SDS 383C: Statistical Modeling I
Fall 2022, Module VI

Abhra Sarkar

Department of Statistics and Data Sciences
The University of Texas at Austin

"All models are wrong, but some are useful."- George E. P. Box

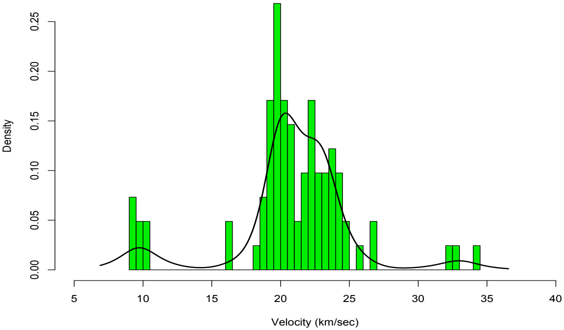
Normal Mixture Models

$$y_1, \ldots, y_n \stackrel{iid}{\sim} \sum_{k=1}^{K} \pi_k \text{Normal}(\mu_k, \sigma_k^2)$$

- **Likelihood:** $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{ -\frac{1}{2\sigma_k^2}(y_i - \mu_k)^2 \right\} \right]$

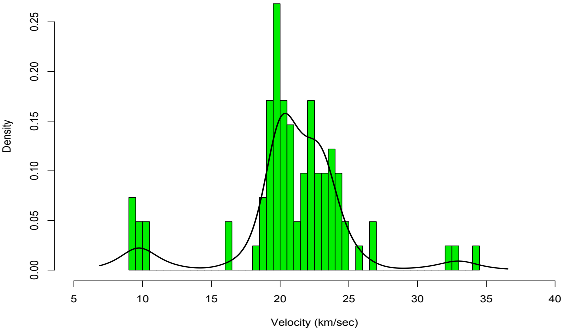
$$y_1, \ldots, y_n \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \text{Normal}(\mu_k, \sigma_k^2)$$

- **Likelihood:** $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{ -\frac{1}{2\sigma_k^2}(y_i - \mu_k)^2 \right\} \right]$



$$y_1, \ldots, y_n \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \text{Normal}(\mu_k, \sigma_k^2)$$

- **Likelihood:** $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{ -\frac{1}{2\sigma_k^2}(y_i - \mu_k)^2 \right\} \right]$

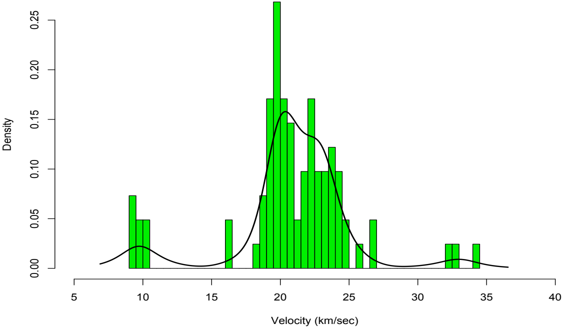


- **Theorem:** Location mixtures of normals can approximate any continuous density.
- Location-scale mixtures are practically much more efficient.

Normal Mixture Models

$$y_1, \ldots, y_n \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \text{Normal}(\mu_k, \sigma_k^2)$$

- **Likelihood:** $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{ -\frac{1}{2\sigma_k^2}(y_i - \mu_k)^2 \right\} \right]$



- **Theorem:** Location mixtures of normals can approximate any continuous density.
- Location-scale mixtures are practically much more efficient.

Multivariate Normal Mixture Models

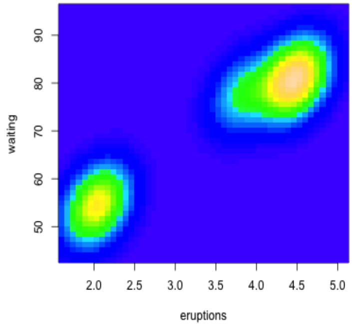
$$\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \text{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

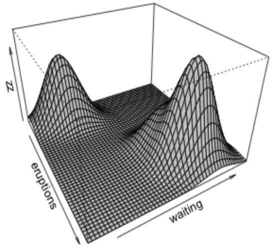
- **Likelihood:**
  $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \frac{1}{(\sqrt{2\pi})^d |\Sigma_k|} \exp\left\{ -\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k) \right\} \right]$

Multivariate Normal Mixture Models

$$\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \mathrm{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- **Likelihood:**
  $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \frac{1}{(\sqrt{2\pi})^d |\Sigma_k|} \exp\left\{ -\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k) \right\} \right]$



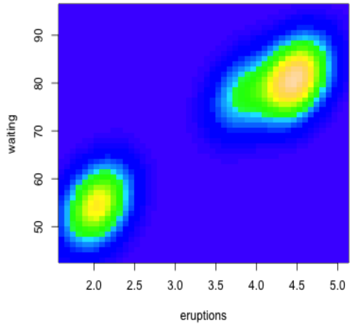


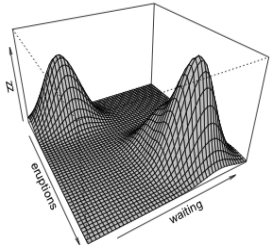
- Theorem: Location mixtures of multivariate normals can approximate any multivariate continuous density.
- Location-scale mixtures are practically much more efficient.

Multivariate Normal Mixture Models

$$\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \text{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- **Likelihood:**
  $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \frac{1}{(\sqrt{2\pi})^d |\Sigma_k|} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k)\right\}\right]$



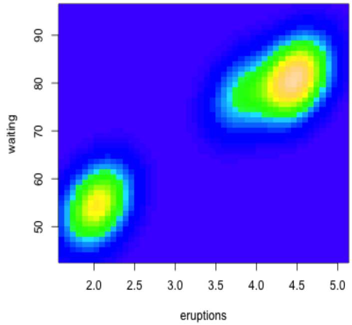


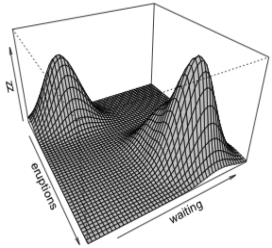
- **Theorem:** Location mixtures of multivariate normals can approximate any multivariate continuous density.
- Location-scale mixtures are practically much more efficient.

