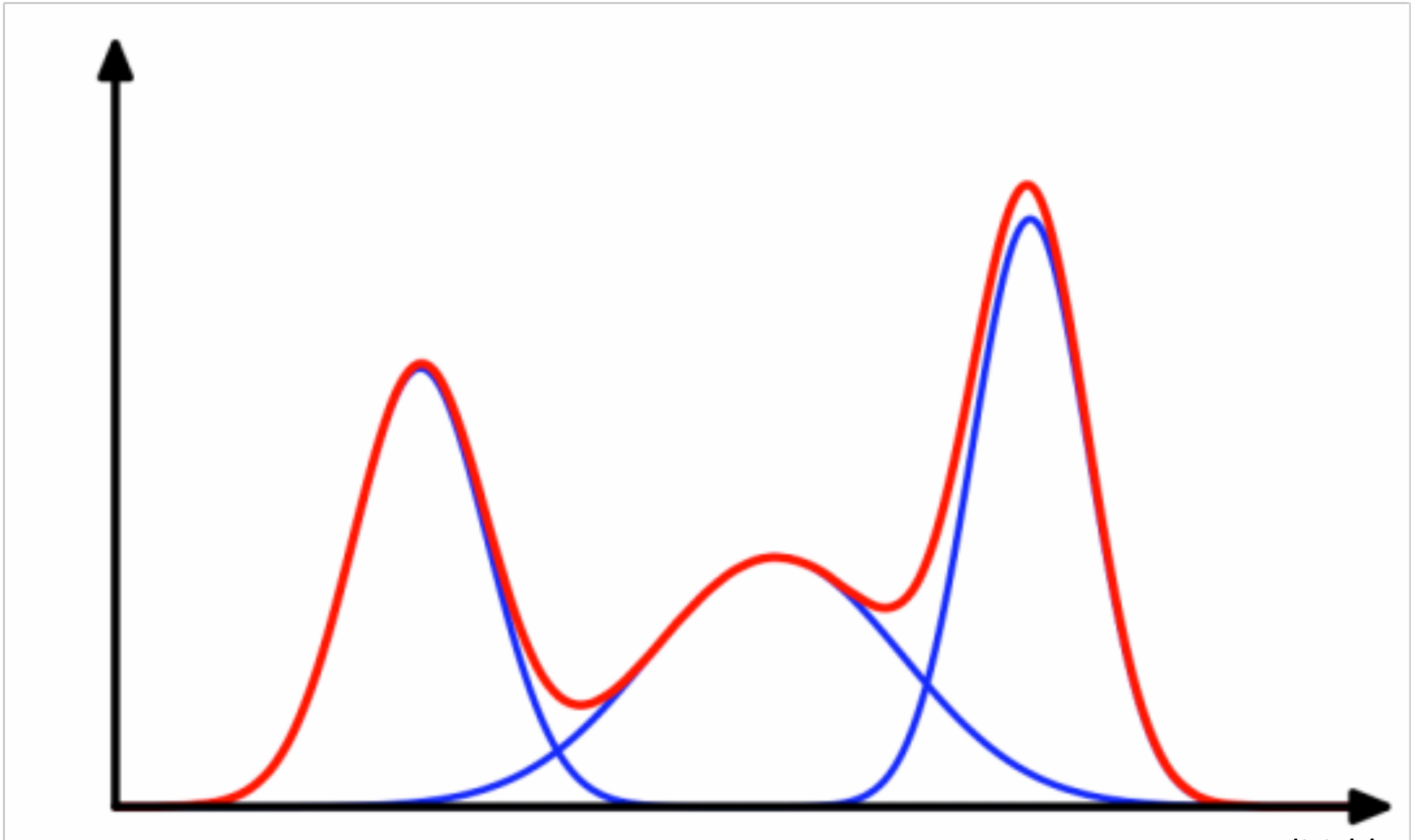
MODULE 2.6

EXPECTATION-MAXIMIZATION ALGORITHM

Gaussian mixture model



Credit: *dirichletprocess.weebly.com*

Overview of E-M Algorithm

- **Iterative algorithm for maximum likelihood estimation**

- **Particularly useful when…**
  - There are `missing' (unobserved) data.
  - The MLE is analytically intractable if missing data is unobserved
  - The MLE is analytically tractable if missing data is observed.

- **Examples include mixture models and censored regression**

- **Popular and highly cited : >50,000 times to date**

The Basic E-M Strategy

- **Observed and unobserved data: *(x, z)***
  - Complete data ***(x, z)*** – what we would like to have
  - Observed data ***x*** – individual observations
  - Missing data ***z*** – hidden/missing variables.

