**Corso di Laurea Magistrale in Statistica e Data Science**
**Statistical Inference and Computational Methods**

**R-project 3**

# EM algorithm

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are all mutually independent random variables, where $Y_i \sim Poisson(\beta\tau_i)$ and $X \sim Poisson(\tau_i)$. Suppose that $Y_i$ models the incidence of a disease, where the underlying rate is a function of an overall effect $\beta$ and an additional factor $\tau_i$, which measures population density in area $i$. We do not see $\tau_i$ but get information on it through $X_i$.

The complete data likelihood is

$$\mathcal{L}(\beta, \tau_1, \ldots, \tau_n; \mathbf{x}^{obs}, \mathbf{y}) \propto f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \beta, \tau_1, \ldots, \tau_n) = \prod_{i=1}^{n} \frac{e^{-\beta\tau_i}\,(\beta\tau_i)^{y_i}}{y_i!} \frac{e^{-\tau_i}\,(\tau_i)^{x_i}}{x_i!}$$

The likelihood estimators which can be found by straightforward differentiation, are

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \qquad \text{and} \qquad \widehat{\tau_i} = \frac{x_i + y_i}{\widehat{\beta} + 1} \qquad i = 1, \ldots, n$$

Suppose that value $x_1$ was missing. The observed-data likelihood with $x_1$ missing is

$$\mathcal{L}_{obs}(\beta, \tau_1, \ldots, \tau_n; \mathbf{x}^{obs}, \mathbf{y}) = \sum_{x_1=0}^{\infty} f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \beta, \tau_1, \ldots, \tau_n)$$

with $\left(\mathbf{x}^{obs}, \mathbf{y}\right) = (y_1, (x_2, y_2), \ldots, (x_n, y_n))$. Differentiation leads to the MLE equantions:

$$\begin{aligned}
\widehat{\beta} &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \widehat{\tau}_i} \\
y_1 &= \widehat{\tau}_1 \widehat{\beta} \\
x_i + y_i &= \widehat{\tau}_i \left(\widehat{\beta} + 1\right) \qquad i = 2, \ldots, n
\end{aligned}$$

which we can solve with the EM algorithm using the expected complete data log likelihood:

$$\begin{aligned}
&Q(\boldsymbol{\theta} \mid \widetilde{\boldsymbol{\theta}}) = \\
&\quad \sum_{i=1}^{n} \left[-\beta\tau_i + y_i \left(\log(\beta) + \log(\tau_i)\right)\right] + \sum_{i=2}^{n} \left[-\tau_i + x_i \log(\tau_i)\right] + \sum_{x_1=0}^{\infty} \left[-\tau_1 + x_1 \log(\tau_1)\right] \frac{e^{-\widetilde{\tau}_1}\,(\widetilde{\tau}_1)^{x_1}}{x_1!} \\
&\quad - \left(\sum_{i=1}^{n} \log(y_i!) + \sum_{i=2}^{n} \log(x_i!) + \sum_{x_1=0}^{\infty} \log(x_1!) \frac{e^{-\widetilde{\tau}_1}\,(\widetilde{\tau}_1)^{x_1}}{x_1!}\right)
\end{aligned}$$

where $\boldsymbol{\theta} = (\beta, \boldsymbol{\tau}) = (\beta, \tau_1, \ldots, \tau_n)$, and

$$\arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}^{(j)}) = \begin{cases}
\widehat{\beta}^{j+1} = \dfrac{\sum_{i=1}^{n} y_i}{\widehat{\tau}_1^{(j)} + \sum_{i=2}^{n} x_i} \\[2ex]
\widehat{\tau}_1^{j+1} = \dfrac{\widehat{\tau}_1^{(j)} + y_1}{\widehat{\beta}^{(j+1)} + 1} \\[2ex]
\widehat{\tau}_i^{j+1} = \dfrac{x_i + y_i}{\widehat{\beta}^{(j+1)} + 1} \qquad i = 2, \ldots, n
\end{cases}$$

Consider the following data on $n = 18$ leukemia counts and the associated populations for a number of areas in New York State (Lange et al. 1994)

| Population | Number of cases | Population | Number of cases |
|---|---|---|---|
| ? | 3 | 948 | 0 |
| 3560 | 4 | 1172 | 1 |
| 3739 | 1 | 1047 | 3 |
| 2784 | 1 | 3138 | 5 |
| 2571 | 3 | 5485 | 4 |
| 2729 | 1 | 5554 | 6 |
| 3952 | 2 | 2943 | 2 |
| 993 | 0 | 4969 | 5 |
| 1908 | 2 | 4828 | 4 |

1. Develop an EM algorithm to fit the Poisson model to these data.

2. A direct solution of the original incomplete-data likelihood equations is possible and is given by

$$\begin{aligned}
\widehat{\beta} &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=2}^{n} x_i} \\
\widehat{\tau}_1 &= \frac{y_1}{\widehat{\beta}} \\
\widehat{\tau}_i &= \frac{x_i + y_i}{\widehat{\beta} + 1} \qquad i = 2, \ldots, n
\end{aligned}$$

Compare the direct solution with the EM solution.