Multivariate Normal Mixture Models

$$\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \mathrm{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- **Likelihood:**
  $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \frac{1}{(\sqrt{2\pi})^d |\Sigma_k|} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k)\right\}\right]$





- **Theorem:** Location mixtures of multivariate normals can approximate any multivariate continuous density.
- Location-scale mixtures are practically much more efficient.

Mixtures of Random Variables vs Mixtures of Distributions

- Sum of independent normal random variables

$$y = \sum_{k=1}^{K} a_k z_k, \quad a_k \in \mathbb{R}\ \forall k, \quad z_k \overset{ind}{\sim} \text{Normal}(z \mid \mu_k, \sigma_k^2)$$

$$y \sim \text{Normal}\left(y \mid \sum_k a_k \mu_k, \sum_k a_k^2 \sigma_k^2\right)$$

- Mixtures of normals

$$y \sim \sum_{k=1}^{K} \pi_k \text{Normal}(y \mid \mu_k, \sigma_k^2), \quad \pi_k \geq 0\ \forall k,\ \sum_{k=1}^{K} \pi_k = 1$$

Mixtures of Random Variables vs Mixtures of Distributions

- Sum of independent normal random variables

$$y = \sum_{k=1}^{K} a_k z_k, \quad a_k \in \mathbb{R}\ \forall k, \qquad z_k \overset{ind}{\sim} \text{Normal}(z \mid \mu_k, \sigma_k^2)$$

$$y \sim \text{Normal}\left(y \mid \sum_k a_k \mu_k, \sum_k a_k^2 \sigma_k^2\right)$$

- Mixtures of normals

$$y \sim \sum_{k=1}^{K} \pi_k \text{Normal}(y \mid \mu_k, \sigma_k^2), \quad \pi_k \geq 0\ \forall k,\ \sum_{k=1}^{K} \pi_k = 1$$

Mixtures of Random Variables vs Mixtures of Distributions

- Sum of independent normal random variables

$$y = \sum_{k=1}^{K} a_k z_k, \quad a_k \in \mathbb{R}\ \forall k, \quad z_k \overset{ind}{\sim} \text{Normal}(z \mid \mu_k, \sigma_k^2)$$

$$y \sim \text{Normal}\left(y \mid \sum_k a_k \mu_k, \sum_k a_k^2 \sigma_k^2\right)$$

- Mixtures of normals

$$y \sim \sum_{k=1}^{K} \pi_k \text{Normal}(y \mid \mu_k, \sigma_k^2), \quad \pi_k \geq 0\ \forall k,\ \sum_{k=1}^{K} \pi_k = 1$$

General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Likelihood:** $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$

$$(z_i \mid \boldsymbol{\pi}) \overset{iid}{\sim} \mathrm{Mult}(1, \boldsymbol{\pi})$$

$$(\mathbf{y}_i \mid z_i = k, \boldsymbol{\xi}) \overset{ind}{\sim} p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Conditional Likelihood:** $p(\mathbf{y}_{1:n} \mid \mathbf{z}_{1:n}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{y}_i \mid \boldsymbol{\xi}_{z_i})$

General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Likelihood:** $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$

$$(z_i \mid \boldsymbol{\pi}) \overset{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$
$$(\mathbf{y}_i \mid z_i = k, \boldsymbol{\xi}) \overset{ind}{\sim} p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Conditional Likelihood:** $p(\mathbf{y}_{1:n} \mid \mathbf{z}_{1:n}, \theta) = \prod_{i=1}^{n} p(\mathbf{y}_i \mid \xi_{z_i})$

General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \textstyle\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Likelihood:** $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k) \right]$

$$(z_i \mid \boldsymbol{\pi}) \overset{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$
$$(\mathbf{y}_i \mid z_i = k, \boldsymbol{\xi}) \overset{ind}{\sim} p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Conditional Likelihood:** $p(\mathbf{y}_{1:n} \mid \mathbf{z}_{1:n}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{y}_i \mid \boldsymbol{\xi}_{z_i})$

General Mixture Models

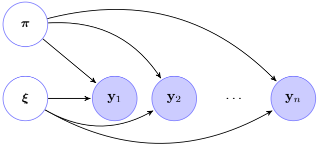
$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

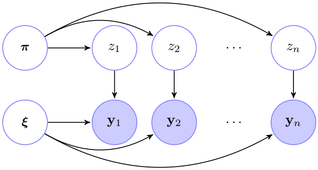
- **Likelihood:** $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$

$$(z_i \mid \boldsymbol{\pi}) \overset{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

$$(\mathbf{y}_i \mid z_i = k, \boldsymbol{\xi}) \overset{ind}{\sim} p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Conditional Likelihood:** $p(\mathbf{y}_{1:n} \mid \mathbf{z}_{1:n}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{y}_i \mid \boldsymbol{\xi}_{z_i})$





General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Likelihood:** $L(\boldsymbol{\theta}) = p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$
  Log-likelihood: $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$
- Some statistical issues
  - Label switching of mixture components
  - Non-identifiability in overfitted models
- **Likelihood equations:**

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} \frac{\pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)}{\left[\sum_{j=1}^{K} \pi_j p(\mathbf{y}_i \mid \boldsymbol{\xi}_j)\right]} \frac{\partial \log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} w_{ik} \frac{\partial \log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = 0$$

  This is a weighted likelihood function but with weights depending on unknown parameters.
- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) Using the current parameters, calculate new weights $\mathbf{w}$ (E-step).
  (b) Using the current weights, maximize the weighted likelihood to get new parameter estimates $\boldsymbol{\theta}$ (M-step).

