
Answers must be given in a pdf file (coming from LaTeX, Word, OpenOffice or similar. In that file you must include everything you consider relevant: **explanations, comments, clarifications,** *R instructions, graphics, parts of the outputs provided R, etc. In particular you should include in the response file, as an Appendix, the R code that you use to solve problems.*

After finishing the exam, upload your file at ATENEA.

A scientific is interested in estimating the value of a parameter μ. She develops an experiment that provides random results distributed as $N(\mu, \sigma^2)$ and then she repeats the experiment n times independently. Let $X_1, \ldots, X_n$ be the results of the experiments.

Unfortunately, the measurement device used by the scientific has not enough precision to capture the exact values of X_i when they are close to 0. Instead of recording X_i, the device provides

$$Y_i = \begin{cases} 0 & \text{if} \quad -1 \le X_i \le 1, \\ X_i & \text{if} \quad |X_i| > 1. \end{cases}$$

The scientific wants to estimate the bidimensional parameter $\theta = (\mu, \sigma)$ using the EM algorithm. She works with $X_1, \ldots, X_n$ as the *complete data* and with $Y_1, \ldots, Y_n$ as the *observed data*. Let $y_1, \ldots, y_n$ be the data she finally obtains from the experimentation.

It can be useful to use this notation:

- $\phi(z)$ and $\Phi(z)$, the density function and the distribution functions of a $N(0, 1)$, respectively.
- $\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$, the density function of a $N(\mu, \sigma^2)$.
- $\Phi\left(\frac{x-\mu}{\sigma}\right)$, the distribution function of a $N(\mu, \sigma^2)$.

Answer the following questions.

1. What is the contribution to the likelihood function of the i-th experiment result when the observation y_i is equal to 0? And what is this contribution when $|y_i| > 1$?

   Solution: *When the observation y_i is equal to 0:*

$$\Pr(|X_i| \le 1; \mu, \sigma) = \Phi\left(\frac{1-\mu}{\sigma}\right) - \Phi\left(\frac{-1-\mu}{\sigma}\right).$$

   When $|y_i| > 1$:

$$f_X(y_i; \mu, \sigma) = \frac{1}{\sigma}\phi\left(\frac{y_i - \mu}{\sigma}\right).$$

2. Write down the log-likelihood function for the complete data, and the log-likelihood for the observed data.

   Solution: *Log-likelihood function for the complete data:*

$$l_c(\mu,\sigma;x_1,\dots,x_n)=\sum_{i=1}^{n}\log\left(\frac{1}{\sigma}\phi\left(\frac{x_i-\mu}{\sigma}\right)\right)=$$

$$-n\log(\sigma)-\frac{n}{2}\log 2\pi-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2.$$

   Log-likelihood for the observed data: Let δ_i be equal to 0 when the observation y_i is equal to 0, and 1 otherwise, then

$$l_o(\mu,\sigma;y_1,\dots,y_n)=$$

$$\sum_{i=1}^{n}(1-\delta_i)\log\left(\Phi\left(\frac{1-\mu}{\sigma}\right)-\Phi\left(\frac{-1-\mu}{\sigma}\right)\right)+\sum_{i=1}^{n}\delta_i\frac{1}{\sigma}\phi\left(\frac{y_i-\mu}{\sigma}\right)=$$

$$n_0\log\left(\Phi\left(\frac{1-\mu}{\sigma}\right)-\Phi\left(\frac{-1-\mu}{\sigma}\right)\right)+\sum_{i:y_i\neq 0}\log\left(\frac{1}{\sigma}\phi\left(\frac{y_i-\mu}{\sigma}\right)\right)=$$

$$n_0\log\left(\Phi\left(\frac{1-\mu}{\sigma}\right)-\Phi\left(\frac{-1-\mu}{\sigma}\right)\right)-(n-n_0)\log(\sigma)-(n-n_0)\frac{1}{2}\log 2\pi-\frac{1}{2}\sum_{i:y_i\neq 0}\left(\frac{y_i-\mu}{\sigma}\right)^2.$$

   where n_0 is the number of observed zeros.

3. **E step in the EM algorithm.** Give the expression of $Q(\mu,\sigma|\mu_m,\sigma_m)$.

   Indication: Truncated normal distribution. Given a r.v. X with density function $f(x)$ and distribution function $F(x)$, the density function of X conditional to $a<X<b$ is

$$f_{X|a<X<b}(x)=\begin{cases}\frac{f(x)}{F(b)-F(a)} & \text{if } a\le x\le b,\\ 0 & \text{otherwise.}\end{cases}$$

   If $X\sim N(\mu,\sigma^2)$, the distribution of X conditional to $a<X<b$ is called *truncated normal distribution in the interval* $[a,b]$ In this case, it can be proved that

$$\mathrm{E}_{\mu,\sigma}(X\mid a<X<b)=\mu+\frac{\phi(\frac{a-\mu}{\sigma})-\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}\sigma,$$

$$\mathrm{Var}_{\mu,\sigma}(X\mid a<X<b)=\sigma^2\left[1+\frac{\frac{a-\mu}{\sigma}\phi(\frac{a-\mu}{\sigma})-\frac{b-\mu}{\sigma}\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}-\left(\frac{\phi(\frac{a-\mu}{\sigma})-\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}\right)^2\right].$$

Moreover, for any $\tau \in \mathbb{R}$,

$$\mathrm{E}_{\mu,\sigma}((X-\tau)^2 \mid a < X < b) = \mathrm{Var}_{\mu,\sigma}(X \mid a < X < b) + (\mathrm{E}_{\mu,\sigma}(X \mid a < X < b) - \tau)^2 .$$

