Variational Bayes for an infinite GMM

Y.Cem Sübakan

April 18, 2014

1 Introduction

In this report we derive the variational Bayes procedure described in Blei, Jordan 2006 (Variational Inference for Dirichlet Process Mixtures) for an infinite mixture model. We will use a Gaussian observation model with known covariances for the sake of simplicity.

1.1 Generative Model

The generative model we'll be working on is a Gaussian mixture model (GMM). We have priors on the cluster means and mixing proportions. The generative model is defined as follows:

$$\begin{aligned}
\pi \sim & \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_K) \\
\mu_k \sim & \mathcal{N}(\mu_k; \mu_0, \Sigma_0), \quad \forall k \in \{1, \dots, K\} \\
z_n \sim & \text{Discrete}(\pi), \quad \forall n \in \{1, \dots, N\} \\
x_n \sim & \mathcal{N}(x_n; \mu_k, \Sigma), \quad \forall n \in \{1, \dots, N\}
\end{aligned}$$

2 Variational Bayes Algorithm for the Finite Case

The log-joint probability distribution is decomposed as follows:

$$\log p(x_{1:N}, z_{1:N}, \pi, \mu_{1:K}) = \sum_{n=1}^{N} \sum_{k=1}^{K} [z_n = k] \log p(x_n|\mu_k) + \sum_{n=1}^{N} \log p(z_n|\pi) + \sum_{k=1}^{K} \log p(\mu_k) + \log p(\pi) \tag{1}$$

With variational Bayes we approximate the joint distribution with the following factorized distribution:

$$q(z_{1:N}, \pi, \mu_{1:K}) = \left(\prod_{n=1}^{N} q(z_n) \right) \left(\prod_{k=1}^{K} q(\mu_k) \right) q(\pi) \tag{2}$$

Then we derive the variational distributions.

2.1 Means $\mu_{1:K}$

Let us start with means $\mu_{1:K}$:

$$\begin{aligned}\log q(\mu_k) =& \mathbb{E}_{q(z_{1:N})q(\pi)}\left[\sum_{n=1}^{N}[z_n = k]\log p(x_n|\mu_k) + \sum_{k=1}^{K}\log p(\mu_k)\right] \\ =& \sum_{n=1}^{N}\mathbb{E}[z_n = k]\log p(x_n|\mu_k) + \sum_{k=1}^{K}\log p(\mu_k) \end{aligned} \tag{3}$$

Let us define $N_k := \sum_{n=1}^{N}\mathbb{E}[z_n = k]$. Then, for Gaussian observation model we have the following expression:

$$\begin{aligned}\log q(\mu_k) =^{+}& -\sum_{n=1}^{N}\mathbb{E}[z_n = k]\left\{(x_n - \mu_k)^T\Sigma^{-1}(x_n - \mu_k)\right\} - (\mu_k - \mu_0)^T\Sigma_0^{-1}(\mu_k - \mu_0) \\ =^{+}& \sum_{n=1}^{N}\mathbb{E}[z_n = k]\left\{-\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}x_n\right\} - \frac{1}{2}\mu_k^T\Sigma_0^{-1}\mu_k + \mu_k^T\Sigma_0^{-1}\mu_0 \\ =& -\frac{1}{2}\mu_k^T\left(N_k\Sigma^{-1} + \Sigma_0^{-1}\right)\mu_k + \mu_k^T\left(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\Big(\sum_{n=1}^{N}\mathbb{E}[z_n = k]x_n\Big)\right) \end{aligned} \tag{4}$$

So, the variational distribution on μ_k is the following:

$$q(\mu_k) = \mathcal{N}(\mu_k; \bar{\mu}, \left(N_k\Sigma^{-1} + \Sigma_0^{-1}\right)^{-1}) \tag{5}$$

where, $\bar{\mu} = \left(N_k\Sigma^{-1} + \Sigma_0^{-1}\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\Big(\sum_{n=1}^{N}\mathbb{E}[z_n = k]x_n\Big)\right)$. If we assume that $\Sigma = \sigma^2 I$, and $\Sigma_0 = \sigma_0^2 I$, then $\bar{\mu}$ becomes the following:

$$\bar{\mu} = \frac{\sigma^2\mu_0 + \sigma_0^2\sum_{n=1}^{N}\mathbb{E}[z_n = k]x_n}{N_k\sigma_0^2 + \sigma^2} \tag{6}$$

Therefore, the resulting variational distribution is the following:

$$q(\mu_k) = \mathcal{N}(\mu_k; \frac{\sigma^2\mu_0 + \sigma_0^2\sum_{n=1}^{N}\mathbb{E}[z_n = k]x_n}{N_k\sigma_0^2 + \sigma^2}, \frac{\sigma^2\sigma_0^2}{N_k\sigma_0^2 + \sigma^2}I) \tag{7}$$