General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Likelihood:** $L(\boldsymbol{\theta}) = p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$

  Log-likelihood: $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$

- Some statistical issues
  - Label switching of mixture components
  - Non-identifiability in overfitted models

- **Likelihood equations:**

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} \frac{\pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)}{\left[\sum_{j=1}^{K} \pi_j p(\mathbf{y}_i \mid \boldsymbol{\xi}_j)\right]} \frac{\partial \log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} w_{ik} \frac{\partial \log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = 0$$

  This is a weighted likelihood function but with weights depending on unknown parameters.

- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) Using the current parameters, calculate new weights $\mathbf{w}$ (E-step).
  (b) Using the current weights, maximize the weighted likelihood to get new parameter estimates $\boldsymbol{\theta}$ (M-step).

General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Likelihood:** $L(\boldsymbol{\theta}) = p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$
  Log-likelihood: $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$
- Some statistical issues
  - Label switching of mixture components
  - Non-identifiability in overfitted models
- **Likelihood equations:**

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} \frac{\pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)}{\left[\sum_{j=1}^{K} \pi_j p(\mathbf{y}_i \mid \boldsymbol{\xi}_j)\right]} \frac{\partial \log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} w_{ik} \frac{\partial \log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = 0$$

  This is a weighted likelihood function but with weights depending on unknown parameters.
- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) Using the current parameters, calculate new weights $\mathbf{w}$ (E-step).
  (b) Using the current weights, maximize the weighted likelihood to get new parameter estimates $\boldsymbol{\theta}$ (M-step).

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Likelihood:** $L(\boldsymbol{\theta}) = p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$
  Log-likelihood: $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$
- Some statistical issues
  - Label switching of mixture components
  - Non-identifiability in overfitted models
- **Likelihood equations:**

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} \frac{\pi_k p(\mathbf{y}_i | \boldsymbol{\xi}_k)}{\left[\sum_{j=1}^{K} \pi_j p(\mathbf{y}_i | \boldsymbol{\xi}_j)\right]} \frac{\partial \log p(\mathbf{y}_i | \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} w_{ik} \frac{\partial \log p(\mathbf{y}_i | \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = \mathbf{0}$$

This is a weighted likelihood function but with weights depending on unknown parameters.

- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) Using the current parameters, calculate new weights $\mathbf{w}$ (E-step).
  (b) Using the current weights, maximize the weighted likelihood to get new parameter estimates $\boldsymbol{\theta}$ (M-step).

General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Likelihood:** $L(\boldsymbol{\theta}) = p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$

  Log-likelihood: $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$

- Some statistical issues
  - Label switching of mixture components
  - Non-identifiability in overfitted models

- **Likelihood equations:**

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} \frac{\pi_k p(\mathbf{y}_i | \boldsymbol{\xi}_k)}{\left[\sum_{j=1}^{K} \pi_j p(\mathbf{y}_i | \boldsymbol{\xi}_j)\right]} \frac{\partial \log p(\mathbf{y}_i | \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} w_{ik} \frac{\partial \log p(\mathbf{y}_i | \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = \mathbf{0}$$

  This is a weighted likelihood function but with weights depending on unknown parameters.

- **Iterative algorithm:**

  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.

  (a) Using the current parameters, calculate new weights $\mathbf{w}$ (E-step).

  (b) Using the current weights, maximize the weighted likelihood to get new parameter estimates $\boldsymbol{\theta}$ (M-step).

General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Likelihood:** $L(\boldsymbol{\theta}) = p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$
  Log-likelihood: $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\right]$
- Some statistical issues
  - Label switching of mixture components
  - Non-identifiability in overfitted models
- **Likelihood equations:**

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} \frac{\pi_k p(\mathbf{y}_i | \boldsymbol{\xi}_k)}{\left[\sum_{j=1}^{K} \pi_j p(\mathbf{y}_i | \boldsymbol{\xi}_j)\right]} \frac{\partial \log p(\mathbf{y}_i | \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = \sum_{i=1}^{n} w_{ik} \frac{\partial \log p(\mathbf{y}_i | \boldsymbol{\xi}_k)}{\partial \boldsymbol{\xi}_k} = \mathbf{0}$$

  This is a weighted likelihood function but with weights depending on unknown parameters.
- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) Using the current parameters, calculate new weights $\mathbf{w}$ (E-step).
  (b) Using the current weights, maximize the weighted likelihood to get new parameter estimates $\boldsymbol{\theta}$ (M-step).

Expectation-Maximization Algorithm - the General Case

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}), \quad \mathbf{y} \rightarrow \text{observed}, \ \mathbf{z} \rightarrow \text{latent}$$

$$p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{y} \mid \boldsymbol{\theta}) \ \Rightarrow \ p(\mathbf{y} \mid \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$$

- **Log-likelihood:** $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}$
- $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \{\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}$

  $= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$

  $= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$
- $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$

Expectation-Maximization Algorithm - the General Case

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}), \quad \mathbf{y} \to \text{observed}, \ \mathbf{z} \to \text{latent}$$

$$p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{y} \mid \boldsymbol{\theta}) \ \Rightarrow \ p(\mathbf{y} \mid \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$$

- **Log-likelihood:** $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$

- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}$

- $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \{\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}$

  $= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$

  $= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

- $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$

Expectation-Maximization Algorithm - the General Case

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}), \quad \mathbf{y} \rightarrow \text{observed}, \ \mathbf{z} \rightarrow \text{latent}$$

$$p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{y} \mid \boldsymbol{\theta}) \ \Rightarrow \ p(\mathbf{y} \mid \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$$

- **Log-likelihood:** $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}$
- $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \{\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}$
  $= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
  $= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$
- $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \textstyle\sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}), \quad \mathbf{y} \to \text{observed}, \; \mathbf{z} \to \text{latent}$$

$$p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{y} \mid \boldsymbol{\theta}) \;\Rightarrow\; p(\mathbf{y} \mid \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$$

- **Log-likelihood:** $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$

- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \left\{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \right\}$

