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

$X_1, \dots X_n \sim \Gamma(\alpha, \beta)$, 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. $\left(\sum_{i=1}^n \log x_i, \sum_{i=1}^n x_i\right)$ is a sufficient statistic for (α, β).

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

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

Example: Gamma distribution (cont'd)

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):

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

$$\rightarrow \begin{bmatrix} \alpha^{\text{new}} \\ \beta^{\text{new}} \end{bmatrix} = \begin{bmatrix} \alpha^{\text{old}} \\ \beta^{\text{old}} \end{bmatrix} - \mathbf{A}^{-1} \begin{bmatrix} \sum_{i=1}^{n} \log x_i + n \log \beta^{\text{old}} - n\Psi(\alpha^{\text{old}}) \\ \frac{n\alpha^{\text{old}}}{\beta^{\text{old}}} - \sum_{i=1}^{n} x_i \end{bmatrix}$$

Example: Gamma distribution (cont'd)

$X \sim \Gamma(\alpha, \beta)$, $\mathbb{E}[X] = \alpha/\beta$, $\mathbb{E}[\log X] =$? We have

$$\frac{\Gamma(\alpha)}{\beta^\alpha} = \int_0^\infty x^{\alpha-1} e^{-\beta x} \mathrm{d}x$$

$$\overset{\text{derivative w.r.t. } \alpha}{\Rightarrow} \quad \frac{\Gamma'(\alpha)\beta^\alpha - \beta^\alpha \log \beta \Gamma(\alpha)}{(\beta^\alpha)^2} = \int_0^\infty \log x \; x^{\alpha-1} e^{-\beta x} \mathrm{d}x$$

$$\Rightarrow \quad \frac{\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}\beta^\alpha - \beta^\alpha \log \beta}{\beta^\alpha} = \int_0^\infty \log x \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathrm{d}x$$

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

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 $\rightarrow$ grouping of continuous r.v.'s
  e.g. age, income, etc. $\rightarrow$ confidential data

Missing data examples (cont'd)

- Mixtures, e.g. mixed effects models
  e.g.
  $$\left.\begin{array}{l} X \sim P(\lambda) \\ \lambda \sim \Gamma(a, b) \end{array}\right\} \rightarrow \text{Negative Binomial}$$
  where λ 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:

$$\mathbf{Y} = (\mathbf{X}, \mathbf{Z}),$$

where $\mathbf{Y}$: complete data, $\mathbf{X}$: observed data and $\mathbf{Z}$: latent data

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

E-step: Estimate $\mathbf{Z}$ from $\mathbf{X}$ and current $\boldsymbol{\theta}$

M-step: $\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})$ (using current $\mathbf{Z}$)

EM algorithm in detail

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

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}) = \int_{\mathbf{Z}} \log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) f(\mathbf{Z} | \boldsymbol{\theta}^{(r)}, \mathbf{X}) \mathrm{d}\mathbf{Z} \equiv \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(r)}} [\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})]$$

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

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

M-step: $\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)})$

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.

$$\boldsymbol{\theta} = (\theta_1, \ldots \theta_p)$$

$$\max_j \left(\left| \theta_j^{(r+1)} - \theta_j^{(r)} \right| \right) \leq \text{tolerance} \quad (j = 1, 2, \ldots p)$$

or

$$\sum_{j=1}^{p} \left(\theta_j^{(r+1)} - \theta_j^{(r)} \right)^2 \leq \text{tolerance}$$

EM theory

$\mathbf{Y} = (\mathbf{X}, \mathbf{Z}) \equiv (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\theta}) &= f(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}) f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta}) \overset{\log}{\Rightarrow} \\ l(\boldsymbol{\theta}; \mathbf{y}) &= l(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}}) + \log f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta}) \Rightarrow \\ l(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}}) &= l(\boldsymbol{\theta}; \mathbf{y}) - \log f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta}) \quad (*) \end{aligned}$$

We would like to estimate $\boldsymbol{\theta}$ by maximizing $l(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}})$.

