Variational Algorithms for Approximate Bayesian Inference: Mixture of Gaussians

Nicholas Zabaras
Warwick Centre of Predictive Modelling (WCPM)
University of Warwick

March 22, 2015
Contents

- Variational Mixture of Gaussians, Computing \( q^*(Z) \), Computing \( q^*(\pi, \mu, \Lambda) \), Computing the responsibilities, Role of the prior \( q(\pi) \), Summary of the algorithm, Example, Automatic Pruning, Bayesian treatment versus MLE

- Variational Lower Bound, Re-estimation equations using the variational lower bound, Predictive distribution, Case of large data set

- Determining the number of mixture components, MAP Estimate versus MLE

- Induced factorizations

- Variational Message Passing

Following:
- Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 10
- Machine Learning: A Probabilistic Perspective, Kevin Murphy, Chapter 21.
Variational Mixture of Gaussians

Suppose we have a set of observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ and let the latent variables be $\mathbf{Z} = \{z_1, \ldots, z_N\}$. Each $z_n$ is a 1-of-$K$ binary vector with elements $z_{nk}$, $k=1,\ldots,K$.

The number of latent variables increases with the size of the data set whereas the number of all other parameters in the model below remain constant.

The starting point is the likelihood function for the Gaussian mixture model (see graphical model below).

In order to formulate a variational treatment of this model, we write down the joint distribution of all random variables

$$p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda) = p(\mathbf{X}|\mathbf{Z}, \mu, \Lambda)p(\mathbf{Z}|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda)$$

In order to formulate a variational treatment of this model, it is first convenient to write down the joint distribution of all random variables

\[
p(X, Z, \pi, \mu, \Lambda) = p(X|Z, \mu, \Lambda)p(Z, \mu, \Lambda) = p(X|Z, \mu, \Lambda)p(Z|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda)
\]

\(p(X|Z, \mu, \Lambda)\) is the likelihood for observations \(X\), given the model parameters:

\[
p(X|Z, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1})^{z_{nk}}
\]

\(p(Z|\pi)\) is the conditional distribution of \(Z\), given the mixing coefficients \(\pi:\)

\[
p(Z|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}
\]

\(p(\pi)\) is the prior distribution for the mixing coefficients \(\pi:\)

\[
p(\pi) = \text{Dir}(\pi|\alpha_0) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_0-1}
\]

This choice of prior gives a conjugate prior (\(\alpha_0\) here is interpreted as the effective prior number of observations associated with each component).
Variational Mixture of Gaussians

\[ p(\mu, \Lambda) = p(\mu|\Lambda)p(\Lambda) \]

is prior distribution for the parameters governing the mean and precision of each Gaussian component.

Using a Gaussian-Wishart prior for each component governing the mean and precision (conjugate prior when both the mean and precision are unknown):

\[
p(\mu|\Lambda)p(\Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_0, \nu_0)
\]

We do not know \( \alpha_0, m_0, W_0, \beta_0, \nu_0 \) parameters, although a sensible choice would be as follows:

\[
m_0 = 0 \\
W_0 = I \\
\nu_0 = D + 1 \text{ (where } D \text{ is the dimensionality of the data)} \\
0 < \alpha_0, \beta_0 << 1
\]

This choice of initial parameters has been chosen with the mind set of placing more importance on the observed data rather than the prior beliefs.
Variational Mixture of Gaussians

We now consider a variational distribution that factorizes between the latent variables and model parameters:

\[ q(Z, \pi, \mu, \Lambda) = q(Z)q(\pi, \mu, \Lambda) \]

This is the only assumption we make in order to solve the Bayesian mixture model.
Computing $q^*(Z)$

$$q(Z, \pi, \mu, \Lambda) = q(Z)q(\pi, \mu, \Lambda)$$

These factors will be determined by optimization of the variational distribution.

**We start with $q(Z)$**

From our general results, the log of the distribution minimizing the KL divergence of the factorized distribution above for latent variables $Z$ is given by:

$$\ln q^*(Z) = \mathbb{E}_{\pi, \mu, \Lambda}[\ln p(X, Z, \pi, \mu, \Lambda)] + \text{const}$$

This can be decomposed:

$$\ln q^*(Z) = \mathbb{E}_{\pi}[\ln p(Z|\pi)] + \mathbb{E}_{\mu, \Lambda}[\ln p(X|Z, \mu, \Lambda)] + \text{const}$$

By substituting $p(Z|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}$ and $p(X|Z, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}$, we obtain:

$$\ln q^*(Z) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const}$$

Where we define: $\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2} \mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2} \ln(2\pi) - \frac{1}{2} \mathbb{E}_{\mu_k, \Lambda_k}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)]$
Computing \( q^*(Z) \)

\[
\ln q^*(Z) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const}
\]

Taking the exponential and requiring that the resulting distribution is normalized, we obtain

\[
q^*(Z) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}
\]

\[
q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}
\]

where

\[
r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}
\]

Note that the form of \( q^*(Z) \) is the same as the form of the prior \( p(Z|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \).

Also note that for the discrete distribution, we have the standard result \( \mathbb{E}[z_{nk}] = r_{nk} \) from which it can be see that \( r_{nk} \) are playing the roles of responsibilities.
Computing \( q^*(\pi, \mu, \Lambda) \)

Since \( q^*(Z) \) depends on the distribution of other variables we have coupled update equations and therefore as before we have to solve the update equations iteratively.

Before proceeding to consider \( q^*(\pi, \mu, \Lambda) \) it is instructive to define the following quantities:

- **Effective number of observations**
  \[
  N_k = \sum_{n=1}^{N} r_{nk}
  \]

- **Effective sample mean**
  \[
  \bar{x}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n
  \]

- **Effective sample variance**
  \[
  S_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T
  \]

These are analogous to what we used in the EM algorithm for Gaussian mixtures.

A useful result from these definitions (expand the lhs in the definition of \( S_k \)) follows:

\[
N_k S_k = \sum_{n=1}^{N} r_{nk} x_n x_n^T - N_k \bar{x}_k \bar{x}_k^T
\]
Computing $q^*(\pi)$

$$q(Z, \pi, \mu, \Lambda) = q(Z)q(\pi, \mu, \Lambda)$$

We now consider $q(\pi, \mu, \Lambda)$

The log of the distribution minimizing the KL divergence of the factorized distribution above for parameters $\pi, \mu, \Lambda$ is given by:

$$\ln q^*(\pi, \mu, \Lambda) = \mathbb{E}_Z[\ln p(X, Z, \pi, \mu, \Lambda)] + \text{const} = \mathbb{E}_Z[\ln(p(X|Z, \mu, \Lambda)p(Z|\pi)p(\pi)p(\mu, \Lambda))] + \text{const}$$

$$= \ln p(\pi) + \sum_{k=1}^K \ln p(\mu_k, \Lambda_k) + \mathbb{E}_Z[\ln p(Z|\pi)] + \sum_{k=1}^K \sum_{n=1}^N \mathbb{E}_Z[z_{nk}] \ln \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}) + \text{const}$$

The variational posterior clearly factorizes into the following form

$$q(\pi, \mu, \Lambda) = q(\pi)q(\mu, \Lambda) = q(\pi) \prod_{k=1}^K q(\mu_k, \Lambda_k)$$

Using, $p(\pi) = \mathcal{C}(\alpha_0) \prod_{k=1}^K \pi_k^{\alpha_0-1}$ and $p(Z|\pi) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}}$, and using $\mathbb{E}[z_{nk}] = r_{nk}$, keeping the terms with $\pi$, for $q(\pi)$ we have

$$\ln q^*(\pi) = (\alpha_0 - 1) \sum_{k=1}^K \ln \pi_k + \sum_{k=1}^K \sum_{n=1}^N r_{nk} \ln \pi_k + \text{const} = (\alpha_0 - 1) \sum_{k=1}^K \ln \pi_k + \sum_{k=1}^K N_k \ln \pi_k + \text{const}$$

i.e. $q^*(\pi) = \text{Dir}(\pi|\alpha)$ where $\alpha_k = \alpha_0 + N_k$
Computing \( q^*(\mu, \Lambda) \)

Let us now keep the terms that depend on \( \mu_k, \Lambda_k \):

\[
\ln q^*(\pi, \mu, \Lambda) = \mathbb{E}_Z[\ln p(X, Z, \pi, \mu, \Lambda)] + \text{const} = \mathbb{E}_Z[\ln(p(X|Z, \mu, \Lambda)p(Z|\pi)p(\pi)p(\mu, \Lambda))] + \text{const}
\]

\[
= \ln p(\pi) + \sum_{k=1}^{K} \ln p(\mu_k, \Lambda_k) + \mathbb{E}_Z[\ln p(Z|\pi)] + \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] \ln \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}) + \text{const}
\]

Using \( p(\mu|\Lambda)p(\Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|m_0, (\beta_0\Lambda_k)^{-1}) \mathcal{W}(\Lambda_k|W_0, \nu_0) \), we derive:

\[
\ln q^*(\mu_k, \Lambda_k) = \ln \mathcal{N}(\mu_k|m_0, (\beta_0\Lambda_k)^{-1}) + \ln \mathcal{W}(\Lambda_k|W_0, \nu_0) + \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] \ln \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}) + \text{const}
\]

\[
= -\frac{\beta_0}{2} (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) + \frac{1}{2} \ln|\Lambda_k| - \frac{1}{2} \text{tr}(\Lambda_k W_0^{-1}) + \frac{\nu_0-D-1}{2} \ln|\Lambda_k|
\]

\[-\frac{1}{2} \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) + \frac{1}{2} \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] \ln|\Lambda_k| + \text{const}
\]

Using the product rule of probability, we express \( \ln q^*(\mu_k, \Lambda_k) = \ln q^*(\mu_k|\Lambda_k) + \ln q^*(\Lambda_k) \). We look first for terms that depend on \( \mu_k \).

\[
\ln q^*(\mu_k|\Lambda_k) = -\frac{1}{2} \mu_k^T (\beta_0 + \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] \Lambda_k) \mu_k + \mu_k^T \Lambda_k (\beta_0 m_0 + \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] x_n) + \text{const} =
\]

\[-\frac{1}{2} \mu_k^T (\beta_0 + N_k \Lambda_k) \mu_k + \mu_k^T \Lambda_k (\beta_0 m_0 + N_k \bar{x}_k) + \text{const}
\]

\[
q^*(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k|m_k, (\beta_k\Lambda_k)^{-1}), \text{ where } \beta_k = \beta_0 + N_k, m_k = \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k)
\]
Computing $q^*(\Lambda)$

To compute $\ln q^*(\Lambda_k)$, we use $\ln q^*(\Lambda_k) = \ln q^*(\mu_k, \Lambda_k) - \ln q^*(\mu_k | \Lambda_k)$. We use: $q^*(\mu_k | \Lambda_k) = \mathcal{N}(\mu_k | m_k, (\beta_k \Lambda_k)^{-1})$ to derive:

$$
\ln q^*(\Lambda_k) = -\frac{\beta_0}{2} (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) + \frac{1}{2} \ln |\Lambda_k| - \frac{1}{2} \text{tr}(\Lambda_k W_0^{-1}) + \frac{\nu_0 - D - 1}{2} \ln |\Lambda_k| \\
- \frac{1}{2} \sum_{n=1}^N \mathbb{E}_Z [z_{nk}] (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) + \frac{1}{2} \sum_{n=1}^N \mathbb{E}_Z [z_{nk}] \ln |\Lambda_k| + \text{const} \\
+ \frac{\beta_k}{2} (\mu_k - m_k)^T \Lambda_k (\mu_k - m_k) - \frac{1}{2} \ln |\Lambda_k| + \text{const}
$$