2.2 Cluster Indicators $z_{1:N}$

The logarithm of the variational distribution $q(z_n)$ can be written as follows:

$$\begin{aligned}\log q(z_n) =^{+}& \sum_{k=1}^{K}[z_n = k]\{-\mathbb{E}_{q(\mu_k)}[\mu_k^T\Sigma^{-1}\mu_k] + \mathbb{E}_{q(\mu_k)}[\mu_k]^T\Sigma^{-1}x_n\} + \sum_{k=1}^{K}[z_n = k]\mathbb{E}_{q(\pi_k)}[\log\pi_k] \\ =& \sum_{k=1}^{K}[z_n = k]\underbrace{\left(-\mathbb{E}_{q(\mu_k)}[\mu_k^T\Sigma^{-1}\mu_k] + 2\mathbb{E}_{q(\mu_k)}[\mu_k]^T\Sigma^{-1}x_n + \mathbb{E}_{q(\pi_k)}[\log\pi_k]\right)}_{:=^{+}\log\bar{p}_k} \end{aligned} \tag{8}$$

So, this expression looks like a discrete distribution distribution with parameters $\widetilde{\mu}_{1:K}$:

$$q(z_n) = \text{Discrete}(z_n; \bar{p}_1, \bar{p}_2, \ldots, \bar{p}_K) \tag{9}$$

2.3 Mixing Proportions $\pi_{1:K}$

The logarithm of the variational distribution $q(\pi_{1:K})$ is as follows:

$$\begin{aligned}\log q(\pi) &=^{+} \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_n = k]\log\pi_k + \sum_{k=1}^{K}(\alpha-1)\log\pi_k \\ &= \sum_{k=1}^{K}\underbrace{\left(\alpha - 1 + \sum_{n=1}^{N}\mathbb{E}[z_n = k]\right)}_{\widetilde{\alpha}_k}\log\pi_k \end{aligned} \tag{10}$$

This looks like a Dirichlet distribution. So, the variational distribution is:

$$q(\pi) = \text{Dirichlet}(\pi; \widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_K) \tag{11}$$

3 Variational Bayes Algorithm for the Infinite Case

The main trick for deriving the variational Bayes algorithm for the infinite case is to have the stick-breaking prior on the mixing proportions. According to the stick breaking prior, the k'th mixing proportion π_k is generated as follows:

$$\pi_k \sim \delta\left(\pi_k - v_k\prod_{j=1}^{K-1}(1-v_j)\right) \tag{12}$$

$$v_k \sim \mathcal{B}(\alpha, 1),\ \forall k \in \{1, \ldots, K\} \tag{13}$$

We then alter the derivation in the previous section for this prior. Notice that $q(\mu_{1:K})$ remains unaltered. We derive the novel $q(v_{1:K})$ and $q(z_{1:N})$ in the next section. Note that we integrate out $\pi_{1:K}$ since there is a deterministic relationship between $\pi_{1:K}$ and $v_{1:K}$.

3.1 For cluster indicators $z_{1:N}$:

$$\begin{aligned}\log q(z_n) =^{+} \sum_{k=1}^{K'}[z_n = k]\left(\{-\mathbb{E}_{q(\mu_k)}[\mu_k^T\Sigma^{-1}\mu_k] + 2\mathbb{E}_{q(\mu_k)}[\mu_k]^T\Sigma^{-1}x_n\} + \mathbb{E}_{q(v_k)}[\log v_k]\right) \\ + \sum_{k=1}^{K'}[z_n > k]\mathbb{E}[\log(1-v_k)]\end{aligned} \tag{14}$$

3.2 For stick breaking variables $v_{1:K'}$:

$$\begin{aligned}\log q(v_{1:K'}) &=^{+} \mathbb{E}\left[\sum_{n=1}^{N} \log p(z_n|v_{1:K'})\right] + \log p(v_{1:K'}) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K'} \mathbb{E}[z_n = k]\log v_k + \sum_{n=1}^{N}\sum_{k=1}^{K'} \mathbb{E}[z_n > k]\log(1-v_k) + \sum_{k=1}^{K'}(\alpha-1)\log(v_k) \\ &= \sum_{k=1}^{K'}\left\{\left(\alpha - 1 + \sum_{n=1}^{N}\mathbb{E}[z_n = k]\right)\log v_k + \left(\sum_{n=1}^{N}\mathbb{E}[z_n > k]\right)\log(1-v_k)\right\} \qquad (15)\end{aligned}$$

Notice that this looks like the log of a Beta pdf. Thus, we conclude that:

$$q(v_k) = \mathcal{B}(\alpha + \sum_{n=1}^{N}\mathbb{E}[z_n = k], \sum_{n=1}^{N}\mathbb{E}[z_n > k] + 1) \qquad (16)$$

4 Expressions for expectations: (Infinite Case)

The expectations we need to compute are: $\mathbb{E}_{q(\mu_k)}[\mu_k^T\Sigma^{-1}\mu_k]$, $\mathbb{E}_{q(\mu_k)}[\mu_k]$, $\mathbb{E}_{q(v_k)}[\log v_k]$, $\mathbb{E}[\log(1-v_k)]$, $\mathbb{E}[z_n = k]$ and $\mathbb{E}[z_n > k]$.

