Maximum Likelihood Estimators

JACKSON BARTH, CHENYU (DEVON) YANG, MING ZHANG

NOVEMBER 19, 2019

MLE - Introduction

- Motivation: Determining the distribution of observation
- Best all-purpose approach for statistical analysis
- Consistent, asymptotically unbiased and efficient (Under mild conditions)
- Focused on computational performance
- Several different methods can be used

Contents

Traditional MLE

What is MLE?

- Say we have $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ IID observations from a distribution with an unknown parameter θ (scalar or vector)
- f(**y**|θ) , where θ* is the true value

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{Y}_i \mid \boldsymbol{\theta});$$

- To simplify, take the log

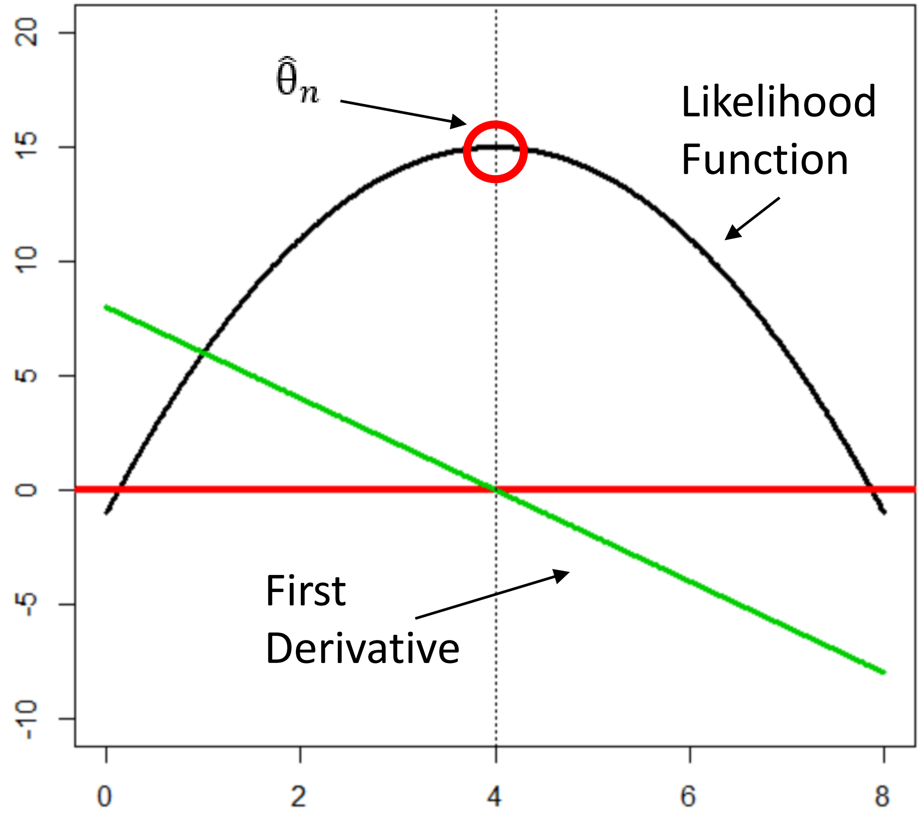
$$\ell_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{Y}_i \mid \boldsymbol{\theta})$$

What is MLE?

- Take the first derivative of the log likelihood function
- Solve for the roots

  Newton's Method, Scoring

- We refer to this as the Traditional MLE method



Example 1 – Analytical Solution

- We have observations 0,0,1,1,1,2,2,2,3,3 from a distribution with the below pmf

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

- For what value of θ is the below likelihood function maximized?

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2$$

*Example from http://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/MLE.pdf

What is MLE?

- Let $\mathrm{E}_{\theta}[\mathrm{g}(\boldsymbol{Y})]$ denote the expectation of a function of the variable with respect to θ

$$E_{\theta}[\ell_n(\boldsymbol{\theta})] = n E_{\theta}[\log f(\mathbf{Y} \mid \boldsymbol{\theta})]$$

Law of Large Numbers

$$n^{-1}\ell_n(\boldsymbol{\theta}) \to \boxed{E_{\theta*}[\log f(\mathbf{Y} \mid \boldsymbol{\theta})] \equiv \ell_*(\boldsymbol{\theta})}$$

- θ* maximizes the true value of this Log-Likelihood function

Assumptions for MLE

Density

The distribution must be either discrete or continuous – not mixed.