We keep terms that depend only on $\Lambda_k$.

$$
\ln q^*(\Lambda_k) = -\frac{\beta_0}{2} (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) - \frac{1}{2} \text{tr}(\Lambda_k W_0^{-1}) + \frac{\nu_0 - D - 1}{2} \ln |\Lambda_k| - \\
\frac{1}{2} \sum_{n=1}^N \mathbb{E}_Z [z_{nk}] (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) + \frac{1}{2} \sum_{n=1}^N \mathbb{E}_Z [z_{nk}] \ln |\Lambda_k| + \frac{\beta_k}{2} (\mu_k - m_k)^T \Lambda_k (\mu_k - m_k) + \text{const}
$$

$$
= \frac{v_0 + \sum_{n=1}^N \mathbb{E}_Z [z_{nk}] - D - 1}{2} \ln |\Lambda_k| - \frac{1}{2} \text{tr}(\Lambda_k W_0^{-1} + \Lambda_k \beta_0 (\mu_k - m_0)(\mu_k - m_0)^T) + \\
+ \Lambda_k \sum_{n=1}^N \mathbb{E}_Z [z_{nk}] (x_n - \mu_k)(x_n - \mu_k)^T - \Lambda_k \beta_k (\mu_k - m_k)(\mu_k - m_k)^T) \\
= \frac{v_k - D - 1}{2} \ln |\Lambda_k| - \frac{1}{2} \text{tr}(\Lambda_k W_k^{-1}) + \text{const}
$$

where $W_k^{-1} = W_0^{-1} + \beta_0 m_0 m_0^T + \sum_{n=1}^N r_{nk} x_n x_n^T - \beta_k m_k m_k^T$ and $v_k = v_0 + N_k$. 

Bayesian Scientific Computing, Spring 2013 (N. Zabaras)
Computing $q^*(\Lambda)$

Thus: $q^*(\Lambda_k) = W(\Lambda_k | W_k, \nu_k)$.

Here we defined

$$\nu_k = \nu_0 + N_k$$

$$W_k^{-1} = W_0^{-1} + \beta_0 m_0 m_0^T + \sum_{n=1}^N r_{nk} x_n x_n^T - \beta_k m_k m_k^T$$

$$= W_0^{-1} + \beta_0 m_0 m_0^T + N_k S_k + N_k \bar{x}_k \bar{x}_k^T - \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k) (\beta_0 m_0 + N_k \bar{x}_k)^T$$

$$= W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0) (\bar{x}_k - m_0)^T$$

We used $m_k = \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k)$ and $N_k S_k = \sum_{n=1}^N r_{nk} x_n x_n^T - N_k \bar{x}_k \bar{x}_k^T$, $\beta_k = \beta_0 + N_k$. 
Computing $q^*(\pi, \mu, \Lambda)$

Finally for $q^*(\mu_k, \Lambda_k)$ we have shown that it takes a Gaussian-Wishart distribution of the form:

$$q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k | m_k, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_k, \nu_k)$$

where we have defined:

$$\beta_k = \beta_0 + N_k$$
$$m_k = \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k)$$
$$W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T$$
$$\nu_k = \nu_0 + N_k$$

These equations are analogous to the M-step in the EM algorithm (involve the same sums over the data set).
Computing the responsibilities

Let us return terms in the expression for the responsibilities. Its terms need to be evaluated:

\[
\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2} \mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2} \ln \pi - \frac{1}{2} \mathbb{E}_{\mu_k,\Lambda_k} [(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)]
\]

Using our results \( q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k | m_k, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_k, \nu_k) \), we can write:

\[
\mathbb{E}_{\mu_k,\Lambda_k} \left[ (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right] = \int \int \left( \text{Tr} \left[ \Lambda_k (x_n - \mu_k)(x_n - \mu_k)^T \right] q^*(\mu_k | \Lambda_k) d\mu_k \right) q^*(\Lambda_k) d\Lambda_k
\]

From the moments of \( q^*(\mu_k | \Lambda_k) = \mathcal{N}(\mu_k | m_k, (\beta_k \Lambda_k)^{-1}) \), we can compute:

\[
\mathbb{E}[\mu_k] = m_k, \quad \mathbb{E}[\mu_k \mu_k^T] = m_k m_k^T + \beta_k^{-1} \Lambda_k
\]

Thus:

\[
\mathbb{E}_{\mu_k} \left[ (x_n - \mu_k)(x_n - \mu_k)^T \right] = x_n x_n^T - x_n m_k^T - m_k x_n^T + m_k m_k^T + \beta_k^{-1} \Lambda_k^{-1} = (x_n - m_k)(x_n - m_k)^T + \beta_k^{-1} \Lambda_k^{-1}
\]

Using the results for the mean of the Wishart:

\[
\mathbb{E}_{\mu_k,\Lambda_k} \left[ (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right] = \int \text{Tr} \left\{ \Lambda_k \left[ (x_n - m_k)(x_n - m_k)^T + \beta_k^{-1} \Lambda_k^{-1} \right] \right\} \mathcal{W}(\Lambda_k | W_k, \nu_k) d\Lambda_k
\]

\[
= \int \left\{ \text{Tr} \left\{ \Lambda_k \left[ (x_n - m_k)(x_n - m_k)^T \right] \right\} + \beta_k^{-1} D \right\} \mathcal{W}(\Lambda_k | W_k, \nu_k) d\Lambda_k
\]

\[
= \nu_k (x_n - m_k)^T W_k (x_n - m_k) + \beta_k^{-1} D
\]
Computing the responsibilities

We now look at the term $\mathbb{E}[\ln \pi_k]$ :

$$\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2} \mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2} \ln \pi - \frac{1}{2} \mathbb{E}_{\mu_k, \Lambda_k} [(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)]$$

$$\ln \tilde{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha})$$

Recall that $q^*(\pi) = \text{Dir}(\pi | \alpha)$ where $\alpha_k = \alpha_0 + N_k$ and a useful expression for the mean of the log. Here $\psi(\alpha_k)$ is the digamma function defined as:

$$\psi(\alpha) = \frac{d}{d\alpha} \ln \Gamma(\alpha)$$

The proof for this can also be found in an earlier lecture on the Dirichlet distribution.
Computing the responsibilities

\[ \ln \rho_{nk} = \mathbb{E} [\ln \pi_k] + \frac{1}{2} \mathbb{E} [\ln |\Lambda_k|] - \frac{D}{2} \ln \pi - \frac{1}{2} \mathbb{E}_{\mu_k, \Lambda_k} [(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] \]

Finally we need to compute the term \( \mathbb{E} [\ln |\Lambda_k|] \):

\[ \mathbb{E}_{\Lambda_k} [\ln |\Lambda_k|] = \int \ln |\Lambda_k| \mathcal{W}(\Lambda_k | W_k, \nu_k) d\Lambda_k \]

We use the following interesting result from the reference below:

Using this we obtain:

\[ \mathbb{E} [\ln |\Lambda_k|] = \sum_{i=1}^{D} \mathbb{E} [\ln \chi^2_{\nu_k+1-i}] + \ln |W_k| = \sum_{i=1}^{D} \left\{ \ln 2 + \psi \left( \frac{\nu_k+1-i}{2} \right) \right\} + \ln |W_k| = \sum_{i=1}^{D} \psi \left( \frac{\nu_k+1-i}{2} \right) + D \ln 2 + \ln |W_k| \]

Here we used that \( \mathbb{E} [\ln \chi^2_{\nu_k}] = \ln 2 + \psi \left( \frac{\nu_k}{2} \right) \). This can be derived directly by recalling that

\[ \chi^2_{\nu} \equiv \text{Gamma} \left( \frac{\nu}{2}, 2 \right) \text{ and } \mathbb{E} [\ln X] = \psi \left( \frac{\nu}{2} \right) + \ln 2 \text{ where } X \sim \text{Gamma} \left( \frac{\nu}{2}, 2 \right) \]

Note here in the relation of \( \chi^2_{\nu} \) and \( \text{Gamma} \left( \frac{\nu}{2}, 2 \right) \), the parametrization of Gamma is on shape and scale parameters.

Computing the responsibilities

Finally the various terms in our expression for the (unnormalized) responsibilities:

\[
\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2} \mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2} \ln \pi - \frac{1}{2} \mathbb{E}_{\mu_k, \Lambda_k}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)]
\]

are as follows:

\[
\ln \tilde{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}), \hat{\alpha}_k = \sum_k \alpha_k
\]

\[
\ln \tilde{\Lambda}_k \equiv \mathbb{E}[\ln |\Lambda_k|] = \sum_{i=1}^{D} \left( \psi \left( \nu_k + 1 - i \right) \frac{1}{2} \right) + D \ln 2 + \ln |W_k|
\]

\[
\mathbb{E}_{\mu_k, \Lambda_k}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] = D \beta_k^{-1} + \nu_k (x_n - m_k)^T W_k (x_n - m_k)
\]

Here \(\psi(\alpha_k)\) is the digamma function defined as:

\[
\psi(\alpha) = \frac{d}{d\alpha} \ln \Gamma(\alpha)
\]

The following result is finally obtained for the normalized responsibilities:

\[
r_{nk} \propto \tilde{\pi}_k \tilde{\Lambda}_k^{1/2} \exp \left( -\frac{D}{2\beta_k} - \frac{\nu_k}{2} (x_n - m_k)^T W_k (x_n - m_k) \right)
\]

This is similar to the corresponding result in maximum likelihood EM:

\[
r_{nk} \propto \pi_k |\Lambda_k|^{1/2} \exp \left( -\frac{1}{2} (x_n - m_k)^T \Lambda_k (x_n - m_k) \right)
\]
Role of the prior $q^*(\pi) = \text{Dir}(\pi|\alpha)$

Note that the variational posterior has the same functional form as the joint distribution:

$$p(X, Z, \pi, \mu, \Lambda) = p(X|Z, \mu, \Lambda)p(Z|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda)$$

$$q(Z, \pi, \mu, \Lambda) = q(Z)p(\pi)p(\mu|\Lambda)p(\Lambda)$$

Using $q^*(\pi) = \text{Dir}(\pi|\alpha)$ where $\alpha_k = \alpha_0 + N_k$, the expected values of the mixing coefficients are given below:

$$\mathbb{E}[\pi_k] = \frac{\alpha_0 + N_k}{K\alpha_0 + N}$$

Components that take essentially no responsibility for explaining the data points have $r_{nk} \approx 0$ and hence $N_k \approx 0$. From $\alpha_k = \alpha_0 + N_k$, we see that $\alpha_k \approx \alpha_0$ and the other parameters ($\beta_k, m_k, W_k, v_k$) revert to their prior values.

