Finite Mixture Models

Mixture distribution is described by

\[
p(x) = \sum_{j=1}^{K} p(x | \omega_j) = \sum_{j=1}^{K} p(x | \omega_j) p(\omega_j).
\]

- \(p(x | \omega_j)\) are component densities and are normalized, i.e.,
  \[
  \int p(x | \omega_j) dx = 1.
  \]
- \(p(\omega_j) = \pi_j\) are mixing parameters which are prior probabilities of data points generated from the component \(j\) of the mixture and satisfy
  \[
  \sum_{j=1}^{K} p(\omega_j) = 1, \quad 0 \leq p(\omega_j) \leq 1.
  \]

Multinomial Parameterization

Introduce a \(K\)-dimensional binary random variable \(z\) having a 1-of-\(K\) representation in which a particular element \(z_j\) is equal to 1 and all other elements are equal to 0, i.e.,

\[
z = [0, \ldots, 1, \ldots, 0]^T \in \mathbb{R}^K.
\]

The marginal distribution over \(z\) is specified in terms of the mixing coefficients \(\pi_j\) such that

\[
\pi_j = p(z_j = 1).
\]

Then we have

\[
p(z) = \prod_{j=1}^{K} \pi_j^z.
\]
Finite Mixture Models: Revisted

We rewrite the mixture distribution as

\[
p(x) = \sum_z p(x, z) = \sum_{j=1}^{K} p(x | z_j = 1) p(z_j = 1)
= \sum_{j=1}^{K} \pi_j p(x | z_j = 1).
\]

Mixture of Gaussians (MoG): Graphical Model

- Multinomial parameterization for indicator variables
  \[ z_t = [0, \ldots, 1, \ldots, 0]^T \in \mathbb{R}^K \]
  such that \( \pi_j = P(z_R = 1) \):
  \[
p(z_t) = \prod_{j=1}^{K} \pi_j^{z_{jt}}.
\]
- The distribution over observed variables conditioned on latent variables is
  \[
p(x_t | z_t) = \prod_{j=1}^{K} N(x_t | \mu_j, \Sigma_j)^{z_{jt}}.
\]

Learning MoG

- Compute maximum likelihood estimates of parameters
  \[ \{ \pi, \{ \mu_j \}, \{ \Sigma_j \} \}. \]
- Compute the posterior distribution over latent variables
  \[ p(z_t | x_t) \]
  which corresponds to cluster membership probabilities.
- We will do this using expectation maximization.

MoG: An Example
EM for MoG

- Introduce indicating variables $z_t \in \{1, \ldots, K\}$ specifying which component of the mixture generated data points.
- Complete-data log-likelihood is given by

$$L_c = \sum_{t=1}^{N} \log p(x_t, z_t) = \sum_{t=1}^{N} \log [p(x_t|z_t)p(z_t)].$$

Multinomial Parameterization

- Introduce a binary vector $z_t = [0, \ldots, 1, \ldots, 0]^T \in \mathbb{R}^K$ such that $\pi_j = P(z_t = 1)$.
- Then we have

$$p(z_t) = \prod_{j=1}^{K} \pi_j^{z_{jt}}, \quad p(x_t|z_t) = \prod_{j=1}^{K} \mathcal{N}(x_t|\mu_j, \Sigma_j)^{z_{jt}}.$$

MoG: E-Step (Cont’d)

Then, the expected complete-data log-likelihood is calculated as

$$\langle L_c \rangle_{z_t|x_t} = \sum_{t=1}^{N} \sum_{j=1}^{K} \langle z_{jt} \log [\pi_j \mathcal{N}(x_t|\mu_j, \Sigma_j)] \rangle$$

$$= \sum_{t=1}^{N} \sum_{j=1}^{K} r_{jt} \log [\pi_j \mathcal{N}(x_t|\mu_j, \Sigma_j)],$$

where $r_{jt} = \langle z_{jt} \rangle_{z_t|x_t} = p(z_{jt}|x_t)$.