- $\mathbb{E}_{q(\mu_k)}[\mu_k^T\Sigma^{-1}\mu_k]$: We choose $\Sigma = \sigma^2 I$ for simplicity. Then,

  $$\mathbb{E}_{q(\mu_k)}[\mu_k^T\Sigma^{-1}\mu_k] = \frac{1}{\sigma^2}\mathbb{E}_{q(\mu_k)}[\mu_k^T\mu_k] = \frac{1}{\sigma^2}(\bar{\mu}_k^T\bar{\mu}_k + L\bar{\sigma}_k^2)$$

  where, $\bar{\mu}_k = \frac{\sigma^2\mu_0 + \sigma_0^2\sum_{n=1}^{N}\mathbb{E}[z_n=k]x_n}{N_k\sigma_0^2+\sigma^2}$, $\bar{\sigma}_k^2 = \frac{\sigma^2\sigma_0^2}{N_k\sigma_0^2+\sigma^2}$, and L is the dimensionality of the observations.

- $\mathbb{E}_{q(\mu_k)}[\mu_k] = \bar{\mu}_k$.

- $\mathbb{E}_{q(v_k)}[\log v_k] = \psi(\bar{\alpha}_k) - \psi(\bar{\alpha}_k + \bar{\beta}_k)$
  where, $\bar{\alpha}_k = \alpha + \sum_{n=1}^{N}\mathbb{E}[z_n = k]$ and $\bar{\beta}_k = \sum_{n=1}^{N}\mathbb{E}[z_n > k] + 1$.

- $\mathbb{E}_{q(v_k)}[\log(1-v_k)] = \psi(\bar{\beta}_k) - \psi(\bar{\alpha}_k + \bar{\beta}_k)$

- $\mathbb{E}[z_n = k] = \bar{p}_{k,n}$
  where, $\bar{p}_{k,n} \propto \exp\left(-\mathbb{E}_{q(\mu_k)}[\mu_k^T\Sigma^{-1}\mu_k] + 2\mathbb{E}_{q(\mu_k)}[\mu_k]^T\Sigma^{-1}x_n + \mathbb{E}_{q(v_k)}[\log v_k] + \sum_{j=1}^{k-1}\mathbb{E}_{q(v_j)}[\log(1-v_j)]\right)$.

- $\mathbb{E}[z_n > k] = \sum_{j=k+1}^{K'}\bar{p}_{k,n}$.

5 Pseudocode (Infinite Case)

All this fuss boils down to the following straightforward algorithm:

Algorithm 1 Variational Bayes for infinite GMM

Initialize $\bar{p}_{1:K,1:N}$
for $e = 1 : maxiter$ **do**
  for $k = 1 : K'$ **do**
    $N_k = \sum_{n=1}^{N} \bar{p}_{k,n}$ (update pseudo counts)
    $\bar{\mu}_k = \frac{\sigma^2 \mu_0 + \sigma_0^2 \sum_{n=1}^{N} \bar{p}_{k,n}}{N_k \sigma_0^2 + \sigma^2}$ (update the means)
    $\bar{\sigma}_k^2 = \frac{\sigma^2 \sigma_0^2}{N_k \sigma_0^2 + \sigma^2}$ (update cluster variances)
    $\bar{\alpha}_k = \alpha + \sum_{n=1}^{N} \bar{p}_{k,n}$ and $\bar{\beta}_k = \sum_{n=1}^{N} \bar{p}_{k,n} + 1$ (update stick breaking parameters)
  end for

  for $n = 1 : N$ **do**
    for $k = 1 : K'$ **do**
      $\bar{p}_{k,n} \propto \exp\Big(-\frac{1}{2\sigma^2}(\bar{\mu}_k^T \bar{\mu}_k + L\bar{\sigma}_k^2) + \frac{1}{\sigma^2} \bar{\mu}_k^T x_n + \psi(\bar{\beta}_k) - \psi(\bar{\alpha}_k + \bar{\beta}_k) + \sum_{j=1}^{k-1} \psi(\bar{\alpha}_j) - \psi(\bar{\alpha}_j + \bar{\beta}_j)\Big)$.
    end for
  end for
end for

Please note that the updating order of the parameters is arbitrary, and one can use a different updating order.