We will be fitting a mixture of Gaussians to a set of samples $y_i$. The model is

$$
\mu_k \sim \mathcal{N}(0, \sigma^2_\mu); \quad z_i \sim \text{Mult}(\pi); \quad y_i \sim \mathcal{N}(\mu_{zi}, \sigma^2_y).
$$

We could fit this model using maximum marginal likelihood using the expectation-maximization (EM) algorithm to marginalize out the $z$ variables. However, if we tried to optimize the hyperparameter $\sigma^2_\mu$ we might encounter a problem: the joint probability $p(y, \mu|\sigma^2_\mu, \sigma^2_y)$ is infinite if we set $\sigma^2_\mu = 0, \mu_k = 0, \sigma^2_y > 0$ for all $k$. This is the maximum marginal likelihood solution, but it’s not a desirable solution—we haven’t learned anything at all about our data!

As an alternative, we can do approximate maximum marginal likelihood using variational inference, approximately integrating out $\mu$ as well as $z$. We will use a variational distribution $q$ over $\mu$ and $z$ of the form

$$
q(\mu, z) = (\prod_k \mathcal{N}(\mu_k; \nu_k, \lambda_k))(\prod_i \phi_{i,z_i}),
$$

where $\nu, \lambda > 0$, and $\phi > 0$ are free parameters subject to $\sum_k \phi_{i,k} = 1$. We can use $q$ to lower bound the marginal probability log $p(y|\pi, \sigma^2_\mu, \sigma^2_y)$ as

$$
\log p(y|\pi, \sigma^2_\mu, \sigma^2_y) \geq \mathcal{L} \equiv \mathbb{E}_q[\log p(y, \pi, \sigma^2_\mu, \sigma^2_y)] - \mathbb{E}_q[\log q(z) - \log q(\mu)] = \sum_i \mathbb{E}_q[\log p(y_i|z_i, \mu, \sigma^2_y)]
$$

$$
+ \sum_i \mathbb{E}_q[\log p(z_i|\pi)] - \mathbb{E}_q[\log q(z_i)]
$$

$$
+ \sum_k \mathbb{E}_q[\log p(\mu_k|\sigma^2_\mu)] - \mathbb{E}_q[\log q(\mu_k)]
$$

$$
= \sum_i \left[ \frac{1}{2\sigma^2_y} \left( y_i^2 + \sum_k \phi_{i,k} (-2y_i\nu_k + \nu_k^2 + \lambda_k) \right) - \frac{1}{2} \log(2\pi\sigma^2_y) \right]
$$

$$
+ \sum_i \sum_k \phi_{i,k} \left( \log \pi_k - \log \phi_{i,k} \right)
$$

$$
+ \sum_k \left[ \frac{\nu_k^2 + \lambda_k}{2\sigma^2_\mu} - \frac{1}{2} \log(2\pi\sigma^2_\mu)^{-1} \right] + \frac{1}{2} \log(2\pi\lambda_k).
$$

We can optimize $\mathcal{L}$ by coordinate ascent on the hyperparameters $\pi$, $\sigma^2_\mu$, and $\sigma^2_y$ and the variational parameters $\phi$, $\nu$, and $\lambda$. Taking partial derivatives of $\mathcal{L}$ with respect to each of these parameters yields the updates

$$
\phi_{i,k} \propto \pi_k \exp \left\{ -\frac{1}{2\sigma^2_y} \left( \nu_k^2 + \lambda_k - 2y_i\nu_k \right) \right\};
$$

$$
\nu_k = \frac{1}{\sum_i \phi_{i,k}} \sum_i \phi_{i,k} y_i; \quad \lambda_k = \left( \frac{1}{\sigma^2_\mu} + \frac{\sum_i \phi_{i,k}}{2\sigma^2_y} \right)^{-1};
$$

$$
\sigma^2_y = \frac{\sum_i \phi_{i,k} (y_i^2 - 2\nu_k y_i + \nu_k^2 + \lambda_k)}{\sum_i \phi_{i,k}}; \quad \sigma^2_\mu = \frac{\sum_k \nu_k^2 + \lambda_k}{K}; \quad \pi_k \propto \sum_i \phi_{i,k}.
$$

Repeatedly applying these updates will produce a local optimum of the variational objective function $\mathcal{L}$.