

# **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

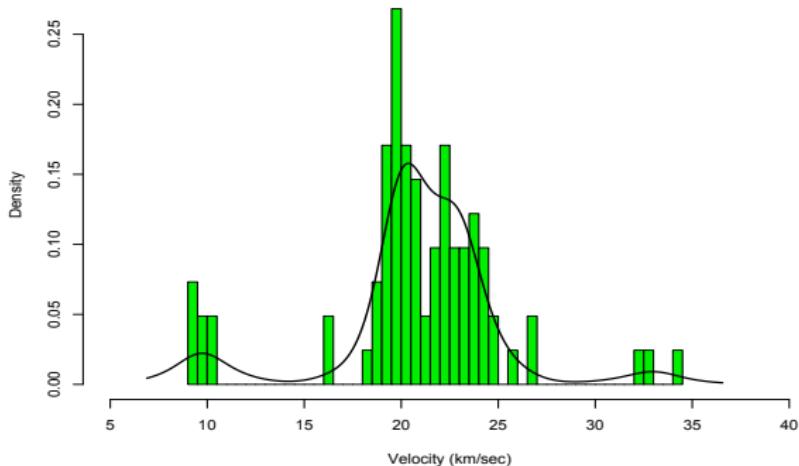
$$y_1, \dots, y_n \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k \text{Normal}(\mu_k, \sigma_k^2)$$

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

## Normal Mixture Models

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

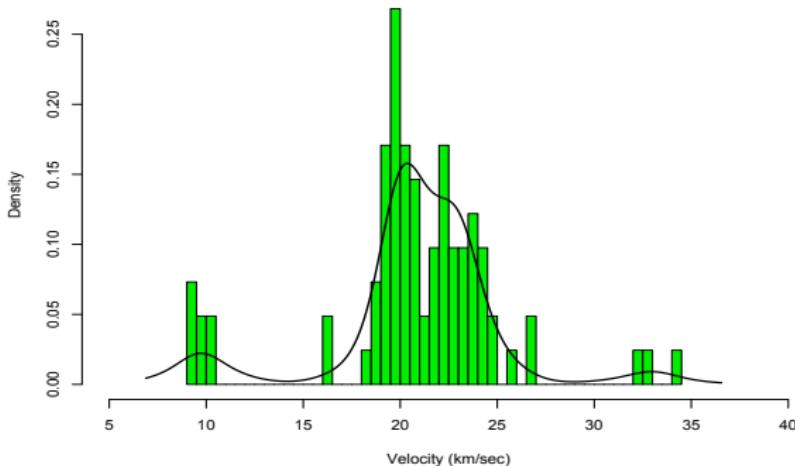
► **Likelihood:**  $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^n \left[ \sum_{k=1}^K \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp \left\{ -\frac{1}{2\sigma_k^2} (y_i - \mu_k)^2 \right\} \right]$



## Normal Mixture Models

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

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

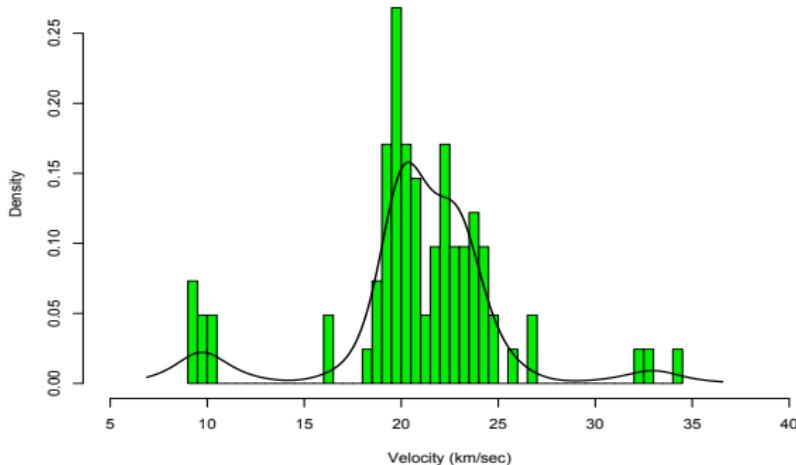


- **Theorem:** Location mixtures of normals can approximate any continuous density.
- Location-scale mixtures are practically much more efficient.

## Normal Mixture Models

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

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



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

# Multivariate Normal Mixture Models

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

## ► Likelihood:

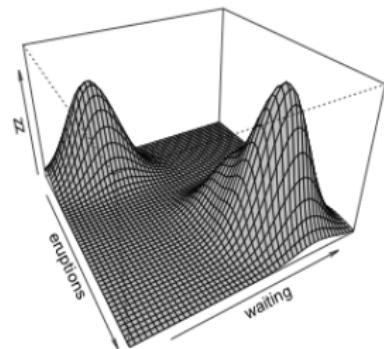
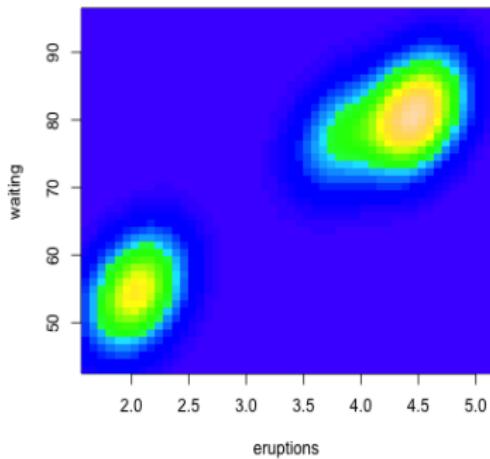
$$p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}) = \prod_{i=1}^n \left[ \sum_{k=1}^K \pi_k \frac{1}{(\sqrt{2\pi})^d |\boldsymbol{\Sigma}_k|} \exp \left\{ -\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k) \right\} \right]$$

# Multivariate Normal Mixture Models

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

## ► Likelihood:

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

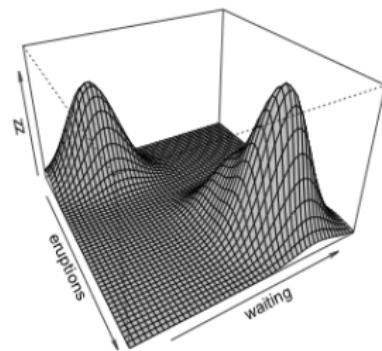
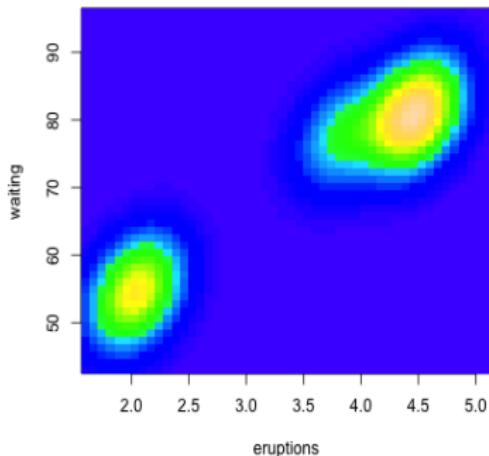


# Multivariate Normal Mixture Models

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

## ► Likelihood:

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



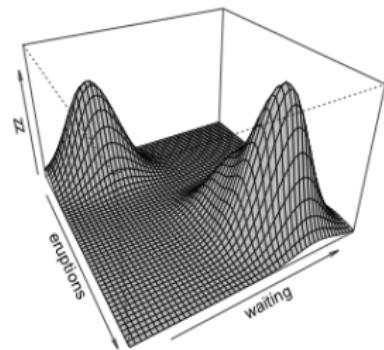
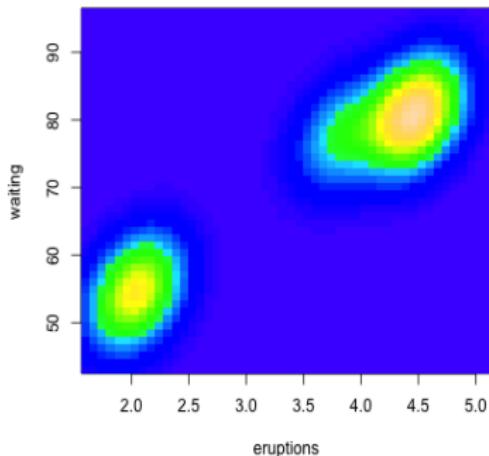
- **Theorem:** Location mixtures of multivariate normals can approximate any multivariate continuous density.
- Location-scale mixtures are practically much more efficient.

