}{2\pi}}\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \quad (23)$$

One can see that this function in equation 22 will naturally complement the likelihood described in equation 17.

2.2 Posterior

From Bayes' theorem we have:

$$
\begin{aligned}
p(\mu, \lambda|x) &\propto p(x|\mu, \lambda)p(\mu, \lambda) && (24)\\
&= \left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\lambda}{2}\left[n(\mu-\bar{x})^2 + \sum_{i=1}^{n}(x_i-\bar{x})^2\right]\right) && (25)\\
&\quad \times \lambda^{\alpha_0-\frac{1}{2}} \exp\left(-\frac{\lambda}{2}\left[\kappa_0(\mu-\mu_0)^2 + 2\beta_0\right]\right) && (26)\\
&\propto \lambda^{\alpha_0+\frac{n}{2}-\frac{1}{2}} \exp\left\{-\frac{\lambda}{2}\left[n(\mu-\bar{x})^2 + \kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2\beta_0\right]\right\} && (27)
\end{aligned}
$$

We try and anticipate the parameters of our posterior. The following piece of logic might be slightly misleading. If one considers our prior as originating from previously observed data, we can consider the κ_0 parameter as the number of observations forming our prior and μ_0 as the mean observed in our prior. In this case we would now expect that the observed mean, μ_n, should correspond to a weighted average of combining the currently observed mean ($\bar{x}$), the number of observations in the current dataset (n) and these variables. Similarly, the "total" number of observations between the prior data and the current data, κ_n, should be the sum of the number of samples in each, i.e.:

$$
\begin{aligned}
\kappa_n &= \kappa_0 + n && (28)\\
\mu_n &= \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n} && (29)
\end{aligned}
$$

If this is the case we would expect a term in the exponent:

$$
\begin{aligned}
\kappa_n(\mu-\mu_n)^2 &= (\kappa_0+n)\left[\mu - \left(\frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0+n}\right)\right]^2 && (30)\\
&= (\kappa_0+n)\left(\mu^2 - 2\mu(\kappa_0\mu_0 + n\bar{x}) - \left(\frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0+n}\right)^2\right) && (31)
\end{aligned}
$$

Returning to equation 27 and considering the part of the exponent containing μ:

$$
\begin{aligned}
\kappa_0(\mu-\mu_0)^2 + n(\mu-\bar{x})^2 &= \mu^2(\kappa_0+n) + n\bar{x}^2 + \kappa_0\mu_0^2 - 2n\mu\bar{x} - 2\kappa_0\mu\mu_0 && (32)\\
&= (\kappa_0+n)\left[\mu^2 + \frac{n\bar{x}^2}{\kappa_0+n} + \frac{\kappa_0\mu_0^2}{\kappa_0+n} - 2\mu\left(\frac{n\bar{x}+\kappa_0\mu_0}{\kappa_0+n}\right)\right] && (33)
\end{aligned}
$$

Notice that several of the desired terms are present, focusing on the components of this equation that do not fit our expected form:

$$\frac{n\bar{x}^2}{\kappa_0+n}+\frac{\kappa_0\mu_0^2}{\kappa_0+n} = \frac{n(\kappa_0+n)\bar{x}^2+\kappa_0(\kappa_0+n)\mu_0^2}{(\kappa_0+n)^2} \quad (34)$$

$$= \frac{n^2\bar{x}^2+2\kappa_0 n\mu_0\bar{x}+\kappa_0^2\mu_0^2}{(\kappa_0+n)^2}2\frac{\kappa_0 n\bar{x}^2+\kappa_0 n\mu_0^2-2\kappa_0 n\mu_0\bar{x}}{(\kappa_0+n)^2} \quad (35)$$

$$(36)$$

$$= \left(\frac{n\bar{x}+\kappa_0\mu_0}{\kappa_0+n}\right)^2+\frac{n\kappa_0\left(\bar{x}-\mu_0\right)^2}{(\kappa_0+n)^2} \quad (37)$$

Substituting this result into equation 33 gives us:

$$\kappa_0(\mu-\mu_0)^2+n(\mu-\bar{x})^2 = (\kappa_0+n)\left[\mu+\left(\frac{n\bar{x}+\kappa_0\mu_0}{\kappa_0+n}\right)\right]^2+\frac{n\kappa_0\left(\bar{x}-\mu_0\right)^2}{\kappa_0+n} \quad (38)$$

Returning to equation 27:

$$p(\mu,\lambda|x) \propto \lambda^{\alpha_0+\frac{n}{2}-\frac{1}{2}}\exp\left\{-\frac{\lambda}{2}\left[(\kappa_0+n)\left(\mu+\left(\frac{n\bar{x}+\kappa_0\mu_0}{\kappa_0+n}\right)\right)^2+\frac{n\kappa_0\left(\bar{x}-\mu_0\right)^2}{\kappa_0+n}+\sum_{i=1}^{n}(x_i-\bar{x})^2+2\beta_0\right]\right\} \quad (39)$$

This is the probability density function of a Normal-Gamma distribution.

$$p(\mu,\lambda|x) = NG(\mu,\lambda|\mu_n,\kappa_n,\alpha_n,\beta_n) \quad (40)$$

$$\mu_n = \frac{n\bar{x}+\kappa_0\mu_0}{\kappa_0+n} \quad (41)$$

$$\kappa_n = \kappa_0+n \quad (42)$$

$$\alpha_n = \alpha_0+\frac{n}{2} \quad (43)$$

$$\beta_n = \beta_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{n\kappa_0\left(\bar{x}-\mu_0\right)^2}{2(\kappa_0+n)} \quad (44)$$

As this posterior distribution is $NG(\mu_n,\kappa_n,\alpha_n,\beta_n)$, we know that the associated normalising constant is:

$$Z_n = \frac{\Gamma(\alpha_n)}{\beta_n^{\alpha_n}}\sqrt{\frac{2\pi}{\kappa_n}}. \quad (45)$$

2.3 Marginal likelihood

Consider now the posterior but track the normalising constants (such as the $(2\pi)^{-n/2}$ in the likelihood). Denote the prior, likelihood and posterior less their normalising constants by $p'(\mu,\lambda)$, $p(x|\mu,\lambda)$ and $p'(\mu,\lambda|x)$ respectively:

$$\frac{1}{Z_n}p'(\mu,\lambda|x) = \frac{1}{p(x)}\frac{1}{Z_0}p'(\mu,\lambda)\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}p'(x|\mu,\lambda) \tag{46}$$

We know that the product of the unnormalised prior and the unnormalised likelihood give the right hand side of equation 39, and that this is our unnormalised posterior, $p'(\mu,\lambda|x)$. Thus:

$$\frac{1}{Z_n}p'(\mu,\lambda|x) = \frac{1}{p(x)}\frac{1}{Z_0}\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}p'(\mu,\lambda|x) \tag{47}$$

$$\implies p(x) = \frac{Z_n}{Z_0}(2\pi)^{-\frac{n}{2}} \tag{48}$$

Thus our marginal likelihood is the ratio of the posterior normalising constants to the product of those of the likelihood and prior. Expanding this we finally have:

$$p(x) = \frac{\Gamma(\alpha_n)}{\beta_n^{\alpha_n}}\sqrt{\frac{2\pi}{\kappa_n}}\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\sqrt{\frac{\kappa_0}{2\pi}}(2\pi)^{-\frac{n}{2}} \tag{49}$$

$$= \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)}\frac{\beta_0^{\alpha_0}}{\beta_n^{\alpha_n}}\sqrt{\frac{\kappa_0}{\kappa_n}}(2\pi)^{-\frac{n}{2}} \tag{50}$$