Solution: *Let $\tilde{y}_i$ be defined as in the following point, and let $\tilde{\sigma}^2 = Var_{\mu_m,\sigma_m}(X| -1 < X < 1)$. Then*

$$Q(\mu,\sigma|\mu_m,\sigma_m) = E_{\mu_m,\sigma_m}(l_c(\mu,\sigma;X_1,\ldots,X_n)|Y_1 = y_1,\ldots,Y_n = y_n) =$$

$$E_{\mu_m,\sigma_m}\left[\left(-n\log(\sigma) - \frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\right)|Y_1 = y_1,\ldots,Y_n = y_n\right] =$$

$$-n\log(\sigma) - \frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i:y_i\neq 0}\left(\frac{y_i-\mu}{\sigma}\right)^2 -$$

$$\frac{1}{2\sigma^2}\sum_{i:y_i=0} E_{\mu_m,\sigma_m}([(X_i-\tilde{y}_i)+(\tilde{y}_i-\mu)]^2| -1 < X_i < 1) =$$

$$-n\log(\sigma) - \frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\tilde{y}_i-\mu}{\sigma}\right)^2 - \frac{n_0}{2}\frac{\tilde{\sigma}^2}{\sigma^2}.$$

4. **M step in the EM algorithm.** Prove that maximizing $Q(\mu,\sigma|\mu_m,\sigma_m)$ in (μ,σ) is equivalent to maximizing the complete log-likelihood calculated from a sample $\tilde{y}_1,\ldots,\tilde{y}_n$ with

$$\tilde{y}_i = \begin{cases} y_i & \text{if } |y_i| > 1 \\ \mathrm{E}_{\mu_m,\sigma_m}(X \mid -1 < X < 1) & \text{otherwise.} \end{cases}$$

Deduce that

$$\mu_{m+1} = \frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i,\ \sigma^2_{m+1} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i-\mu_{m+1})^2.$$

Solution: There was a mistake in this question. It should be as follows:
Prove that maximizing $Q(\mu,\sigma|\mu_m,\sigma_m)$ in μ ...
and

$$\sigma^2_{m+1} = \frac{1}{n}\left(\sum_{i=1}^{n}(\tilde{y}_i-\mu_{m+1})^2 + n_0\tilde{\sigma}^2\right)$$

Solution:
It must be noted that $Q(\mu,\sigma|\mu_m,\sigma_m)$, as a function of μ, has the same expression as $l_c(\mu,\sigma)$. The second part is deduced from taken derivatives with respect to σ.

5. Write an R code implementing this EM algorithm.
6. File `results.txt` contains the data obtained by the scientific. Read them by

   ```
   y <- read.table("results.txt",col.names=FALSE)[,1]
   ```

   Use your EM algorithm for estimating (μ, σ) by maximum likelihood.

```
# EM algorithm

# \mbox{E}_{\mu,\sigma} (X \mid a<X<b) =
# \mu +  \frac{\phi(\frac{a-\mu}{\sigma})-\phi(\frac{b-\mu}{\sigma})}
#             {\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}\sigma,
E.trunc.normal <- function(mu=0,sigma=1,a=-1,b=1){
z.a<-(a-mu)/sigma
z.b<-(b-mu)/sigma
num <- dnorm(z.a)-dnorm(z.b)
den <- pnorm(z.b)-pnorm(z.a)
return(mu + sigma*num/den)
}

# \mbox{Var}_{\mu,\sigma} (X \mid a<X<b) =
# \sigma^2
#  \left[1+\frac{\frac{a-\mu}{\sigma}\phi(\frac{a-\mu}{\sigma})-
#               \frac{b-\mu}{\sigma}\phi(\frac{b-\mu}{\sigma})}
#              {\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}
#   -\left(\frac{\phi(\frac{a-\mu}{\sigma})-\phi(\frac{b-\mu}{\sigma})}
#              {\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}
#    \right)^2
# \right].
V.trunc.normal <- function(mu=0,sigma=1,a=-1,b=1){
z.a<-(a-mu)/sigma
z.b<-(b-mu)/sigma
num1 <- z.a*dnorm(z.a)-z.b*dnorm(z.b)
num2 <- dnorm(z.a)-dnorm(z.b)
den <- pnorm(z.b)-pnorm(z.a)
return(sigma*(1 + num1/den - (num2/den)^2))
}
```

```
##### Without correcting the mistake in the question
n.iter <- 10#00
I.0 <- which(y==0)
y.tilde <- y

mu.m <- mean(y)
sigma.m <- sd(y)

for (m in (1:n.iter)){
print(c(m-1,mu.m,sigma.m))
y.tilde[I.0] <- E.trunc.normal(mu=mu.m,sigma=sigma.m)
mu.m <- mean(y.tilde)
sigma.m <- sqrt(sum((y.tilde-mu.m)^2)/n)
}
print(c(m,mu.m,sigma.m))
#[1] 1000.0000000    0.2670108    1.9649811

##### With the right statement for the question
n.iter <- 10#00
I.0 <- which(y==0)
n0 <- length(I.0)
y.tilde <- y

mu.m <- mean(y)
sigma.m <- sd(y)

for (m in (1:n.iter)){
print(c(m-1,mu.m,sigma.m))
y.tilde[I.0] <- E.trunc.normal(mu=mu.m,sigma=sigma.m)
mu.m <- mean(y.tilde)
sigma.m <- sqrt(sum((y.tilde-mu.m)^2)/n +
                (n0/n)*V.trunc.normal(mu=mu.m,sigma=sigma.m))
}
print(c(m,mu.m,sigma.m))
#[1] 1000.000000    0.266878    1.980648
```