Compactness

The parameter space for θ is closed and bounded

Identifiability

For any $\theta_1 \neq \theta_2$, there must exist a set A such that

$\Pr(\mathbf{Y} \in \mathrm{A} \mid \theta = \theta_1) \neq \Pr(\mathbf{Y} \in \mathrm{A} \mid \theta = \theta_2)$

Assumptions (cont.)

Boundness

The expectation of the Likelihood function will not diverge to infinity

Continuity

The density is continuous in θ

Asymptotic Normality

Under appropriate conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)$ is asymptotically normal, with mean Vector **0** and covariance matrix $\mathbf{J}(\theta^*)^{-1}$

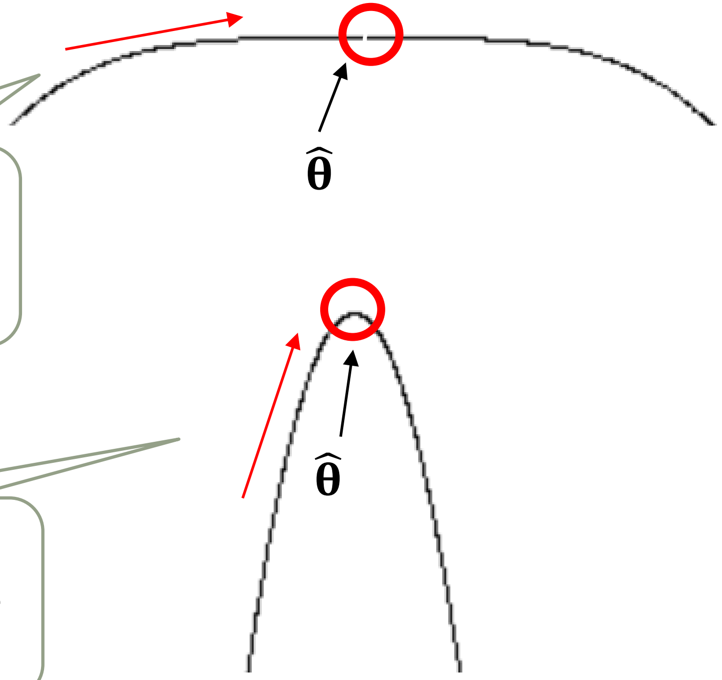
J = Fisher's Information Matrix

2 ways we can estimate Information:

"Expected" -> Expected value of the First derivative squared

"Observed" -> Negative second derivative of Log Likelihood function

Fisher's Information Matrix



Flat slope near $\hat{\theta}$, so second derivative has low magnitude

Steep slope near $\hat{\theta}$, so second derivative has high magnitude

High Variance

Low Information: Unreliable Estimate

High Information: Reliable Estimate

Low Variance

Estimating Variance with Information

What’s the Variance Estimate of our original example?

Optimization - Scoring

Recall Newton's Method of Optimization:

$$x_{n+1} = x_n - f'(x_n)/f''(x_n)$$

Using the Expected Information Matrix, we can adapt the above

$$\hat{\boldsymbol{\theta}}^{(2)} = \hat{\boldsymbol{\theta}}^{(1)} - [-\mathbf{J}_n]^{-1} \boldsymbol{\nabla} \ell_n$$

Computationally efficient, since the second derivatives are not needed

Traditional MLE - Drawbacks

Drawbacks to traditional MLE approach:

1) We assume that all observations are IID

2) Computation time can be costly, especially if the parameter has many dimensions

3) Potential for human error is significant

EM Algorithm

EM - Introduction

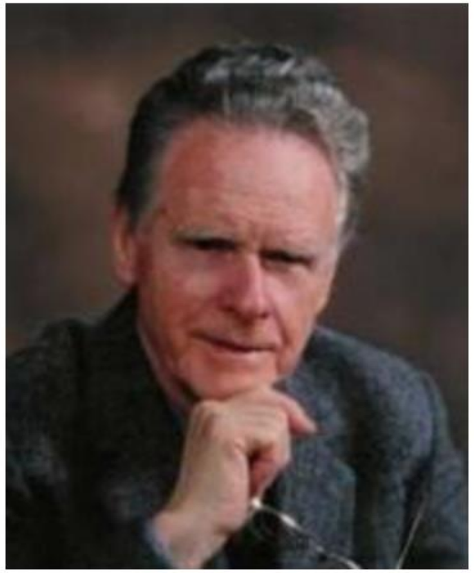
- EM stands for “Expectation Maximization”.

