STAT 428 Statistical Computing

Homework 7 Solutions

Department of Statistics, University of Illinois at Urbana-Champaign

April 11, 2017

Ex.1

Observed data:

$$\mathbf{y} = (y_1, ..., y_n) \quad y_i \in (-\infty, \infty)$$

Missing data:

$$\mathbf{z} = (z_1, ..., z_n) \quad z_i \in \{0, 1\}$$

τ and μ are parameters need to be estimated.

Y_i follows Mixture Gaussian Distribution conditioning on Z_i.

$$p(y|z, \tau, \mu) = [\tau \times N(\mu, 1)]^z [(1-\tau) \times N(0, 1)]^{(1-z)}$$

Z_i follows Bernoulli Distribution with parameter τ.

$$Z_i \sim Bernoulli(\tau) \Longrightarrow \begin{cases} P(Z_i = 0) = 1 - \tau \\ P(Z_i = 1) = \tau \end{cases}$$

E Step

$$E_{\tau,\mu}[Z_i|Y_i = y_i] = P(Z_i = 1|y_i) = \frac{P(Z_i = 1, Y_i = y_i)}{P(Y_i = y_i)} = \frac{P(Y_i = y_i|Z_i = 1)P(Z_i = 1)}{P(Y_i = y_i|Z_i = 1)P(Z_i = 1) + P(Y_i = y_i|Z_i = 0)P(Z_i = 0)}$$

M Step

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

Substitute τ with $\bar{z}$ and Log-Likelihood function becomes

$$l(\mu; \mathbf{y}, \mathbf{z}) = \sum z_i(-\frac{(y_i - \mu)^2}{2}) + (1 - z_i)(\frac{{y_i}^2}{2}) + C = \sum z_i y_i \mu - \frac{\sum z_i}{2}\mu^2 + C,$$

where C is a constant.

$\Longrightarrow$

$$l'(\mu; \mathbf{y}, \mathbf{z}) = \sum z_i y_i - \sum z_i \mu = 0$$

$\Longrightarrow$

$$\hat{\mu} = \frac{\sum z_i y_i}{\sum z_i}$$

```r
###Log-likelihood function
###Input:
#y: data
#z: latent variable ('missing data')
#mu, tau: unknown parameters
###Output: value of log-likelihood function
loglike <- function(y,z,mu,tau){
    n = length(y)
    l1 = sum(z*(log(tau/sqrt(2*pi))-0.5*(y-mu)^2))
    l2 = sum((1-z)*(log((1-tau)/sqrt(2*pi))-0.5*y^2))
    return(l1+l2)
}
```

```r
###Function to implement EM-algorithm to estimate the parameters
###Input:
#y: data
#iter: maximum number of iterations
#mu0: initial value for mu
#tau0: initial value for tau
#threshold: when change in log-likelihood function < threshold, stop the iteration
###Output:
#mu,tau: the estimated parameters
#i: number of iterations before convergence
EM.mix.gaussian <- function(y,iter=1000,mu0,tau0,threshold=1e-6){
    n = length(y) #sample size
    #Initialize
    mu = mu0
    tau = tau0
    l = 1
    #iterations
    for(i in 1:iter){
        #E-step
        p1 = tau*dnorm(y,mu,1)
        p0 = (1-tau)*dnorm(y,0,1)
        z = p1/(p1+p0)
        #M-step: update parameters
        mu = sum(y*z)/sum(z)
        tau = mean(z)
        lnew = loglike(y,z,mu,tau)
        delta = abs(l - lnew)
        if(delta>threshold){
            l = lnew
        }
        else break
    }
    return(list(mu = mu, tau = tau, iterations = i))
}
```

Ex.2

```r
tau = 0.5
mu = 1
```

```r
n = 1000

y = c(rnorm(n*tau,mu,1),rnorm((1-tau)*n,0,1))
print(EM.mix.gaussian(y = y, mu0 = 0.1, tau0 = 0.1))
```

```
## $mu
## [1] 0.9149007
##
## $tau
## [1] 0.479055
##
## $iterations
## [1] 300
```