Thus, we have

$$\langle L_c \rangle = \sum_{t=1}^{N} \sum_{j=1}^{K} r_{jt} \left\{ \log \pi_j - \frac{D}{2} \log \sigma_j^2 - \frac{1}{2\sigma_j^2} \|x_t - \mu_j\|^2 \right\} + \text{const}$$

MoG: M-Step

In the M-step, we update parameters $\theta = \{\pi_j, \mu_j, \sigma_j^2\}$ such that the expected complete-data log-likelihood is maximized:

$$\theta \leftarrow \arg \max_{\theta} \langle L_c \rangle.$$

Thus, we have

$$\frac{\partial \langle L_c \rangle}{\partial \mu_j} = -\frac{1}{\sigma_j^2} \sum_{t} r_{jt} (\mu_j - x_t) = 0$$

$$\Rightarrow \mu_j^{\text{new}} = \frac{\sum_{t} r_{jt} x_t}{\sum_{t} r_{jt}}.$$

$$\frac{\partial \langle L_c \rangle}{\partial \sigma_j^2} = \sum_{t} r_{jt} \left\{ -\frac{D}{\sigma_j^2} + \frac{1}{2\sigma_j^2} \|x_t - \mu_j^{\text{new}}\|^2 \right\} = 0$$

$$\Rightarrow \sigma_j^{2 \text{ new}} = \frac{\sum_{t} r_{jt} \|x_t - \mu_j^{\text{new}}\|^2}{\sum_{t} r_{jt}}.$$
MoG: M-Step (Cont’d)

Consider the Lagrangian
\[
\langle \mathcal{L}_c \rangle = \langle \mathcal{L} \rangle + \lambda \left( 1 - \sum_j \pi_j \right).
\]

\[
\frac{\partial \langle \mathcal{L}_c \rangle}{\partial \pi_j} = \sum_t r_{jt} \left\{ \frac{1}{\pi_j} \right\} - \lambda = 0.
\]

The optimal Lagrangian multiplier \( \lambda \) is given by \( \lambda = N \). Thus we have an updating rule that has the form
\[
\pi_j^{\text{new}} = \frac{1}{N} \sum_{t=1}^N r_{jt}.
\]

Algorithm Outline: EM-MoG (Isotropic)

E-step: Compute responsibilities
\[
r_{jt} = \frac{p(x_t|\omega_j)\pi_j}{\sum_{j=1}^M p(x_t|\omega_j)\pi_j}.
\]

M-step: Update parameters
\[
\mu_j^{\text{new}} = \frac{\sum_{t=1}^N r_{jt}x_t}{R_j}, \quad \sigma_j^{2\text{ new}} = \frac{1}{D} \frac{\sum_{t=1}^N r_{jt}||x_t - \mu_j^{\text{new}}||^2}{R_j},
\]
\[
\pi_j^{\text{new}} = \frac{1}{N} \sum_{t=1}^N r_{jt}.
\]

Algorithm Outline: EM-MoG (General Case)

E-step: Compute responsibilities
\[
R_{jt} = \frac{p(x_t|\omega_j)\pi_j}{\sum_{j=1}^M p(x_t|\omega_j)\pi_j}.
\]

M-step: Update parameters
\[
\mu_j^{\text{new}} = \frac{\sum_{t=1}^N R_{jt}x_t}{R_j}, \quad \Sigma_j^{\text{new}} = \frac{\sum_{t=1}^N R_{jt}x_t x_t^T}{R_j} - \mu_j^{\text{new}} [\mu_j^{\text{new}}]^T, \quad \pi_j^{\text{new}} = \frac{1}{N} \sum_{t=1}^N R_{jt}.
\]

Model Selection

How to decide \( K \) (the number of clusters)?

- Akaike Information Criterion (AIC)
  \[
  \text{AIC} = -2 \log p(X|\theta) + 2M.
  \]
  \( M \) is the number of parameters.

- Bayesian Information Criterion (BIC)
  \[
  \text{BIC} = -2 \log p(X|\theta) + M \log N.
  \]
Variational Bayesian Mixture of Gaussians: Graphical Model

- We treat parameters $\pi$, $\mu$, $\Lambda$ as random variables, where we choose a Dirichlet distribution over $\pi$ and an independent Gaussian-Wishart prior governing the mean and precision of each Gaussian component.
- Compute variational posterior distribution over $z_t$.