- A **parameter estimation** method: it falls into the general framework of **maximum-likelihood estimation** (MLE).

- The general form was given in (Dempster, Laird, and Rubin, 1977), although essence of the algorithm appeared previously in various forms.

EM - Introduction

Generalized by Arthur Dempster, Nan Laird, and Donald Rubin in a classic 1977 JRSSB paper, which is widely known as the “DLR” paper.



Arthur Dempster
Harvard University
Emeritus Professor of Statistics

Nan Laird
Harvard School of Public Health
professor in Biostatistics

Donald Rubin
Harvard University
Emeritus Professor of Statistics

What’s EM used for?

- Some random variables are not observed
- Directly maximizing the target likelihood function is very difficult
- Typical applications:
  - Discovering the value of latent variables
  - Estimate parameters for finite mixtures (Example 2)
  - Estimating parameters of HMMS
  - Filling in missing data in a sample
  - Other applications

Basic setting in EM

- y_{obs} denotes **observed** data ("incomplete" data)
- $\mathrm{y}_{\mathrm{mis}}$ denotes **hidden** data ("missing" data)
- Θ denotes a parameter vector
- Objective： find $\theta_{\mathrm{MLE}} = \underset{\theta}{\operatorname{argmax}}\, \mathrm{P}(Y_{obs} \mid \theta)$
- EM is a method to find θ_{MLE} where
  - Maximizing $\mathrm{P}(Y_{obs} \mid \theta)$ directly is hard.
  - Working with $\mathrm{P}(Y_{obs}, Y_{mis} \mid \theta)$ is much simpler

EM Algorithm

Step 1 : initialization of the parameters $as\ \theta^0$

Step 2: <u>E-step</u> $Q\left(\theta|\theta^k\right) = E_{Y_{mis}}(logf(Y_{obs}, Y_{mis}|\theta)|\mathrm{y}_{obs}, \theta^k)$ *k=0,1,2,3…*

Step 3: <u>M-step</u> $\underset{\theta}{\operatorname{argmax}} Q\left(\theta|\theta^k\right) \rightarrow \theta^{k+1}$

Step 4: if stop condition is reached, stop; otherwise, let *k =k+1 , go back to step 2*

EM Algorithm (informally)

1. Consider a set of starting parameters
2. Use these to “estimate” the missing data
3. Use “complete” data to update parameters
4. Repeat until convergence

EM Why it works.

1. $P(Y_{obs}|\theta) = \dfrac{P(Y_{obs},\ Y_{\text{mis}}|\theta)}{P(Y_{\text{mis}}|Y_{obs},\ \theta)}$
2. $\log(P(Y_{obs}|\theta)) = \log(P(Y_{obs},\ Y_{\text{mis}}|\theta) - \log(P(Y_{\text{mis}}|Y_{obs},\ \theta)$
3. *Take expectation w.r.t* $Y_{mis}|y_{obs},\ \theta^{\mathrm{k}}$
4. $\log(P(Y_{obs}|\theta) = \sum_{Y_{\text{mis}}} \log(P(Y_{obs},\ Y_{\text{mis}}|\theta)\, P(Y_{\text{mis}}|y_{obs},\ \theta^{\mathrm{k}}) - \sum_{Y\text{mis}} \log(P(Y_{\text{mis}}|y_{obs},\ \theta)P(Y_{\text{mis}}|y_{obs},\ \theta^{\mathrm{k}})$
   $= Q\left(\theta|\theta^{k}\right) + \mathrm{H}\left(\theta|\theta^{k}\right)$
5. $\log(P\left(Y_{obs}|\theta^{k}\right) = Q\left(\theta^{k}|\theta^{k}\right) + \mathrm{H}\left(\theta^{k}|\theta^{k}\right)$

Next step: (4)-(5)

