

University of Edinburgh, School of Mathematics

Incomplete Data Analysis, 2020/2021

EM algorithm: (toy) examples

Vanda Inácio

In this supplementary file we implement the EM algorithm for the examples seen in the lecture.

Toy example (2 exponential observations)

Remember that we have the following updating equation

$$\theta^{(t+1)} = \frac{2\theta^{(t)}}{5\theta^{(t)} + 1}.$$

Note that at convergence, $\theta^{(t)} \rightarrow \theta^{(t+1)} \rightarrow \hat{\theta}$, and so a fixed point of these iterations is $\hat{\theta} = 1/5$. Nevertheless, let us code it. The function takes as input a starting value for θ , say $\theta^{(0)}$, and the value ϵ used for the stopping criterion: $|\theta^{(t)} - \theta^{(t-1)}| < \epsilon$. The variable `diff` in the code below is just to control whether the convergence criterion is met or not. Although we have started setting `diff=1`, any value greater than ϵ would work. The variable `theta.old` stores the value from the previous iteration, so that we can compute $|\theta^{(t+1)} - \theta^{(t)}|$.

```
toyex <- function(theta0, eps){  
  
  diff <- 1  
  theta <- theta0  
  while(diff > eps){  
  
    theta.old <- theta  
    theta <- 2*theta/(5*theta+1)  
    diff <- abs(theta - theta.old)  
  }  
  return(theta)  
}  
  
toyex(10, 0.00001)
```

```
## [1] 0.200006
```

Genetic linkage model

Remember that we have for the E-step

$$Q(\theta \mid \theta^{(t)}) = (y_1 - z^{(t)} + y_4) \log \theta + (y_2 + y_3) \log(1 - \theta),$$
$$z^{(t)} = y_1 \times \frac{1/2}{1/2 + \theta^{(t)}/4},$$

while for the M-step we have

$$\theta^{(t+1)} = \frac{y_1 - z^{(t)} + y_4}{n - z^{(t)}}, \quad n = y_1 + y_2 + y_3 + y_4.$$

```
multi <- function(y, theta0, eps){
  n <- sum(y); diff <- 1
  theta <- theta0

  while(diff>eps){
    theta.old <- theta

    #E step
    zt <- y[1]*0.5/(0.5 + 0.25*theta)

    #M step
    theta <- (y[1] + y[4] - zt)/(n - zt)

    diff <- abs(theta - theta.old)
  }
  return(theta)
}

y <- c(125, 18, 20, 34)
multi(y = y, 0.5, 0.00001)
```

```
## [1] 0.6268207
```

Incomplete univariate (normal) data

Remember that we have the following updating equation

$$\mu^{(t+1)} = \frac{\sum_{i=1}^m y_i + (n - m)\mu^{(t)}}{n}.$$

Again, at convergence, $\mu^{(t)} \rightarrow \mu^{(t+1)} \rightarrow \hat{\mu}$ and so a fixed point of these iterations is $\hat{\mu} = \frac{1}{m} \sum_{i=1}^m y_i$, exactly what we would have obtained by maximising the log likelihood of the observed data.

For this example I have simulated $n = 200$ observations and the true value of μ is 3. The missing data mechanism is MCAR and I have simply sampled 20 individuals to exclude from the analysis. So, in the notation of our example we have $n = 200$ and $m = 180$.

```
n <- 200
mu <- 3
set.seed(1)
y <- rnorm(n, mu, 1)
ind <- sample(x = 1:n, size = 20, replace = FALSE)

#observed data
y_obs <- y[-ind]

toyex <- function(mu0, eps, y, n){

  diff <- 1
  mu <- mu0
```

```

m <- length(y)

while(diff > eps){
  mu.old <- mu
  mu <- (sum(y) + (n-m)*mu)/n
  diff <- abs(mu - mu.old)
}
return(mu)
}

toyex(15, 0.00001, y = y_obs, n = n)

```

```
## [1] 3.033178
```

```
sum(y_obs)/(n-length(ind))
```

```
## [1] 3.033177
```