Consider a component for which $N_k \approx 0$ and $\alpha_k \approx \alpha_0$. If the prior is broad so that $\alpha_0 \to 0$, then $\mathbb{E}[\pi_k] \to 0$ and the component plays no role in the model, whereas if the prior tightly constrains the mixing coefficients, $\alpha_0 \to \infty$, then $\mathbb{E}[\pi_k] \to 1/K$. 
Summary of the algorithm

1) Initialize parameters $m_0, W_0, \nu_0, \alpha_0, \beta_0$

2) E-step: Update the responsibilities $r_{nk}$ needed in the approximate posterior $q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$

3) M-step: Update the parameters $N_k, \bar{x}_k, S_k, \alpha_k, \beta_k, m_k, W_k^{-1}, \nu_k$ needed in $q^*(\pi, \mu, \Lambda)$

4) Cycle between the E and M steps until convergence (Next slide)

Summary of the algorithm

- **E-Step**
  
  Compute the following:

  \[ r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nk}} \]

  \[ \ln \hat{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}) , \hat{\alpha} = \sum_k \alpha_k \]

  \[ \equiv \mathbb{E}[\ln |\Lambda_k|] = \sum_{i=1}^{D} \left( \psi \frac{\nu_k + 1 - i}{2} \right) + D \ln 2 + \ln |W_k| \]

  \[ \mathbb{E}_{\mu_k, \Lambda_k} [(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] = D \beta_k^{-1} + \nu_k (x_n - m_k)^T W_k (x_n - m_k) \]

  \[ r_{nk} \propto \hat{\pi}_k \Lambda_k^{\frac{1}{2}} \exp \left( -\frac{D}{2\beta_k} - \frac{\nu_k}{2} (x_n - m_k)^T W_k (x_n - m_k) \right) \]

- **M-Step**
  
  Compute the following:

  \[ N_k = \sum_{n=1}^{N} r_{nk} \]

  \[ \bar{x}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n \]

  \[ S_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \bar{x}_k) (x_n - \bar{x}_k)^T \]

  \[ \alpha_k = \alpha_0 + N_k \]

  \[ \beta_k = \beta_0 + N_k \]

  \[ m_k = \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k) \]

  \[ W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0) (\bar{x}_k - m_0)^T \]

  \[ \nu_k = \nu_0 + N_k \]
Variational mixture of $K = 6$ Gaussians (Old Faithful data set).

The ellipses denote the one standard-deviation density contours for each of the components, and the density of red ink inside each ellipse corresponds to the mean value of the mixing coefficient for each component.

The number in the top left of each diagram shows the number of iterations of VI.

Here, $\alpha_0 = 10^{-3}$ (favors sparsity)
Variational Mixture of Gaussians: Example
Variational Mixture of Gaussians: Example
Variational Mixture of Gaussians: Example
Variational Mixture of Gaussians: Example
We initialize using 6 clusters with K-means.
We use $\alpha_0 = 0.001$ to promote sparsity.
The posterior values of $\alpha_k$ are shown.
Unnecessary components are killed off.

Run `mixGaussVbDemoFaithful` in the PMTK3 toolbox.
Automatic Pruning

Using $q^*(\pi) = \text{Dir}(\pi|\alpha)$ where $\alpha_k = \alpha_0 + N_k$, we write $p(\pi|\alpha) = \exp(\sum_k \alpha_k \ln \pi_k - A(\alpha))$, $A(\alpha) = \sum_k \ln \Gamma(\alpha_k) - \ln \Gamma(\sum_k \alpha_k)$ and using the definition of the digamma function,

$$E[\ln \pi_k] = \frac{\partial A(\alpha)}{\partial \alpha_k} = \psi(\alpha_k) - \psi\left(\sum_{k'} \alpha_{k'}\right)$$

In VBEM, we use $\hat{\pi}_k \equiv \frac{\exp(\psi(\alpha_k))}{\exp(\psi(\sum_k \alpha_k))}$, $\alpha_k=\alpha_0+N_k$. This is better than using the mode of $\hat{\pi}_k \propto \alpha_k - 1$ that can be negative for $\alpha_0 = 0$ and $N_k = 0$. Using the approximation $\exp(\psi(x)) \approx x - 0.5, x > 1$, we simplify as:

$$\hat{\pi}_k \propto \alpha_k - 0.5$$

i.e. we remove 0.5 from each posterior count. So clusters with few weighted members become more empty with iterations while popular clusters get more members (the rich get richer).

This approach is more efficient than performing a discrete search over the number of clusters and comparing the marginal likelihood.

Bayesian treatment Vs. MLE

- For $N \to \infty$, the Bayesian treatment converges to the MLE EM algorithm.

- There is little computational overhead in using the Bayesian approach as compared to the traditional MLE approach.

- However, there are substantial advantages.
  - The singularities that arise in MLE when a Gaussian component ‘collapses’ onto a specific data point are absent in the Bayesian treatment. These singularities are removed if we simply introduce a prior and then use a MAP estimate instead of MLE.
  - There is no over-fitting for large number $K$ of components in the mixture.
  - The variational treatment allows determining the optimal number of components in the mixture without cross validation.
Variational Lower Bound

It is useful to monitor the bound during the re-estimation in order to test for convergence.

At each step of the iterative re-estimation procedure the value of this bound should not decrease.

Use this to verify the derivation of the update equations and their implementation.

For the variational mixture of Gaussians, the lower bound is given by

\[
\mathcal{L} = \sum_z \int \int \int q(Z, \pi, \mu, \Lambda) \ln \frac{p(X, Z, \pi, \mu, \Lambda)}{q(Z, \pi, \mu, \Lambda)} d\pi d\mu d\Lambda =
\]

\[
= \mathbb{E} \left[ \ln p(X, Z, \pi, \mu, \Lambda) \right] - \mathbb{E} \left[ \ln q(Z, \pi, \mu, \Lambda) \right]
\]

\[
= \mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] + \mathbb{E} \left[ \ln p(Z | \pi) \right] + \mathbb{E} \left[ \ln p(\pi) \right] + \mathbb{E} \left[ \ln p(\mu, \Lambda) \right]
\]

\[
- \mathbb{E} \left[ \ln q(Z) \right] - \mathbb{E} \left[ \ln q(\pi) \right] - \mathbb{E} \left[ \ln q(\mu, \Lambda) \right]
\]

Variational Lower Bound

The various terms are given below (for proof see following Appendix):

\[
\mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] =
\frac{1}{2} \sum_{k=1}^{K} N_k \left\{ \ln \Lambda_k - D \beta_k^{-1} - \nu_k \text{Tr}(S_k W_k) - \nu_k \left( \bar{x}_k - m_k \right) W_k \left( \bar{x}_k - m_k \right)^T - D \ln(2\pi) \right\}
\]

\[
\mathbb{E} \left[ \ln p(Z | \pi) \right] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln \pi_k
\]

\[
\mathbb{E} \left[ \ln p(\pi) \right] = \ln C(\alpha_0) + (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k
\]

\[
\mathbb{E} \left[ \ln p(\mu, \Lambda) \right] = \frac{1}{2} \sum_{k=1}^{K} \left\{ D \ln \left( \beta_0 / 2\pi \right) + \ln \Lambda_k - D \frac{\beta_0}{\beta_k}
\right.
\]

\[
- \beta_0 \nu_k \left( m_k - m_0 \right)^T W_k \left( m_k - m_0 \right) + K \ln B(W_0, \nu_0)
\]

\[
+ \frac{\nu_0 - D - 1}{2} \sum_{k=1}^{K} \ln \Lambda_k - \frac{1}{2} \sum_{k=1}^{K} \nu_k \text{Tr}(W_0^{-1} W_k)
\]
Variational Lower Bound

The remaining terms are:

\[
\mathbb{E} \left[ \ln q(Z) \right] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln r_{nk}
\]

\[
\mathbb{E} \left[ \ln q(\pi) \right] = \sum_{k=1}^{K} (\alpha_k - 1) \ln \pi_k + \ln C(\alpha)
\]

\[
\mathbb{E} \left[ \ln q(\mu, \Lambda) \right] = \sum_{k=1}^{K} \left\{ \frac{1}{2} \ln \Lambda_k + \frac{D}{2} \ln \frac{\beta_k}{2\pi} - \frac{D}{2} - H[q(\Lambda_k)] \right\}
\]

\(D\) is the dimensionality of \(x\), \(H[q(\Lambda_k)]\) is the entropy of the Wishart, and \(C(\alpha)\) and \(B(W, \nu)\) are normalization factors for the Dirichlet and Wishart distributions.

Note that the terms involving expectations of the logs of the \(q\) distributions simply represent the negative entropies of those distributions.
Appendix: Variational Lower Bound

Using \( p(X|Z, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1})^z_{nk} \) we derive:

\[
\mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{r_{nk}} \left\{ \mathbb{E} \left[ \ln | \Lambda_k | \right] - \mathbb{E} \left[ (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right] - D \ln(2\pi) \right\}
\]

We now use \( \ln \widetilde{\Lambda}_k \equiv \mathbb{E} [\ln | \Lambda_k |] = \sum_{i=1}^{D} \left( \psi \frac{\nu_k + 1 - i}{2} \right) + D \ln 2 + \ln |W_k| \)

\[
\mathbb{E}_{\mu_k, \Lambda_k} [(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] = D \beta_k^{-1} + \nu_k (x_n - m_k)^T W_k (x_n - m_k)
\]

Substitution gives

\[
\mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{r_{nk}} \left\{ \ln \Lambda_k - D \beta_k^{-1} - \nu_k (x_n - m_k)^T W_k (x_n - m_k) - D \ln(2\pi) \right\}
\]

Make use of the following: \( N_k = \sum_{n=1}^{N} r_{nk}, \bar{x}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n \)

\[
S_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T, \quad N_k S_k = \sum_{n=1}^{N} r_{nk} x_n x_n^T - N_k \bar{x}_k \bar{x}_k^T
\]

\[
\mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{r_{nk}} \left\{ \ln \Lambda_k - D \beta_k^{-1} - \nu_k Tr \left( W_k \left( x_n x_n^T - 2 x_n m_k^T + m_k m_k^T \right) \right) - D \ln(2\pi) \right\} = \frac{1}{2} \sum_{k=1}^{K} N_k \left\{ \ln \Lambda_k - D \beta_k^{-1} - \nu_k Tr \left( W_k S_k + \bar{x}_k \bar{x}_k^T - 2 \bar{x}_k m_k^T + m_k m_k^T \right) \right\} - D \ln(2\pi) \right\} = \frac{1}{2} \sum_{k=1}^{K} N_k \left\{ \ln \Lambda_k - D \beta_k^{-1} - \nu_k Tr \left( W_k S_k \right) - \nu_k \left( \bar{x}_k - m_k \right)^T W_k \left( \bar{x}_k - m_k \right) \right\} - D \ln(2\pi) \right\}
Appendix: Variational Lower Bound

Using \( p(Z|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \) we derive:

\[
\mathbb{E} \left[ \ln p(Z | \pi) \right] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln \pi_k
\]

where we used: \( \ln \tilde{\pi}_k \equiv \mathbb{E} [\ln \pi_k] = \psi(\alpha_k) - \psi(\tilde{\alpha}), \tilde{\alpha} = \sum_k \alpha_k \)

Using \( p(\pi) = \text{Dir}(\pi|\alpha_0) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \) and the equation above we derive:

\[
\mathbb{E} \left[ \ln p(\pi) \right] = \ln C(\alpha_0) + (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k
\]