EM: a short derivation

6. $\log(P(Y_{obs}|\theta) - \log(P(Y_{obs}|\theta^k) = Q(\theta|\theta^k) - Q(\theta^k|\theta^k) + \{H(\theta|\theta^k) - H(\theta^k|\theta^k)\}^*$

can show "*" ≥ 0 using Gibbs inequality

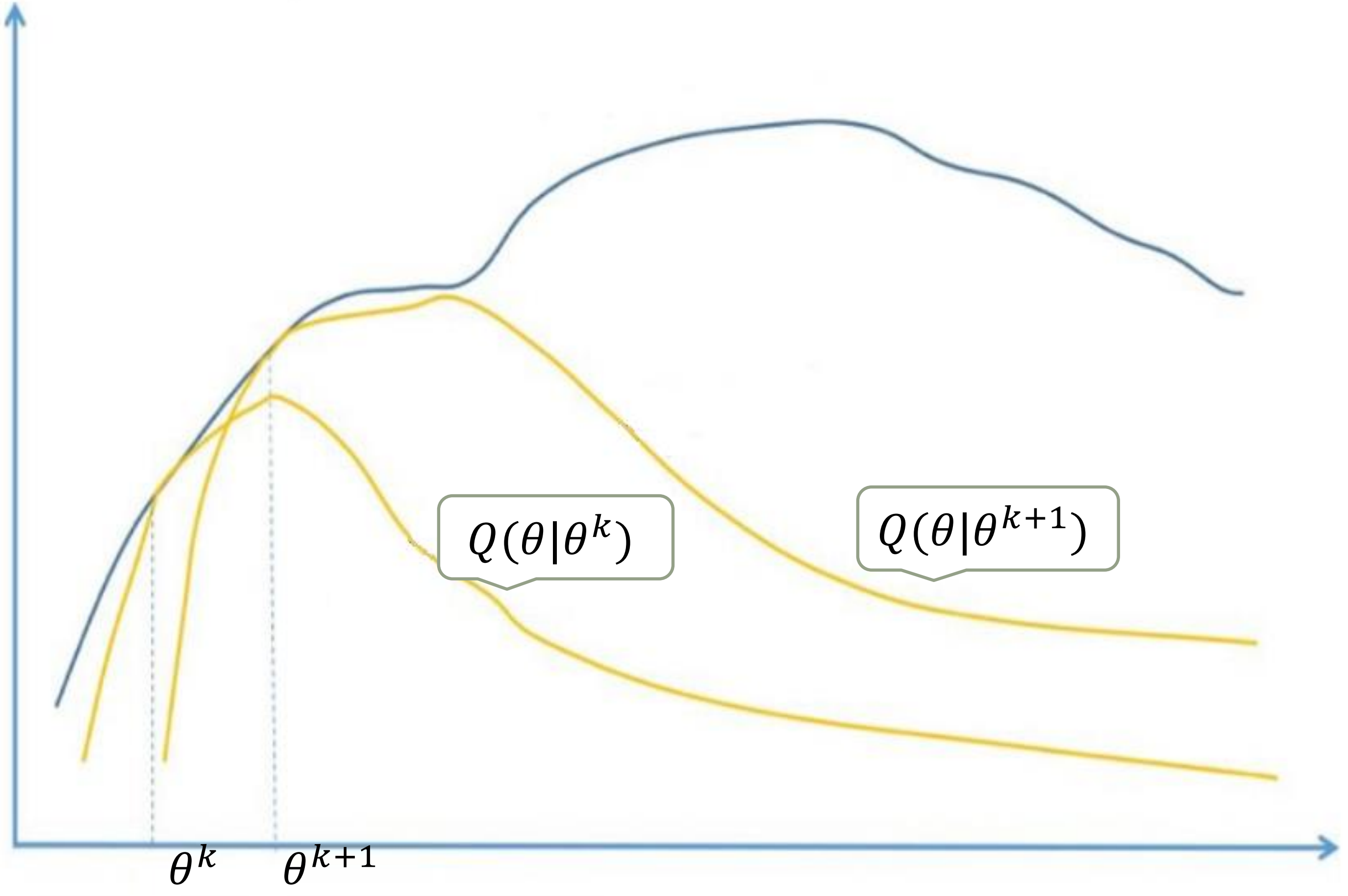
7. $\log(P(Y_{obs}|\theta) - \log(P(Y_{obs}|\theta^k) \geq Q(\theta|\theta^k) - Q(\theta^k|\theta^k)$

choose θ such that $Q(\theta|\theta^k) - Q(\theta^k|\theta^k) \geq 0$

$\longrightarrow \log(P(Y_{obs}|\theta) - \log(P(Y_{obs}|\theta^k) \geq 0$

$\longrightarrow$ θ that increases Q function increases logP

Graphic representation of one iteration of EM algorithm



Monte Carlo EM

Step1 : initialization of the parameters $as\ \theta^0$

Step2*: E-step $Q(\theta|\theta^k) = \frac{1}{M}\sum_{m=1}^{M} logP(y_{obs}, y_{mis}{}^m|\theta)$ *k=0,1,2,3…*

where $y_{mis}{}^m$ *are M's sample drawn from* $P_{Y_{mis}|y_{obs},\theta^k}(\cdot\,|y_{obs},\theta^k)$

