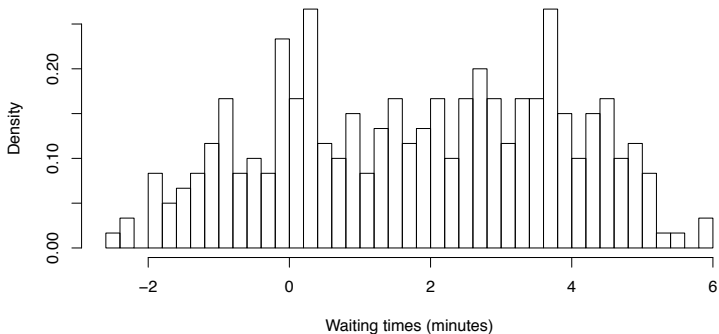


# The expectation-maximisation (EM) algorithm

- The expectation maximization (EM) algorithm (Dempster et al., 2007) is an **alternative procedure for the computation of maximum likelihood estimators**.
- In certain models – particularly **missing data and data augmentation problems** – the EM algorithm appears naturally and simplifies the maximum likelihood problem.

# Motivation: mixture models

Histogram of data from a mixture of two gaussian distribution.



→ Mixture of two univariate normal distributions

# Statistical model

- We model the data as coming from a normal mixture

$$y_i \sim \pi_1 \mathcal{N}(\mu_1, 1) + \pi_2 \mathcal{N}(\mu_2, 1),$$

$i = 1, \dots, n$  with  $\pi_1 \in (0, 1)$ ,  $\pi_1 + \pi_2 = 1$

- The vector of unknowns is  $\theta = (\pi_1, \mu_1, \mu_2)^\top$
- What is the likelihood function?

See Blackboard

## EM for gaussian mixture

1. Take initial guesses for  $\hat{\pi}_j, \hat{\mu}_j$ .
2. **E-Step:** Compute

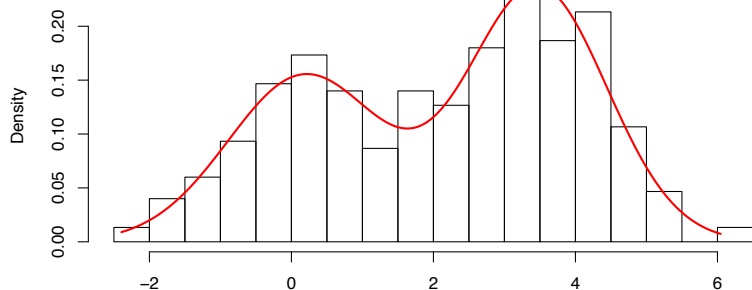
$$\hat{\delta}_{z_i}(k) = E\{I(z_i = k)|x_i, \theta\} = \frac{\hat{\pi}_k f_k(x_i|\hat{\mu}_k, 1)}{\sum_{k=1}^2 \hat{\pi}_k f_k(x_i|\hat{\mu}_k, 1)}$$

3. **M-Step:** Compute the mixing parameters and weighted means:

$$\hat{\pi}_k = \frac{\sum_i \delta_{z_i}(k)}{n}$$
$$\hat{\mu}_k = \frac{\sum_i \delta_{z_i}(k) x_i}{\sum_i \delta_{z_i}(k)}$$

4. Iterate steps 2 and 3 until convergence.

## EM for gaussian mixtures



We obtain  $\hat{\pi}_1 = 0.439$ ,  $\hat{\mu}_1 = -0.08$ ,  $\hat{\mu}_2 = 3.534$ .

## Direct Maximization

- Possible with the R function `optim()`
- Advantage: Standard errors are directly available
- Important: Choose good starting values

# EM algorithm in general

Situation:

- Observed variables  $X$
- Unobserved/hidden variables  $Z$
- Complete dataset  $Y = (X, Z)$

## The EM algorithm to maximize $\log L(\boldsymbol{\theta}; \mathbf{x})$

We would like to maximize  $L(\boldsymbol{\theta}; \mathbf{x})$  regarding  $\boldsymbol{\theta}$ , but we use  $L(\boldsymbol{\theta}; \mathbf{y})$  or rather  $\ell(\boldsymbol{\theta}; \mathbf{y}) = \log L(\boldsymbol{\theta}; \mathbf{y})$ .

**Input:** Function  $\ell(\boldsymbol{\theta}; \mathbf{y})$  and start value  $\boldsymbol{\theta}^{(0)}$

$i \leftarrow 0$ ;

**while** *not converged* **do**

*E-Step:* Compute the conditional expectation

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)}) = E(\ell(\boldsymbol{\theta}; \mathbf{y})|\mathbf{x}, \boldsymbol{\theta}^{(i)}),$$

where  $\ell(\boldsymbol{\theta}; \mathbf{y})$  is the complete data loglikelihood ;

*M-Step:* Determine  $\boldsymbol{\theta}^{(i+1)} = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta})$

*Update iteration number:*  $i \leftarrow i + 1$ ;

**end**



## Gene Example

See Blackboard

## Properties of the EM algorithm

- + In each iteration step of the EM algorithm the (incomplete) likelihood is increased:

$$L(\boldsymbol{\theta}^{(i+1)}; \mathbf{y}) \geq L(\boldsymbol{\theta}^{(i)}; \mathbf{y})$$

- + Parameter restrictions are (mostly) automatically fulfilled
- Convergence can be very slow – this especially depends on the “amount” of missing data
- Standard errors are not directly available. Some methods exist to try to approximate it, but they are not so easy to use in practice. Much easier just to bootstrap your data!

## Frequent applications of the EM algorithm

- Mixture models, cluster analysis
- Hidden Markov models
- Likelihood-based parameter estimation with missing data