Starting from \( p(\mu|\Lambda)p(\Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k|W_0, \nu_0) \), and denoting with \( B(W_0, \nu_0) \) the normalization of the Wishart, we can write:

\[
\mathbb{E} \left[ \ln p(\mu, \Lambda) \right] = \frac{1}{2} \sum_{k=1}^{K} \left\{ D \ln (\beta_0 / 2\pi) + \mathbb{E} \left[ \ln |\Lambda_k| \right] \right. \\
- \beta_0 \mathbb{E} \left[ (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) \right] + K \ln B(W_0, \nu_0) \\
+ \sum_{k=1}^{K} \left\{ \frac{\nu_0 - D - 1}{2} \mathbb{E} \left[ \ln |\Lambda_k| \right] - \frac{1}{2} Tr(W_0^{-1} \mathbb{E} \left[ \Lambda_k \right]) \right. \\
\left. \left. - \frac{1}{2} Tr(W_0^{-1} \mathbb{E} \left[ \Lambda_k \right]) \right] \right\}
\]
Appendix: Variational Lower Bound

We compute $\mathbb{E}\left[ (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) \right]$ as follows:

$$
\mathbb{E}_{\mu_k, \Lambda_k} \left[ (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) \right] = \int \int \left( \text{Tr} \left[ \Lambda_k (\mu_k - m_0)(\mu_k - m_0)^T \right] q^*(\mu_k | \Lambda_k) d\mu_k \right) q^*(\Lambda_k) d\Lambda_k
$$

Using $\mathbb{E}[\mu_k] = m_k$, $\mathbb{E}[\mu_k \mu_k^T] = m_k m_k^T + \beta_k^{-1} \Lambda_k^{-1}$, as well as the mean of the Wishart, we simplify:

$$
\mathbb{E}_{\mu_k, \Lambda_k} \left[ (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) \right] = \int \text{Tr} \left[ \Lambda_k \left( m_k m_k^T + \beta_k^{-1} \Lambda_k^{-1} - 2m_k m_0^T + m_0 m_0^T \right) \right] q^*(\Lambda_k) d\Lambda_k = D \beta_k^{-1} + \mathbb{E}_{\Lambda_k} \left[ (m_k - m_0)^T \Lambda_k (m_k - m_0) \right] = D \beta_k^{-1} + (m_k - m_0)^T \mathbb{E}[\Lambda_k](m_k - m_0)
$$

Finally using $ln\Lambda_k \equiv \mathbb{E}[\ln |\Lambda_k|]$ our expression for $\mathbb{E}[\ln p(\mu, \Lambda)]$ becomes:

$$
\mathbb{E}[\ln p(\mu, \Lambda)] = \frac{1}{2} \sum_{k=1}^K \left\{ D \ln \left( \frac{\beta_0}{2\pi} \right) + \ln \Lambda_k - \frac{\beta_0}{\beta_k} D - \beta_0 \nu_k (m_k - m_0)^T W_k (m_k - m_0) \right\} + K \ln B(W_0, \nu_0) + \frac{\nu_0 - D - 1}{2} \sum_{k=1}^K \ln \Lambda_k - \frac{1}{2} \sum_{k=1}^K \nu_k \text{Tr}(W_0^{-1}W_k)
$$
Using $q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$ we derive as before:

$$
\mathbb{E}[\ln q(Z)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}[z_{nk}] \ln r_{nk} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln r_{nk}
$$

Note that $\mathbb{E}[\ln q(\pi)]$ is the negative entropy of a Dirichlet distribution. Thus:

$$
\mathbb{E}[\ln q(\pi)] = \sum_{k=1}^{K} (\alpha_k - 1) \ln \pi_k + \ln C(\alpha)
$$

c(\alpha) is the normalization constant.

Note that here we used $\ln \tilde{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}), \hat{\alpha}_k = \sum_k \alpha_k$

Finally we need to show

$$
\mathbb{E}[\ln q(\mu, \Lambda)] = \sum_{k=1}^{K} \left\{ \frac{1}{2} \ln \Lambda_k + \frac{D}{2} \ln \frac{\beta_k}{2\pi} - \frac{D}{2} - H[q(\Lambda_k)] \right\}
$$

Recall that

$$
q(\mu, \Lambda) = \prod_{k=1}^{K} q(\mu_k, \Lambda_k) = \prod_{k=1}^{K} q(\mu_k | \Lambda_k) q(\Lambda_k) \Rightarrow \ln q(\mu, \Lambda) = \sum_{k=1}^{K} q(\mu_k | \Lambda_k) + \sum_{k=1}^{K} q(\Lambda_k)
$$

Use the entropy of the multivariate Gaussian

$$
\frac{D}{2} (1 + \ln(2\pi)) - \frac{D}{2} \ln \beta_k - \frac{1}{2} \ln |\Lambda_k|
$$

and Wishart distributions

$$
H[q(\Lambda_k)] = -\ln(B(W_k, \nu_k)) - \frac{\nu_k - D - 1}{2} \mathbb{E}[\ln |\Lambda_k|] + \frac{\nu_k D}{2}
$$
Appendix: Variational Lower Bound

Thus using $q^* (\mu_k, \Lambda_k) = \mathcal{N}(\mu_k | m_k, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_k, \nu_k)$ we can write:

$$\mathbb{E} \left[ \ln q(\mu, \Lambda) \right] = \sum_{k=1}^{K} \mathbb{E}_{\Lambda_k, \mu_k} \left[ q(\mu_k | \Lambda_k) \right] + \sum_{k=1}^{K} \mathbb{E}_{\Lambda_k} \left[ q(\Lambda_k) \right]$$

$$= \sum_{k=1}^{K} \left\{ \frac{1}{2} \mathbb{E}_{\Lambda_k} \left[ \ln \Lambda_k \right] + \frac{D}{2} \ln \beta_k - \frac{D}{2} \ln 2\pi - \frac{D}{2} - H \left[ q(\Lambda_k) \right] \right\}$$

$$= \sum_{k=1}^{K} \left\{ \frac{1}{2} \ln \Lambda_k + \frac{D}{2} \ln \frac{\beta_k}{2\pi} - \frac{D}{2} - H \left[ q(\Lambda_k) \right] \right\}$$

where

$$H \left[ q(\Lambda_k) \right] = -\ln \left( B(W_k, \nu_k) \right) - \frac{\nu_k - D - 1}{2} \ln |\Lambda_k| + \frac{\nu_k D}{2}$$
Re-Estimation Eqs. Using the Variational Lower Bound

The lower bound provides an alternative approach for deriving the re-estimation eqs using direct differentiation.

With conjugate priors, the functional form of the factors in the posterior is known: discrete for $Z$, Dirichlet for $\pi$, and Gaussian-Wishart for $(\mu_k, \Lambda_k)$. By substitution, we can then derive the lower bound as a function of the parameters of these distributions. Maximizing wrt these parameters gives the re-estimation eqs.

$$
\mathcal{L} = \mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] + \mathbb{E} \left[ \ln p(Z | \pi) \right] + \mathbb{E} \left[ \ln p(\pi) \right] + \mathbb{E} \left[ \ln p(\mu, \Lambda) \right] - \mathbb{E} \left[ \ln q(Z) \right] - \mathbb{E} \left[ \ln q(\pi) \right] - \mathbb{E} \left[ \ln q(\mu, \Lambda) \right]
$$

$$
\mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] = \frac{1}{2} \sum_{k=1}^{K} N_k \left\{ \ln \Lambda_k - D \beta_k^{-1} - \nu_k Tr \left( W_k S_k \right) - \nu_k \left( \bar{x}_k - m_k \right)^T W_k \left( \bar{x}_k - m_k \right) - D \ln(2\pi) \right\}
$$

$$
\mathbb{E} \left[ \ln p(\mu, \Lambda) \right] = \frac{1}{2} \sum_{k=1}^{K} \left\{ D \ln \left( \beta_0 / 2\pi \right) + \ln \Lambda_k - \frac{\beta_0}{\beta_k} D - \beta_0 \nu_k \left( m_k - m_0 \right)^T W_k \left( m_k - m_0 \right) \right\} + K \ln B(W_0, \nu_0)
$$

$$
\mathbb{E} \left[ \ln q(\mu, \Lambda) \right] = \sum_{k=1}^{K} \left\{ \frac{1}{2} \ln \Lambda_k + \frac{D}{2} \ln \frac{\beta_k}{2\pi} - \frac{D}{2} - H \left[ q(\Lambda_k) \right] \right\} \quad \mathbb{E} \left[ \ln p(\pi) \right] = \ln C(\alpha_0) + (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k
$$

$$
q(Z, \pi, \mu, \Lambda) = q(Z)p(\pi)p(\mu|\Lambda)p(\Lambda) \quad q(\pi, \mu, \Lambda) = q(\pi)q(\mu, \Lambda) = q(\pi) \prod_{k=1}^{K} q(\mu_k, \Lambda_k)
$$

$$
q(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} \gamma_{nk}^{z_{nk}} \quad q(\pi) = \text{Dir}(\pi | \alpha) \quad q(\mu, \Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | m_k, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_k, \nu_k)
$$
Re-Estimation Eqs. Using the Variational Lower Bound

The re-estimation Equations to show (in order) are as follows:

\[ \beta_k = \beta_0 + N_k \]
\[ m_k = \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k) \]
\[ W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0) (\bar{x}_k - m_0)^T \]
\[ \nu_k = \nu_0 + N_k \]
\[ \alpha_k = \alpha_0 + N_k \]
\[ r_{nk} \propto \pi_k \Lambda_k^2 \exp \left( - \frac{D}{2\beta_k} - \frac{\nu_k}{2} (x_n - m_k)^T W_k (x_n - m_k) \right) \]
Re-Estimatioin Eqs. Using the Variational Lower Bound

Set the derivative wrt $\beta_k^{-1}$ equal to zero:

$$
\frac{d}{d\beta_k^{-1}} \mathcal{L} = \frac{D}{2} (-N_k - \beta_0 + \beta_k) = 0 \Rightarrow \beta_k = \beta_0 + N_k
$$

Then set:

$$
\frac{dm_k}{dm_k} \mathcal{L} = -N_k \nu_k \left( W_k m_k - W_k \bar{x}_k \right) - \beta_0 \nu_k \left( W_k m_k - W_k m_0 \right) = 0 \Rightarrow m_k = \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k)
$$

$$
\mathcal{L} = \mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] + \mathbb{E} \left[ \ln p(Z | \pi) \right] + \mathbb{E} \left[ \ln p(\pi) \right] + \mathbb{E} \left[ \ln p(\mu, \Lambda) \right] - \mathbb{E} \left[ \ln q(Z) \right] - \mathbb{E} \left[ \ln q(\pi) \right] - \mathbb{E} \left[ \ln q(\mu, \Lambda) \right]
$$

$$
\mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] = \frac{1}{2} \sum_{k=1}^{K} N_k \left\{ \ln \Lambda_k - D \beta_k^{-1} - \nu_k \text{Tr} \left( W_k S_k \right) - \nu_k \left( \bar{x}_k - m_k \right)^T W_k \left( \bar{x}_k - m_k \right) - D \ln(2\pi) \right\}
$$

$$
\mathbb{E} \left[ \ln p(\mu, \Lambda) \right] = \frac{1}{2} \sum_{k=1}^{K} \left\{ D \ln \left( \beta_0 / 2\pi \right) + \ln \Lambda_k - \frac{\beta_0}{\beta_k} D - \beta_0 \nu_k \left( m_k - m_0 \right)^T W_k \left( m_k - m_0 \right) \right\} + K \ln B(W_0, \nu_0)
$$

$$
+ \frac{\nu_0 - D - 1}{2} \sum_{k=1}^{K} \ln \Lambda_k - \frac{1}{2} \sum_{k=1}^{K} \nu_k \text{Tr} \left( W_0^{-1} W_k \right)
$$

$$
\mathbb{E} \left[ \ln q(\mu, \Lambda) \right] = \sum_{k=1}^{K} \left\{ \frac{1}{2} \ln \Lambda_k + \frac{D}{2} \ln \frac{\beta_k}{2\pi} - \frac{D}{2} - H \left[ q(\Lambda_k) \right] \right\}
$$
Re-Estimation Eqs. Using the Variational Lower Bound