Multinomial Parameterization: Once Again

- We denote the latent variables by $Z = [z_1, \ldots, z_N]$, where each $z_t$ comprises a 1-of-$K$ binary vector, i.e., $z_t = [0, \ldots, 1, \ldots, 0]^T \in \mathbb{R}^K$.
- The conditional distribution of $Z$ given the mixing coefficients $\pi$ is of the form
  
  $p(Z|\pi) = \prod_{t=1}^N \prod_{j=1}^K \pi_j^{z_{jt}}.$

Conjugate Prior Distributions over Parameters

- Prior over $\pi$: Choose a Dirichlet distribution over $\pi$
  
  $p(\pi) = \text{Dir}(\pi|\alpha_0) = \frac{1}{Z(\alpha_0)} \prod_{j=1}^K \pi_j^{\alpha_0_j-1},$

  where the same hyperparameter $\alpha_0$ is chosen for each of the components.

- Prior over $\mu$ and $\Lambda$: Choose independent Gaussian-Wishart distribution
  
  $p(\mu, \Lambda) = p(\mu|\Lambda)p(\Lambda)$
  
  $= \prod_{j=1}^K \mathcal{N}(\mu_j | \mu_0, (\beta \Lambda_j)^{-1}) \mathcal{W}(\Lambda_j | W_0, \nu_0),$

  which is the conjugate prior when both the mean and precision are unknown.

Dirichlet Distribution

- The Dirichlet is a multivariate distribution over $K$ random variables $\pi_j \in [0, 1]$, where $j = 1, \ldots, K$, subject to the sum-to-one constraint $\sum_{j=1}^K \pi_j = 1$.
- Denote $\pi = [\pi_1, \ldots, \pi_K]^T$ and $\alpha = [\alpha_1, \ldots, \alpha_K]^T$. Then the Dirichlet distribution is described by

  $\text{Dir}(\pi | \alpha) = \frac{1}{Z(\alpha)} \prod_{j=1}^K \pi_j^{\alpha_j-1},$

  where

  \[
  Z(\alpha) = \int_0^1 \cdots \int_0^1 \pi_1^{\alpha_1-1} \cdots \pi_K^{\alpha_K-1} d\alpha_1 \cdots d\alpha_K
  = \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)}{\Gamma(\alpha_1 + \cdots + \alpha_K)}.
  \]
Wishart Distribution

- The Wishart distribution is the conjugate prior for the precision matrix of a multivariate Gaussian.

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

where \( W \in \mathbb{R}^{D \times D} \) is symmetric and positive definite and

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

- The parameter \( \nu \) is called the number of degrees of freedom of the distribution and is restricted to \( \nu > D - 1 \) to ensure that the Gamma function in the normalization factor is well-defined.

- In one dimension, the Wishart reduces to the Gamma distribution \( \text{Gam}(\lambda \mid a, b) \) with parameters \( a = \nu/2 \) and \( b = 1/2W \).

Algorithm Outline: VBEM

- **VBE-step**: Update \( q_\nu(S) \)

\[ q^{(k+1)}_\nu(S) = \frac{1}{Z_\nu} \exp \left\{ \mathbb{E}_{q_\nu} \log p(X, S, \theta) \right\}. \]

- **VBM-step**: Update \( q_\theta(\theta) \)

\[ q^{(k+1)}_\theta(\theta) = \frac{1}{Z_\theta} \exp \left\{ \mathbb{E}_{q^{(k+1)}_\nu} \log p(X, S, \theta) \right\}. \]

Gaussian-Wishart Distribution

- The Gaussian-Wishart is the conjugate prior for a multivariate Gaussian \( \mathcal{N}(x \mid \mu, \Lambda) \) in which both the mean \( \mu \) and the precision \( \Lambda \) are unknown.

