Expectation-Maximization Algorithm

Dimitris Fouskakis

Department of Mathematics
School of Applied Mathematical and Physical Sciences
National Technical University of Athens

fouskakis@math.ntua.gr

Spring Semester
Example: Gamma distribution

\[ X_1, \ldots X_n \sim \Gamma(\alpha, \beta), \text{ i.e. } f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \]

\[ L(\alpha, \beta) = \frac{\beta^\alpha n}{\Gamma(\alpha)^n} \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\beta \sum_{i=1}^{n} x_i} \]

\[ l(\alpha, \beta) = n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \beta \sum_{i=1}^{n} x_i, \]

i.e. \((\sum_{i=1}^{n} \log x_i, \sum_{i=1}^{n} x_i)\) is a sufficient statistic for \((\alpha, \beta)\).

- \[ \frac{\partial l(\alpha, \beta)}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\beta} = \frac{\alpha}{\bar{x}} \]

\[ l(\alpha, \hat{\beta}) = n\alpha \log \frac{\alpha}{\bar{x}} - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{\alpha}{\bar{x}} \sum_{i=1}^{n} x_i \]

- \[ \frac{\partial l(\alpha, \hat{\beta})}{\partial \alpha} = -n \log \bar{x} + n \log \alpha + n - n[\log \Gamma(\alpha)]' + \sum_{i=1}^{n} \log x_i - n, \]

since \(\sum_{i=1}^{n} x_i / \bar{x} = n\)

Newton-Raphson (1 dimension):

\[ \alpha^{\text{new}} = \alpha^{\text{old}} - \frac{\sum_{i=1}^{n} \log x_i - n \log \bar{x} + n \log \alpha^{\text{old}} - n \Psi(\alpha^{\text{old}})}{n/\alpha^{\text{old}} - n \Psi'(\alpha^{\text{old}})} \]

where \(\Psi(\alpha) := [\log \Gamma(\alpha)]':\) digamma function and \(\Psi_3(\alpha) := \Psi'(\alpha):\) trigamma function.
Alternatively,

\[ l(\alpha, \beta) = n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \beta \sum_{i=1}^{n} x_i. \]

\[ \frac{\partial l(\alpha, \beta)}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i = 0 \]

\[ \frac{\partial l(\alpha, \beta)}{\partial \alpha} = n \log \beta - n\Psi(\alpha) + \sum_{i=1}^{n} \log x_i = 0 \]

Newton-Raphson (2 dimensions):

\[ \mathbf{A} = \begin{bmatrix} -n\Psi_3(\alpha) & \frac{n}{\beta} \\ \frac{n}{\beta} & -\frac{n\alpha}{\beta^2} \end{bmatrix} \text{ (Hessian matrix)} \]

\[ \begin{bmatrix} \alpha_{\text{new}} \\ \beta_{\text{new}} \end{bmatrix} = \begin{bmatrix} \alpha_{\text{old}} \\ \beta_{\text{old}} \end{bmatrix} - \mathbf{A}^{-1} \left[ \sum_{i=1}^{n} \log x_i + n \log \beta_{\text{old}} - n\Psi(\alpha_{\text{old}}) \right] \]
Example: Gamma distribution (cont’d)

\[ X \sim \Gamma(\alpha, \beta), \quad \mathbb{E}[X] = \frac{\alpha}{\beta}, \quad \mathbb{E}[\log X] =? \]

We have

\[
\frac{\Gamma(\alpha)}{\beta^\alpha} = \int_0^\infty x^{\alpha-1} e^{-\beta x} \, dx
\]

derivative w.r.t. \( \alpha \)

\[
\Rightarrow \quad \frac{\Gamma'(\alpha) \beta^\alpha - \beta^\alpha \log \beta \Gamma(\alpha)}{(\beta^\alpha)^2} = \int_0^\infty \log x \cdot x^{\alpha-1} e^{-\beta x} \, dx
\]

\[
\Rightarrow \quad \frac{\Gamma'(\alpha) \beta^\alpha - \beta^\alpha \log \beta}{\beta^\alpha} = \int_0^\infty \log x \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \, dx
\]

\[
\Rightarrow \quad \mathbb{E}[\log X] = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log \beta ,
\]

where \( \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = [\log \Gamma(\alpha)]' \): digamma function.
Missing data examples