- $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \left\{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \right\}$
  $= \textstyle\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \textstyle\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
  $= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

- $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - H(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$
  $= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) + \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log \dfrac{p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})}{p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})}$
  $= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) + D_{KL} \left\{ p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}), p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \right\}$
  $\geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$

Expectation-Maximization Algorithm - the General Case

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}), \quad \mathbf{y} \to \text{observed}, \ \mathbf{z} \to \text{latent}$$

$$p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{y} \mid \boldsymbol{\theta}) \ \Rightarrow \ p(\mathbf{y} \mid \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$$

- **Log-likelihood:** $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}$
- $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \{\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}$
  $= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
  $= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$
- $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$
- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) **E-step:** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$.
  (b) **M-step:** Compute $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$.
- $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) \geq Q(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) \geq 0$

Expectation-Maximization Algorithm - the General Case

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}), \quad \mathbf{y} \rightarrow \text{observed}, \ \mathbf{z} \rightarrow \text{latent}$$
$$p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{y} \mid \boldsymbol{\theta}) \ \Rightarrow \ p(\mathbf{y} \mid \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) / p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$$

- **Log-likelihood:** $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \left\{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \right\}$
- $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \left\{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \right\}$
  $= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
  $= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$
- $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$
- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) **E-step:** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$.
  (b) **M-step:** Compute $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$.
- $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) \geq Q(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) \geq 0$

EM Algorithm - General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^K \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) **E-step:** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$.
  (b) **M-step:** Compute $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$.

- $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{y}_i, z_i \mid \boldsymbol{\theta}) = \prod_{i=1}^n \{p(\mathbf{y}_i \mid z_i, \boldsymbol{\theta}) p(z_i \mid \boldsymbol{\theta})\}$

$$= \prod_{i=1}^n \prod_{k=1}^K \{p(\mathbf{y}_i \mid z_i = k, \boldsymbol{\theta}) p(z_i = k \mid \boldsymbol{\theta})\}^{1(z_i = k)} = \prod_{i=1}^n \prod_{k=1}^K \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k) \pi_k\}^{1(z_i = k)}$$

- $\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K 1(z_i = k) \log \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k) \pi_k\}$

- $\pi_{i,k}^{(m)} = p(z_i = k \mid \mathbf{y}_i, \boldsymbol{\theta}^{(m)}) = \frac{\pi_k^{(m)} p(\mathbf{y}_i \mid \boldsymbol{\xi}_k^{(m)})}{\sum_{j=1}^K \pi_j^{(m)} p(\mathbf{y}_i \mid \boldsymbol{\xi}_j^{(m)})}$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$

$$= \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} 1(z_i = k) \log \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k) \pi_k\}$$

$$= \sum_{i=1}^n \sum_{k=1}^K \pi_{i,k}^{(m)} \{\log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k) + \log \pi_k\}$$

EM Algorithm - General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^K \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) **E-step:** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$.
  (b) **M-step:** Compute $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$.

- $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = \prod_{i=1}^n \prod_{k=1}^K \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}^{1(z_i = k)}$

- $\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K 1(z_i = k) \log \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}$

- $\pi_{i,k}^{(m)} = p(z_i = k \mid \mathbf{y}_i, \boldsymbol{\theta}^{(m)}) = \frac{\pi_k^{(m)} p(\mathbf{y}_i \mid \boldsymbol{\xi}_k^{(m)})}{\sum_{j=1}^K \pi_j^{(m)} p(\mathbf{y}_i \mid \boldsymbol{\xi}_j^{(m)})}$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$

$$= \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} 1(z_i = k) \log \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}$$

$$= \sum_{i=1}^n \sum_{k=1}^K \pi_{i,k}^{(m)} \{\log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k) + \log \pi_k\}$$

- **M-step:** $\boldsymbol{\theta}^{(m+1)} = (\boldsymbol{\pi}^{(m+1)}, \boldsymbol{\xi}^{(m+1)}) = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

EM Algorithm - General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) **E-step:** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$.
  (b) **M-step:** Compute $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$.
- $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}^{1(z_i = k)}$
- $\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} 1(z_i = k) \log \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}$
- $\pi_{i,k}^{(m)} = p(z_i = k \mid \mathbf{y}_i, \boldsymbol{\theta}^{(m)}) = \frac{\pi_k^{(m)} p(\mathbf{y}_i \mid \boldsymbol{\xi}_k^{(m)})}{\sum_{j=1}^{K} \pi_j^{(m)} p(\mathbf{y}_i \mid \boldsymbol{\xi}_j^{(m)})}$
- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} 1(z_i = k) \log \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \{\log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k) + \log \pi_k\}$$

- **M-step:** $\boldsymbol{\theta}^{(m+1)} = (\boldsymbol{\pi}^{(m+1)}, \boldsymbol{\xi}^{(m+1)}) = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

EM Algorithm - General Mixture Models

$$\mathbf{y}_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) **E-step:** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$.
  (b) **M-step:** Compute $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$.
- $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}^{1(z_i = k)}$
- $\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} 1(z_i = k) \log \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}$
- $\pi_{i,k}^{(m)} = p(z_i = k \mid \mathbf{y}_i, \boldsymbol{\theta}^{(m)}) = \frac{\pi_k^{(m)} p(\mathbf{y}_i \mid \boldsymbol{\xi}_k^{(m)})}{\sum_{j=1}^{K} \pi_j^{(m)} p(\mathbf{y}_i \mid \boldsymbol{\xi}_j^{(m)})}$
- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$
  $$= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} 1(z_i = k) \log \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}$$
  $$= \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \{\log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k) + \log \pi_k\}$$
- **M-step:** $\boldsymbol{\theta}^{(m+1)} = (\boldsymbol{\pi}^{(m+1)}, \boldsymbol{\xi}^{(m+1)}) = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

EM Algorithm - General Mixture Models

$$\mathbf{y}_i \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)$$

- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) **E-step:** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$.
  (b) **M-step:** Compute $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$.