# Multivariate Normal Mixture Models

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

## ► Likelihood:

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



- **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

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

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

- Mixtures of normals

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

## Mixtures of Random Variables vs Mixtures of Distributions

- Sum of independent normal random variables

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

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

- Mixtures of normals

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

## Mixtures of Random Variables vs Mixtures of Distributions

- Sum of independent normal random variables

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

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

- Mixtures of normals

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

## General Mixture Models

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

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

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

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

- 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

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

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

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

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

- 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

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

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

$$(z_i | \boldsymbol{\pi}) \stackrel{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

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

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

# General Mixture Models

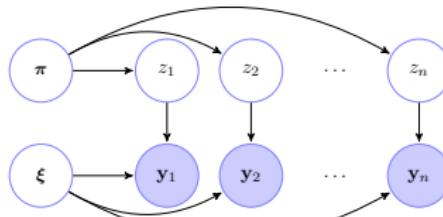
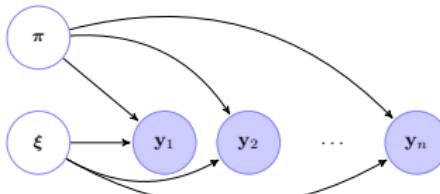
$$\mathbf{y}_i \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k p(\mathbf{y}_i | \boldsymbol{\xi}_k)$$

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

$$(z_i | \boldsymbol{\pi}) \stackrel{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

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

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



# General Mixture Models

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

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

- Some statistical issues

- Label switching of mixture components
- Non-identifiability in overfitted models

- Likelihood equations:

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

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.

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

# General Mixture Models

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

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

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

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

# General Mixture Models

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

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

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

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

# General Mixture Models

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

- **Likelihood:**  $L(\boldsymbol{\theta}) = p(\mathbf{y}_{1:n} | \boldsymbol{\theta}) = \prod_{i=1}^n \left[ \sum_{k=1}^K \pi_k p(\mathbf{y}_i | \boldsymbol{\xi}_k) \right]$   
**Log-likelihood:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}_{1:n} | \boldsymbol{\theta}) = \sum_{i=1}^n \log \left[ \sum_{k=1}^K \pi_k p(\mathbf{y}_i | \boldsymbol{\xi}_k) \right]$
- 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  $w$  (E-step).
  - (b) Using the current weights, maximize the weighted likelihood to get new parameter estimates  $\boldsymbol{\theta}$  (M-step).

# General Mixture Models

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

- **Likelihood:**  $L(\boldsymbol{\theta}) = p(\mathbf{y}_{1:n} | \boldsymbol{\theta}) = \prod_{i=1}^n \left[ \sum_{k=1}^K \pi_k p(\mathbf{y}_i | \boldsymbol{\xi}_k) \right]$   
**Log-likelihood:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}_{1:n} | \boldsymbol{\theta}) = \sum_{i=1}^n \log \left[ \sum_{k=1}^K \pi_k p(\mathbf{y}_i | \boldsymbol{\xi}_k) \right]$
- 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 w (E-step).
  - (b) Using the current weights, maximize the weighted likelihood to get new parameter estimates  $\boldsymbol{\theta}$  (M-step).

# General Mixture Models

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

- **Likelihood:**  $L(\boldsymbol{\theta}) = p(\mathbf{y}_{1:n} | \boldsymbol{\theta}) = \prod_{i=1}^n \left[ \sum_{k=1}^K \pi_k p(\mathbf{y}_i | \boldsymbol{\xi}_k) \right]$   
**Log-likelihood:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}_{1:n} | \boldsymbol{\theta}) = \sum_{i=1}^n \log \left[ \sum_{k=1}^K \pi_k p(\mathbf{y}_i | \boldsymbol{\xi}_k) \right]$
- 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:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \}$
- $$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \} \\ &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \\ &= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \end{aligned}$$
- $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$

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

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

## 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:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \}$
- $$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \} \\ &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \\ &= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \end{aligned}$$
- $$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) &= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - H(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) \\ &= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) + \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log \frac{p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})}{p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})} \\ &= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) + D_{KL} \left\{ p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}), p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \right\} \\ &\geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) \end{aligned}$$

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

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

- **Iterative algorithm:**

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

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

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

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

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

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

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

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

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

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

## EM Algorithm - General Mixture Models

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

- **Iterative algorithm:**

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

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

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

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

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

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

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

- 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

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

- **Iterative algorithm:**

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

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

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

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

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

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

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

- 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

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

- **Iterative algorithm:**

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

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

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

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

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

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

- **E-step:**  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} | \boldsymbol{\theta})$   
 $= \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(m)})} 1(z_i = k) \log \{p(\mathbf{y}_i | \boldsymbol{\xi}_k)\pi_k\}$   
 $= \sum_{i=1}^n \sum_{k=1}^K \pi_{i,k}^{(m)} \{\log p(\mathbf{y}_i | \boldsymbol{\xi}_k) + \log \pi_k\}$

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

## EM Algorithm - General Mixture Models

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

- **Iterative algorithm:**

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

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

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

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

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

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

- **E-step:** 
$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} | \boldsymbol{\theta})$$
$$= \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(m)})} 1(z_i = k) \log \{p(\mathbf{y}_i | \boldsymbol{\xi}_k)\pi_k\}$$
$$= \sum_{i=1}^n \sum_{k=1}^K \pi_{i,k}^{(m)} \{\log p(\mathbf{y}_i | \boldsymbol{\xi}_k) + \log \pi_k\}$$

- **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

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

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

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

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

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

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

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

## EM Algorithm - Normal Location-Scale Mixtures

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

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

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

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

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

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

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

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

## EM Algorithm - Normal Location-Scale Mixtures

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

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

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

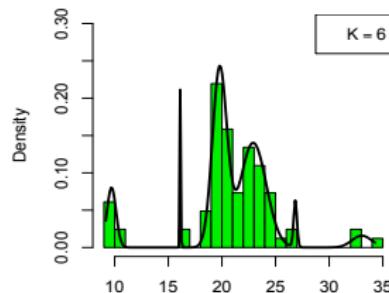
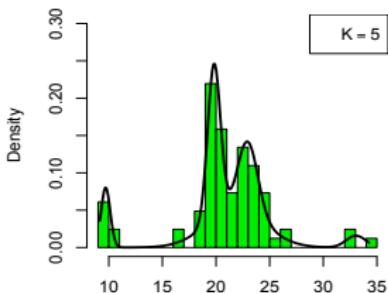
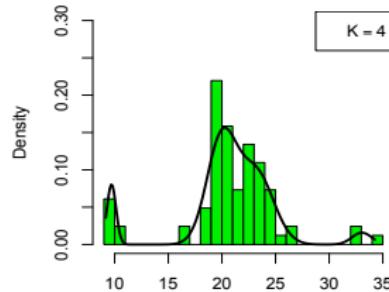
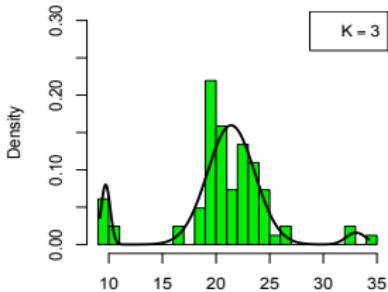
$$\Rightarrow \pi_k^{(m+1)} = \frac{\sum_{i=1}^n \pi_{i,k}^{(m)}}{\sum_{j=1}^K \sum_{i=1}^n \pi_{i,j}^{(m)}} \text{ with } \pi_{i,k}^{(m)} = \frac{\pi_k^{(m)} \text{Normal}(y_i | \mu_k^{(m)}, \sigma_k^{2(m)})}{\sum_{j=1}^K \pi_j^{(m)} \text{Normal}(y_i | \mu_j^{(m)}, \sigma_j^{2(m)})},$$

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

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

## EM Algorithm - Normal Location-Scale Mixtures

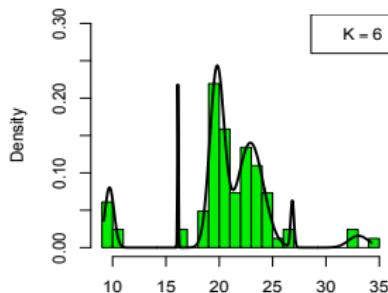
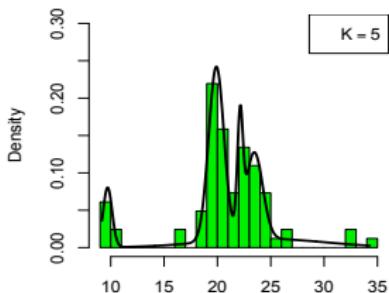
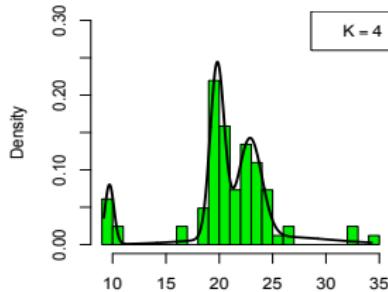
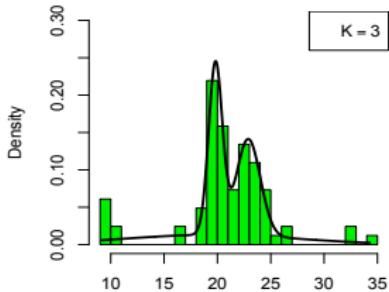
$$y_i \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k \text{Normal}(y_i | \mu_k, \sigma_k^2)$$