Step3: M-step $\underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^k) \rightarrow \theta^{k+1}$

Step4: if stop condition is reached, stop; otherwise, let *k =k+1 , go back to* Step 2*

EM for Multinomial case

Discover the value for latent variables (Multinomial)

$\boldsymbol{Y} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \boldsymbol{y_3}, \boldsymbol{y_4})$*has a multinomical distribution*

with probability of $(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4})$

$$L(\theta|Y) \equiv \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left(\frac{1-\theta}{4}\right)^{y_2}\left(\frac{1-\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}$$

$\boldsymbol{Y_{obs}}$

Assume that $\mathbf{X} = (\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{y_2}, \boldsymbol{y_3}, \boldsymbol{y_4})$*has a multinomical distribution*

with probability of $(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}) \rightarrow \boldsymbol{y_1} = \boldsymbol{x_0} + \boldsymbol{x_1}$

$$L(\theta|X) \equiv \frac{(x_0+x_1+y_2+y_3+y_4)!}{x_0!x_1!y_2!y_3!y_4!}\left(\frac{1}{2}\right)^{x_0}\left(\frac{\theta}{4}\right)^{x_1}\left(\frac{1-\theta}{4}\right)^{y_2}\left(\frac{1-\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}$$

$\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}$

*Example from http://web1.sph.emory.edu/users/hwu30/teaching/statcomp/Notes/Lecture3_EM.pdf

Multinomial case

Model1: $L(\theta|Y) \equiv \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left(\frac{1-\theta}{4}\right)^{y_2}\left(\frac{1-\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}$

Model2 : $L(\theta|X) \equiv \frac{(x_0+x_1+y_2+y_3+y_4)!}{x_0!x_1!y_2!y_3!y_4!}\left(\frac{1}{2}\right)^{x_0}\left(\frac{\theta}{4}\right)^{x_1}\left(\frac{1-\theta}{4}\right)^{y_2}\left(\frac{1-\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}$

Goal:

- Given data y1,y2,y3,y4 (but no x0, x1 observed)
- Find maximum likelihood estimates of θ

EM Algorithm

Step 1: Guess a parameter θ^0

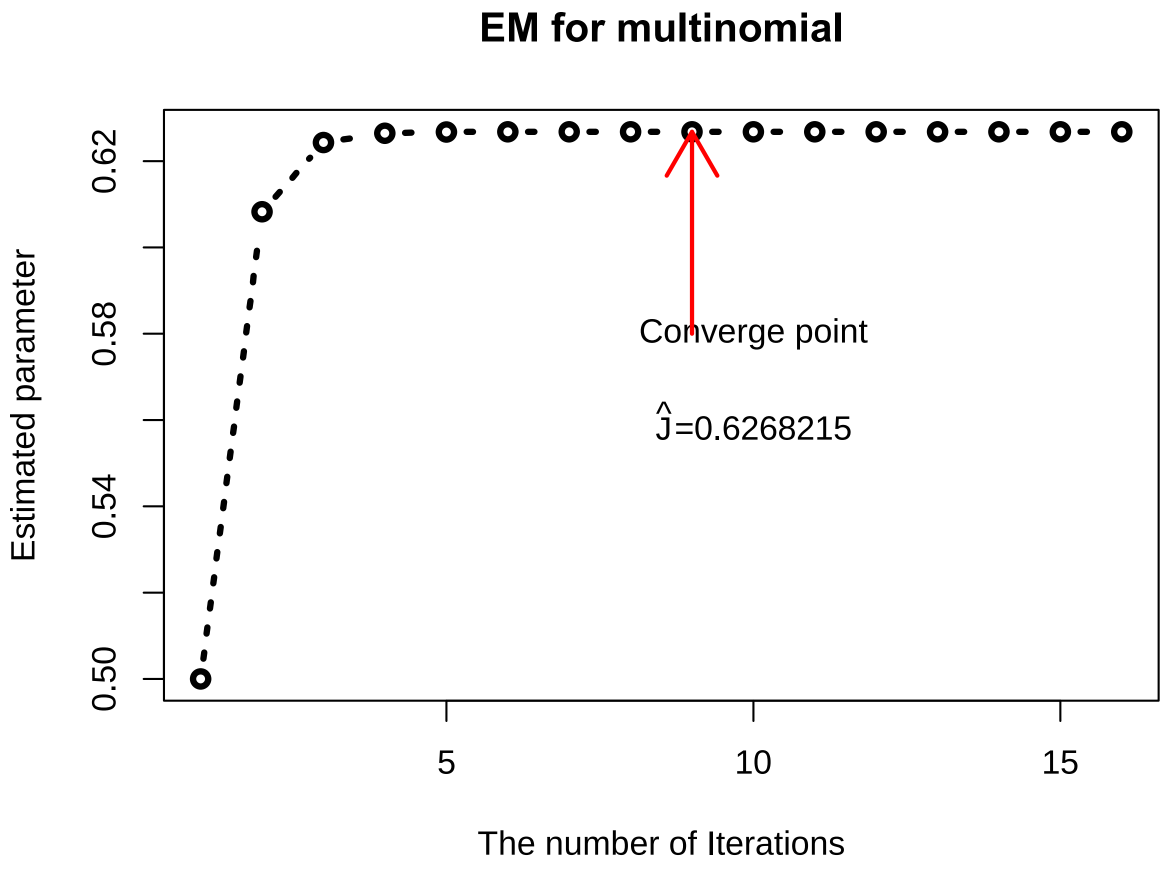
$$Q(\theta, \theta^0) = E(logf(Y_{obs}, Y_{mis}|\theta)|Y_{obs}, \theta^0)$$

