University of Edinburgh, School of Mathematics

Incomplete Data Analysis, 2020/2021

EM algorithm: bivariate normal data with one variable subject to missigness

Vanda Inácio

We will simulate one dataset, of size 1000, from a bivariate normal distribution with the following structure

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 5 \\ -1 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1.3 & 0.4 \\ 0.4 & 0.9 \end{pmatrix}.$$

Then, Y_2 values will be missing if the corresponding Y_1 value is above 6. The below function takes as input the data on Y_1 and Y_2, the missing data (on Y_2) indicator, a vector of initial values and ε (which controls the convergence criterion).

```
require(MASS)
n <- 1000
mu <- c(5, -1)
Sigma <- matrix(c(1.3, 0.4, 0.4, 0.9), nrow = 2)
set.seed(1)
y <- mvrnorm(n, mu = mu, Sigma = Sigma)
y1 <- y[, 1]
y2 <- y[, 2]
r <- ifelse(y1 > 6, 0, 1)
# percentage of observed Y2 values
sum(r)/n
```

```
## [1] 0.801
```

```
em.bivariate.normal <- function(y1, y2, r, theta0, eps){

  n <- length(y1)
  m <- sum(r)
  theta <- theta0

  mu1 <- theta[1]
  sigma11 <- theta[2]
  mu2 <- theta[3]
  sigma22 <- theta[4]
  sigma12 <- theta[5]

  diff <- 1
  while(diff > eps){
    theta.old <- theta
```

```r
    beta1 <- sigma12/sigma11
    beta0 <- mu2 - beta1*mu1
    sigma12prime <- sigma22 - ((sigma12^2)/sigma11)

    # E-step
    t1 <- sum(y1)
    t11 <- sum(y1^2)
    t2 <- sum(y2[r == 1]) + (n-m)*beta0 + beta1*sum(y1[r == 0])
    t22 <- sum(y2[r == 1]^2) + (n-m)*sigma12prime + (n-m)*(beta0^2) + 2*beta0*beta1*sum(y1[r == 0]) + (
    t12 <- sum(y1[r == 1]*y2[r == 1]) + beta0*sum(y1[r == 0]) + beta1*sum(y1[r == 0]*y1[r == 0])

    # M-step
    mu1 <- t1/n
    sigma11 <- (t11/n) - (mu1^2)
    mu2 <- t2/n
    sigma22 <- (t22/n) - (mu2^2)
    sigma12 <- (t12/n) - mu1*mu2

    theta <- c(mu1, sigma11, mu2, sigma22, sigma12)
    diff <- sum(abs(theta - theta.old))
  }
  return(theta)
}

theta0 <- c(2, 0.5, 1, 0.5, 0.1)
em.bivariate.normal(y1 = y1, y2 = y2, r = r, theta0 = theta0, eps = 0.00001)
```

```
## [1]  5.0054174  1.3866814 -0.9951727  0.9422207  0.4040600
```

As it was emphasised in the lecture, $t_1^{(t)}$, $t_{11}^{(t)}$, $\mu_1^{(t+1)}$, and $(\sigma_1^{(t+1)})^2$ are constant across iterations as we do not have missing values on Y_1. As can be appreciated, we are able to recover the true values of the parameters well.