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

## EM Algorithm - Normal Location-Scale Mixtures

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



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

## EM Algorithm - Normal Location Mixtures

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

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

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

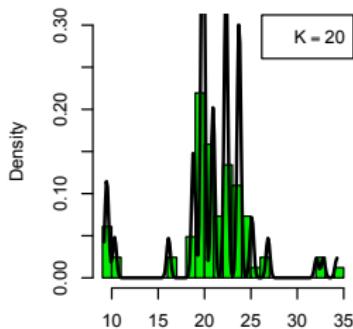
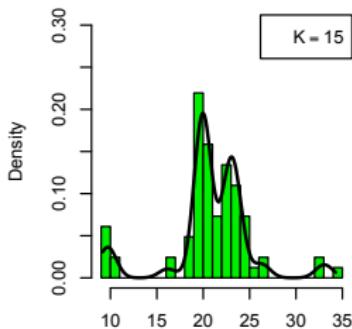
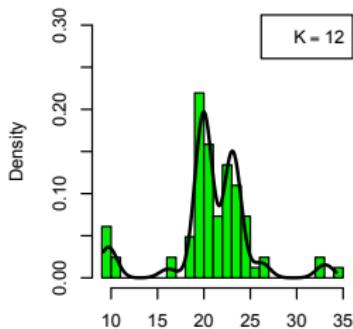
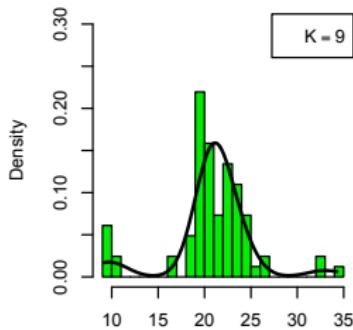
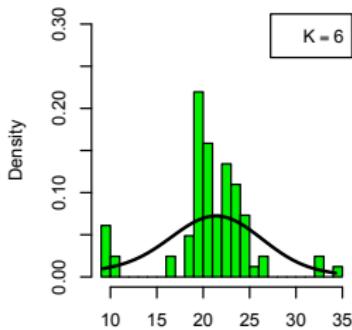
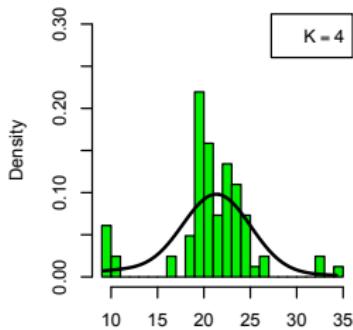
$$\Rightarrow \pi_k^{(m+1)} = \frac{\sum_{i=1}^n \pi_{i,k}^{(m)}}{\sum_{j=1}^K \sum_{i=1}^n \pi_{i,j}^{(m)}} \quad \text{with} \quad \pi_{i,k}^{(m)} = \frac{\pi_k^{(m)} \text{Normal}(y_i | \mu_k^{(m)}, \sigma^{2(m)})}{\sum_{j=1}^K \pi_j^{(m)} \text{Normal}(y_i | \mu_j^{(m)}, \sigma^{2(m)})},$$

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

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

# EM Algorithm - Normal Location Mixtures

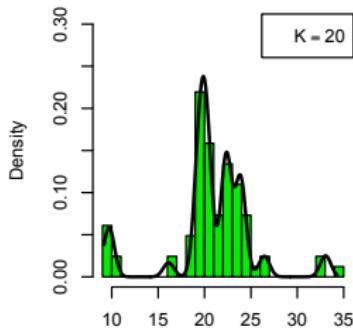
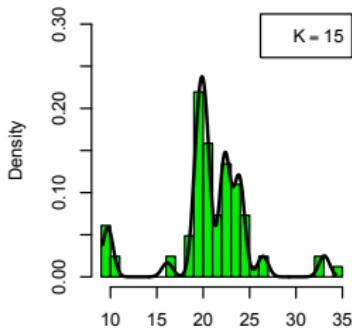
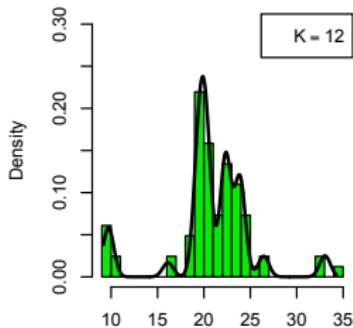
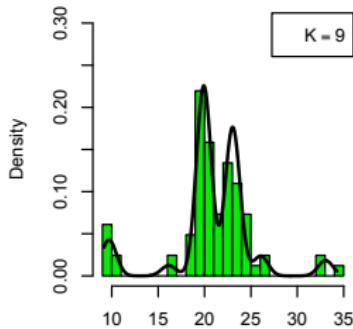
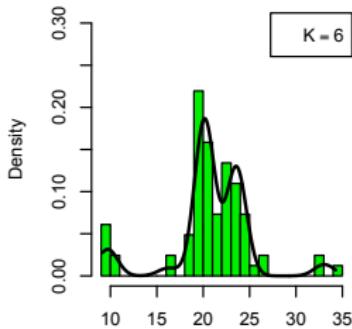
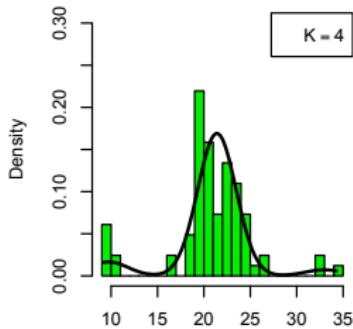
$$y_i \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k \text{Normal}(y_i | \mu_k, \sigma^2)$$



- 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

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



- 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

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

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

- **Posterior:**  $p(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\boldsymbol{\theta})p(\mathbf{y} \mid \boldsymbol{\theta})$
- $\tilde{\mathcal{L}}(\boldsymbol{\theta} \mid \mathbf{y}) = \log p(\boldsymbol{\theta}) + \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\tilde{\mathcal{L}}(\boldsymbol{\theta} \mid \mathbf{y}) = \log p(\boldsymbol{\theta}) + \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \}$
- $$\begin{aligned} \tilde{\mathcal{L}}(\boldsymbol{\theta} \mid \mathbf{y}) &= \log p(\boldsymbol{\theta}) + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \} \\ &= \log p(\boldsymbol{\theta}) + \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \\ &= \log p(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \tilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \end{aligned}$$
- $\tilde{\mathcal{L}}(\boldsymbol{\theta} \mid \mathbf{y}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}^{(m)} \mid \mathbf{y}) \geq \tilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) - \tilde{Q}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)})$
- **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  $\boldsymbol{\theta}^{(m+1)} = \arg \max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}) + Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) \right\}$ .
- $\tilde{\mathcal{L}}(\boldsymbol{\theta}^{(m+1)} \mid \mathbf{y}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}^{(m)} \mid \mathbf{y}) \geq \tilde{Q}(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\theta}^{(m)}) - \tilde{Q}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m)}) \geq 0$

## EM Algorithm - MAP Estimation - Normal Location-Scale Mixtures

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

- ▶ **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:**  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto 1$

►  $\hat{\boldsymbol{\theta}}_{MAP} = \hat{\boldsymbol{\theta}}_{MLE}$

- ▶ Proper Prior:  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto p(\boldsymbol{\pi}) \prod_{k=1}^K \{p(\mu_k)p(\sigma_k^2)\}$   
 $\propto \pi_1^{\alpha_1-1} \cdots \pi_K^{\alpha_K-1} \prod_{k=1}^K \left[ \exp \left\{ -\frac{1}{2\sigma_0^2} (\mu_k - \mu_0)^2 \right\} (\sigma_k^2)^{-a_0-1} \exp \left\{ -\frac{b_0}{\sigma_k^2} \right\} \right]$

MAP vs MLE

## EM Algorithm - MAP Estimation - Normal Location-Scale Mixtures

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

- ▶ **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:**  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto 1$ 
  - ▶  $\hat{\boldsymbol{\theta}}_{MAP} = \hat{\boldsymbol{\theta}}_{MLE}$

