An Introduction to the EM Algorithm and its Applications

Ruben H. Zamar

Department of Statistics

University of British Columbia

Introduction

- The EM algorithm is used to compute MLE estimates when data are missing.
- In the case of mixture models, missing data can take at least three forms:
  - *Missing labels*
  - *Unknown number of components*
  - *Missing data in the data vectors*



Goal of the EM Algorithm

Find the MLE when direct maximization

of the likelihood function is difficult.

One difficult maximization problem → Sequence of simpler maximization problems

Some Applications of the EM Algorithm

- Estimation in the presence of missing data
- Estimation of the parameters of a mixture distribution

Some Notation

$\mathbf{Y}$	$=$	Incomplete data
$\mathbf{X}$	$=$	Augmented data
$(\mathbf{Y}, \mathbf{X})$	$=$	Complete data
θ	$=$	Parameters to be estimated

The EM algorithm has two main steps:

(i) The Expectation Step (E-step)

(ii) The Maximization Step (M-step)



Expectation Step

$$l\left(\theta|\mathbf{y}\right) = \log\left(f\left(\mathbf{y}|\theta\right)\right)$$

Incomplete log-likelihood

$$l\left(\theta|\mathbf{y},\mathbf{x}\right) = \log\left(f\left(\mathbf{y},\mathbf{x}|\theta\right)\right)$$

Complete log-likelihood

$$\tilde{l}\left(\theta|\mathbf{y},\theta^{(k)}\right) = E_{\mathbf{X}|\mathbf{y},\theta^{(k)}}\left\{l\left(\theta|\mathbf{y},\mathbf{X}\right)\right\} =$$

$$\int\cdots\int \log f\left(\mathbf{y},\mathbf{x}|\theta\right) h\left(\mathbf{x}|\mathbf{y},\theta^{(k)}\right) \mathbf{dx}$$

Expected log-likelihood



The Conditional Density

$$h\left(\mathbf{x}|\mathbf{y},\theta^{(k)}\right) = \frac{f\left(\mathbf{y},\mathbf{x}|\theta^{(k)}\right)}{f\left(\mathbf{y}|\theta^{(k)}\right)}$$

Note that h is a fully specified density.

Maximization Step

$$\theta^{(k+1)} = \arg\max_{\theta} \tilde{l}\left(\theta|\mathbf{y}, \theta^{(k)}\right)$$

Each iteration (EM-step) increases the value of the incomplete-data log-likelihood:

$$l\left(\theta^{(k+1)}|\mathbf{y}\right) \geq l\left(\theta^{(k)}|\mathbf{y}\right)$$

Overview of the EM algorithm

Step 1: Obtain the **complete-data** log-likelihood function.

Step 2: Take expectation using the **conditional distribution** of the augmented data, given the incomplete data and the current values of the parameters. This is called the **E-step**.

Step 3: **Maximize** the resulting expected log-likelihood function. This is called the **M-step**.

Step 4: Repeat steps 2 and 3 until **convergence**.

A Simple Example: Mixture Distribution

Incomplete-Data: $Y_1, Y_2, ..., Y_n \quad iid$ with common density

$$f(y) = (1-p) f_0(y) + p f_1(y)$$

Incomplete-Data Log-Likelihood:

$$l(p|\mathbf{y}) = \sum_{i=1}^{n} \log\left[(1-p) f_0(y_i) + p f_1(y_i)\right]$$



Complete-Data Log-Likelihood

Augmented Data: $X_1, X_2, ..., X_n, \quad iid \quad Bin(1, p)$

Complete Data: $(Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n)$, independent, with common density

$$f(y, x) = [(1-p) f_0(y)]^{1-x} [p f_1(y)]^x$$

Complete-Data Log-Likelihood:

$$l(p|\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} (1-x_i) \log[(1-p) f_0(y_i)] + x_i \log[p f_1(y_i)]$$

A key observation (for the E-Step)

$$
\begin{aligned}
E\left[X_i|y_i,p^{(k)}\right] &= P\left[X_i=1|y_i,p^{(k)}\right] \\
&= \frac{f\left(y_i,1|p^{(k)}\right)}{f\left(y_i,0|p^{(k)}\right)+f\left(y_i,1|p^{(k)}\right)} \\
&= \frac{f_1\left(y_i\right)p^{(k)}}{f_0\left(y_i\right)\left(1-p^{(k)}\right)+f_1\left(y_i\right)p^{(k)}} \\
&= \tilde{p}_i
\end{aligned}
$$

