Computing Intensive Methods (MESIO UPC-UB) **2014-15**

Exam 2015. Part 2

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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ as the *complete data* and with $Y_1, \dots, Y_n$ as the *observed data*. Let $y_1, \dots, y_n$ 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 $|y_i| > 1$?
2. Write down the log-likelihood function for the complete data, and the log-likelihood for the observed data.
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

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

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 $\tau \in \mathbb{R}$,

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

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

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

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.