▶ Proper Prior: 
$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto p(\boldsymbol{\pi}) \prod_{k=1}^K \{p(\mu_k)p(\sigma_k^2)\}$$
$$\propto \pi_1^{\alpha_1-1} \cdots \pi_K^{\alpha_K-1} \prod_{k=1}^K \left[ \exp \left\{ -\frac{1}{2\sigma_0^2} (\mu_k - \mu_0)^2 \right\} (\sigma_k^2)^{-a_0-1} \exp \left\{ -\frac{b_0}{\sigma_k^2} \right\} \right]$$

MAP

## EM Algorithm - MAP Estimation - Normal Location-Scale Mixtures

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

- ▶ **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:**  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto 1$ 
  - ▶  $\hat{\boldsymbol{\theta}}_{MAP} = \hat{\boldsymbol{\theta}}_{MLE}$
- ▶ **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

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

► **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:**  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto 1$

►  $\hat{\boldsymbol{\theta}}_{MAP} = \hat{\boldsymbol{\theta}}_{MLE}$

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

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

## EM Algorithm - MAP Estimation - Normal Location-Scale Mixtures

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

► **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:**  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto 1$

►  $\hat{\boldsymbol{\theta}}_{MAP} = \hat{\boldsymbol{\theta}}_{MLE}$

► **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:**

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

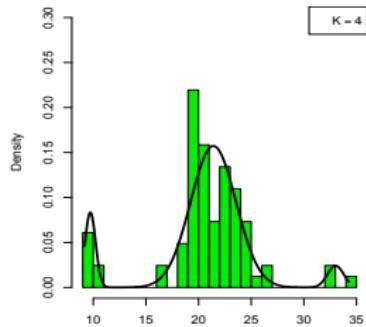
$$\mu_k^{(m+1)} = \left( \frac{\sum_{i=1}^n \pi_{i,k}^{(m+1)}}{\sigma_k^{2(m)}} + \frac{1}{\sigma_0^2} \right)^{-1} \left( \frac{\sum_{i=1}^n \pi_{i,k}^{(m+1)} y_i}{\sigma_k^{2(m)}} + \frac{\mu_0}{\sigma_0^2} \right),$$

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

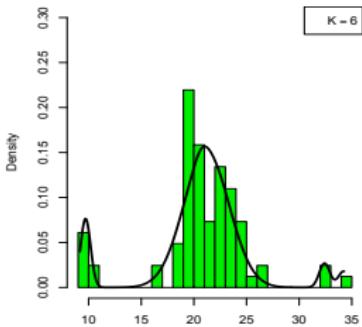
# EM Algorithm - MAP Estimation - Normal Location-Scale Mixtures

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

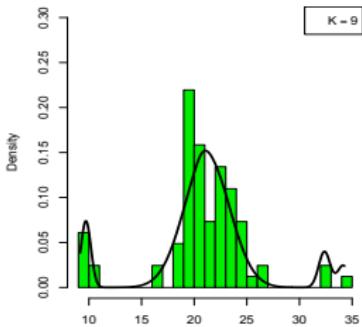
$$\boldsymbol{\pi} \sim \text{Dir}(2, \dots, 2), \quad \mu_k \stackrel{iid}{\sim} \text{Normal}(\bar{y}, 5s_y^2), \quad \sigma_k^2 \stackrel{iid}{\sim} \text{Inv-Ga}(1, 1)$$



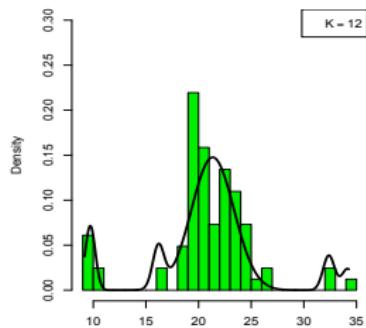
K = 4



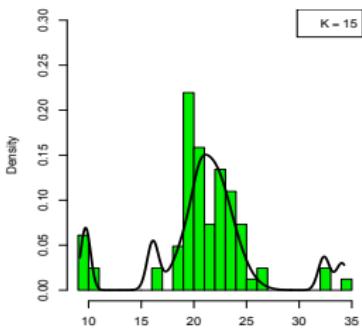
K = 6



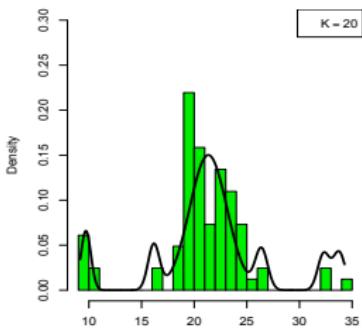
K = 9



K = 12



K = 15



K = 20

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

## Normal Location-Scale Mixtures - Gibbs Sampling

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

- **Equivalent representation:**

$$(z_i | \boldsymbol{\pi}) \stackrel{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

$$(y_i | z_i = k, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \stackrel{ind}{\sim} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

- Priors:  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}^2)$

$$= \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \cdot \prod_{k=1}^K \{\text{Normal}(\mu_k | \mu_0, \sigma_0^2) \cdot \text{Inv-Ga}(\sigma_k^2 | a_0, b_0)\}$$

- Full Conditionals:

## Normal Location-Scale Mixtures - Gibbs Sampling

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

- ▶ **Equivalent representation:**

$$(z_i | \boldsymbol{\pi}) \stackrel{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

$$(y_i | z_i = k, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \stackrel{ind}{\sim} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

- ▶ **Priors:**  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}^2)$

$$= \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \cdot \prod_{k=1}^K \{\text{Normal}(\mu_k | \mu_0, \sigma_0^2) \cdot \text{Inv-Ga}(\sigma_k^2 | a_0, b_0)\}$$

- ▶ **Full Conditionals:**

## Normal Location-Scale Mixtures - Gibbs Sampling

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

- ▶ **Equivalent representation:**

$$(z_i | \boldsymbol{\pi}) \stackrel{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

$$(y_i | z_i = k, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \stackrel{ind}{\sim} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

- ▶ **Priors:**  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}^2)$

$$= \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \cdot \prod_{k=1}^K \left\{ \text{Normal}(\mu_k | \mu_0, \sigma_0^2) \cdot \text{Inv-Ga}(\sigma_k^2 | a_0, b_0) \right\}$$

- ▶ **Full Conditionals:**

- ▶  $p(z_i = k | -) \propto \pi_k \times \text{Normal}(y_i | \mu_k, \sigma_k^2)$

- ▶  $p(\mu_k | -) \propto \text{Normal}(\mu_k | \mu_0, \sigma_0^2) \times \prod_{\{i:z_i=k\}} \text{Normal}(y_i | \mu_k, \sigma_k^2)$

- ▶  $p(\sigma_k^2 | -) \propto \text{Inv-Ga}(\sigma_k^2 | a_0, b_0) \times \prod_{\{i:z_i=k\}} \text{Normal}(y_i | \mu_k, \sigma_k^2)$

# Normal Location-Scale Mixtures - Gibbs Sampling

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

- ▶ **Equivalent representation:**

$$(z_i | \boldsymbol{\pi}) \stackrel{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

$$(y_i | z_i = k, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \stackrel{ind}{\sim} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

- ▶ **Priors:**  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}^2)$

$$= \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \cdot \prod_{k=1}^K \left\{ \text{Normal}(\mu_k | \mu_0, \sigma_0^2) \cdot \text{Inv-Ga}(\sigma_k^2 | a_0, b_0) \right\}$$

- ▶ **Full Conditionals:**

$$\blacktriangleright p(z_i = k | -) \propto \pi_k \times \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

$$\blacktriangleright p(\mu_k | -) \propto \text{Normal}(\mu_k | \mu_0, \sigma_0^2) \times \prod_{\{i:z_i=k\}} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

$$\blacktriangleright p(\sigma_k^2 | -) \propto \text{Inv-Ga}(\sigma_k^2 | a_0, b_0) \times \prod_{\{i:z_i=k\}} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

## Normal Location-Scale Mixtures - Gibbs Sampling

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

- ▶ **Equivalent representation:**

$$(z_i | \boldsymbol{\pi}) \stackrel{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

$$(y_i | z_i = k, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \stackrel{ind}{\sim} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

- ▶ **Priors:**  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}^2)$

$$= \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \cdot \prod_{k=1}^K \left\{ \text{Normal}(\mu_k | \mu_0, \sigma_0^2) \cdot \text{Inv-Ga}(\sigma_k^2 | a_0, b_0) \right\}$$

- ▶ **Full Conditionals:**

$$\blacktriangleright p(z_i = k | -) \propto \pi_k \times \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

$$\blacktriangleright p(\mu_k | -) \propto \text{Normal}(\mu_k | \mu_0, \sigma_0^2) \times \prod_{\{i:z_i=k\}} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

$$\blacktriangleright p(\sigma_k^2 | -) \propto \text{Inv-Ga}(\sigma_k^2 | a_0, b_0) \times \prod_{\{i:z_i=k\}} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