E-Step

$$
\begin{aligned}
l\left(p|\mathbf{y},\mathbf{X}\right) &= \log\left(1-p\right)\sum_{i=1}^{n}\left(1-X_i\right)+\sum_{i=1}^{n}\left(1-X_i\right)\log\left(f_0\left(y_i\right)\right) \\
&\quad +\log\left(p\right)\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}X_i\log\left(f_1\left(y_i\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
\tilde{l}\left(p|\mathbf{y}\right) &= \log\left(1-p\right)\sum_{i=1}^{n}\left(1-\tilde{p}_i\right)+\sum_{i=1}^{n}\left(1-\tilde{p}_i\right)\log\left(f_0\left(y_i\right)\right) \\
&\quad +\log\left(p\right)\sum_{i=1}^{n}\tilde{p}_i+\sum_{i=1}^{n}\tilde{p}_i\log\left(f_1\left(y_i\right)\right)
\end{aligned}
$$

The M-Step

$$\frac{\partial}{\partial p}\tilde{l}\,(p|\mathbf{y}) \;=\; -\frac{\sum_{i=1}^{n}(1-\tilde{p}_i)}{1-p}+\frac{\sum_{i=1}^{n}\tilde{p}_i}{p}=0$$

$$p^{(k+1)}=\frac{\sum\limits_{i=1}^{n}\tilde{p}_i}{n}=\frac{1}{n}\sum_{i=1}^{n}\frac{f_1(y_i)p^{(k)}}{f_0(y_i)\left(1-p^{(k)}\right)+f_1(y_i)p^{(k)}}$$

Extensions

- Mixture distribution has more than two components
- Distributions include unknown parameters
- Data are vector valued



More Than Two Components

In this case the complete-data log-likelihood

$$l\left(\mathbf{p}|y_1,...,y_1,\mathbf{x}_1,...,\mathbf{x}_n\right) = \log\left\{\prod_{i=1}^{n}\prod_{j=1}^{m}\left[p_j f_j\left(y_i\right)\right]^{x_{ij}}\right\}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\left[\log\left(p_j\right)+\log\left(f_j\left(y_i\right)\right)\right]$$

where $\mathbf{x}_i=(x_{i1},x_{i2},...,x_{im})$ is $Multinomial\left(1,p_1,p_2,...,p_m\right)$

The conditional probability that the i^{th} observation came from the j^{th} mixture component, given the observation value and the current parameter estimates, are

$$\begin{aligned}\tilde{p}_{ij} &= E\left(X_{ij}|y_i, \mathbf{p}^{(k)}\right) \\ &= P\left(X_{ij}=1|y_i, \mathbf{p}^{(k)}\right) = \frac{p_j^{(k)} f_j\left(y_i\right)}{\sum_{\alpha=1}^{m} p_\alpha^{(k)} f_\alpha\left(y_i\right)}\end{aligned}$$

The updated probabilities for the j^{th} mixture component is

$$\boxed{p_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{p}_{ij}, \;\; j=1,...,m}$$

Distributions Include Unknown Parameters

- Densities include unknown parameters. For example:

$$f_0(y) = N(\mu_0, \sigma_0^2)$$

$$f_1(y) = N(\mu_1, \sigma_1^2)$$

In this case

$$\theta = (p, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$$

"Independent Pieces"

$$\begin{aligned}
\tilde{l}\left(\theta|\mathbf{y}\right) &= \sum_{i=1}^{n}\left(1-\tilde{p}_i\right)\log\left(f_0\left(y_i\right)\right) \\
&\quad +\sum_{i=1}^{n}\tilde{p}_i\log\left(f_1\left(y_i\right)\right) \\
&\quad +\log\left(1-p\right)\sum_{i=1}^{n}\left(1-\tilde{p}_i\right)+\log\left(p\right)\sum_{i=1}^{n}\tilde{p}_i \\
&= \tilde{l}_0\left(\mu_0,\sigma_0^2|\mathbf{y}\right)+\tilde{l}_1\left(\mu_1,\sigma_1^2|\mathbf{y}\right)+\tilde{l}_2\left(p|\mathbf{y}\right)
\end{aligned}$$