- It comprises the product of a Gaussian distribution for \( \mu \), whose precision is proportional to \( \Lambda \) and a Wishart distribution over \( \Lambda \).

\[ p(\mu, \Lambda \mid m_0, \beta, W, \nu) = \mathcal{N}(\mu \mid m_0, (\beta \Lambda)^{-1}) \mathcal{W}(\Lambda \mid W, \nu). \]

Variational Inference: General Case

- Denote by \( Y \) the set of all latent variables and parameters and assume that the variational distribution factorizes:

\[ q(Y) = \sum_{i=1}^{M} q_i(Y_i). \]

- Variational inference determines the optimal solution \( q_j^*(Y_j) \) given by

\[ \log q_j^*(Y_j) = \mathbb{E}_{q_j} \left\{ \log p(X, Y_1, \ldots, Y_M) \right\}, \]

or

\[ q_j^*(Y_j) = \frac{\exp \left\{ \mathbb{E}_{q_j} \left[ \log p(X, Y_1, \ldots, Y_M) \right] \right\}}{\int \exp \left\{ \mathbb{E}_{q_j} \left[ \log p(X, Y_1, \ldots, Y_M) \right] \right\} dY_j}. \]
Variational Distribution

- The joint distribution of all of the random variables factorizes as follows.

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

- Consider a variational distribution which factorizes between the latent variables and the parameters so that

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

\[
= q(Z)q(\pi)\prod_{j=1}^{K} q(\mu_j|\Lambda)q(\Lambda).
\]

- No assumption on the functional form of \(q(Z)\) and \(q(\pi, \mu, \Lambda)\). The functional form of \(q(Z)\) and \(q(\pi, \mu, \Lambda)\) is determined automatically by optimization of the variational distribution. (free-form optimization)

Optimized \(q^*(Z)\) (Cont’d)

Taking the exponential of both sides of this equation, we have

\[
q^*(Z) \propto \prod_{t=1}^{N} \prod_{j=1}^{K} \rho_{jt}^{z_{jt}}.
\]

\[
\log \rho_{jt} = \mathbb{E}_\pi \{\log \pi_j\} + \frac{1}{2} \mathbb{E}_\Lambda \{\log |\Lambda_j|\} - \frac{D}{2} \log 2\pi - \frac{1}{2} \mathbb{E}_\mu A \{(x_t - \mu_j)^T \Lambda_j (x_t - \mu_j)\}.
\]

Note that for each value of \(t\), \(z_{jt}\) are binary and sum to 1 over all values of \(j\), leading to

\[
q^*(Z) = \prod_{t=1}^{N} \prod_{j=1}^{K} \rho_{jt}^{z_{jt}}, \quad \rho_{jt} = \frac{\rho_j}{\sum_{k=1}^{K} \rho_{kt}}.
\]

Optimized \(q^*(Z)\) (Cont’d)

The log of the optimized factor \(q^*(Z)\) is given by

\[
\log q^*(Z) = \mathbb{E}_\pi, \mu, \Lambda \{\log p(X, Z, \pi, \mu, \Lambda)\}\]

\[
+ \mathbb{E}_\pi \{\log p(Z|\pi)\} + \mathbb{E}_\mu, \Lambda \{p(X|Z, \mu, \Lambda)\} + \text{const}
\]

\[
= \mathbb{E}_\pi \left\{ \sum_{t=1}^{N} \sum_{j=1}^{K} z_{jt} \log \pi_j \right\} + \text{const}
\]

\[
+ \mathbb{E}_\mu, \Lambda \left\{ \sum_{t=1}^{N} \sum_{j=1}^{K} z_{jt} \left( \frac{1}{2} \log |\Lambda_j| - \frac{1}{2} (x_t - \mu_j)^T \Lambda_j (x_t - \mu_j) \right) \right\}
\]

\[
= \sum_{t=1}^{N} \sum_{j=1}^{K} z_{jt} \left[ \mathbb{E}_\pi \{\log \pi_j\} + \frac{1}{2} \mathbb{E}_\Lambda \{\log |\Lambda_j|\} - \frac{D}{2} \log 2\pi - \frac{1}{2} \mathbb{E}_\mu, \Lambda \{(x_t - \mu_j)^T \Lambda_j (x_t - \mu_j)\} \right] + \text{const}.
\]

Note that the optimized \(q^*(Z) = \prod_{t=1}^{N} \prod_{j=1}^{K} r_{jt}^{z_{jt}}\) takes the same functional form as the prior \(p(Z|\pi) = \prod_{t=1}^{N} \prod_{j=1}^{K} \pi_{jt}^{z_{jt}}\).