- **The E-M algorithm**
  1. E-step : Infer distribution of ***z*** using current data and parameters
  2. M-step : Update parameters using the inferred distribution.
  3. Repeat step 1-2 until convergence

The E-M Algorithm

- **Notations**
  - Complete data likelihood : $L(\theta|\mathbf{x},\mathbf{z}) = f(\mathbf{x},\mathbf{z}|\theta)$
  - Observed data likelihood : $L(\theta|\mathbf{x}) = g(\mathbf{x}|\theta) = \int_{\mathcal{X}} f(\mathbf{x},\mathbf{z}|\theta)d\mathbf{x}$
  - Expected log-likelihood: $Q(\theta|\theta^{(t)}) = \mathbf{E}\left[\log L\left(\theta|\mathbf{x},\mathbf{Z}\right)|\theta^{(t)},\mathbf{x}\right]$
- **E-step calculates** $\Pr(\mathbf{Z}|\mathbf{x},\theta^{(t)})$ **to evaluate** $Q(\theta|\hat{\theta}^{(t)})$
- **M-step finds** $\hat{\theta}^{(t+1)} = \arg\max_{\theta} Q(\theta|\hat{\theta}^{(t)})$

Key Theorem for E-M Algorithm

The E-M sequence $\{\hat{\theta}^{(t)}\}$ defined as $\hat{\theta}^{(t+1)} = \arg\max_{\theta} Q(\theta|\hat{\theta}^{(t)})$ satisfies

$$L(\hat{\theta}^{(t+1)}|\mathbf{x}) \geq L(\hat{\theta}^{(t)}|\mathbf{x})$$

with equality holding if and only if success iterations yield the same value of maximized expected complete-data log-likelihood, i.e.

$$Q(\hat{\theta}^{(t+1)}|\hat{\theta}^{(t)}) = Q(\hat{\theta}^{(t)}|\hat{\theta}^{(t)})$$

(Casella and Berger Theorem 7.20)

Why E-M algorithm works

- The E-step constructs a surrogate function such that

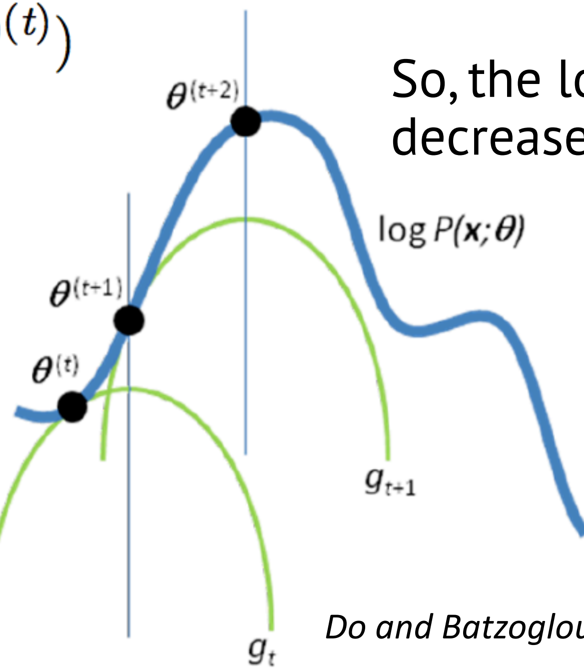
$$g^{(t)}(\theta) \leq \log p(\mathbf{x}|\theta)$$

$$g^{(t)}(\theta^{(t)}) = \log p(\mathbf{x}|\theta^{(t)})$$

- The M-step maximize the surrogate function

$$\theta^{(t+1)} = \text{argmax}_{\theta} g^{(t)}(\theta)$$

So, the log-likelihood never decreases.



Do and Batzoglou (2008) Nat Biotechnol.

Convergence of E-M Algorithm

- **The E-M sequences monotonically increases the observed data likelihood (which we would like to maximize) because expected log-likelihood will monotonically increase.**

- **With infinite iteration, the E-M sequences will reach to its local maximum.**
  - However, it does not guarantee that it reaches to the global maximum likelihood.