Initialization: $p^{(0)}$, $\hat{\mu}_0^{(0)}$, $\hat{\sigma}_0^{(0)}$, $\hat{\mu}_1^{(0)}$, $\hat{\sigma}_1^{(0)}$

$$\tilde{p}_i = \frac{p^{(k)} f\left(y_i ; \hat{\mu}_1^{(k)}, \hat{\sigma}_1^{(k)}\right)}{p^{(k)} f\left(y_i ; \hat{\mu}_1^{(k)}, \hat{\sigma}_1^{(k)}\right) + \left(1 - p^{(k)}\right) f\left(y_i ; \hat{\mu}_0^{(k)}, \hat{\sigma}_0^{(k)}\right)}$$

$$p^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \tilde{p}_i$$

The Iteration Steps

$$\hat{\mu}_0^{(k+1)} = \frac{\sum_{i=1}^{n}(1-\tilde{p}_i)\,y_i}{\sum_{i=1}^{n}(1-\tilde{p}_i)}$$

$$\hat{\mu}_1^{(k+1)} = \frac{\sum_{i=1}^{n}\tilde{p}_i y_i}{\sum_{i=1}^{n}\tilde{p}_i}$$

$$\left(\hat{\sigma}_0^2\right)^{(k+1)} = \frac{\sum_{i=1}^{n}(1-\tilde{p}_i)\left(y_i-\hat{\mu}_0^{(k+1)}\right)^2}{\sum_{i=1}^{n}(1-\tilde{p}_i)}$$

$$\left(\hat{\sigma}_1^2\right)^{(k+1)} = \frac{\sum_{i=1}^{n}\tilde{p}_i\left(y_i-\hat{\mu}_1^{(k+1)}\right)^2}{\sum_{i=1}^{n}\tilde{p}_i}$$



Example 2

Estimating the overall infection rate.

We observe $Y_1, Y_2, \ldots, Y_n$ and $Z_1, Z_2, \ldots, Z_n$ (mutually independent)

$Y_i =$	Number of infected people in the ith region	Poisson$(\beta\tau_i)$
$Z_i =$	Measure of population density in the ith region	Poisson(τ_i)
$\tau_i =$	Factor influencing population density in the ith region	Unknown
$\beta =$	Overall infection rate	constants

Complete-Data Likelihood

The complete-data likelihood is:

$$f\left(\mathbf{y},\mathbf{z}|\beta,\tau_1,\ldots,\tau_n\right)=\prod_{i=1}^{n}\frac{e^{-\beta\tau_i}\left(\beta\tau_i\right)^{y_i}}{y_i!}\times\frac{e^{-\tau_i}\left(\tau_i\right)^{z_i}}{z_i!}$$

The MLE estimates are:

$$\hat{\beta} = \frac{\bar{y}}{\bar{z}}$$

$$\hat{\tau}_i = \frac{z_i+y_i}{\hat{\beta}+1},\quad i=1,...,n$$

$$l_n(\tau_1, \tau_2, \dots, \tau_n, \beta) = -\beta \sum_{i=1}^{n} \tau_i + \log(\beta) \sum y_i + \sum y_i \log(\tau_i)$$

$$- \sum_{i=1}^{n} \tau_i + \sum_{i=1}^{n} z_i \log(\tau_i)$$

$$\frac{\partial l_n}{\partial \beta} = 0 \Rightarrow -\sum_{i=1}^{n} \tau_i + \frac{\sum y_i}{\beta} = 0 \Rightarrow \boxed{\hat{\beta} = \frac{\sum y_i}{\sum \hat{\tau}_i}}$$

$$\frac{\partial l_n}{\partial \tau_i} = 0 \Rightarrow -\beta + \frac{y_i}{\tau_i} - 1 + \frac{z_i}{\tau_i} = 0 \quad , \quad i = 1, \dots, n$$