We minimize now wrt $W_k, \nu_k$ (jointly). Dropping terms independent of $W_k, \nu_k$:

$$\mathcal{L} = \frac{1}{2} \sum_{k=1}^{K} \left( N_k \ln \Lambda_k - N_k \nu_k \left( \text{Tr}(W_k S_k) + \text{Tr} \left( W_k (\bar{x}_k - m_k)(\bar{x}_k - m_k)^T \right) \right) \right)$$

$$+ \ln \Lambda_k - \beta_0 \nu_k (m_k - m_0)^T W_k (m_k - m_0) + (\nu_0 - D - 1) \ln \Lambda_k - \nu_k \text{Tr} (W_0^{-1} W_k) - \ln \Lambda_k + 2H [q(\Lambda_k)]$$

where $\ln \Lambda_k \equiv \mathbb{E} [\ln |\Lambda_k|] = \sum_{i=1}^{D} \psi \left( \frac{\nu_k + 1 - i}{2} \right) + D \ln 2 + \ln \left| W_k \right|$ and

$$H [q(\Lambda_k)] = -\ln \left( B(W_k, \nu_k) \right) - \frac{\nu_k - D - 1}{2} \ln \Lambda_k + \frac{\nu_k D}{2}$$

$$\ln \left( B(W_k, \nu_k) \right) = -\frac{\nu_k}{2} \ln \left| W_k \right| - \frac{\nu_k D}{2} \ln 2 - \sum_{i=1}^{D} \ln \Gamma \left( \frac{\nu_k + 1 - i}{2} \right) - \frac{D(D - 1)}{4} \ln \pi$$

$$\mathcal{L} = \mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] + \mathbb{E} \left[ \ln p(Z | \pi) \right] + \mathbb{E} \left[ \ln p(\pi) \right] + \mathbb{E} \left[ \ln p(\mu, \Lambda) \right] - \mathbb{E} \left[ \ln q(Z) \right] - \mathbb{E} \left[ \ln q(\pi) \right] - \mathbb{E} \left[ \ln q(\mu, \Lambda) \right]$$

$$\mathbb{E} \left[ \ln p(X | Z, \mu, \Lambda) \right] = \sum_{k=1}^{K} \frac{N_k}{2} \left( \ln \Lambda_k - D \beta_k^{-1} - \nu_k \text{Tr}(W_k S_k) - \nu_k (\bar{x}_k - m_k)^T W_k (\bar{x}_k - m_k) \right)$$

$$\mathbb{E} \left[ \ln p(\mu, \Lambda) \right] = \frac{1}{2} \sum_{k=1}^{K} \left( D \ln \left( \frac{\beta_0}{2\pi} \right) + \ln \Lambda_k - \beta_0 \nu_k (m_k - m_0)^T W_k (m_k - m_0) \right) + K \ln B(W_0, \nu_0)$$

$$+ \frac{\nu_0 - D - 1}{2} \sum_{k=1}^{K} \ln \Lambda_k - \frac{1}{2} \sum_{k=1}^{K} \nu_k \text{Tr}(W_0^{-1} W_k)$$

$$\mathbb{E} \left[ \ln q(\mu, \Lambda) \right] = \sum_{k=1}^{K} \left( \frac{1}{2} \ln \Lambda_k + \frac{D}{2} \ln \beta_k - \frac{D}{2} - H [q(\Lambda_k)] \right)$$
Re-Estimation Eqs. Using the Variational Lower Bound

We minimize now wrt \( W_k, \nu_k \) (jointly). Dropping terms independent of \( W_k, \nu_k \):

\[
\mathcal{L} = \frac{1}{2} \sum_{k=1}^{K} \left( N_k \ln \Lambda_k - N_k \nu_k \left( \text{Tr}(W_k S_k) + \text{Tr}(W_k (\bar{x}_k - m_k)(\bar{x}_k - m_k)^T) \right) \right) + \ln \Lambda_k - \beta_0 \nu_k (m_k - m_0)^T W_k (m_k - m_0) + (\nu_0 - D - 1) \ln \Lambda_k - \nu_k \text{Tr}(W_0^{-1}W_k) - \ln \Lambda_k + 2H[q(\Lambda_k)]
\]

where \( \ln \tilde{\Lambda}_k \equiv \mathbb{E}[\ln |\Lambda_k|] = \sum_{i=1}^{D} \psi \left( \frac{\nu_k + 1 - i}{2} \right) + D \ln 2 + \ln |W_k| \)

and

\[
H[q(\Lambda_k)] = -\ln \left( B(W_k, \nu_k) \right) - \frac{\nu_k - D - 1}{2} \ln \Lambda_k + \frac{\nu_k D}{2} \ln(\pi) - \frac{D(D - 1)}{4} \ln \pi
\]

For a single component:

\[
\mathcal{L} = \frac{1}{2} \left( N_k + \nu_0 - \nu_k \right) \ln \Lambda_k - \frac{\nu_k}{2} \text{Tr}(W_k F_k) - \ln B(W_k, \nu_k) + \frac{\nu_k D}{2}
\]

\[
F_k = W_0^{-1} + N_k S_k + N_k \left( \bar{x}_k - m_k \right) (\bar{x}_k - m_k)^T + \beta_0 (m_k - m_0)^T (m_k - m_0)^T
\]

\[
\frac{d}{dv_k} \mathcal{L} = \frac{1}{2} \left( (N_k + \nu_0 - \nu_k) \frac{d \ln \Lambda_k}{dv_k} - \ln \Lambda_k - \text{Tr}(W_k F_k) + \ln |W_k| + D \ln 2 + \sum_{i=1}^{D} \psi \left( \frac{\nu_k + 1 - i}{2} \right) + D \right)
\]

Here use \( \ln \tilde{\Lambda}_k \equiv \sum_{i=1}^{D} \psi \left( \frac{\nu_k + 1 - i}{2} \right) + D \ln 2 + \ln |W_k| \)
Re-Estimation Eqs. Using the Variational Lower Bound

\[ \mathcal{L} = \frac{1}{2} (N_k + \nu_0 - \nu_k) \ln \Lambda_k - \frac{\nu_k}{2} \text{Tr}(W_k F_k) - \ln B(W_k, \nu_k) + \frac{\nu_k D}{2}, \quad F_k = W_0^{-1} + N_k S_k + \frac{N_k \beta_0}{N_k + \beta_0} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \]

\[ \ln \left( B(W_k, \nu_k) \right) = -\frac{\nu_k}{2} \ln |W_k| - \frac{\nu_k D}{2} \ln 2 - \sum_{i=1}^D \ln \Gamma \left( \frac{\nu_k + 1 - i}{2} \right) - \frac{D(D-1)}{4} \ln \pi \]

\[ \ln \tilde{\Lambda}_k \equiv \mathbb{E}[\ln |\Lambda_k|] = \sum_{i=1}^D \psi \left( \frac{\nu_k + 1 - i}{2} \right) + D \ln 2 + \ln |W_k| \]

\[ F_k = W_0^{-1} + N_k S_k + \frac{N_k \beta_0}{N_k + \beta_0} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \]

Similarly differentiation wrt \( W_k \) gives (use differentiation formulas \( \frac{\partial}{\partial A} \ln |A| = (A^{-1})^T, \frac{\partial}{\partial A} \text{tr}(A^T B) = B \)):

\[ \frac{d}{dW_k} \mathcal{L} = \frac{1}{2} ((N_k + \nu_0 - \nu_k)W_k^{-1} - \nu_k(F_k - W_k^{-1})) = 0 \]

From the above Eq. simultaneously with:

\[ \frac{d}{dv_k} \mathcal{L} = \frac{1}{2} \left( (N_k + \nu_0 - \nu_k) \frac{d}{dv_k} \ln \Lambda_k - \text{Tr}(W_k F_k) + D \right) = 0 \]

we see that the only solution is:

\[ 0 = N_k + \nu_0 - \nu_k \Rightarrow \nu_k = N_k + \nu_0 \]

\[ F_k - W_k^{-1} = 0 \Rightarrow W_0^{-1} + N_k S_k + \frac{N_k \beta_0}{N_k + \beta_0} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T - W_k^{-1} = 0 \]

from which:

\[ W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \]

\[ \nu_k = \nu_0 + N_k \]
Re-Estimation Eqs. Using the Variational Lower Bound

We now differentiate wrt $\alpha_k$. They appear in the following terms:

$$\mathbb{E} \left[ \ln p(Z | \pi) \right] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln \pi_k$$

$$\mathbb{E} \left[ \ln p(\pi) \right] = \ln C(\alpha_0) + (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k$$

$$\ln \bar{\pi}_k \equiv \mathbb{E} [\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}), \hat{\alpha}_k = \sum_k \alpha_k$$

$$\mathbb{E} \left[ \ln q(\pi) \right] = \sum_{k=1}^{K} (\alpha_k - 1) \ln \pi_k + \ln C(\alpha), C(\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)}, \alpha = \sum_{k=1}^{K} \alpha_k$$

Using the di- and tri-gamma functions $\psi(\cdot), \psi_1(\cdot)$ we finally have:

$$\frac{\partial}{\partial \alpha_k} \mathcal{L} = \left[ N_k + (\alpha_0 - 1) - (\alpha_k - 1) \right] \frac{\partial \ln \pi_k}{\partial \alpha_k} - \ln \pi_k - \frac{\partial \ln C(\alpha)}{\partial \alpha_k}$$

$$= \left[ N_k + (\alpha_0 - 1) - (\alpha_k - 1) \right] \left( \psi_1(\alpha_k) - \psi_1(\alpha) \frac{\partial \alpha}{\partial \alpha_k} \right) + \psi(\alpha) - \psi(\alpha_k) - \psi_1(\alpha) \frac{\partial \alpha}{\partial \alpha_k} + \psi(\alpha_k)$$

$$= \left[ N_k + (\alpha_0 - 1) - (\alpha_k - 1) \right] \psi_1(\alpha_k) = 0 \quad \Rightarrow \alpha_k = N_k + \alpha_0$$

We used here that the tri-gamma function is $>0$ and monotonically decreasing for positive arguments: $\psi_1(z) = \sum_{n=0}^{\infty} \frac{1}{(z+n)^2}$. 