Step 2: E-step : $x_1^1 = E(x_1|Y, \theta^0) = \frac{y_1\left(\frac{\theta^0}{4}\right)}{\frac{1}{2}+\frac{\theta^0}{4}}$

Step 3: M-step : $\theta^1 = \frac{x_1^1+y_4}{x_1^1+y_4+y_2+y_3}$

M step (Handout)

Step :4 Repeat Step2 and Step3 until the difference of $(\theta^{k+1} - \theta^k) \leq 10^{-8}$



R Result for multinomial case

MCEM for Multinomial case

MCEM for Multinomial case

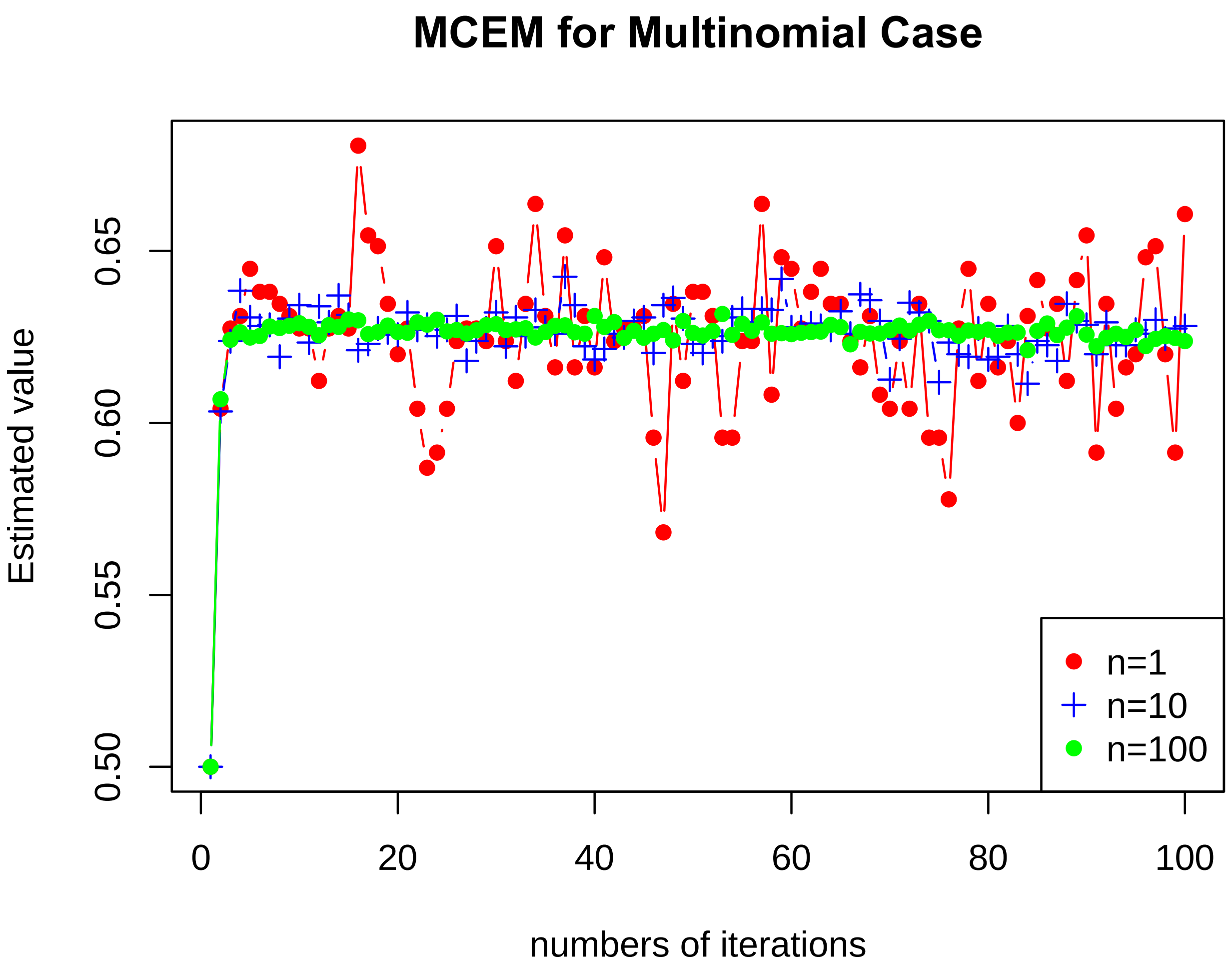
(MC E-step) On the i^{t+1}iteration, draw unobserved data from unobserved data density $f(y_{mis}|y_{obs},\theta^{t})$ to get Q function