- **If the observed likelihood function is concave, E-M will converge to MLE (with unknown convergence speed)**

Likelihoods in Gaussian Mixture

$$\log L(\theta|\mathbf{x}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{k} \frac{\pi_k}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}} \right]$$

$$\log L(\theta|\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} e^{-\frac{(x_i - \mu_{z_i})^2}{2\sigma_{z_i}^2}} \right]$$

$$= -\frac{1}{2} \sum_{i=1}^{n} \log(2\pi\sigma_{z_i}^2) - \sum_{i=1}^{n} \frac{(x_i - \mu_{z_i})^2}{2\sigma_{z_i}^2}$$

E-step : Evaluating Expected Log Likelihood

$$
\begin{aligned}
Q(\theta|\hat{\theta}^{(t)}) &= \mathbf{E}\left[\log L\left(\theta|\mathbf{x},\mathbf{Z}\right)|\theta^{(t)},\mathbf{x}\right] \\
&= \sum_{\mathbf{Z}} \log L\left(\theta|\mathbf{x},\mathbf{Z}\right)\Pr(\mathbf{Z}|\mathbf{x},\hat{\theta}^{(t)}) \\
&= \sum_{i=1}^{n}\left[\sum_{z=1}^{k} w_i(z|x_i,\hat{\theta}^{(t)})\left(-\log(2\pi\sigma_z^2)-\frac{(x_i-\mu_z)^2}{2\sigma_z^2}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
w_i(z|x_i,\hat{\theta}^{(t)}) &= \Pr(Z_i = z|x_i,\hat{\theta}^{(t)}) \\
&= \frac{\pi_z f(x_i, Z_i = z|\hat{\theta}^{(t)}))}{\sum_{j=1}^{k}\pi_j f(x_i, Z_i = j|\hat{\theta}^{(t)}))}
\end{aligned}
$$

M-step : Maximizing Expected log-likelihood

$$Q(\theta|\hat{\theta}^{(t)}) = \sum_{i=1}^{n} \left[\sum_{z=1}^{k} w_i(z|x_i, \hat{\theta}^{(t)}) \left(-\log(2\pi\sigma_z^2) - \frac{(x_i - \mu_z)^2}{2\sigma_z^2} \right) \right]$$

- **Considering *w_i(.)* as given, we want to find parameters that maximizes the expected-log-likelihood**

$$\hat{\theta}^{(t+1)} = \left(\hat{\pi}^{(t+1)}, \hat{\mu}^{(t+1)}, \hat{\sigma^2}^{(t+1)} \right)$$

$$= \arg\max_{\theta} Q(\theta|\hat{\theta}^{(t)})$$

Details of M-step

$$Q(\theta|\hat{\theta}^{(t)}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\sum_{z=1}^{z} w_i(z)\log\sigma_z^2$$

$$-\sum_{i=1}^{n}\sum_{z=1}^{z}\frac{w_i(z)(x_i-\mu_z)^2}{2\sigma_z^2}$$

$$\hat{\mu}_z^{(t+1)} = \frac{\sum_{i=1}^{n} w_i(z)x_i}{\sum_{i=1}^{n} w_i(z)} \qquad \hat{\pi}_z^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} w_i(z)$$

$$\hat{\sigma^2}_z^{(t+1)} = \frac{\sum_{i=1}^{n} w_i(z)(x_i-\mu_z^{(t+1)})^2}{\sum_{i=1}^{n} w_i(z)}$$

Implementation in R (E-step)

```
em <- function(x, k, max.iter = 1000, tol = 1e-8) {
    n <- length(x)
    pis <- rep(1/k, k)    ## start with uniform priors
    mus <- sample(x,k)    ## start with random points as mean
    sds <- rep(sd(x), k) ## start with pooled variance
    W <- t(rmultinom(n,1,pis))  ## W is n x k matrix
    prevLLK <- -1e300
    for(i in 1:max.iter) {
        ## E-step, calculate pi_j * Pr(x_i|mu_j,sd_j)
        W <- matrix(pis,n,k,byrow=TRUE) * dnorm(matrix(x,n,k),
                  matrix(mus,n,k,byrow=TRUE),matrix(sds,n,k,byrow=TRUE));
        Wsum <- rowSums(W)
        W <- W / matrix(Wsum, n, k)  ## calculate Pr(Z|x)
        llk <- sum(log(Wsum))        ## calculate likelihood
        if ( llk - prevLLK < tol ) { break }
        prevLLK <- llk
```

Implementation in R (M-step)

```
        ## M-step
        pis <- colSums(W) / n
        mus <- (x %*% W) / (pis * n)
        sds <- sqrt(colSums((matrix(x, n, k) - matrix(mus, n, k, byrow=TRUE))^2
* W) / (pis * n))
    }
    return(list(llk=llk, pis=pis, mus=mus, sds=sds,iter=i))
}
```