# Normal Location-Scale Mixtures - Gibbs Sampling

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

## ► Equivalent representation:

$$(z_i | \boldsymbol{\pi}) \stackrel{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

$$(y_i | z_i = k, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \stackrel{ind}{\sim} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

## ► Priors: $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}^2)$

$$= \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \cdot \prod_{k=1}^K \left\{ \text{Normal}(\mu_k | \mu_0, \sigma_0^2) \cdot \text{Inv-Ga}(\sigma_k^2 | a_0, b_0) \right\}$$

## ► Full Conditionals:

$$\blacktriangleright p(z_i = k | -) = \text{Mult}(1, \boldsymbol{\pi}_i), \quad \pi_{ik} = \frac{\pi_k \times \text{Normal}(y_i | \mu_k, \sigma_k^2)}{\sum_{j=1}^K \pi_j \times \text{Normal}(y_i | \mu_j, \sigma_j^2)}$$

$$\blacktriangleright p(\mu_k | -) = \text{Normal}(\mu_{k,n}, \sigma_{k,n}^2), \quad \mu_{k,n} = \sigma_{k,n}^2 \left( \frac{\mu_0}{\sigma_0^2} + \frac{n_k \bar{y}_k}{\sigma_k^2} \right),$$

$$\sigma_{k,n}^2 = \left( \frac{1}{\sigma_0^2} + \frac{n_k}{\sigma_k^2} \right)^{-1}, \quad n_k = \sum_i 1(z_i = k), \quad \bar{y}_k = \frac{\sum_i 1(z_i = k) y_i}{n_k}$$

$$\blacktriangleright p(\sigma_k^2 | -) = \text{Inv-Ga}(a_{k,n}, b_{k,n}),$$

$$a_{k,n} = \left( a_0 + \frac{n_k}{2} \right), \quad b_{k,n} = \left\{ b_0 + \frac{(n_k - 1)s_k^2}{2} + \frac{n_k(\bar{y}_k - \mu_k)^2}{2} \right\}.$$

# Normal Location-Scale Mixtures - Gibbs Sampling

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

## ► Equivalent representation:

$$(z_i | \boldsymbol{\pi}) \stackrel{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

$$(y_i | z_i = k, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \stackrel{ind}{\sim} \text{Normal}(y_i | \mu_k, \sigma_k^2)$$

## ► Priors: $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}^2)$

$$= \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \cdot \prod_{k=1}^K \left\{ \text{Normal}(\mu_k | \mu_0, \sigma_0^2) \cdot \text{Inv-Ga}(\sigma_k^2 | a_0, b_0) \right\}$$

## ► Full Conditionals:

$$\blacktriangleright p(z_i = k | -) = \text{Mult}(1, \boldsymbol{\pi}_i), \quad \pi_{ik} = \frac{\pi_k \times \text{Normal}(y_i | \mu_k, \sigma_k^2)}{\sum_{j=1}^K \pi_j \times \text{Normal}(y_i | \mu_j, \sigma_j^2)}$$

$$\blacktriangleright p(\mu_k | -) = \text{Normal}(\mu_{k,n}, \sigma_{k,n}^2), \quad \mu_{k,n} = \sigma_{k,n}^2 \left( \frac{\mu_0}{\sigma_0^2} + \frac{n_k \bar{y}_k}{\sigma_k^2} \right),$$

$$\sigma_{k,n}^2 = \left( \frac{1}{\sigma_0^2} + \frac{n_k}{\sigma_k^2} \right)^{-1}, \quad n_k = \sum_i 1(z_i = k), \quad \bar{y}_k = \frac{\sum_i 1(z_i = k) y_i}{n_k}$$

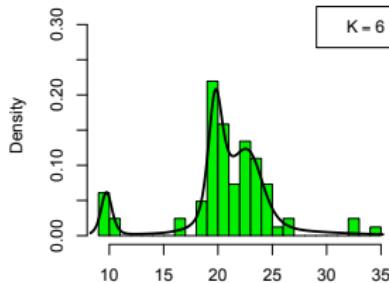
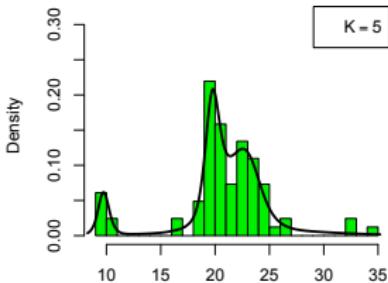
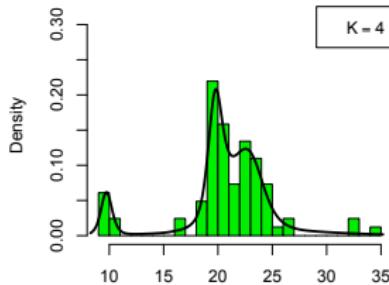
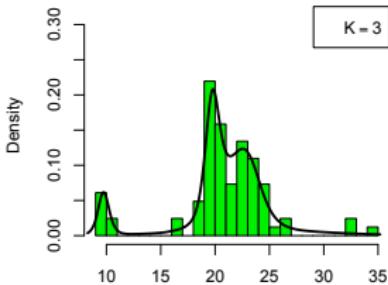
$$\blacktriangleright p(\sigma_k^2 | -) = \text{Inv-Ga}(a_{k,n}, b_{k,n}),$$

$$a_{k,n} = \left( a_0 + \frac{n_k}{2} \right), \quad b_{k,n} = \left\{ b_0 + \frac{(n_k - 1)s_k^2}{2} + \frac{n_k(\bar{y}_k - \mu_k)^2}{2} \right\}.$$

→ Compare with results from Module 5

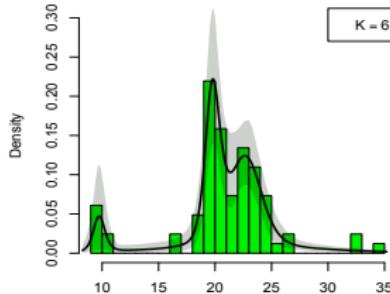
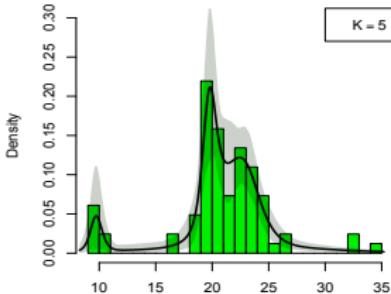
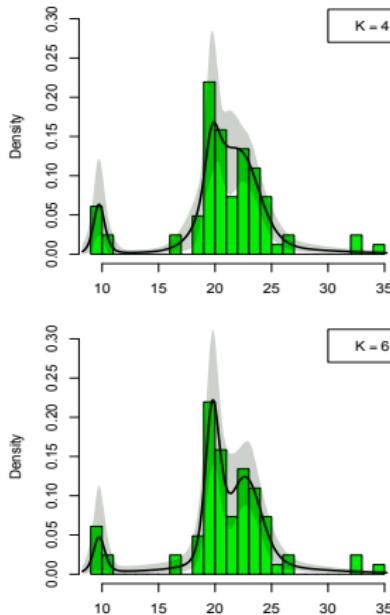
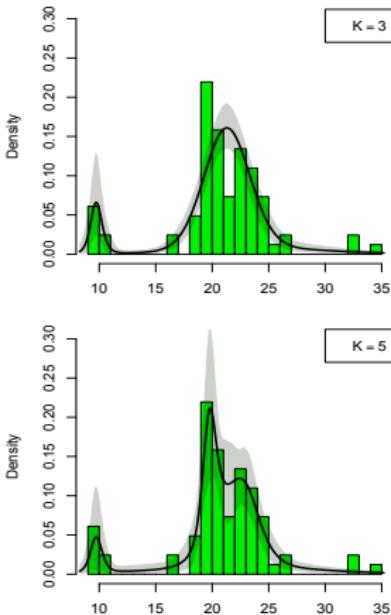
# Normal Location-Scale Mixtures - Gibbs Sampling

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



## Normal Location-Scale Mixtures - Gibbs Sampling

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



- Posterior mean and point-wise credible intervals are straightforwardly computed!