Note that \(r_{jt}\) is given by the exponential of a real quantity, the quantities \(r_{jt}\) are nonnegative and sum to one, as required.

For the discrete distribution \(q^*(Z)\), we have the standard result

\[
\mathbb{E} \{z_{jt}\} = r_{jt},
\]

from which we see that the quantities \(r_{jt}\) are playing the role of responsibilities.
Optimized \( q^*(\pi, \mu, \Lambda) = q^*(\pi) \prod_{j=1}^{K} q^*(\mu_j, \Lambda_j) \)

The log of the optimized factor \( q^*(\pi, \mu, \Lambda) \) is given by

\[
\log q^*(\pi, \mu, \Lambda) = E_Z \{ \log p(X, Z, \pi, \mu, \Lambda) \} + \text{const}
\]

\[
= \log p(\pi) + E_Z \{ \log p(Z|\pi) \} + \sum_{j=1}^{K} \log p(\mu_j, \Lambda_j)
\]

\[
+ \sum_{t=1}^{N} \sum_{j=1}^{K} E \{ z_{jt} \} \left[ \frac{1}{2} \log \frac{1}{(2\pi)^D} - \frac{1}{2} (x_t - \mu_j)^\top \Lambda_j (x_t - \mu_j) \right] + \text{const}
\]

\[
= \log q^*(\pi) + \sum_{j=1}^{K} \log [q^*(\mu_j|\Lambda_j)q^*(\Lambda_j)].
\]

Optimized \( q^*(\pi) \)

\[
\log q^*(\pi) = \log p(\pi) + E_Z \{ \log p(Z|\pi) \} + \text{const}
\]

\[
= (\alpha_0 - 1) \sum_{j=1}^{K} \log \pi_j + \sum_{t=1}^{N} \sum_{j=1}^{K} r_{jt} \log \pi_j + \text{const}
\]

\[
= (\alpha_0 - 1) \sum_{j=1}^{K} \log \pi_j + \sum_{j=1}^{K} N_j \log \pi_j + \text{const}
\]

\[
= \sum_{j=1}^{K} (\alpha_j + N_j - 1) \log \pi_j + \text{const}.
\]

Optimized \( q^*(\mu) \)

\[
\log q^*(\mu) = \log p(\mu) + E_Z \{ \log p(Z|\mu) \} + \text{const}
\]

\[
= \sum_{t=1}^{N} \sum_{j=1}^{K} r_{jt} \sum_{t=1}^{N} r_{jt} x_t + \frac{1}{N_j} \sum_{t=1}^{N} r_{jt} (x_t - \bar{x}_j)(x_t - \bar{x}_j)^\top.
\]

Optimized \( q^*(\Lambda) \)

\[
\log q^*(\Lambda) = \log p(\Lambda) + E_Z \{ \log p(Z|\Lambda) \} + \text{const}
\]

\[
= \sum_{t=1}^{N} \sum_{j=1}^{K} r_{jt} (x_t - \bar{x}_j)(x_t - \bar{x}_j)^\top.
\]

Notations

We define three statistics of the observed data set evaluated with respect to the responsibilities, given by

\[
N_j = \sum_{t=1}^{N} r_{jt},
\]

\[
\bar{x}_j = \frac{1}{N_j} \sum_{t=1}^{N} r_{jt} x_t,
\]

\[
S_j = \frac{1}{N_j} \sum_{t=1}^{N} r_{jt} (x_t - \bar{x}_j)(x_t - \bar{x}_j)^\top.
\]

Thus, the optimized \( q(\pi) \) is given by

\[
q^*(\pi) = \text{Dir}(\pi | \alpha) \propto \prod_{j=1}^{K} \pi_j^{\alpha_j - 1},
\]

where \( \alpha = [\alpha_1, \ldots, \alpha_K]^\top \) and \( \alpha_j = \alpha_0 + N_j \).
Optimized \( q^*(\mu, \Lambda) \)