- $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = \prod_{i=1}^n \prod_{k=1}^K \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}^{1(z_i = k)}$

- $\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K 1(z_i = k) \log \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}$

- $\pi_{i,k}^{(m)} = p(z_i = k \mid \mathbf{y}_i, \boldsymbol{\theta}^{(m)}) = \frac{\pi_k^{(m)} p(\mathbf{y}_i \mid \boldsymbol{\xi}_k^{(m)})}{\sum_{j=1}^K \pi_j^{(m)} p(\mathbf{y}_i \mid \boldsymbol{\xi}_j^{(m)})}$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$

$$= \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} 1(z_i = k) \log \{p(\mathbf{y}_i \mid \boldsymbol{\xi}_k)\pi_k\}$$

$$= \sum_{i=1}^n \sum_{k=1}^K \pi_{i,k}^{(m)} \{\log p(\mathbf{y}_i \mid \boldsymbol{\xi}_k) + \log \pi_k\}$$

- **M-step:** $\boldsymbol{\theta}^{(m+1)} = (\boldsymbol{\pi}^{(m+1)}, \boldsymbol{\xi}^{(m+1)}) = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

EM Algorithm - Normal Location-Scale Mixtures

$$y_i \stackrel{iid}{\sim} \sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \left\{ -\frac{1}{2} \log \sigma_k^2 - \frac{(y_i - \mu_k)^2}{2\sigma_k^2} + \log \pi_k \right\}$
- **M-step:** $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

$$\frac{\partial \left\{ Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + \lambda(\sum_{k=1}^{K} \pi_k - 1) \right\}}{\partial \pi_k / \partial \lambda} = 0, \quad \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})}{\partial \mu_k} = 0, \quad \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})}{\partial \sigma_k^2} = 0$$

$$\Rightarrow \pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,k}^{(m)}}{\sum_{j=1}^{K} \sum_{i=1}^{n} \pi_{i,j}^{(m)}} \quad \text{with} \quad \pi_{i,k}^{(m)} = \frac{\pi_k^{(m)} \text{Normal}(y_i \mid \mu_k^{(m)}, \sigma_k^{2(m)})}{\sum_{j=1}^{K} \pi_j^{(m)} \text{Normal}(y_i \mid \mu_j^{(m)}, \sigma_j^{2(m)})},$$

$$\Rightarrow \mu_k^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)} y_i}{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}} = \sum_{i=1}^{n} \frac{\pi_{i,k}^{(m+1)}}{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}} y_i,$$

$$\Rightarrow \sigma_k^{2(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)} (y_i - \mu_k^{(m+1)})^2}{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}} = \sum_{i=1}^{n} \frac{\pi_{i,k}^{(m+1)}}{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}} (y_i - \mu_k^{(m+1)})^2.$$

$$y_i \overset{iid}{\sim} \textstyle\sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \; \left\{\log p(y_i \mid \mu_k, \sigma_k^2) + \log \pi_k\right\}$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \; \left\{-\tfrac{1}{2}\log \sigma_k^2 - \tfrac{(y_i-\mu_k)^2}{2\sigma_k^2} + \log \pi_k\right\}$$

- **M-step:** $\boldsymbol{\theta}^{(m+1)} = \underset{\boldsymbol{\theta}}{\arg\max}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

$$\frac{\partial\left\{Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + \lambda(\sum_{k=1}^{K}\pi_k - 1)\right\}}{\partial \pi_k / \partial \lambda} = 0, \quad \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})}{\partial \mu_k} = 0, \quad \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})}{\partial \sigma_k^2} = 0$$

$$\Rightarrow \pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,k}^{(m)}}{\sum_{j=1}^{K}\sum_{i=1}^{n} \pi_{i,j}^{(m)}} \quad \text{with} \quad \pi_{i,k}^{(m)} = \frac{\pi_k^{(m)}\text{Normal}(y_i \mid \mu_k^{(m)}, \sigma_k^{2(m)})}{\sum_{j=1}^{K}\pi_j^{(m)}\text{Normal}(y_i \mid \mu_j^{(m)}, \sigma_j^{2(m)})},$$

$$\Rightarrow \mu_k^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)} y_i}{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}} = \sum_{i=1}^{n} \frac{\pi_{i,k}^{(m+1)}}{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}} y_i,$$

$$\Rightarrow \sigma_k^{2(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)} (y_i - \mu_k^{(m+1)})^2}{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}} = \sum_{i=1}^{n} \frac{\pi_{i,k}^{(m+1)}}{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}} (y_i - \mu_k^{(m+1)})^2.$$

EM Algorithm - Normal Location-Scale Mixtures

$$y_i \overset{iid}{\sim} \sum_{k=1}^K \pi_k \,\text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^n \sum_{k=1}^K \pi_{i,k}^{(m)} \left\{ -\frac{1}{2}\log\,\sigma_k^2 - \frac{(y_i - \mu_k)^2}{2\sigma_k^2} + \log\,\pi_k \right\}$
- **M-step:** $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

$$\frac{\partial \left\{ Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + \lambda(\sum_{k=1}^K \pi_k - 1) \right\}}{\partial \pi_k / \partial \lambda} = 0, \quad \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})}{\partial \mu_k} = 0, \quad \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})}{\partial \sigma_k^2} = 0$$