$$\Rightarrow \frac{y_i + z_i}{\tau_i} = 1 + \hat{\beta} \Rightarrow \boxed{\hat{\tau}_i = \frac{y_i + z_i}{1 + \hat{\beta}}}$$

$$\Rightarrow \hat{\beta} = \frac{\sum y_i}{\sum y_i + \sum z_i} (1 + \hat{\beta})$$

$$\Rightarrow \quad \frac{1}{\hat{\beta}} + 1 = 1 + \frac{\sum z_i}{\sum y_i}$$

$$\Rightarrow \quad \boxed{\hat{\beta} = \frac{\bar{y}}{\bar{z}}}$$

and

$$\boxed{\hat{\tau}_i = \frac{y_i + z_i}{\hat{\beta} + 1}}$$

Incomplete-Data Likelihood

Suppose now that z_1 is missing

The incomplete-data likelihood is:

$$f\left(\mathbf{y},\mathbf{z}|\beta,\tau_1,\ldots,\tau_n\right)=\prod_{i=1}^{n}\frac{e^{-\beta\tau_i}\left(\beta\tau_i\right)^{y_i}}{y_i!}\prod_{i=2}^{n}\frac{e^{-\tau_i}\left(\tau_i\right)^{z_i}}{z_i!}$$

Incomplete-Data MLE Equations

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \hat{\tau}_i}$$

$$y_1 = \hat{\tau}_1 \hat{\beta}$$

$$z_i + y_i = \hat{\tau}_i \left(\hat{\beta} + 1 \right) \quad i = 2, ..., n$$

This system of equations is not easy to solve

Notations

Incomplete Data: $\mathbf{y} = (y_1, ..., y_n)\,,\ \mathbf{z}_{(-1)} = (z_2, ..., z_n)$

Augmented Data: z_1

Complete Data: $\mathbf{y} = (\mathbf{y_1}, ..., \mathbf{y_n})\,,\ \mathbf{z} = (z_1, ..., z_n)$

Complete data log-likelihood

$$l\left(\beta,\tau_1,\ldots,\tau_n|\mathbf{y},\mathbf{z}\right)=\log\left[f\left(\mathbf{y},\mathbf{z}|\beta,\tau_1,\ldots,\tau_n\right)\right]$$

$$=\quad C-\left(1+\beta\right)\sum_{i=1}^{n}\tau_i+\sum_{i=1}^{n}y_i\log\left(\beta\tau_i\right)+\sum_{i=1}^{n}z_i\log\left(\tau_i\right)$$

$$=\quad C-\sum_{i=1}^{n}\left[\left(1+\beta\right)\tau_i+y_i\log\left(\beta\tau_i\right)\right]+\sum_{i=2}^{n}z_i\log\left(\tau_i\right)+z_1\log\left(\tau_1\right)$$

The E-Step

$$\tilde{l}\left(\beta,\tau|\mathbf{y},\mathbf{z}_{(-1)}\right) = E_{Z_1|\mathbf{y},\mathbf{z}_{(-1)},\hat{\tau}^{(k)},\hat{\beta}^{(k)}}\left\{l\left(\beta,\tau_1,\ldots,\tau_n|\mathbf{y},\mathbf{z}\right)\right\}$$

$$= C_1 - \sum_{i=1}^{n}\left[\left(1+\beta\right)\tau_i + y_i\log\left(\beta\tau_i\right)\right] + \sum_{i=2}^{n} z_i\log\left(\tau_i\right) + \hat{\tau}_1^{(k)}\log\left(\tau_1\right)$$

$$= C_2 + l\left(\beta,\tau_1,\ldots,\tau_n|\mathbf{y},\hat{\tau}_1^{(k)},z_2,...,z_n\right)$$

The M-Step

$$\hat{\beta}^{(k+1)} = \frac{\sum_{i=1}^{n} y_i}{\hat{\tau}_1^{(k)} + \sum_{i=2}^{n} z_i}$$

$$\hat{\tau}_1^{(k+1)} = \frac{\hat{\tau}_1^{(k)} + y_1}{\hat{\beta}^{(k+1)} + 1}$$

$$\hat{\tau}_i^{(k+1)} = \frac{z_i + y_i}{\hat{\beta}^{(k+1)} + 1}, \qquad i = 2, ..., n$$