- Some variables for certain observations might have not been observed/measured.
- Censored observations, e.g. survival analysis
  The value of a r.v. representing the survival time is larger than a certain value but we do not know its exact value.
- Truncated observations (e.g. truncated Poisson)
  Some specific values cannot be observed and thus appear with zero frequency.
- Grouped data
  Questionnaires → grouping of continuous r.v.’s
  e.g. age, income, etc. → confidential data
Missing data examples (cont’d)

- Mixtures, e.g. mixed effects models
  
e.g.
  \[
  \begin{align*}
  X & \sim P(\lambda) \\
  \lambda & \sim \Gamma(a, b)
  \end{align*}
  \]
  → Negative Binomial

  where \( \lambda \) is a r.v. that we have not observed.

- Convolutions: \( X = Y + Z \),
  where \( X \) is observed while \( Y \) and \( Z \) are not observed.

- Random sums: \( Y = X_1 + \ldots + X_N \),
  where \( N \) is a r.v. (e.g. \( N \sim P(\lambda) \)), \( Y \) is observed, \( X_i \) and \( N \) are not observed.
  e.g. actuarial science → amount of compensation paid by an insurance company

- Hidden Markov Models
  
  Time series → the value at each time point depends on an unobservable state.
Expectation–maximization (EM) algorithm

- Dempster at al. 1977
- Application: datasets with missing values (see previous slides)

IDEA:

\[ Y = (X, Z), \]

where \( Y \): complete data, \( X \): observed data and \( Z \): latent data

Aim: \( \max_\theta L(\theta; X) \), i.e. the likelihood of the parameter \( \theta \), given the observed data \( X \). This maximization has difficulties. We augment the data, to make the problem simpler!

E-step: Estimate \( Z \) from \( X \) and current \( \theta \)

M-step: \( \max_\theta L(\theta; X, Z) \) (using current \( Z \))
EM algorithm in detail

We begin with $\theta^{(0)}$. In iteration $r$

$$Q(\theta, \theta^{(r)}) = \int \log L(\theta; X, Z)f(Z|\theta^{(r)}, X)dZ \equiv \mathbb{E}_{Z|X, \theta^{(r)}} [\log L(\theta; X, Z)]$$

**E-step:** Compute $Q(\theta, \theta^{(r)})$

$\rightarrow$ expected value of the log likelihood of $\theta$ for the complete data w.r.t. the conditional distribution of $Z|X, \theta^{(r)}$, i.e. the log likelihood of $\theta$ for the complete data $Y$ with the conditional expectations of $Z$ (given the actual data $X$ and the current value $\theta^{(r)}$ of the parameter) in the place of $Z$

**M-step:** $\max_{\theta} Q(\theta, \theta^{(r)})$
EM - Termination criteria

1. \[
\left| \frac{l(r+1) - l(r)}{l(r+1)} \right| \leq \text{tolerance},
\]
where \(l(r)\): log likelihood of the complete data after iteration \(r\).