Running example

```
x <- c(rnorm(1000), rnorm(500)+5)
em(x,2)
```

```
$llk
[1] -3055.658

$pis
[1] 0.3306386 0.6693614

$mus
         [,1]       [,2]
[1,] 5.046562 0.01170993

$sds
[1] 0.9502852 1.0247390

$iter
[1] 41
```

Evaluation of E-M Algorithm

- **Advantages**
  - Converges to (local) maximum
  - Does not require much information (e.g. derivatives)
  - Easy to implement and use
  - A very widely used method : cited >40,000 times to date
  - Often faster than alternative algorithms (such as Nelder-Mead)

- **Disadvantages**
  - Convergence to global maximum is not guaranteed
  - Speed of convergence is not guaranteed (could be very slow).
  - For high-dimensional parameters, convergence property is poor
  - Convergence criteria is not so clear

More E-M Algorithm Examples

- Suppose that

$$X_i \sim \pi_1 \text{Poisson}(\lambda_1 \mu_i) + \pi_2 \text{Poisson}(\lambda_2 \mu_i)$$

$$i \in \{1, \cdots, n\}, X_i \in \{0, 1, 2, \cdots\}$$

$$\mu_i, \lambda_1, \lambda_2, \pi_1, \pi_2 > 0, \pi_1 + \pi_2 = 1$$

where μ_i are given

Intuitive guess : Use fractional counts

$$Z_i \in \{1, 2\}$$

$$X_i | Z_i = 1 \sim \text{Poisson}(\lambda_1^{(t)} \mu_i)$$

$$X_i | Z_i = 2 \sim \text{Poisson}(\lambda_2^{(t)} \mu_i)$$

$$w_{ik}^{(t)} = \Pr(Z_i = k | X_i) = \frac{\pi_k^{(t)} \Pr(X_i | Z_i = k)}{\pi_1^{(t)} \Pr(X_i | Z_i = 1) + \pi_2^{(t)} \Pr(X_i | Z_i = 2)}$$

$$\lambda_k^{(t+1)} = \frac{\sum_{i=1}^n w_{ik}^{(t)} x_i}{\sum_{i=1}^n w_{ik}^{(t)} \mu_i} \qquad \pi_k^{(t+1)} = \frac{\sum_{i=1}^n w_{ik}^{(t)}}{\sum_{j=1}^2 \sum_{i=1}^n w_{ij}^{(t)}}$$

Check whether the guess is right

$$l(\theta|\mathbf{x},\mathbf{z}) = \sum_{i=1}^{n} \log\left[\frac{e^{-\lambda_{z_i}\mu_i}\left(\lambda_{z_i}\mu_i\right)^{x_i}}{x_i!}\right]$$

$$= \sum_{i=1}^{n} \left[-\lambda_{z_i}\mu_i + x_i \log(\lambda_{z_i}\mu_i) - \log x_i!\right]$$

$$Q(\theta|\theta^{(t)}) = \mathrm{E}_{\mathbf{Z},\theta^{(t)}}\left[l(\theta|\mathbf{x},\mathbf{Z})\right]$$

$$= \sum_{i=1}^{n} \left[-\mathrm{E}_{\mathbf{Z},\theta^{(t)}}[\lambda_{z_i}]\mu_i + x_i \mathrm{E}_{\mathbf{Z},\theta^{(t)}}[\log \lambda_{z_i}]\right.$$

$$\left. + x_i \log \mu_i - \log x_i!\right]$$

To maximize the expected log-likelihood..

$$f(\lambda) = \sum_{i=1}^{n} \left[\sum_{k=1}^{2} \left(-w_{ik}^{(t)} \lambda_k \mu_i + x_i w_{ik}^{(t)} \log \lambda_k \right) + C_i \right]$$

$$\frac{\partial f(\lambda)}{\partial \lambda_k} = \sum_{i=1}^{n} \left[-w_{ik}^{(t)} \mu_i + \frac{x_i w_{ik}^{(t)}}{\lambda_k} \right] = 0$$

$$\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ik}^{(t)} x_i}{\sum_{i=1}^{n} w_{ik}^{(t)} \mu_i}$$

Recommended Readings

- **Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977).**
  Maximum likelihood from incomplete data via the EM algorithm.
  ***J. R. Stat. Soc.*, B 39, 1–38.**

- **Casella & Berger (2001) Chapter 7.2.4**