Consider

\[
\log q^*(\mu, \Lambda) = \sum_{j=1}^{K} \log p(\mu_j, \Lambda_j) + \sum_{t=1}^{N} \sum_{j=1}^{K} r_{jt} \log \mathcal{N}(x_t | \mu_j, \Lambda_j^{-1}) + \text{const.}
\]

Note that the Gaussian-Wishart prior is given by

\[
p(\mu, \Lambda | m_0, \beta_0, W_0, \nu_0) = \prod_{j=1}^{K} p(\mu_j | \Lambda_j) p(\Lambda_j)
\]

\[
= \prod_{j=1}^{K} \mathcal{N}(\mu_j | m_0, (\beta_0 \Lambda_j)^{-1}) \mathcal{W}(\Lambda_j | W_0, \nu_0),
\]

where

\[
\mathcal{W}(\Lambda_j | W_0, \nu_0) = B(W_0, \nu_0) |\Lambda_j|^{(\nu_0-D-1)/2} \exp \left\{ -\frac{1}{2} \text{tr} (W_0^{-1} \Lambda_j) \right\}.
\]

Optimized \( q^*(\mu, \Lambda) \) (Cont’d)

One can show that (which is your homework!) \( q^*(\mu, \Lambda) \) is again independent Gaussian-Wishart:

\[
q^*(\mu, \Lambda) = \prod_{j=1}^{K} \mathcal{N}(\mu_j | m_j, (\beta_j \Lambda_j)^{-1}) \mathcal{W}(\Lambda_j | \nu_j),
\]

where

\[
\beta_j = \beta_0 + N_j,
\]

\[
m_j = \frac{1}{\beta_j} (\beta_0 m_0 + N_j \bar{x}_j),
\]

\[
W_j^{-1} = W_0^{-1} + N_j S_j + \frac{\beta_0 N_j}{\beta_0 + N_j} (\bar{x}_j - m_0)(\bar{x}_j - m_0)^T,
\]

\[
\nu_j = \nu_0 + N_j.
\]

Optimized \( q^*(\mu, \Lambda) \) (Cont’d)

Then we have

\[
\log q^*(\mu, \Lambda) = \sum_{j=1}^{K} \left[ \log \mathcal{N}(\mu_j | m_0, (\beta_0 \Lambda_j)^{-1}) + \log \mathcal{W}(\Lambda_j | W_0, \nu_0) \right]
\]

\[
+ \sum_{j=1}^{K} \sum_{t=1}^{N} r_{jt} \log \mathcal{N}(x_t | \mu_j, \Lambda_j^{-1}) + \text{const}
\]

\[
= \sum_{j=1}^{K} \log q^*(\mu_j | \Lambda_j) + \sum_{j=1}^{K} \log q^*(\Lambda_j).
\]

Algorithm Outline: Variational Bayesian E-Step

Compute the optimized variational distributions over latent variables using the current distributions over model parameters:

\[
q^*(Z) = \prod_{i=1}^{N} \prod_{j=1}^{K} \frac{p(\tau_{ij})}{\sum_{k=1}^{K} p(\tau_{ik})},
\]

\[
\log p_{\tau_{jt}} = E_\pi \{ \log \tau_{jt} \} + \frac{1}{2} E_\Lambda \{ \log |\Lambda_j| \} - \frac{D}{2} \log 2\pi
\]

\[
- \frac{1}{2} E_\mu \Lambda \left\{ (x_t - \mu_j)^T \Lambda_j (x_t - \mu_j) \right\},
\]

\[
r_{jt} \propto \tau_{jt} \exp \left\{ -\frac{1}{2} (x_t - \mu_j)^T \Lambda_j (x_t - \mu_j) \right\},
\]

\[
E_\pi \{ \log \tau_{jt} \} = \psi(\alpha_j) - \psi(\alpha_1 + \cdots + \alpha_K),
\]

\[
E_\Lambda \{ \log |\Lambda_j| \} = \sum_{i=1}^{D} \psi \left( \frac{\nu_j + \frac{i-1}{2} + \frac{1}{2}}{2} \right) + D \log 2 + \log |W_j|,
\]