2. \[
\theta = (\theta_1, \ldots, \theta_p)
\]
\[
\max_j \left( \left| \theta_j^{(r+1)} - \theta_j^{(r)} \right| \right) \leq \text{tolerance} \quad (j = 1, 2, \ldots, p)
\]
or
\[
\sum_{j=1}^{p} \left( \theta_j^{(r+1)} - \theta_j^{(r)} \right)^2 \leq \text{tolerance}
\]
EM theory

\[ Y = (X, Z) \equiv (Y_{\text{obs}}, Y_{\text{mis}}) \]

\[
f(y|\theta) = f(y_{\text{obs}}|\theta)f(y_{\text{mis}}|y_{\text{obs}}, \theta) \quad \log \Rightarrow \]

\[
l(\theta; y) = l(\theta; y_{\text{obs}}) + \log f(y_{\text{mis}}|y_{\text{obs}}, \theta) \quad \Rightarrow \]

\[
l(\theta; y_{\text{obs}}) = l(\theta; y) - \log f(y_{\text{mis}}|y_{\text{obs}}, \theta) \quad (*) \]

We would like to estimate \( \theta \) by maximizing \( l(\theta; y_{\text{obs}}) \).

The expected value of (*) w.r.t. \( Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r)} \) is:

\[
\mathbb{E}_{Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r)}}[l(\theta; y_{\text{obs}})] = \int l(\theta; y)f(y_{\text{mis}}|y_{\text{obs}}, \theta^{(r)})dy_{\text{mis}} -
\]

\[
- \int \log f(y_{\text{mis}}|y_{\text{obs}}, \theta)f(y_{\text{mis}}|y_{\text{obs}}, \theta^{(r)})dy_{\text{mis}}
\]

We denote by \( Q(\theta, \theta^{(r)}) \) the first term of the right-hand side and by \( H(\theta, \theta^{(r)}) \) the second term, while the expectation on the left-hand side is equal to \( l(\theta; y_{\text{obs}}) \) (constant w.r.t. \( Y_{\text{mis}} \)).

Thus, \( l(\theta^{(r+1)}; y_{\text{obs}}) - l(\theta^{(r)}; y_{\text{obs}}) = \left[ Q(\theta^{(r+1)}, \theta^{(r)}) - Q(\theta^{(r)}, \theta^{(r)}) \right] -
\]

\[
- \left[ H(\theta^{(r+1)}, \theta^{(r)}) - H(\theta^{(r)}, \theta^{(r)}) \right]
\]
EM theory (cont’d)

We need to show that the above is \( \geq 0 \) (thus the log likelihood of \( \theta \) for the observed data is increased in two consecutive iterations). However, in the M-step we maximize \( Q \), so the first term on the right-hand side is \( \geq 0 \).

It suffices thus to show that: \( H(\theta^{(r+1)}, \theta^{(r)}) - H(\theta^{(r)}, \theta^{(r)}) \leq 0 \).

But, \( H(\theta^{(r+1)}, \theta^{(r)}) - H(\theta^{(r)}, \theta^{(r)}) = \)

\[
= \mathbb{E}_{Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r)}} \left[ \log f \left( Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r+1)} \right) \right] - \\
\mathbb{E}_{Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r)}} \left[ \log f \left( Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r)} \right) \right] = \\
= \mathbb{E}_{Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r)}} \left[ \log \frac{f \left( Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r+1)} \right)}{f \left( Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r)} \right)} \right]
\]

Jensen \( \leq \) log concave \( \log \mathbb{E}_{Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r)}} \left[ \frac{f \left( Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r+1)} \right)}{f \left( Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(r)} \right)} \right] = 0 \) \( \leftrightarrow \) \( = 1 \).
$n = 197$ animals divided in 4 categories based on a theoretical model about the genetic linkage. The data for the 4 categories are:

$$\mathbf{x} = (x_1, x_2, x_3, x_4) = (125, 18, 20, 34)$$

with theoretical cell probabilities

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4) = \left( \frac{1}{2} + \frac{\theta}{4}, \frac{1 - \theta}{4}, \frac{1 - \theta}{4}, \frac{\theta}{4} \right)$$

MLE for $\pi$? $\rightarrow$ MLE for $\theta$?

