第３章のデータとプログラムについての説明

関西大学　高井啓二

ここでは，3.6.2 の正規分布の数値例と 3.6.4 の多項分布に対するデータとプログラムの説明を与える．

3.6.2 の例のデータ apple.csv とプログラム emexample.R について

使用するデータは apple.csv である．このデータの列は，変数(２変数)と欠測指標(２変数)からなり，第１，第２列はデータ，第３，第４列は欠測指標である．データにおいて欠測している部分には 1000 を入れてあるが，この値は何でも良い（最終的には０をかけて消してしまうので結果に影響は出ない)．欠測指標の値は，0 のとき対応する変数が観測，1 のとき欠測と本書内での一般的な記法とは逆になっていることに注意されたい．

プログラム emexample.R は 3.6.1 の「EM アルゴリズムの別表現」の通りに作成している．プログラムの主要な部分の説明は以下の通りである．

em.step(R,Y)：２変量正規分布モデルにもとづいて EM アルゴリズムを実行する．引数の R は欠測指標行列，Y は（仮想的完全）データ行列（欠測値には適当なもの代入しておく）である．戻り値は，Heikin，Bunsan，lik，count である．それぞれ，平均，分散，収束までの対数尤度の列，収束までの回数を意味する．

log.likelihood(mu,Sigma,R,Y)：２変量正規分布の対数観測尤度を計算する．引数の mu，Sigma，R，Y はそれぞれ平均，分散，（仮想的完全）データ，欠測指標行列である．戻り値は，尤度の値である．

<使用例>　*現在のフォルダが”Chapter3”であるとする．

```
source(“emexample1.R”)
tmp.dat <- read.csv(“apple.csv”,header=F) #データの読み込み
tmp.dat <- as.matrix(tmp.dat)
result <- em.step(tmp.dat[,3:4],tmp.dat[,1:2])
          #EM アルゴリズムの実行
result #結果を見る
plot(result$lik,type="l",ylab="value of log-likelihood",
      xlab="iteration") # 尤度のプロット
```

3.6.4 の例のデータとプログラム emalgorithm2.R について

この例では，EM アルゴリズムは 3.6.3 で導出されたものを使用している．データ(名前 dat)は，emalgorithm2.R の中で定義されているため，読み込みは不要である．プログラムの主要な部分の説明は以下の通りである．

em.step(dat)：多項分布のパラメータを計算する．引数の dat は 3×3 行列であり，表 3.8 に対応している．本文でも説明した通り，両方の変数が欠測している部分（115 のセル）は何が入っていても結果に影響を及ぼさない．戻り値は，param，lik，count である．それぞれ，多項分布のパラメータ，収束までの対数尤度の列，収束までの回数である．ただし，param はπの添え字が 11，12，21，22 の順番で並んでいる．

log.likelihood(param,dat)：多項分布の対数観測尤度を計算する．引数の param はパラメータ，dat はデータである．両方の引数共に em.step と同じ意味である．

<使用例>　*現在のフォルダが”Chapter3”であるとする．

```
source(“emalgorithm2.R”)
result <- em.step(dat) # EM アルゴリズムの実行
result  #結果を見る
plot(result$lik,type="l",ylab="value of log-likelihood",
     xlab="iteration") #尤度のプロット
```