## Normal Location Mixtures - Gibbs Sampling

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

- ▶ **Equivalent representation:**

$$(z_i | \boldsymbol{\pi}) \stackrel{iid}{\sim} \text{Mult}(1, \boldsymbol{\pi})$$

$$(y_i | z_i = k, \boldsymbol{\mu}, \sigma^2) \stackrel{ind}{\sim} \text{Normal}(y_i | \mu_k, \sigma^2)$$

- ▶ **Priors:**  $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\sigma^2)$

$$= \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \cdot \text{Inv-Ga}\left(\sigma^2 | a_0, b_0\right) \cdot \prod_{k=1}^K \{\text{Normal}(\mu_k | \mu_0, \sigma_0^2)\}$$

- ▶ **Full Conditionals:**

$$\blacktriangleright p(z_i = k | -) = \text{Mult}(1, \boldsymbol{\pi}_i), \quad \pi_{ik} = \frac{\pi_k \times \text{Normal}(y_i | \mu_k, \sigma^2)}{\sum_{j=1}^K \pi_j \times \text{Normal}(y_i | \mu_j, \sigma^2)}$$

$$\blacktriangleright p(\mu_k | -) = \text{Normal}(\mu_{k,n}, \sigma_{k,n}^2), \quad \mu_{k,n} = \sigma_{k,n}^2 \left( \frac{\mu_0}{\sigma_0^2} + \frac{n_k \bar{y}_k}{\sigma^2} \right),$$

$$\sigma_{k,n}^2 = \left( \frac{1}{\sigma_0^2} + \frac{n_k}{\sigma^2} \right)^{-1}, \quad n_k = \sum_i 1(z_i = k), \quad \bar{y}_k = \frac{\sum_i 1(z_i = k) y_i}{n_k}$$

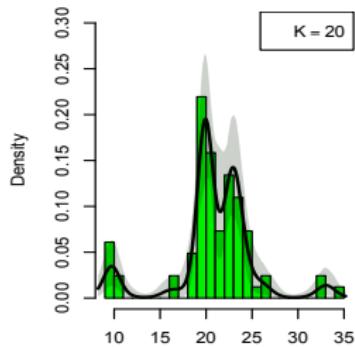
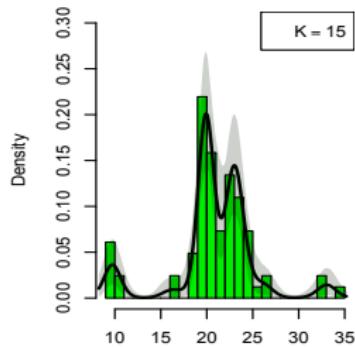
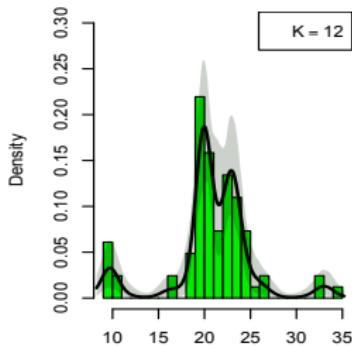
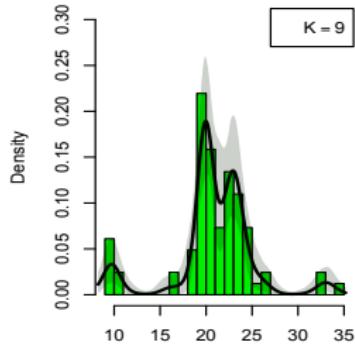
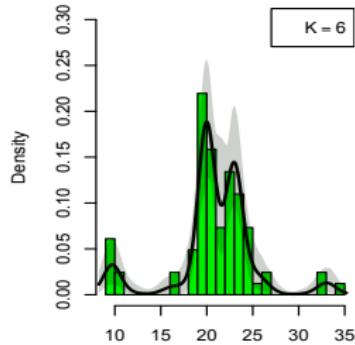
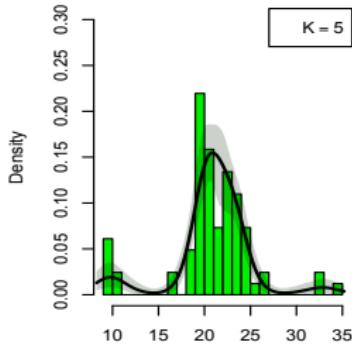
$$\blacktriangleright p(\sigma^2 | -) = \text{Inv-Ga}(a_n, b_n),$$

$$a_n = \left(a_0 + \frac{n}{2}\right), \quad b_n = \left\{ b_0 + \frac{\sum_{k=1}^K \sum_{i:z_i=k} (y_i - \mu_k)^2}{2} \right\}.$$

$$\blacktriangleright p(\boldsymbol{\pi} | -) = \text{Dir}\left(\frac{\alpha}{K} + n_1, \dots, \frac{\alpha}{K} + n_K\right).$$

# Normal Location Mixtures - Gibbs Sampling

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



## General Mixture Models - Asymptotic Behavior of the Posterior

$$y_i \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k p(y_i | \boldsymbol{\xi}_k)$$

- Let the true model be a finite mixture with  $K_0$  components.
  - The fitted model has  $K$  components with  $K > K_0$ .
  - Theorem: When  $\alpha_k < L/2$ , where  $L = |\boldsymbol{\xi}_k|$  denotes the number of parameters specifying the component kernels, the posterior is stable and concentrates in regions with empty redundant components.
  - Markov chains expected to reach steady states with empty redundant components.

## General Mixture Models - Asymptotic Behavior of the Posterior

$$y_i \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k p(y_i | \boldsymbol{\xi}_k)$$

- Let the true model be a finite mixture with  $K_0$  components.
- The fitted model has  $K$  components with  $K > K_0$ .
  - Theorem: When  $\alpha_k < L/2$ , where  $L = |\boldsymbol{\xi}_k|$  denotes the number of parameters specifying the component kernels, the posterior is stable and concentrates in regions with empty redundant components.
  - Markov chains expected to reach steady states with empty redundant components.

$$y_i \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k p(y_i | \boldsymbol{\xi}_k)$$

- Let the true model be a finite mixture with  $K_0$  components.
- The fitted model has  $K$  components with  $K > K_0$ .
- **Theorem:** When  $\alpha_k < L/2$ , where  $L = |\boldsymbol{\xi}_k|$  denotes the number of parameters specifying the component kernels, the posterior is stable and concentrates in regions with empty redundant components.
- Markov chains expected to reach steady states with empty redundant components.

$$y_i \stackrel{iid}{\sim} \sum_{k=1}^K \pi_k p(y_i | \boldsymbol{\xi}_k)$$

- Let the true model be a finite mixture with  $K_0$  components.
- The fitted model has  $K$  components with  $K > K_0$ .
- **Theorem:** When  $\alpha_k < L/2$ , where  $L = |\boldsymbol{\xi}_k|$  denotes the number of parameters specifying the component kernels, the posterior is stable and concentrates in regions with empty redundant components.
- Markov chains expected to reach steady states with empty redundant components.

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
- Expectation-Maximization (EM) algorithm
  - Iterative algorithm for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables
  - The basic idea is to iteratively guess the values of the latent variables and then update the parameter estimates based on those guesses
  - The process repeats until the parameter estimates converge
- EM algorithm for mixture models
  - Iterative algorithm for estimating the parameters of a mixture model
  - The basic idea is to iteratively guess the values of the latent variables (which cluster each data point belongs to) and then update the parameter estimates based on those guesses
  - The process repeats until the parameter estimates converge
- MCMC algorithm for mixture models
  - Monte Carlo Markov Chain (MCMC) algorithm for estimating the parameters of a mixture model
  - The basic idea is to generate a sequence of random samples from the posterior distribution of the parameters given the data
  - The samples are used to estimate the parameters and their uncertainty

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions

• Inference for mixture models

- Bayesian approach
- Maximum likelihood estimation
- EM algorithm

- Likelihood function becomes difficult to directly work with

- EM algorithm

• The EM algorithm iterates between two steps:

- E-step: calculate the expected value of the log-likelihood function given the current parameter estimates
- M-step: calculate the maximum likelihood estimate of the parameters given the expected values from the E-step

• The process repeats until the parameters converge to a local maximum of the log-likelihood function.

• The EM algorithm is widely used in mixture modeling due to its simplicity and efficiency.

- EM algorithm for mixture models

• The EM algorithm for mixture models follows a similar iterative process:

- E-step: calculate the posterior probabilities of each cluster membership for each data point
- M-step: calculate the maximum likelihood estimates of the parameters given the posterior probabilities from the E-step

- MCMC algorithm for mixture models

• The MCMC algorithm for mixture models uses a different approach:

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm

• Maximum likelihood estimation for mixture models is non-convex and has many local optima.

• The EM algorithm provides a way to iteratively find local optima.

• The EM algorithm is guaranteed to converge to a local optimum.

• The EM algorithm is widely used for mixture models and other latent variable models.

• The EM algorithm is a general iterative optimization method.

• The EM algorithm is a general iterative optimization method.

- EM algorithm for mixture models

• The EM algorithm for mixture models is a special case of the general EM algorithm.

• The EM algorithm for mixture models is a special case of the general EM algorithm.

- MCMC algorithm for mixture models

• The MCMC algorithm for mixture models is a special case of the general MCMC algorithm.