Step 1. Draw x1 of sample size = 1,10,100 from x1's density $\sim Bin(y_1,\frac{\frac{\theta}{4}}{\frac{\theta}{4}+\frac{1}{2}})$

Step 2. Calculate the Mean for x1 and use $\overline{x_1}$ in Q function

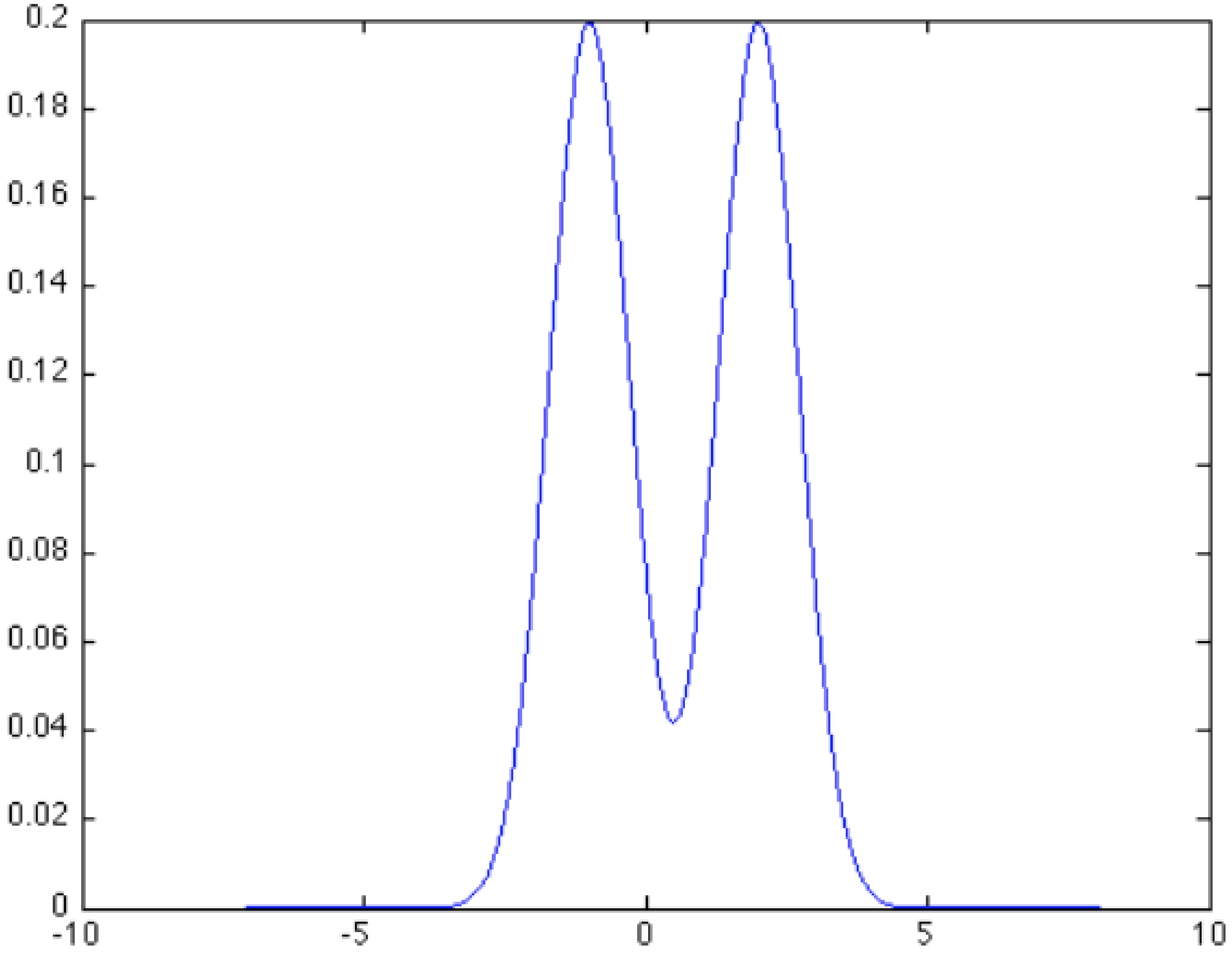
(MC M-step) Maximize the approximate Q function and put θ^{t+1} as the maximizer.