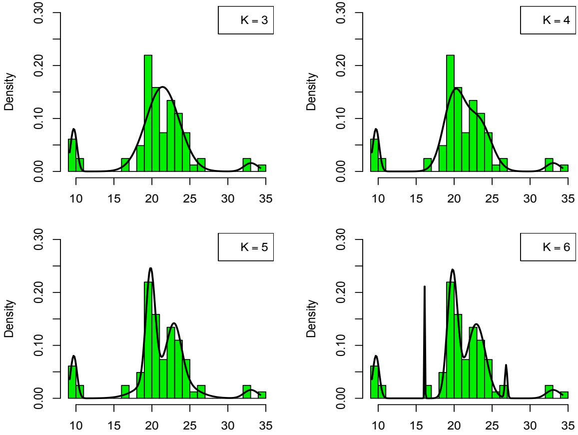
$$\Rightarrow \pi_k^{(m+1)} = \frac{\sum_{i=1}^n \pi_{i,k}^{(m)}}{\sum_{j=1}^K \sum_{i=1}^n \pi_{i,j}^{(m)}} \quad \text{with} \quad \pi_{i,k}^{(m)} = \frac{\pi_k^{(m)} \text{Normal}(y_i \mid \mu_k^{(m)}, \sigma_k^{2(m)})}{\sum_{j=1}^K \pi_j^{(m)} \text{Normal}(y_i \mid \mu_j^{(m)}, \sigma_j^{2(m)})},$$

$$\Rightarrow \mu_k^{(m+1)} = \frac{\sum_{i=1}^n \pi_{i,k}^{(m+1)} y_i}{\sum_{i=1}^n \pi_{i,k}^{(m+1)}} = \sum_{i=1}^n \frac{\pi_{i,k}^{(m+1)}}{\sum_{i=1}^n \pi_{i,k}^{(m+1)}} y_i,$$

$$\Rightarrow \sigma_k^{2(m+1)} = \frac{\sum_{i=1}^n \pi_{i,k}^{(m+1)} (y_i - \mu_k^{(m+1)})^2}{\sum_{i=1}^n \pi_{i,k}^{(m+1)}} = \sum_{i=1}^n \frac{\pi_{i,k}^{(m+1)}}{\sum_{i=1}^n \pi_{i,k}^{(m+1)}} (y_i - \mu_k^{(m+1)})^2.$$

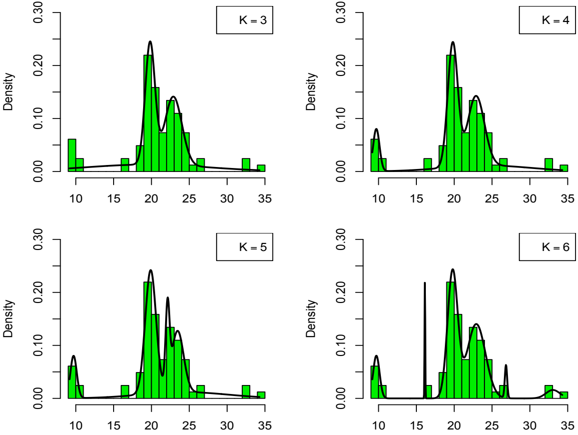
EM Algorithm - Normal Location-Scale Mixtures

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$



- Performance if $\pi_{i,k}$'s are NOT updated with $\pi_k^{(m+1)}$ and $\mu_k^{(m+1)}$ before updating $\sigma_k^{2(m+1)}$.

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$



- Performance when $\pi_{i,k}$'s are updated with $\pi_k^{(m+1)}$ and $\mu_k^{(m+1)}$ before updating $\sigma_k^{2(m+1)}$.

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma^2)$$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \left\{ -\frac{1}{2}\log \sigma^2 - \frac{(y_i - \mu_k)^2}{2\sigma^2} + \log \pi_k \right\}$
- **M-step:** $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$

$$\frac{\partial \left\{ Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + \lambda(\sum_{k=1}^{K} \pi_k - 1) \right\}}{\partial \pi_k / \partial \lambda} = 0, \quad \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})}{\partial \mu_k} = 0, \quad \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})}{\partial \sigma^2} = 0$$

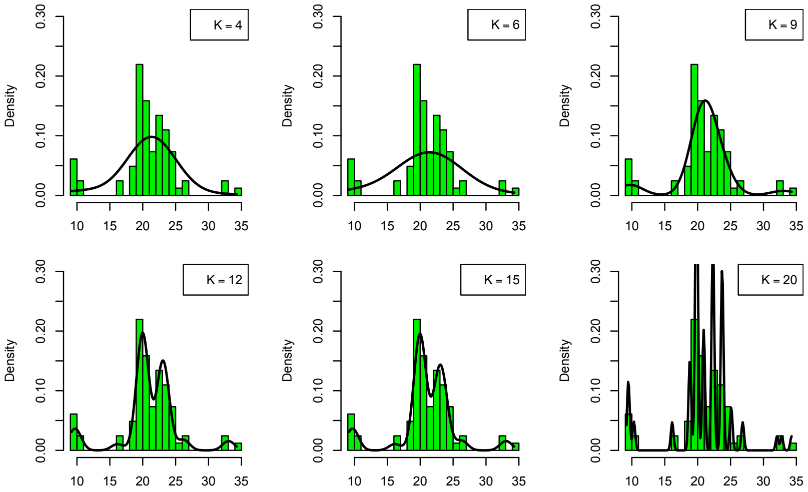
$$\Rightarrow \pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,k}^{(m)}}{\sum_{j=1}^{K} \sum_{i=1}^{n} \pi_{i,j}^{(m)}} \quad \text{with} \quad \pi_{i,k}^{(m)} = \frac{\pi_k^{(m)} \text{Normal}(y_i \mid \mu_k^{(m)}, \sigma^{2(m)})}{\sum_{j=1}^{K} \pi_j^{(m)} \text{Normal}(y_i \mid \mu_j^{(m)}, \sigma^{2(m)})},$$

$$\Rightarrow \mu_k^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)} y_i}{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}} = \sum_{i=1}^{n} \frac{\pi_{i,k}^{(m+1)}}{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}} y_i,$$

$$\Rightarrow \sigma^{2(m+1)} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m+1)} (y_i - \mu_k^{(m+1)})^2}{\sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m+1)}} = \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\pi_{i,k}^{(m+1)}}{n} (y_i - \mu_k^{(m+1)})^2.$$