\[
E_\mu \Lambda \left\{ (x_t - \mu_j)^T \Lambda_j (x_t - \mu_j) \right\} = D \beta_j^{-1} + \nu_j (x_t - \mu_j)^T W_j (x_t - \mu_j).
\]
Algorithm Outline: Variational Bayesian M-Step
Compute the optimized variational distributions over model parameters
using the current distributions over latent variables:

\[ q^*(\mu, \Lambda) = \prod_{j=1}^{K} \mathcal{N}(\mu_j | m_j, (\beta_j \Lambda_j)^{-1}) \mathcal{W}(\Lambda_j | \nu_j), \]

\[ N_j = \sum_{t=1}^{N} r_{jt}, \quad \mathbf{x}_j = \frac{1}{N_j} \sum_{t=1}^{N} r_{jt} \mathbf{x}_t, \]

\[ S_j = \frac{1}{N_j} \sum_{t=1}^{N} r_{jt}(\mathbf{x}_t - \mathbf{x}_j)(\mathbf{x}_t - \mathbf{x}_j)^\top. \]

\[ \beta_j = \beta_0 + N_j, \quad \mathbf{m}_j = \frac{1}{\beta_j} (\beta_0 \mathbf{m}_0 + N_j \mathbf{x}_j), \quad \nu_j = \nu_0 + N_j, \]

\[ W_j^{-1} = W_0^{-1} + N_j S_j + \frac{\beta_0 N_j}{\beta_0 + N_j} (\mathbf{x}_j - \mathbf{m}_0)(\mathbf{x}_j - \mathbf{m}_0)^\top, \]

Predictive Distributions
Predictive distribution for a new data \( \mathbf{x}_* \) is computed by

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

Recall multinomial parameterizations:

\[ p(\mathbf{z}_*) = \prod_{j=1}^{K} \mathcal{P}_{\beta_j}, \]

\[ p(\mathbf{x}_* | \mathbf{z}_*, \mu, \Lambda) = \prod_{j=1}^{K} \mathcal{N}(\mathbf{x}_* | \mu_j, \Lambda_j^{-1})^{2\nu_j}. \]

Perform the summation over \( \mathbf{z}_* \) to give

\[ p(\mathbf{x}_* | \mathbf{X}) = \sum_{j=1}^{K} \int \int \pi_j \mathcal{N}(\mathbf{x}_* | \mu_j, \Lambda_j^{-1}) p(\pi, \mu, \Lambda | \mathbf{X}) d\pi d\mu d\Lambda. \]

Example: Old Faithful Data
Component whose expected mixing coefficients are numerically
indistinguishable from zero are not plotted.

Since the true posterior distribution of the parameters \( p(\pi, \mu, \Lambda | \mathbf{X}) \) is
unknown, we approximate the predictive distribution by replacing the true
posterior distribution with its variational approximation \( q(\pi) q(\mu, \Lambda) \) to
give

\[ p(\mathbf{x}_* | \mathbf{X}) \approx \sum_{j=1}^{K} \int \int \pi_j \mathcal{N}(\mathbf{x}_* | \mu_j, \Lambda_j^{-1}) q(\pi) q(\mu_j, \Lambda_j) d\pi d\mu_j d\Lambda_j, \]

where we used \( q(\pi, \mu, \Lambda) = q(\pi) \prod_{j=1}^{K} q(\mu_j, \Lambda_j) \).
The integrations can be evaluated analytically, leading to a mixture of
Student’s t-distributions:

\[ p(\mathbf{x}_* | \mathbf{X}) \approx \frac{1}{\alpha_1 + \cdots + \alpha_K} \sum_{j=1}^{K} \alpha_j \mathcal{St}(\mathbf{x}_*, \mathbf{m}_j, \mathbf{L}_j, \nu_j + 1 - D), \]

where

\[ \mathbf{L}_j = \frac{(\nu_j + 1 - D)\beta_j}{(1 + \beta_j)} W_j. \]
Determining the Number of Components

- Plot the variational lower bound on the marginal likelihood versus the number \( K \) of components in VBMoG, for the Old Faithful data.
- Peak at \( K = 2 \).
- For each value of \( K \), the model is trained from 100 different random starts.