# 1 Variational Bayes - Normal Mixture Model

We have *iid* observations $y_i$ genrated by a two level Normal Mixture Model with means $\mu_1$ and $\mu_2$ and known variance 1, so

$$p(y_i|\mu_1, \mu_2, k_i) = \mathcal{N}(\mu_1, 1)^{k_i}\mathcal{N}(\mu_2, 1)^{1-k_i}.$$

where the latent variable $k_i = 1$ if $y_i$ is drawn from $\mathcal{N}(\mu_1, 1)$ and $k_i = 0$ otherwise.

Further, $\mathbf{k}$ is modelled as *iid* Bernoulli with parameter $\pi$, so

$$p(k_i|\pi) = \pi^{k_i}(1-\pi)^{1-k_i}.$$

Introducing the priors $p(\pi) \sim U(0, 1)$ and $p(\mu_1, \mu_2) \propto 1$, the joint distribution becomes

$$\begin{aligned}
p(y, k, \mu_1, \mu_2, \pi) &= \prod_{i=1}^{n} p(y_i|k_i, \mu_1, \mu_2, \pi)p(k_i|\pi)p(\pi)p(\mu_1\mu_2) \\
&\propto \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-(y_i-\mu_1)^2}{2}\right\}\right)^{k_i} \left(\frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-(y_i-\mu_2)^2}{2}\right\}\right)^{1-k_i} \\
&\times \pi^{k_i}(1-\pi)^{1-k_i} \\
\ln(p(y, k, \mu_1, \mu_2, \pi)) &= \sum_{i=1}^{N} \left[\ln\left(\exp\left\{\frac{-(y_i-\mu_1)^2}{2}\right\}^{k_i}\right)\right] + \sum_{i=1}^{N} \left[\ln\left(\exp\left\{\frac{-(y_i-\mu_2)^2}{2}\right\}^{1-k_i}\right)\right] \\
&+ \sum_{i=1}^{N} k_i \ln(\pi) + \sum_{i=1}^{N} (1-k_i)\ln(1-\pi) \\
&= \sum_{i=1}^{N} \left[k_i \frac{-(y_i-\mu_1)^2}{2}\right] + \sum_{i=1}^{N} \left[(1-k_i)\frac{-(y_i-\mu_2)^2}{2}\right] \\
&+ \sum_{i=1}^{N} k_i \ln(\pi) + \sum_{i=1}^{N} (1-k_i)\ln(1-\pi) + c. \qquad (1)
\end{aligned}$$

We can take the variational approximation factorisation $q(k_{1:n}, \mu_1, \mu_2, \pi) = \prod_{i=1}^{n} q(k_i)q(\mu_1)q(\mu_2)q(\pi)$, which implies independence of $k_i, k_j$ for $i \neq j$:

It can be shown that the factorisable distribution that minmises the KL Divergence between $q(\theta)$ and $p(\theta|y)$ satisfies

$$q_i \propto \exp(\mathbb{E}_{q_{j\neq i}}(\ln(p(y, x, \theta)))) \qquad (2)$$

for all $q_i$, where $y$ is the observed data, $x$ is a latent variable and $\theta$ is a vector of unknown parameters.

Substituting (1) into (2) yields

$$\begin{aligned}
\ln(q(\pi)) &= \mathbb{E}_{k_{1:n}}\left[\sum_{i=1}^{n} k_i \ln(\pi) + (1-k_i)\ln(1-\pi) + c\right] \\
&= \sum_{i=1}^{n} \mathbb{E}(k_i)\ln(\pi) + (n - \sum_{i=1}^{n} \mathbb{E}(k_i))\ln(1-\pi) + c \\
&= \ln(\pi^{\sum_{i=1}^{n} \mathbb{E}(k_i)}(1-\pi)^{n-\sum_{i=1}^{n} \mathbb{E}(k_i)}) + c \\
q(\pi) &\propto \pi^{\sum_{i=1}^{n} \mathbb{E}(k_i)}(1-\pi)^{n-\sum_{i=1}^{n} \mathbb{E}(k_i)}
\end{aligned}$$

Recognizing the kernel of a Beta distribution, we see that $q(\pi) \sim \mathcal{B}(\alpha = \sum_{i=1}^{n} \mathbb{E}(k_i) + 1, \beta = n - \sum_{i=1}^{n} \mathbb{E}(k_i) + 1)$. Continuing, we can find

$$\begin{aligned}
\ln(q(\mu_1)) &= \mathbb{E}_{k_{1:n}} \sum_{i=1}^{n} -k_i \frac{(y_i - \mu_1)^2}{2} + c \\
&= -\frac{1}{2}\left(\sum_{i=1}^{n} \mathbb{E}(k_i)(y_i - \mu_1)^2\right) + c \\
&= -\frac{1}{2}\left(\sum_{i=1}^{n} \mathbb{E}(k_i)((y_i - \tilde{y}_1) + (\tilde{y}_1 - \mu_1))^2\right) + c \\
&= -\frac{1}{2}\left(\sum_{i=1}^{n} \mathbb{E}(k_i)((y_i - \tilde{y}_1)^2 + (\tilde{y}_1 - \mu_1)^2 - 2(y_i - \tilde{y}_1)(\tilde{y}_1 - \mu_1))\right) + c.
\end{aligned}$$

Where

$$\tilde{y}_1 = \frac{\sum_{i=1}^{n} \mathbb{E}(k_i) y_i}{\sum_{i=1}^{n} \mathbb{E}(k_i)}.$$