The theoretical model is a polynomial distribution with probabilities $\pi$, thus the likelihood for the observations $\mathbf{x}$ is:

$$\propto \left( \frac{1}{2} + \frac{\theta}{4} \right)^{x_1} \left( \frac{1 - \theta}{4} \right)^{x_2} \left( \frac{1 - \theta}{4} \right)^{x_3} \left( \frac{\theta}{4} \right)^{x_4}$$

$$\propto (2 + \theta)^{x_1} (1 - \theta)^{x_2 + x_3} \theta^{x_4}$$

and its logarithm

$$\propto x_1 \log(2 + \theta) + (x_2 + x_3) \log(1 - \theta) + x_4 \log \theta$$

(maximization $\rightarrow$ 2nd degree polynomial with solutions $0.62 \checkmark$ and $-0.55 \times$)
EM - Example (cont’d)

\[ y = (y_0, y_1, y_2, y_3, y_4), \ y_i = x_i \ (i = 2, 3, 4) \text{ and } y_0 + y_1 = x_1 \]

\[ Y \sim \text{Mult} \left( \frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4} \right) \]

The likelihood for the complete data is:

\[ L(\theta; y) \propto (1 - \theta)^{y_2 + y_3} \theta^{y_4 + y_1} \]

\[ \log L(\theta; y) \equiv l(\theta; y) \propto (y_2 + y_3) \log(1 - \theta) + (y_1 + y_4) \log \theta \]

\[ \frac{\partial l(\theta; y)}{\partial \theta} = \frac{y_1 + y_4}{\theta} - \frac{y_2 + y_3}{1 - \theta} = 0 \]

\[ \rightarrow \hat{\theta} = \frac{y_1 + y_4}{y_1 + y_2 + y_3 + y_4} \quad (y_1: \text{unknown}) \]

Note that \( Y_1|\theta, X \sim \text{Bin} \left( 125, \frac{\theta/4}{\theta/4+1/2} = \frac{\theta}{\theta+2} \right) \)

Thus, \( E \)-step

\[ Q(\theta, \theta^{(r)}) = \mathbb{E}_{Y_1|\theta^{(r)}, X} \left[ \log L(\theta; Y) \right] = \]

constant + \( \mathbb{E}_{Y_1|\theta^{(r)}, X} \left[ (y_2 + y_3) \log(1 - \theta) + (Y_1 + y_4) \log \theta \right] = \)

constant + \( (y_2 + y_3) \log(1 - \theta) + (\mathbb{E}[Y_1] + y_4) \log \theta, \ \mathbb{E}[Y_1] = 125\theta/(\theta+2) \)
EM - Example (cont’d)

M-step

\[ \theta^{(r+1)} = \frac{\mathbb{E}[Y_1] + y_4}{\mathbb{E}[Y_1] + y_2 + y_3 + y_4} = \frac{125 \theta^{(r)}}{\theta^{(r)} + 2} + y_4 \]

Application

\[ \theta^{(0)} = 0.4 \rightarrow \]

\[ (0.4, 0.5906643, 0.6218892, 0.6216642, 0.6267342, 0.6268099, 0.626820, 0.6268213, 0.6268215) \]

\[ |\theta^{(r+1)} - \theta^{(r)}| \leq 10^{-6} \]
EM variants

1. Stochastic EM (SEM)
   In the E-step instead of computing the expected value, simply draw a value from the conditional distribution of the missing data $Z|X, \theta^{(r)}$ (using simulation or MCMC)
   (-) The likelihood does not increase at every step but behaves well in general.
   (-) Since the likelihood does not increase at every step, it might skip the local maximum.

2. Monte Carlo EM (MCEM)
   In the E-step, it estimates the expected value through Monte Carlo integration. That is it draws several values (e.g. $M$) from the conditional distribution of the missing data $Z|X, \theta^{(r)}$ and estimates the expected value from the sample mean.
   To increase the likelihood at every step choose large $M$.
   To avoid local minima, begin with a small $M$ and increase it gradually.

3. Generalized EM (GEM)
   When the maximization at the M-step is hard just compute a value which increases the likelihood.