The expected value of (*) w.r.t. $\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}^{(r)}$ is:

$$\mathbb{E}_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}^{(r)}}[l(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}})] = \int l(\boldsymbol{\theta}; \mathbf{y}) f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(r)}) \mathrm{d}\mathbf{y}_{\text{mis}} -$$

$$- \int \log f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta}) f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(r)}) \mathrm{d}\mathbf{y}_{\text{mis}}$$

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

Thus, $l(\boldsymbol{\theta}^{(r+1)}; \mathbf{y}_{\text{obs}}) - l(\boldsymbol{\theta}^{(r)}; \mathbf{y}_{\text{obs}}) = \left[Q(\boldsymbol{\theta}^{(r+1)}, \boldsymbol{\theta}^{(r)}) - Q(\boldsymbol{\theta}^{(r)}, \boldsymbol{\theta}^{(r)})\right] -$

$- \left[H(\boldsymbol{\theta}^{(r+1)}, \boldsymbol{\theta}^{(r)}) - H(\boldsymbol{\theta}^{(r)}, \boldsymbol{\theta}^{(r)})\right]$

EM theory (cont'd)

We need to show that the above is ≥ 0 (thus the log likelihood of $\boldsymbol{\theta}$ 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 ≥ 0.

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

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

$$
\begin{aligned}
& \mathbb{E}_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r)}}\left[\log f\left(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r+1)}\right)\right] - \\
- \quad & \mathbb{E}_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r)}}\left[\log f\left(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r)}\right)\right] = \\
= \quad & \mathbb{E}_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r)}}\left[\log \frac{f\left(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r+1)}\right)}{f\left(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r)}\right)}\right] \\
\underset{\log\ \text{concave}}{\overset{\text{Jensen}}{\leq}} \quad & \log \mathbb{E}_{\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r)}}\underbrace{\left[\frac{f\left(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r+1)}\right)}{f\left(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(r)}\right)}\right]}_{\hookrightarrow\ =1} = 0
\end{aligned}
$$

EM - Example

$n = 197$ 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

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

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

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

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 ✓ and -0.55 ×)

EM - Example (cont'd)

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

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

The likelihood for the complete data is:

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

$$\begin{aligned}
\log L(\theta; \mathbf{y}) \equiv l(\theta; \mathbf{y}) &\propto (y_2 + y_3)\log(1-\theta) + (y_1 + y_4)\log\theta \\
\frac{\partial l(\theta; \mathbf{y})}{\partial\theta} &= \frac{y_1 + y_4}{\theta} - \frac{y_2 + y_3}{1-\theta} = 0 \\
&\rightarrow \hat{\theta} = \frac{y_1 + y_4}{y_1 + y_2 + y_3 + y_4} \quad (y_1 : \text{unknown})
\end{aligned}$$

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

Thus, E-step

$$\begin{aligned}
Q(\theta, \theta^{(r)}) &= \mathbb{E}_{Y_1|\theta^{(r)},\mathbf{X}}[\log L(\theta; \mathbf{Y})] = \\
\text{constant} &+ \mathbb{E}_{Y_1|\theta^{(r)},\mathbf{X}}[(y_2 + y_3)\log(1-\theta) + (Y_1 + y_4)\log\theta] = \\
\text{constant} &+ (y_2 + y_3)\log(1-\theta) + (\mathbb{E}[Y_1] + y_4)\log\theta, \ \mathbb{E}[Y_1] = 125\theta^{(r)}/(\theta^{(r)} + 2)
\end{aligned}$$

EM - Example (cont'd)

M-step

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

Application

$\theta^{(0)} = 0.4 \rightarrow$

$(0.4, 0.5906643, 0.6218892, 0.6216642, 0.6267342,$
$0.6268099, 0.626820, 0.6268213, \underline{\underline{0.6268215}})$

$\left|\theta^{(r+1)} - \theta^{(r)}\right| \leq 10^{-6}$

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 $\mathbf{Z}|\mathbf{X}, \theta^{(r)}$ (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 $\mathbf{Z}|\mathbf{X}, \theta^{(r)}$ and estimates the expected value from the sample mean.
   To increase the likelihood at every step choose large M.
   To avoid local maxima, 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.