WCPM Bayesian Scientific Computing, Spring 2013 (N. Zabaras)
Re-Estimation Eqs. Using the Variational Lower Bound

Finally we maximize $\mathcal{L}$ wrt $r_{nk}$ subject to $1=\sum_k r_{nk}$ for all $n$. $r_{nk}$ appears in

$$
\mathbb{E}[\ln p(Z | \pi)] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln \pi_k \quad \mathbb{E}[\ln q(Z)] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln r_{nk}
$$

$$
\mathbb{E}[\ln p(X | Z, \mu, \Lambda)] = \frac{1}{2} \sum_{k=1}^{K} N_k \left\{ \ln \Lambda_k - D \beta_k^{-1} \right\} - v_k \text{Tr}(W_k S_k) - v_k (\bar{x}_k - m_k)^T W_k (\bar{x}_k - m_k) + D \ln(2\pi)
$$

In the last expression $r_{nk}$ appears via the three terms shown

The last 2 terms in $\mathbb{E}[\ln p(X | Z, \mu, \Lambda)]$ can be written as:

$$
\frac{1}{2} \sum_{k=1}^{K} v_k \text{Tr}(W_k Q_k), Q_k = \sum_{n=1}^{N} r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T + N_k (\bar{x}_k - m_k)(\bar{x}_k - m_k)^T
$$

$$
= \sum_{n=1}^{N} r_{nk} (x_n - m_k)(x_n - m_k)^T
$$

Here we simply expanded the terms and used $N_k = \sum_{n=1}^{N} r_{nk}, \bar{x}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n$. 


Re-Estimation Eqs. Using the Variational Lower Bound

Considering all terms in $\mathcal{L}$ wrt $r_{nk}$ subject to $1=\sum_k r_{nk}$ for all n. $r_{nk}$ appears in

$$
\frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \left( \ln \Delta_k - D \beta_k^{-1} \right) - \frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \nu_k \left( x_n - m_k \right)^T W_k \left( x_n - m_k \right) + \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \ln \pi_k - \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \ln r_{nk} + \sum_{n=1}^{N} \lambda_n \left( 1 - \sum_{k=1}^{K} r_{nk} \right)
$$

Taking derivatives wrt $r_{nk}$:

$$
0 = \frac{1}{2} \ln \Delta_k - \frac{D}{2 \beta_k} - \frac{1}{2} \nu_k \left( x_n - m_k \right)^T W_k \left( x_n - m_k \right) + \ln \pi_k - \ln r_{nk} - 1 - \lambda_n
$$

Moving $\ln r_{nk}$ to the lhs and exponentiating leads to:

$$
r_{nk} \propto \tilde{\pi}_k \Delta_k^2 \exp \left( -\frac{D}{2 \beta_k} - \frac{\nu_k}{2} \left( x_n - m_k \right)^T W_k \left( x_n - m_k \right) \right)
$$

This can then be normalized as usual. This step completes all the update Eqs for the parameters of the approximations to the posterior distributions.
Predictive Distribution

In applications of the Bayesian mixture of Gaussians model we will often be interested in the predictive density for a new value \( \hat{x} \) of the observed variable. Associated with this observation will be a corresponding latent variable \( \hat{z} \), and the predictive density is then given by

\[
p(\hat{x}|X) = \sum_{\hat{z}} \int \int \int p(\hat{x}|\hat{z}, \mu, \Lambda)p(\hat{z}|\pi)p(\pi, \mu, \Lambda|X)d\pi d\mu d\Lambda
\]

where \( p(\pi, \mu, \Lambda|X) \) is the (unknown) true posterior distribution of the parameters. Using \( p(\hat{z}|\pi) = \prod_{k=1}^{K} \pi_k \hat{z}_k \) and \( p(\hat{x}|\hat{z}, \mu, \Lambda) = \prod_{k=1}^{K} \mathcal{N}(\hat{x}|\mu_k, \Lambda_k^{-1}) \hat{z}_k \), we can first perform the summation over \( \hat{z} \) to give

\[
p(\hat{x}|X) = \sum_{k=1}^{K} \int \int \pi_k \mathcal{N}(\hat{x}|\mu_k, \Lambda_k^{-1}) p(\pi, \mu, \Lambda|X)d\pi d\mu d\Lambda
\]

Because the remaining integrations are intractable, we approximate the predictive density by replacing the true posterior distribution \( p(\pi, \mu, \Lambda|X) \) with its variational approximation \( q(\pi, \mu, \Lambda) = q(\pi)q(\mu, \Lambda) = q(\pi) \prod_{k=1}^{K} q(\mu_k, \Lambda_k) \) to give (and in each term we have implicitly integrated out all variables \( \{\mu_j, \Lambda_j\} \) for \( j \neq k \))

\[
p(\hat{x}|X) \approx \sum_{k=1}^{K} \int \int \pi_k \mathcal{N}(\hat{x}|\mu_k, \Lambda_k^{-1}) q(\pi)q(\mu_k, \Lambda_k) d\pi d\mu_k d\Lambda_k
\]
Predictive Distribution

\[ p(\hat{x}|X) = \sum_{k=1}^{K} \int \int \pi_k \mathcal{N}(\hat{x}|\mu_k, \Lambda_k^{-1}) q(\pi) q(\mu_k, \Lambda_k) d\pi d\mu_k d\Lambda_k \]

Performing the integration in \( \pi \) and using \( \mathbb{E}[\pi_k] = \frac{\alpha_k}{\alpha} \), we can simplify as

\[ p(\hat{x}|X) = \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \int \int \mathcal{N}(\hat{x}|\mu_k, \Lambda_k^{-1}) q(\mu_k, \Lambda_k) d\mu_k d\Lambda_k \]

We perform next the integration in \( \mu_k \) exactly using \( q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k|\mathbf{m}_k, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k|\mathbf{W}_k, \nu_k) \) and convolution of Gaussian linear models:

\[ p(\hat{x}|X) = \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \int \int \mathcal{N}(\hat{x}|\mu_k, \Lambda_k^{-1}) \mathcal{N}(\mu_k|\mathbf{m}_k, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k|\mathbf{W}_k, \nu_k) d\mu_k d\Lambda_k \]

\[ = \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \int \mathcal{N}(\mu_k|\mathbf{m}_k, (1 + \beta_k^{-1}) \Lambda_k^{-1}) \mathcal{W}(\Lambda_k|\mathbf{W}_k, \nu_k) d\Lambda_k \]

This is the convolution over \( \Lambda_k \) of a Wishart with a Gaussian.
Predictive Distribution

\[ p(\hat{x}|X) = \sum_{k=1}^{K} \frac{\alpha_k}{\alpha} \int \mathcal{N}\left(\mu_k | m_k, (1 + \beta_k^{-1}) \Lambda_k^{-1}\right) \mathcal{W}(\Lambda_k | W_k, \nu_k) d\Lambda_k \]

\[ \propto \sum_{k=1}^{K} \frac{\alpha_k}{\alpha} \int |\Lambda_k|^{1/2 + (\nu_k - D - 1)/2} \exp\left(-\frac{1}{2(1 + \beta_k^{-1})} Tr[\Lambda_k (\hat{x} - m_k)(\hat{x} - m_k)^T] - \frac{1}{2} Tr[\Lambda_k W_k^{-1}]\right) d\Lambda_k \]

Recall the form of the Wishart distribution:

\[ \mathcal{W}(\Lambda|W, \nu) = \mathcal{B}(W, \nu)|\Lambda|^{(\nu - D - 1)/2} \exp\left(-\frac{1}{2} Tr[W^{-1} \Lambda]\right), \]

\[ \mathcal{B}(W, \nu) = |W|^{-\nu/2} \left(2^{\nu D/2} \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu + 1 - i}{2}\right)\right)^{-1} \]

Using this we can write the predictive distribution as a normalization constant of a Wishart:

\[ \left| W_k^{-1} + \frac{1}{1 + \beta_k^{-1}} (\hat{x} - m_k)(\hat{x} - m_k)^T \right|^{-(\nu_k + 1)/2} \]

Keeping only terms that depend on \( \hat{x} \):

\[ I + \frac{1}{1 + \beta_k^{-1}} W_k (\hat{x} - m_k)(\hat{x} - m_k)^T \]

and finally using \(|I + ab^T| = 1 + a^T b\):

\[ p(\hat{x}|X) \propto \sum_{k=1}^{K} \frac{\alpha_k}{\alpha} \left(1 + \frac{1}{1 + \beta_k^{-1}} (\hat{x} - m_k)^T W_k (\hat{x} - m_k)\right)^{-(\nu_k + 1)/2} \]
Predictive Distribution

\[
p(\hat{x}|X) \approx \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \left(1 + \frac{1}{1 + \beta_k^{-1}}(\hat{x} - m_k)^T W_k (\hat{x} - m_k) \right)^{-(\nu_k+1)/2}
\]

We can recognize in the parenthesis above the Student’s t distribution:

\[
St(x|\mu, \Lambda, \nu) = \frac{\Gamma\left(\frac{\nu + D}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{(\nu \pi)^{D/2}} \left(1 + \frac{\Delta^2}{\nu}\right)^{-\frac{\nu}{2}} \frac{D}{2}, \quad \Delta^2 = (x - \mu)^T \Lambda (x - \mu)
\]

Finally:

\[
p(\hat{x}|X) \approx \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} St(\hat{x}|m_k, L_k, \nu_k + 1 - D)
\]

in which the \(k\)th component has mean \(m_k\), and the precision is given by

\[
L_k = \frac{\nu_k + 1 - D}{1 + \beta_k^{-1}} W_k
\]

with \(\nu_k = \nu_0 + N_k\). We will show next that when the size \(N\) of the data set is large the predictive distribution reduces to a mixture of Gaussians.
Predictive Distribution for Large Data Set

We will show that the variational Bayes solution \( p(\bar{x}|X) \cong \sum_{k=1}^{K} \frac{\alpha_k}{\bar{\alpha}} St(\bar{x}|m_k, L_k, \nu_k + 1 - D) \) for the mixture of Gaussians model when the size \( N \) of the data set is large reduces (as we would expect) to the MLE solution based on the EM.

\[
p(\bar{x}|X) \cong \sum_{k=1}^{K} \frac{N_k}{N} N(\bar{x}|\bar{x}_k, S_k)
\]

1. We first show that the posterior distribution \( q^*(\Lambda_k) \) of the precisions becomes sharply peaked around the maximum likelihood solution.

2. We do the same for the posterior distribution of the means \( q^*(\mu_k|\Lambda_k) \).

3. Next consider the posterior distribution \( q(\pi) \) for the mixing coefficients and we show that this too becomes sharply peaked around the maximum likelihood solution.

4. Similarly, we show that the responsibilities become equal to the corresponding maximum likelihood values for large \( N \), by making use of the following asymptotic result for the digamma function for large \( x \), \( \psi(x) = \ln x + O(1/x) \)

Finally, we show that for large \( N \), the predictive distribution becomes a mixture of Gaussians.
Predictive Distribution for Large Data Set

1. We first show that the posterior distribution \( q^*(\Lambda_k) \) of the precisions becomes sharply peaked around the maximum likelihood solution.