EM Algorithm - Normal Location Mixtures

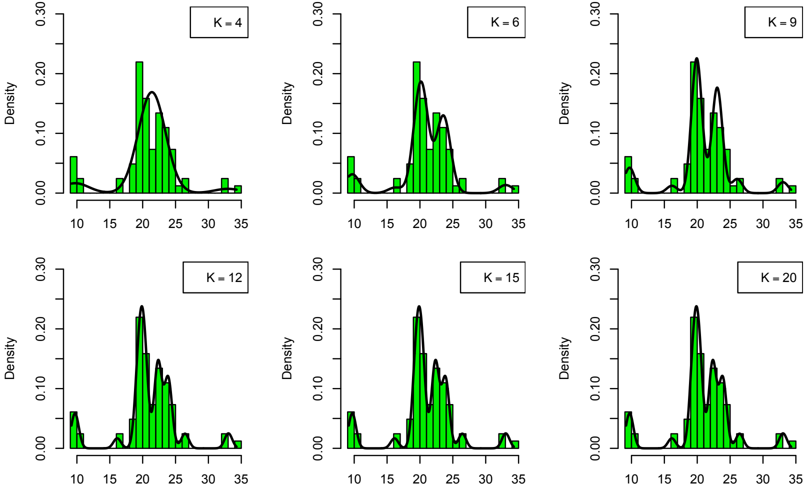
$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma^2)$$



- Performance if $\pi_{i,k}$'s are NOT updated with $\pi_k^{(m+1)}$ and $\mu_k^{(m+1)}$ before updating $\sigma^{2(m+1)}$.

EM Algorithm - Normal Location Mixtures

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma^2)$$



- Performance when $\pi_{i,k}$'s are updated with $\pi_k^{(m+1)}$ and $\mu_k^{(m+1)}$ before updating $\sigma^{2(m+1)}$.

EM Algorithm - MAP Estimation - the General Case

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \textstyle\sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}), \quad \mathbf{y} \rightarrow \text{observed}, \ \mathbf{z} \rightarrow \text{latent}$$
$$p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta})/p(\mathbf{y} \mid \boldsymbol{\theta}) \ \Rightarrow \ p(\mathbf{y} \mid \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta})/p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$$

- **Posterior:** $p(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta})$
- $\widetilde{\mathcal{L}}(\boldsymbol{\theta} \mid \mathbf{y}) = \log p(\boldsymbol{\theta}) + \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\widetilde{\mathcal{L}}(\boldsymbol{\theta} \mid \mathbf{y}) = \log p(\boldsymbol{\theta}) + \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}$
- $\widetilde{\mathcal{L}}(\boldsymbol{\theta} \mid \mathbf{y}) = \log p(\boldsymbol{\theta}) + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \{\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})\}$
  $= \log p(\boldsymbol{\theta}) + \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
  $= \log p(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \widetilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$
- $\widetilde{\mathcal{L}}(\boldsymbol{\theta} \mid \mathbf{y}) - \widetilde{\mathcal{L}}(\boldsymbol{\theta}^{(m)} \mid \mathbf{y}) \geq \widetilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - \widetilde{Q}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$
- **Iterative algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively update $\boldsymbol{\theta}^{(m)}$ until convergence.
  (a) **E-step:** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$.
  (b) **M-step:** Compute $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \right\}$.
- $\widetilde{\mathcal{L}}(\boldsymbol{\theta}^{(m+1)} \mid \mathbf{y}) - \widetilde{\mathcal{L}}(\boldsymbol{\theta}^{(m)} \mid \mathbf{y}) \geq \widetilde{Q}(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\theta}^{(m)}) - \widetilde{Q}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) \geq 0$

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \,\text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \left\{ -\frac{1}{2} \log \sigma_k^2 - \frac{(y_i - \mu_k)^2}{2\sigma_k^2} + \log \pi_k \right\}$
- **M-step:** $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \right\}$
- **Non-informative Improper Prior:** $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto 1$
  - $\widehat{\theta}_{MAP} = \widehat{\theta}_{MLE}$
- **Proper Prior:** $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto p(\boldsymbol{\pi}) \prod_{k=1}^{K} \{p(\mu_k) p(\sigma_k^2)\}$

  $\propto \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1} \prod_{k=1}^{K} \left[\exp\left\{ -\frac{1}{2\sigma_0^2} (\mu_k - \mu_0)^2 \right\} (\sigma_k^2)^{-a_0 - 1} \exp\left\{ -\frac{b_0}{\sigma_k^2} \right\} \right]$
  - M-step

EM Algorithm - MAP Estimation - Normal Location-Scale Mixtures

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \left\{ -\frac{1}{2} \log \sigma_k^2 - \frac{(y_i - \mu_k)^2}{2\sigma_k^2} + \log \pi_k \right\}$
- **M-step:** $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \right\}$
- **Non-informative Improper Prior:** $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto 1$
  - $\widehat{\boldsymbol{\theta}}_{MAP} = \widehat{\boldsymbol{\theta}}_{MLE}$
- **Proper Prior:** $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto p(\boldsymbol{\pi}) \prod_{k=1}^{K} \{p(\mu_k) p(\sigma_k^2)\}$

  $\propto \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1} \prod_{k=1}^{K} \left[\exp\left\{ -\frac{1}{2\sigma_0^2} (\mu_k - \mu_0)^2 \right\} (\sigma_k^2)^{-a_0 - 1} \exp\left\{ -\frac{b_0}{\sigma_k^2} \right\} \right]$
  - M-step:

EM Algorithm - MAP Estimation - Normal Location-Scale Mixtures

$$y_i \stackrel{iid}{\sim} \sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \left\{ -\frac{1}{2} \log \sigma_k^2 - \frac{(y_i - \mu_k)^2}{2\sigma_k^2} + \log \pi_k \right\}$
- **M-step:** $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \right\}$
- **Non-informative Improper Prior:** $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto 1$
  - $\widehat{\boldsymbol{\theta}}_{MAP} = \widehat{\boldsymbol{\theta}}_{MLE}$
- **Proper Prior:** $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto p(\boldsymbol{\pi}) \prod_{k=1}^{K} \{ p(\mu_k) p(\sigma_k^2) \}$
  $\propto \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1} \prod_{k=1}^{K} \left[\exp\left\{ -\frac{1}{2\sigma_0^2} (\mu_k - \mu_0)^2 \right\} (\sigma_k^2)^{-a_0 - 1} \exp\left\{ -\frac{b_0}{\sigma_k^2} \right\} \right]$
  - **M-step:**