Note that

$$\sum_{i=1}^{n} \mathbb{E}(k_i)(y_i - \tilde{y}_1) = \sum_{i=1}^{n} \mathbb{E}(k_i)\left(y_i - \frac{\sum_{i=1}^{n} \mathbb{E}(k_i) y_i}{\sum_{i=1}^{n} \mathbb{E}(k_i)}\right) = 0,$$

hence

$$\ln(q(\mu_1)) = -\frac{\sum_{i=1}^{n} \mathbb{E}(k_i)(\tilde{y}_1 - \mu_1)^2}{2} + c.$$

Recognizing the kernel of a Gaussian distribution, we can see that $q(\mu_1) \sim \mathcal{N}(\bar{\mu}_1 = \tilde{y}_1, \lambda_1 = (\sum_{i=1}^n \mathbb{E}(k_i)^{-1})$. Similarly, $q(\mu_2) \sim \mathcal{N}(\bar{\mu}_2 = \tilde{y}_2, \lambda_2 = \sum_{i=1}^n \mathbb{E}(1-k_i)^{-1}))$ with

$$\tilde{y_2} = \frac{\sum_{i=1}^n \mathbb{E}(1-k_i) y_i}{\sum_{i=1}^n \mathbb{E}(1-k_i)}.$$

Through independence, all $q(k_i)$ have the same form,

$$\begin{aligned}
\ln(q(k_i)) &= \mathbb{E}_{\mu_1,\mu_2,\pi}\left[k_i \frac{-(y_i-\mu_1)^2}{2} + (1-k_i)\frac{-(y_i-\mu_2)^2}{2} + k_i \ln(\pi) + (1-k_i)\ln(1-\pi) + c\right] \\
&= k_i \frac{\mathbb{E}_{\mu_1} - (y_i-\mu_1)^2}{2} + (1-k_i)\frac{\mathbb{E}_{\mu_2} - (y_i-\mu_2)^2}{2} + k_i \mathbb{E}_\pi \ln(\pi) + (1-k_i)\mathbb{E}_\pi \ln(1-\pi) + c \\
&= k_i \frac{2\tilde{\pi}_1 - ((y_i-\bar{\mu}_1)^2 + \lambda_1)}{2} + (1-k_i)\frac{2\tilde{\pi}_2 - ((y_i-\bar{\mu}_2)^2 + \lambda_2)}{2} + c \\
q(k_i) &\propto \exp\left\{\frac{2\tilde{\pi}_1 - ((y_i-\bar{\mu}_1)^2 + \lambda_1)}{2}\right\}^{k_i} \exp\left\{\frac{2\tilde{\pi}_2 - ((y_i-\bar{\mu}_2)^2 + \lambda_2)}{2}\right\}^{1-k_i}
\end{aligned}$$

The quantity $\tilde{\pi}_1 = \mathbb{E}_\pi \ln(\pi) = \psi(\alpha) - \psi(\alpha+\beta)$, and $\tilde{\pi}_2 = \mathbb{E}_\pi \ln(1-\pi) = \psi(\beta) - \psi(\alpha+\beta)$, where $\psi(\cdot)$ is the digamma function (Archambeau and Verleysen 2007).

Each $k_i$ has a Bernoulli distribution with parameters $p_i = \exp\left\{\frac{2\tilde{\pi}_1 - ((y_i-\bar{\mu}_1)^2+\lambda_1)}{2}\right\}$, and $q_i = \exp\left\{\frac{2\tilde{\pi}_2 - ((y_i-\bar{\mu}_2)^2+\lambda_2)}{2}\right\}$.

This gives us the update rules for the Variational Bayes iterations:

$$
\begin{aligned}
\alpha &= \sum_{i=1}^{n} \frac{p_i}{p_i + q_i} + 1 \\
\beta &= \sum_{i=1}^{n} \frac{q_i}{p_i + q_i} + 1 \\
\bar{\mu}_1 &= \frac{\sum_{i=1}^{n} y_i p_i/(p_i + q_i)}{\sum_{i=1}^{n} p_i/(p_i + q_i)} \\
\lambda_1 &= \left( \sum_{i=1}^{n} \frac{p_i}{p_i + q_i} \right)^{-1} \\
\bar{\mu}_2 &= \frac{\sum_{i=1}^{n} y_i q_i/(p_i + q_i)}{\sum_{i=1}^{n} q_i/(p_i + q_i)} \\
\lambda_2 &= \left( \sum_{i=1}^{n} \frac{q_i}{p_i + q_i} \right)^{-1} \\
p_i &= \exp\left\{ \frac{2(\psi(\alpha) - \psi(\alpha + \beta)) - ((y_i - \bar{\mu}_1)^2 + \lambda_1)}{2} \right\} \\
q_i &= \exp\left\{ \frac{2(\psi(\beta) - \psi(\alpha + \beta)) - ((y_i - \bar{\mu}_2)^2 + \lambda_2)}{2} \right\}
\end{aligned}
$$

250 draws were simulated with parameters $\mu_1 = 3, \mu_2 = 6, \pi = 0.6$ and th e variational algorithm was ran. After manually correcting mislabeling, $y_i$ was allocated to distribution 1 if $p_i > q_i$ and to distribution 2 if $p_i < q_i$, resulting in the successful classification of 235/250 draws. A more trivial decision rule to allocate $y_i$ to distribution 1 if $y_i < \bar{y}$ and to distribution 2 if $y_i > \bar{y}$ successfully classified 234/250 draws.

250 draws were simulated with parameters $\mu_1 = 5.5, \mu_2 = 6, \pi = 0.6$ to try and force an overlap in the data. $y_i$ was allocated to distribution 1 if $p_i > q_i$ and to distribution 2 if $p_i < q_i$, resulting in the successful classification of 158/250 draws. The trivial decision rule had identical classifications.