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm

• Maximum likelihood estimation for mixture models

• Bayesian approach to mixture models

• Inference for mixture models

• Model selection for mixture models

• Applications of mixture models

- EM algorithm for mixture models

• Maximum likelihood estimation for mixture models

• Bayesian approach to mixture models

- MCMC algorithm for mixture models

• Inference for mixture models

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - **Location-scale mixtures of normals are practically more efficient**
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm

• Maximum likelihood estimation for mixture models

• Bayesian approach to mixture models

• Inference for mixture models

• Model selection for mixture models

• Applications of mixture models

• Nonparametric Bayesian mixture models

• Finite mixture models

• Mixture models for discrete data

• Mixture models for survival analysis

• Mixture models for multivariate data

• Mixture models for time series data

• Mixture models for spatial data

• Mixture models for network data

• Mixture models for text data

• Mixture models for image data

• Mixture models for audio data

• Mixture models for video data

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm

• Maximum likelihood estimation for mixture models

• Expectation-Maximization (EM) algorithm for mixture models

• Bayesian approach to mixture models

• Nonparametric Bayesian approach to mixture models

• Dirichlet process mixture models

• Gaussian process mixture models

• Hierarchical Bayesian approach to mixture models

• Bayesian nonparametric approach to mixture models

- EM algorithm for mixture models

• Expectation-Maximization (EM) algorithm for mixture models

• Bayesian approach to mixture models

• Nonparametric Bayesian approach to mixture models

- MCMC algorithm for mixture models

• Markov chain Monte Carlo (MCMC) algorithm for mixture models

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm

- EM algorithm for mixture models

• The EM algorithm iterates between two steps:

- MCMC algorithm for mixture models

• The MCMC algorithm iterates between two steps:

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some 'missing' data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of 'missing' data
  - Suffers from typical optimization algorithm issues
- EM algorithm for mixture models
  - Iteratively updates the parameters of the mixture components
  - The E-step computes the posterior probabilities of each component given the current parameter estimates
  - The M-step updates the parameter estimates based on the posterior probabilities and the observed data
- MCMC algorithm for mixture models
  - Monte Carlo Markov Chain (MCMC) is used to sample from the posterior distribution of the mixture parameters
  - The MCMC algorithm iteratively samples from the conditional posterior distributions of the parameters given the current state of the other parameters
  - The MCMC algorithm provides a way to estimate the posterior distribution of the mixture parameters without having to compute the full likelihood function

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
  - Suffers from typical optimization algorithm issues
- EM algorithm for mixture models
- MCMC algorithm for mixture models

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
- EM algorithm for mixture models
- MCMC algorithm for mixture models

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
      - MAP: iteratively maximizes the posterior
    - Suffers from typical optimization algorithm issues
- EM algorithm for mixture models
- MCMC algorithm for mixture models

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
- EM algorithm for mixture models
- Variational Bayes for mixture models
- MCMC algorithm for mixture models

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
    - Can be very sensitive to initial conditions and get stuck at local optima
    - Can be very sensitive to small variations in updating rules
- EM algorithm for mixture models
- MCMC algorithm for mixture models

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
    - Can be very sensitive to initial conditions and get stuck at local optima
    - Can be very sensitive to small variations in updating rules
- EM algorithm for mixture models
- MCMC algorithm for mixture models

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
    - Can be very sensitive to initial conditions and get stuck at local optima
    - Can be very sensitive to small variations in updating rules
- EM algorithm for mixture models
- MCMC algorithm for mixture models

# Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
    - Can be very sensitive to initial conditions and get stuck at local optima
    - Can be very sensitive to small variations in updating rules
- EM algorithm for mixture models
  - Updating equations are typically straightforward
  - MLE often leads to singularity problems
  - Singularity problems can be avoided with MAP estimation under proper priors
- MCMC algorithm for mixture models

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
    - Can be very sensitive to initial conditions and get stuck at local optima
    - Can be very sensitive to small variations in updating rules
- EM algorithm for mixture models
  - **Updating equations are typically straightforward**
  - MLE often leads to singularity problems
  - Singularity problems can be avoided with MAP estimation under proper priors
- MCMC algorithm for mixture models

## Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
    - Can be very sensitive to initial conditions and get stuck at local optima
    - Can be very sensitive to small variations in updating rules
- EM algorithm for mixture models
  - Updating equations are typically straightforward
  - **MLE often leads to singularity problems**
    - Singularity problems can be avoided with MAP estimation under proper priors
- MCMC algorithm for mixture models

# Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
    - Can be very sensitive to initial conditions and get stuck at local optima
    - Can be very sensitive to small variations in updating rules
- EM algorithm for mixture models
  - Updating equations are typically straightforward
  - MLE often leads to singularity problems
  - Singularity problems can be avoided with MAP estimation under proper priors
- MCMC algorithm for mixture models

# Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
    - Can be very sensitive to initial conditions and get stuck at local optima
    - Can be very sensitive to small variations in updating rules
- EM algorithm for mixture models
  - Updating equations are typically straightforward
  - MLE often leads to singularity problems
  - Singularity problems can be avoided with MAP estimation under proper priors
- MCMC algorithm for mixture models
  - Gibbs sampling straightforward in conjugate/semi-conjugate models
  - Unlike EM, approximates the whole posterior not just a point estimate

# Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
    - Can be very sensitive to initial conditions and get stuck at local optima
    - Can be very sensitive to small variations in updating rules
- EM algorithm for mixture models
  - Updating equations are typically straightforward
  - MLE often leads to singularity problems
  - Singularity problems can be avoided with MAP estimation under proper priors
- MCMC algorithm for mixture models
  - Gibbs sampling straightforward in conjugate/semi-conjugate models
    - Unlike EM, approximates the whole posterior not just a point estimate

# Module VI Summary

- Mixture models
  - Identifiability issues - label switching and overfitting
  - High flexibility - can approximate large classes of distributions
    - Location mixtures of normals can approximate essentially any continuous density
    - Location-scale mixtures of normals are practically more efficient
    - Mixtures of Poissons are somewhat limited for modeling distributions on counts
  - Likelihood function becomes difficult to directly work with
- EM algorithm
  - Applicable to a broad class of problems when the likelihood conditional on some ‘missing’ data is easier to handle compared to the marginal likelihood based only on observed data
  - Iteratively updates an optimization function making clever use of ‘missing’ data
    - MLE: iteratively maximizes the likelihood
    - MAP: iteratively maximizes the posterior
  - Suffers from typical optimization algorithm issues
    - Can be very sensitive to initial conditions and get stuck at local optima
    - Can be very sensitive to small variations in updating rules
- EM algorithm for mixture models
  - Updating equations are typically straightforward
  - MLE often leads to singularity problems
  - Singularity problems can be avoided with MAP estimation under proper priors
- MCMC algorithm for mixture models
  - Gibbs sampling straightforward in conjugate/semi-conjugate models
  - Unlike EM, approximates the whole posterior not just a point estimate

## Additional topics

- EM algorithm - an alternative view
- Stochastic EM algorithm

## EM Algorithm - an Alternative View of 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:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \}$
- $$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \} \\ &= \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \\ &= \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z}) + \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})} \\ &= \mathcal{J}(q, \boldsymbol{\theta}) + D_{KL} \{ q(\mathbf{z}), p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \} \end{aligned}$$
- (a) After the E-step:  $\mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m)}) \geq \mathcal{J}(q^{(m)}, \boldsymbol{\theta}^{(m)})$   
(b) After the M-step:  $\mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m+1)}) \geq \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m)})$
- $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) = \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m+1)}) - \mathcal{J}(q^{(m)}, \boldsymbol{\theta}^{(m)}) + (0 - 0) \geq 0$

## EM Algorithm - an Alternative View of 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:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \}$
- $\mathcal{J}(q, \boldsymbol{\theta}) = \underbrace{\mathcal{L}(\boldsymbol{\theta})}_{no \ q} - D_{KL} \{ q(\mathbf{z}), p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \} = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})}_{no \ \boldsymbol{\theta}}$
- **EM algorithm:**

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

(a) **E-step:**  $q^{(m+1)} = \arg \max_q \mathcal{J}(q, \boldsymbol{\theta}^{(m)})$

(b) **M-step:**  $\boldsymbol{\theta}^{(m+1)} = \arg \max_{\boldsymbol{\theta}} \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta})$

- (a) After the E-step:  $\mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m)}) \geq \mathcal{J}(q^{(m)}, \boldsymbol{\theta}^{(m)})$
- (b) After the M-step:  $\mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m+1)}) \geq \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m)})$
- $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) = \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m+1)}) - \mathcal{J}(q^{(m)}, \boldsymbol{\theta}^{(m)}) + (0 - 0) \geq 0$