EM Algorithm - MAP Estimation - Normal Location-Scale Mixtures

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \left\{ -\frac{1}{2} \log \sigma_k^2 - \frac{(y_i - \mu_k)^2}{2\sigma_k^2} + \log \pi_k \right\}$
- **M-step:** $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \right\}$
- **Non-informative Improper Prior:** $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto 1$
  - $\widehat{\boldsymbol{\theta}}_{MAP} = \widehat{\boldsymbol{\theta}}_{MLE}$
- **Proper Prior:** $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto p(\boldsymbol{\pi}) \prod_{k=1}^{K} \{ p(\mu_k) p(\sigma_k^2) \}$

$$\propto \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1} \prod_{k=1}^{K} \left[\exp\left\{ -\frac{1}{2\sigma_0^2} (\mu_k - \mu_0)^2 \right\} (\sigma_k^2)^{-a_0 - 1} \exp\left\{ -\frac{b_0}{\sigma_k^2} \right\} \right]$$

$$\frac{\partial \left\{ Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + \sum_{k=1}^{K} (\alpha_k - 1) \log \pi_k + \lambda (\sum_{k=1}^{K} \pi_k - 1) \right\}}{\partial \pi_k / \partial \lambda} = 0,$$

- **M-step:** $$\frac{\partial \left\{ Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - (\mu_k - \mu_0)^2 / (2\sigma_0^2) \right\}}{\partial \mu_k} = 0,$$

$$\frac{\partial \left\{ Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - (a_0 + 1) \log \sigma_k^2 - b_0 / \sigma_k^2 \right\}}{\partial \sigma_k^2} = 0$$

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \, \text{Normal}(y_i \mid \mu_k, \sigma_k^2)$$

- **E-step:** $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{i,k}^{(m)} \left\{ -\frac{1}{2} \log \sigma_k^2 - \frac{(y_i - \mu_k)^2}{2\sigma_k^2} + \log \pi_k \right\}$
- **M-step:** $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \right\}$
- **Non-informative Improper Prior:** $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto 1$
  - $\widehat{\boldsymbol{\theta}}_{MAP} = \widehat{\boldsymbol{\theta}}_{MLE}$
- **Proper Prior:** $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto p(\boldsymbol{\pi}) \prod_{k=1}^{K} \{ p(\mu_k) p(\sigma_k^2) \}$

$$\propto \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1} \prod_{k=1}^{K} \left[\exp\left\{ -\frac{1}{2\sigma_0^2} (\mu_k - \mu_0)^2 \right\} (\sigma_k^2)^{-a_0 - 1} \exp\left\{ -\frac{b_0}{\sigma_k^2} \right\} \right]$$

  - **M-step:**

$$\pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,k}^{(m)} + (\alpha_k - 1)}{\sum_{j=1}^{K} \{ \sum_{i=1}^{n} \pi_{i,j}^{(m)} + (\alpha_j - 1) \}},$$

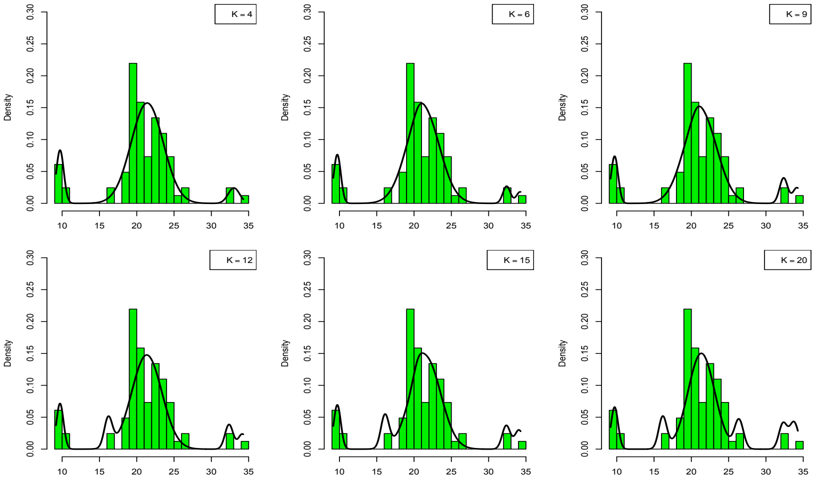
$$\mu_k^{(m+1)} = \left(\frac{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)}}{\sigma_k^{2(m)}} + \frac{1}{\sigma_0^2} \right)^{-1} \left(\frac{\sum_{i=1}^{n} \pi_{i,k}^{(m+1)} y_i}{\sigma_k^{2(m)}} + \frac{\mu_0}{\sigma_0^2} \right),$$

$$\sigma_k^{2(m+1)} = \left(a_0 + 1 + \frac{1}{2} \sum_{i=1}^{n} \pi_{i,k}^{(m+1)} \right)^{-1} \left\{ b_0 + \frac{1}{2} \sum_{i=1}^{n} \pi_{i,k}^{(m+1)} (y_i - \mu_k^{(m+1)})^2 \right\}.$$

EM Algorithm - MAP Estimation - Normal Location-Scale Mixtures

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \ \text{Normal}(y_i \mid \mu_k, \sigma_k^2),$$

$$\boldsymbol{\pi} \sim \text{Dir}(2, \ldots, 2), \qquad \mu_k \overset{iid}{\sim} \text{Normal}(\overline{y}, 5 s_y^2), \qquad \sigma_k^2 \overset{iid}{\sim} \text{Inv-Ga}(1, 1)$$



- MAP estimation with proper priors usually does NOT lead to singularities!