Consider first the posterior distribution over the precision of component \( k \) given by

\[
q^*(\Lambda_k) = \mathcal{W}(\Lambda_k | W_k, \nu_k)
\]

From \( \nu_k = \nu_0 + N_k \) we see that for large \( N \) we have \( \nu_k \to N_k \), and similarly from \( W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \) we see that \( W_k \to N_k^{-1} S_k^{-1} \)

The mean of the distribution over \( \Lambda_k \) is

\[
\mathbb{E}[\Lambda_k] = \nu_k W_k \to S_k^{-1}, \quad S_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T
\]

which is the maximum likelihood value (this assumes that the quantities \( r_{nk} \) reduce to the corresponding EM values, which is indeed the case as we shall show shortly).

In order to show that this posterior is also sharply peaked, we consider the differential entropy, \( H[\Lambda_k] \) given by

\[
H[\Lambda] = -\ln B(W, \nu) - \frac{\nu - D - 1}{2} \mathbb{E}[\ln |\Lambda|] + \frac{\nu D}{2}
\]

\[
\mathbb{E}[\ln |\Lambda|] = \sum_{i=1}^{D} \psi \left( \frac{\nu + 1 + i}{2} \right) + D \ln 2 + \ln |W|
\]

and show that, as

\[
N_k \to \infty \Rightarrow H[\Lambda_k] \to 0
\]
Predictive Distribution for Large Data Set

\[ H[\Lambda] = -\ln B(W, \nu) - \frac{\nu - D - 1}{2} \mathbb{E}[\ln |\Lambda|] + \frac{\nu D}{2}, \quad \mathbb{E}[\ln |\Lambda|] = \sum_{i=1}^{D} \psi\left(\frac{\nu + 1 - i}{2}\right) + D \ln 2 + \ln |W| \]

To show \( N_k \to \infty \Rightarrow H[\Lambda_k] \to 0 \), consider first the normalizing factor. Since \( W_k \to N_k^{-1}S_k^{-1} \) and \( \nu_k \to N_k \):

\[
B(W, \nu) = |W|^{-\nu/2} \left( 2^{\nu D/2} \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu + 1 - i}{2}\right) \right)^{-1}
\]

\[-\ln B(W_k, \nu_k) \to -\frac{N_k}{2}(D \ln N_k + \ln |S_k| - D \ln 2) + \sum_{i=1}^{D} \ln \Gamma\left(\frac{N_k + 1 - i}{2}\right) \]

Stirling’s approximation: \( \ln \Gamma\left(\frac{N_k + 1 - i}{2}\right) \approx \frac{N_k}{2}(\ln N_k - \ln 2 - 1) \)

This leads to:

\[-\ln B(W_k, \nu_k) \to -\frac{N_k D}{2}(\ln N_k - \ln 2 - \ln N_k + \ln 2 + 1) - \frac{N_k}{2} \ln |S_k| = -\frac{N_k}{2}(\ln |S_k| + D) \]

Now using \( \mathbb{E}[\ln |\Lambda_k|] = \sum_{i=1}^{D} \psi\left(\frac{\nu + 1 - i}{2}\right) + D \ln 2 + \ln |W| \), \( W_k \to N_k^{-1}S_k^{-1} \) and \( \psi(x) = \ln x + O(1/x) \):

\[ \mathbb{E}[\ln |\Lambda_k|] \to D \ln \frac{N_k}{2} + D \ln 2 - D \ln N_k - \ln |S_k| \]

Finally with \( \nu_k \to N_k \) and \( N_k \to \infty \)

\[ H[\Lambda_k] = -\ln B(W_k, \nu_k) - \frac{\nu_k - D - 1}{2} \mathbb{E}[\ln |\Lambda_k|] + \frac{\nu_k D}{2} \to -\frac{N_k}{2}(\ln |S_k| + D) + \frac{N_k - D - 1}{2} \ln |S_k| + \frac{N_k D}{2} \to 0 \Rightarrow \Lambda_k \to \delta(\Lambda_k - S_k^{-1}) \]
Predictive Distribution for Large Data Set

We do the same for the posterior distribution of the means $q^*(\mu_k|\Lambda_k)$.

$$q^*(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k|m_k, (\beta_k \Lambda_k)^{-1})$$

From $m_k = \frac{1}{\beta_0 + N_k} (\beta_0 m_0 + N_k \bar{x}_k)$ we see that for large $N$, the mean of this distribution reduces to $\bar{x}_k$ which is the corresponding MLE value.

From $\beta_k = \beta_0 + N_k$, we see that $\beta_k \rightarrow N_k$ and thus $\beta_k \Lambda_k \rightarrow N_k S_k^{-1}$ which is large for large $N$ and hence this distribution is sharply peaked around its mean. Thus $q^*(\mu_k|\Lambda_k) \rightarrow \delta(\mu_k - m_k)$

Now consider the posterior distribution $q^*(\pi)$ given by $q(\pi) = \text{Dir}(\pi|\alpha)$. For large $N$ we have $\alpha_k = \alpha_0 + N_k \rightarrow N_k$ and so from $\mathbb{E}[\pi_k] = \frac{\alpha_k}{\alpha} \rightarrow \frac{N_k}{N}$, $\text{var}[\pi_i] = \frac{\alpha_i (\alpha - \alpha_i)}{\alpha^2 (\alpha + 1)} \rightarrow 0$, $\alpha = \sum_i \alpha_i$ we see that the posterior distribution becomes sharply peaked around its mean $\mathbb{E}[\pi_k] = \frac{\alpha_k}{\alpha} \rightarrow \frac{N_k}{N}$ which is the maximum likelihood solution.
Predictive Distribution for Large Data Set

For the distribution $q^*(\mathbf{z})$ we consider the responsibilities given by

$$r_{nk} \propto \tilde{\pi}_k \tilde{\Lambda}_k^{\frac{1}{2}} \exp\left(-\frac{D}{2\beta_k} - \frac{v_k}{2} (\mathbf{x}_n - \mathbf{m}_k)^T \mathbf{W}_k (\mathbf{x}_n - \mathbf{m}_k)\right).$$

Using

$$\ln \tilde{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\bar{\alpha}) \to \ln N_k - \ln N, \Rightarrow \tilde{\pi}_k \to \frac{N_k}{N}$$

$$\ln \tilde{\Lambda}_k \equiv \mathbb{E}[\ln |\Lambda_k|] = \sum_{i=1}^D \psi\left(\frac{v_k+1-i}{2}\right) + D \ln 2 + \ln|\mathbf{W}_k| \to -\ln |\mathbf{S}_k|, \mathbf{m}_k \to \bar{x}_k, \frac{D}{2\beta_k} = \frac{D}{2(\beta_0+N_k)} \to 0 \text{ and }$$

$$v_k \mathbf{W}_k \to N_k N_k^{-1} \mathbf{S}_k^{-1} = \mathbf{S}_k^{-1}.$$ We again obtain the MLE expression for the responsibilities for large $N$:

$$r_{nk} \propto \pi_k \mathcal{N}(\mathbf{x}_n | \bar{x}_k, \mathbf{S}_k).$$

Finally for the predictive distribution after the integration over $\pi$, we have:

$$p(\hat{x} | \mathbf{X}) \approx \sum_{k=1}^K \frac{\alpha_k}{\bar{\alpha}} \int \int \mathcal{N}(\hat{x} | \bar{x}_k, \mathbf{A}_k^{-1}) q(\mu_k, \Lambda_k) d\mu_k d\Lambda_k$$

$$\Lambda_k \to \delta(\Lambda_k - \mathbf{S}_k^{-1})$$

$$\mathbf{m}_k \to \bar{x}_k$$

$$q^*(\mu_k | \Lambda_k) \to \delta(\mu_k - \mathbf{m}_k)$$

The integration is simple as the arguments are now delta functions. We obtain:

$$p(\hat{x} | \mathbf{X}) \equiv \sum_{k=1}^K \frac{N_k}{N} \mathcal{N}(\hat{x} | \bar{x}_k, \mathbf{S}_k).$$
Determining the number of components

Consider a mixture of two Gaussians and a single observed variable $x$.

*Case A:* the parameters have the values $\pi_1 = a$, $\pi_2 = b$, $\mu_1 = c$, $\mu_2 = d$, $\sigma_1 = e$, $\sigma_2 = f$.

*Case B:* the parameter values $\pi_1 = b$, $\pi_2 = a$, $\mu_1 = d$, $\mu_2 = c$, $\sigma_1 = f$, $\sigma_2 = e$, in which the two components have been exchanged,

Both cases by symmetry give rise to the same value of $p(x)$.

If we have a mixture model comprising $K$ components, then each parameter setting will be a member of a family of $K!$ equivalent settings.

The lower bound needs to be modified somewhat to take into account the lack of identifiability of the parameters.

Although VB will approximate the volume occupied by the parameter posterior, it will only do so around one of the local modes.

With $K$ components, there are $K!$ equivalent modes, which differ merely by permuting the labels. Therefore we should use $\log p(X|K) \approx \mathcal{L}(K) + \log(K!)$.
Determining the number of components

In the context of MLE this redundancy is irrelevant because the parameter optimization algorithm (e.g. EM) will, depending on the initialization of the parameters, find one specific solution, and the other equivalent solutions play no role.

In a Bayesian setting, however, we marginalize over all possible parameter values. Variational inference based on the minimization of $\text{KL}(q//p)$ approximates the distribution in the neighbourhood of one of the modes and ignores the others.

Again, because equivalent modes have equivalent predictive densities, this is of no concern provided we are considering a model having a specific number $K$ of components. If, however, we wish to compare different values of $K$, then we need to take account of this multimodality.

A simple approximate solution is to add a term $\ln K!$ onto the lower bound when used for model comparison and averaging.

$$\log p(X|K) \approx \mathcal{L}(K) + \log(K!)$$
Determining the number of components

We have seen that the variational lower bound can be used to determine a posterior distribution over the number $K$ of components in the mixture model.

The simplest way to select $K$ when using VB is to fit several models, and then to use the variational lower bound to the log marginal likelihood, $\mathcal{L}(K) \leq \log p(X | K)$, to approximate

$$q(k) \propto p(k) \exp(\mathcal{L}(k))$$

where:

$$\mathcal{L}(k) = \sum_z q(Z | k) \ln \frac{p(Z, X | k)}{q(Z | k)}$$

or without the contribution from the prior (all models with the same prior probability):

$$p(K|X) \approx \frac{e^{\mathcal{L}(K)}}{\sum_{K'} e^{\mathcal{L}(K')}}$$

The lower bound should be modified $\mathcal{L}(K) + \log(K!)$ to take into account the lack of identifiability of the parameters.
Determining the number of components

The Fig. shows a plot of the lower bound, including the multimodality factor, versus the number $K$ of components for the Old Faithful data set. $K=2$ is the maximum of the lower bound.

MLE leads to values of the likelihood function that increase monotonically with $K$ (assuming the singular solutions have been avoided, and discounting the effects of local maxima) and so cannot be used to determine an appropriate model complexity.

By contrast, Bayesian inference automatically makes the trade-off between model complexity and fitting the data.

For each $K$, the model is trained from 100 different random starts (sample from $p(\pi) = \text{Dir}(\pi | \alpha_0)$), and the results are shown as ‘+’ with small random horizontal perturbations so that they can be distinguished.