## EM Algorithm - an Alternative View of 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:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \}$
- $\mathcal{J}(q, \boldsymbol{\theta}) = \underbrace{\mathcal{L}(\boldsymbol{\theta})}_{no \ q} - D_{KL} \{ q(\mathbf{z}), p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \} = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})}_{no \ \boldsymbol{\theta}}$
- **EM algorithm:**

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

(a) **E-step:**  $q^{(m+1)} = \arg \min_q D_{KL} \left\{ q(\mathbf{z}), p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \right\} = p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})$

(b) **M-step:**  $\boldsymbol{\theta}^{(m+1)} = \arg \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$

- (a) After the E-step:  $\mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m)}) \geq \mathcal{J}(q^{(m)}, \boldsymbol{\theta}^{(m)})$
- (b) After the M-step:  $\mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m+1)}) \geq \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m)})$
- $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) = \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m+1)}) - \mathcal{J}(q^{(m)}, \boldsymbol{\theta}^{(m)}) + (0 - 0) \geq 0$

## EM Algorithm - an Alternative View of 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:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \}$
- $\mathcal{J}(q, \boldsymbol{\theta}) = \underbrace{\mathcal{L}(\boldsymbol{\theta})}_{no \ q} - D_{KL} \{ q(\mathbf{z}), p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \} = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})}_{no \ \boldsymbol{\theta}}$
- **EM algorithm:**

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

(a) **E-step:**  $q^{(m+1)} = \arg \min_q D_{KL} \left\{ q(\mathbf{z}), p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \right\} = p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})$

(b) **M-step:**  $\boldsymbol{\theta}^{(m+1)} = \arg \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$

- (a) **After the E-step:**  $\mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m)}) \geq \mathcal{J}(q^{(m)}, \boldsymbol{\theta}^{(m)})$
- (b) **After the M-step:**  $\mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m+1)}) \geq \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m)})$

•  $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) = \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m+1)}) - \mathcal{J}(q^{(m)}, \boldsymbol{\theta}^{(m)}) + (0 - 0) \geq 0$

## EM Algorithm - an Alternative View of 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:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \{ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \}$
- $\mathcal{J}(q, \boldsymbol{\theta}) = \underbrace{\mathcal{L}(\boldsymbol{\theta})}_{no \ q} - D_{KL} \{ q(\mathbf{z}), p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) \} = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})}_{no \ \boldsymbol{\theta}}$
- **EM algorithm:**

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

(a) **E-step:**  $q^{(m+1)} = \arg \min_q D_{KL} \left\{ q(\mathbf{z}), p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}) \right\} = p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})$

(b) **M-step:**  $\boldsymbol{\theta}^{(m+1)} = \arg \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$

- (a) **After the E-step:**  $\mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m)}) \geq \mathcal{J}(q^{(m)}, \boldsymbol{\theta}^{(m)})$
- (b) **After the M-step:**  $\mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m+1)}) \geq \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m)})$
- $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) = \mathcal{J}(q^{(m+1)}, \boldsymbol{\theta}^{(m+1)}) - \mathcal{J}(q^{(m)}, \boldsymbol{\theta}^{(m)}) + (0 - 0) \geq 0$

## Stochastic EM 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:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- If  $\mathbf{y}_i \stackrel{iid}{\sim} p(\mathbf{y}_i \mid \boldsymbol{\theta})$ , then  $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathcal{L}_i(\boldsymbol{\theta}) = \sum_{i=1}^n \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$   
 $= \sum_{i=1}^n \{\log p(\mathbf{y}_i, \mathbf{z}_i \mid \boldsymbol{\theta}) - \log p(\mathbf{z}_i \mid \mathbf{y}_i, \boldsymbol{\theta})\}.$
- **EM algorithm:**

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

(a) **E-step:** Compute  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^n \mathbb{E}_{\mathbf{z}_i \sim p(\mathbf{z}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}_i, \mathbf{z}_i \mid \boldsymbol{\theta}).$

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

## Stochastic EM 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:**  $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$
- If  $\mathbf{y}_i \stackrel{iid}{\sim} p(\mathbf{y}_i \mid \boldsymbol{\theta})$ , then  $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathcal{L}_i(\boldsymbol{\theta}) = \sum_{i=1}^n \log p(\mathbf{y}_i \mid \boldsymbol{\theta})$   
 $= \sum_{i=1}^n \{\log p(\mathbf{y}_i, \mathbf{z}_i \mid \boldsymbol{\theta}) - \log p(\mathbf{z}_i \mid \mathbf{y}_i, \boldsymbol{\theta})\}.$
- **EM algorithm:**

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

(a) **E-step:** Compute  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^n \mathbb{E}_{\mathbf{z}_i \sim p(\mathbf{z}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(m)})} \log p(\mathbf{y}_i, \mathbf{z}_i \mid \boldsymbol{\theta}).$

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

- **Stochastic EM algorithm:**

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

(a) **E-step:** Compute  $Q_S(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^n \log p(\mathbf{y}_i, \mathbf{z}_i^{(m)} \mid \boldsymbol{\theta}),$   
where  $\mathbf{z}_i^{(m)} \sim p(\mathbf{z}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(m)}).$

(b) **M-step:** Compute  $\boldsymbol{\theta}^{(m+1)} = \arg \max_{\boldsymbol{\theta}} Q_S(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}).$

## Stochastic EM Algorithm - Normal Location-Scale Mixtures

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

► **E-step:**  $Q_S(\theta, \theta^{(m)}) = \sum_{i=1}^n \left\{ -\frac{1}{2} \log \sigma_{z_i^{(m)}}^2 - \frac{(y_i - \mu_{z_i^{(m)}})^2}{2\sigma_{z_i^{(m)}}^2} + \log \pi_{z_i^{(m)}} \right\},$

$$z_i^{(m)} \sim p(z_i | y_i, \theta^{(m)}) \equiv \text{Mult}(1, \pi_i^{(m)}), \quad \pi_{ik}^{(m)} = \frac{\pi_k^{(m)} \times \text{Normal}(y_i | \mu_k^{(m)}, \sigma_k^{(m)2})}{\sum_{j=1}^K \pi_j^{(m)} \times \text{Normal}(y_i | \mu_j^{(m)}, \sigma_j^{(m)2})}$$

► **M-step:**  $\theta^{(m+1)} = \arg \max_{\theta} Q_S(\theta, \theta^{(m)})$

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

$$\Rightarrow \pi_k^{(m+1)} = \frac{\sum_{i=1}^n 1\{z_i^{(m)} = k\}}{n},$$

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

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

## Stochastic EM Algorithm - Normal Location-Scale Mixtures

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

- ▶ **E-step:**  $Q_S(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^n \left\{ -\frac{1}{2} \log \sigma_{z_i^{(m)}}^2 - \frac{(y_i - \mu_{z_i^{(m)}})^2}{2\sigma_{z_i^{(m)}}^2} + \log \pi_{z_i^{(m)}} \right\},$
- $z_i^{(m)} \sim p(z_i | y_i, \boldsymbol{\theta}^{(m)}) \equiv \text{Mult}(1, \boldsymbol{\pi}_i^{(m)}), \quad \pi_{ik}^{(m)} = \frac{\pi_k^{(m)} \times \text{Normal}(y_i | \mu_k^{(m)}, \sigma_k^{(m)2})}{\sum_{j=1}^K \pi_j^{(m)} \times \text{Normal}(y_i | \mu_j^{(m)}, \sigma_j^{(m)2})}$

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

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

$$\Rightarrow \pi_k^{(m+1)} = \frac{\sum_{i=1}^n 1\{z_i^{(m)} = k\}}{n},$$

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

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

## Stochastic EM Algorithm - Normal Location-Scale Mixtures

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

► **E-step:**  $Q_S(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^n \left\{ -\frac{1}{2} \log \sigma_{z_i^{(m)}}^2 - \frac{(y_i - \mu_{z_i^{(m)}})^2}{2\sigma_{z_i^{(m)}}^2} + \log \pi_{z_i^{(m)}} \right\},$

$$z_i^{(m)} \sim p(z_i | y_i, \boldsymbol{\theta}^{(m)}) \equiv \text{Mult}(1, \boldsymbol{\pi}_i^{(m)}), \quad \pi_{ik}^{(m)} = \frac{\pi_k^{(m)} \times \text{Normal}(y_i | \mu_k^{(m)}, \sigma_k^{(m)2})}{\sum_{j=1}^K \pi_j^{(m)} \times \text{Normal}(y_i | \mu_j^{(m)}, \sigma_j^{(m)2})}$$

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

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

$$\Rightarrow \pi_k^{(m+1)} = \frac{\sum_{i=1}^n 1\{z_i^{(m)} = k\}}{n},$$

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

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