Some solutions find suboptimal local maxima.
Determining the number of components

An alternative approach to determining a suitable value for $K$ is to treat the mixing coefficients $\pi$ as parameters and make point estimates of their values by maximizing the lower bound with respect to $\pi$ instead of maintaining a probability distribution over them as in the fully Bayesian approach.

When we are treating $\pi$ as a parameter, there is neither a prior, nor a variational posterior distribution, over $\pi$. Therefore, the only term remaining from the lower bound

$$L = \mathbb{E}[\ln p(X | Z, \mu, \Lambda)] + \mathbb{E}[\ln p(Z | \pi)] + \mathbb{E}[\ln p(\pi)] + \mathbb{E}[\ln p(\mu, \Lambda)] - \mathbb{E}[\ln q(Z)] - \mathbb{E}[\ln q(\pi)] - \mathbb{E}[\ln q(\mu, \Lambda)]$$

that involves $\pi$ is the second term. Note however, that $\mathbb{E}[\ln p(Z | \pi)] = \sum_n \sum_k r_{nk} \ln \pi_k$ involves the log of $\pi_k$ under $q(\pi)$, whereas here, we operate directly with $\pi_k$, yielding

$$\mathbb{E}_{q(Z)}[\ln p(Z | \pi)] = \sum_n \sum_k r_{nk} \ln \pi_k$$

Using a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$ leads to $N_k/\pi_k + \lambda = 0$ or $N_k + \lambda \pi_k = 0$ and summing over $k$ to $-\lambda = N$ and thus: $\pi_k = N_k/N$.

Determining the number of components

This leads to the re-estimation equation

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} r_{nk}$$

This maximization is interleaved with the variational updates for the $q$ distribution over the remaining parameters.

Components that provide insufficient contribution to explaining the data will have their mixing coefficients driven to zero during the optimization, and so they are effectively removed from the model through automatic relevance determination.

This allows us to make a single training run in which we start with a relatively large initial value of $K$, and allow surplus components to be pruned out of the model.

Selecting the number of mixture components

The factorized assumption causes the variance of the posterior distribution to be under-estimated.

As the number of mixture components $K$ grows, so does the number of variables that may be correlated (which are treated as independent under a variational approximation)

Thus, the proportion of probability mass under the true distribution, $p(Z, \pi, \mu, \Sigma | X)$, that the variational approximation $q(Z, \pi, \mu, \Sigma)$ does not capture, will grow.

The consequence will be that the KL divergence between $q(Z, \pi, \mu, \Sigma)$ and $p(Z, \pi, \mu, \Sigma | X)$ will increase.

Thus the lower bound must decrease compared to the true log marginal. Thus choosing the number of components based on the lower bound will tend to underestimate the optimal number of components.

The dashed arrows emphasize the typical increase in the difference between the true log marginal likelihood and the bound.

As a consequence, the bound tends to have its peak at a lower value of $K$ than the true log marginal likelihood.
MAP Estimate vs MLE

The singularities arising in the MLE treatment of Gaussian mixture models do not arise if the Bayesian model were solved using maximum posterior (MAP) estimation.

Recall that the singularities that may arise in maximum likelihood estimation are caused by a mixture component, $k$, collapsing on a data point, $x_n$, i.e., $r_{kn} = 1$, $\mu_k = x_n$ and $|\Lambda_k| \rightarrow \infty$.

However, the prior distribution $p(\mu, \Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_0, v_0)$ will prevent this from happening, also in the case of MAP estimation.

Let us compute the MAP $(\hat{\mu}, \hat{\Lambda})$ by maximizing the expected log of the product of the complete-likelihood and prior $p(\mu, \Lambda)$ as a function of $\Lambda_k$:

\[
\mathbb{E}_{q(Z)} \left[ \ln p(X | Z, \mu, \Lambda) p(\mu, \Lambda) \right] = \frac{1}{2} \sum_{n=1}^{N} r_{kn} \left( \ln |\Lambda_k| - (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right) + \frac{1}{2} \left( \ln |\Lambda_k| - \beta_0 (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) + (v_o - D - 1) \ln |\Lambda_k| - TR |W_0^{-1} \Lambda_k| \right) + \text{const}
\]

where we have used $p(X|Z, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1}) z_{nk}$, $p(\mu, \Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_0, v_0)$ and $\mathbb{E} \left[ z_{nk} \right] = r_{nk}$ together with the definitions for the Gaussian and Wishart distributions; the last term summarizes terms independent of $\Lambda_k$. Using $N_k = \sum_{n=1}^{N} r_{nk}$, $\bar{x}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n$, $S_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T$, we can rewrite this as shown next.
MAP Estimate Vs MLE

\[
\mathbb{E}_{q(Z)} \left[ \ln p(X \mid Z, \mu, \Lambda) p(\mu, \Lambda) \right] = \frac{1}{2} \sum_{n=1}^{N} r_{kn} \left( \ln |\Lambda_k| - (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right) \
\]

\[
+ \frac{1}{2} \left( \ln |\Lambda_k| - \beta_0 (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) + (v_o - D - 1) \ln |\Lambda_k| - Tr |W_0^{-1} \Lambda_k| \right) + \text{const}
\]

\[(v_o + N_k - D) \ln |\Lambda_k| - Tr \left[ \left( W_0^{-1} + \beta_0 (\mu_k - m_0)(\mu_k - m_0)^T + N_k (\bar{x}_k - \mu_k)(\bar{x}_k - \mu_k)^T + N_k S_k \right) \Lambda_k \right] + \text{const}
\]

Here we used \( N_k S_k = \sum_{n=1}^{N} r_{nk} x_n x_n^T - N_k \bar{x}_k \bar{x}_k^T \). Using \( \frac{\partial}{\partial A} \text{Tr} (AB) = B^T, \frac{\partial}{\partial A} \ln |A| = (A^{-1})^T \) we can compute the derivative of this w.r.t. \( \Lambda_k \) and setting the result equal to zero, we find the MAP estimate for \( \Lambda_k \) to be

\[
\Lambda_k^{-1} = \frac{1}{(v_o + N_k - D)} \left( W_0^{-1} + \beta_0 (\mu_k - m_0)(\mu_k - m_0)^T + N_k (\bar{x}_k - \mu_k)(\bar{x}_k - \mu_k)^T + N_k S_k \right)
\]

Here we use for the MLE estimate \( \mu_k = m_k \) and \( W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \)

From this we see that \( |\Lambda_k^{-1}| \) can never become 0, because of the presence of \( W_0^{-1} \) (which we must chose to be positive definite) in the expression on the r.h.s. Note this result is not the same as the mode of

\[
\text{mode} \left[ q^*(\Lambda_k) \right] = (\nu_k - D - 1) W_k
\]
Induced Factorizations

In deriving variational update equations for the Gaussian mixture model, we assumed a particular factorization of the variational posterior distribution given by

\[ q(Z, \pi, \mu, \Lambda) = q(Z) q(\pi, \mu, \Lambda) \]

However, the optimal solutions for the various factors exhibit additional factorizations.

In particular, \( q(\mu, \Lambda) = \prod_k q(\mu_k, \Lambda_k) \), whereas \( q^*(Z) = \prod_{n=1}^N \prod_{k=1}^K r_{nk} z_{nk} \) factorizes into an independent distribution \( q(z_n) \) for each observation \( n \).

Note that \( q^*(Z) \) does not further factorize with respect to \( k \) because, for each value of \( n \), the \( z_{nk} \) are constrained to sum to one over \( k \).

These additional factorizations are a consequence of the interaction between the assumed factorization and the conditional independence properties of the true distribution.
Induced Factorizations

*Induced factorizations* arise from an interaction between the factorization assumed in the variational posterior distribution and the conditional independence properties of the true joint distribution.

It is important to account for such additional factorizations.

E.g., it would be very inefficient to maintain a full precision matrix for the Gaussian distribution over a set of variables if the optimal form for that distribution always had a diagonal precision matrix (corresponding to a factorization with respect to the individual variables described by that Gaussian).

Such induced factorizations can easily be detected using d-separation. Partition the latent variables into three disjoint Groups $A,B,C$ and then let us assume a factorization between $C$ and $A,B$ so that

$$q(A,B,C) = q(A,B)q(C).$$

Using the general result $\ln q^*_j(Z_j) = \mathbb{E}_{i \neq j}[\ln p(X,Z)] + \text{const}$ together with the product rule for probabilities, we see that the optimal solution for $q(A,B)$ is given by

$$\ln q(A,B) = \mathbb{E}_C[\ln p(X,A,B,C)] + \text{const}$$

$$= \mathbb{E}_C[\ln p(A,B/X,C)] + \text{const}.$$  

We now ask whether this resulting solution will factorize between $A$ and $B$, in other words whether $q(A,B) = q(A)q(B)$. This will happen if, and only if, $\ln p(A,B/X,C) = \ln p(A/X,C) + \ln p(B/X,C)$, that is, if the conditional independence relation $A \perp \perp B / X,C$ holds.
Induced Factorizations for the Bayesian Mixture of Gaussians

In the Bayesian mixture of Gaussians we are assuming a factorization given by

\[ q(Z, \pi, \mu, \Lambda) = q(Z)q(\pi, \mu, \Lambda) \]

The variational posterior distribution over the parameters must factorize between \( \pi \) and the remaining parameters \( \mu \) and \( \Lambda \) because all paths connecting \( \pi \) to either \( \mu \) or \( \Lambda \) must pass through one of the nodes \( z_n \) all of which are in the conditioning set for our conditional independence test and all of which are head-to-tail with respect to such paths:

\[
p(X, Z, \pi, \mu, \Lambda) = p(X, \mu, \Lambda|Z, \pi) \ p(Z, \pi) = p(X|Z, \pi, \mu, \Lambda) \ p(\mu, \Lambda|Z, \pi) \ p(Z, \pi) = p(X|Z, \mu, \Lambda) \ p(\mu, \Lambda) \ p(Z|\pi) \ p(\pi)
\]

\[
\ln q^*(\pi, \mu, \Lambda) = \mathbb{E}_Z[\ln p(X, Z, \pi, \mu, \Lambda)] + \text{const} = \mathbb{E}_Z[\ln(p(X|Z, \mu, \Lambda)p(Z|\pi)p(\pi)p(\mu, \Lambda))] + \text{const}
\]

The variational posterior clearly factorizes into the following form

\[
q(\pi, \mu, \Lambda) = q(\pi)q(\mu, \Lambda) = q(\pi) \prod_{k=1}^{K} q(\mu_k, \Lambda_k)
\]
**Variational Message Passing (VMP)**

We have seen that mean field methods are all very similar: just compute each node’s full conditional, and average out the neighbors.

This is similar to Gibbs sampling except the derivation of the eqs is usually a bit more work.

Fortunately it is possible to derive a general purpose set of update equations that work for any DGM for which all CPDs are in the exponential family, and for which all parent nodes have conjugate distributions. (See (Wand et al. 2011) for a recent extension to handle non-conjugate priors.)

One can then sweep over the graph, updating nodes one at a time, in a manner similar to Gibbs sampling. This is known as VMP and has been implemented in the open-source program VIBES5.

This is a VB analog to BUGS, which is a popular generic program for Gibbs sampling. VMP/ mean field is best-suited to inference where one or more of the hidden nodes are continuous (e.g., when performing “Bayesian learning”).