Step 3. $\theta^{t+1} = \frac{\overline{x_1^t}+y_4}{\overline{x_1^t}+y_4+y_2+y_3}$



R Result for MCEM multinomial case

EM – Two component GMM



Estimate parameters for finite mixtures (Two-component Normal Mixture Model)

What is GMM?

Two – Component Mixture model

$$X_1 \sim N(\mu_1, \sigma_1^2)\ , X_2 \sim N(\mu_2, \sigma_2^2)$$

$$X = (1 - Z) * X_1 + Z * X_2, where\ Z = 0{,}1\ and\ P(Z = 1) = \alpha_1$$
$$Let\ f_\theta(x)\ denotes\ the\ normal\ density\ with\ parameter\ \theta = \{\mu, \sigma^2\}$$

$$Thus, it\ can\ write\ in\ \boldsymbol{X} = \boldsymbol{\alpha_1} * \boldsymbol{f_{\theta_1}(x)} + (\boldsymbol{1} - \boldsymbol{\alpha_1}) * \boldsymbol{f_{\theta_2}(x)}$$
$$Therefore, the\ MLE\ of this\ gmm\ for\ n\ training\ data\ is:$$
$$l(\theta) = \sum_{i=1}^{n} log(\alpha_1 * f_{\theta_1}(x_i) + (1 - \alpha_1) * f_{\theta_2}(x_i))$$

However, to get the MLE estimators for all parameters $\{\theta, \sigma, \alpha\}$ is very hard!

EM

Model: $P(Z = 1) = \alpha_1$

$l(\theta|x) = \sum_{i=1}^{n} log(\alpha_1 * \varphi_{\theta_1}(x_i) + (1 - \alpha_1) * \varphi_{\theta_2}(x_i))$

Goal: Given data $x_1, x_2, \dots, x_n$ (but no z_i observed)

Find maximum likelihood estimates of $\mu_1, \mu_2, \sigma_1, \sigma_2, \alpha_1$

Algorithm 8.1 *EM Algorithm for Two-component Gaussian Mixture.*

1. Take initial guesses for the parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$ (see text).

2. *Expectation Step*: compute the responsibilities

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, \quad i = 1, 2, \ldots, N. \tag{8.42}$$

3. *Maximization Step*: compute the weighted means and variances:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)},$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i},$$

   and the mixing probability $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$.

4. Iterate steps 2 and 3 until convergence.

From https://web.stanford.edu/~hastie/Papers/ESLII.pdf P275

R Result for Two component GMM case

Initial value ->

p=0.5

$\mu_1 = 1, \mu_2 = 5$

$\sigma_1 = \sigma_2 =$

$sd(X) = 2.08$

Values after converge->

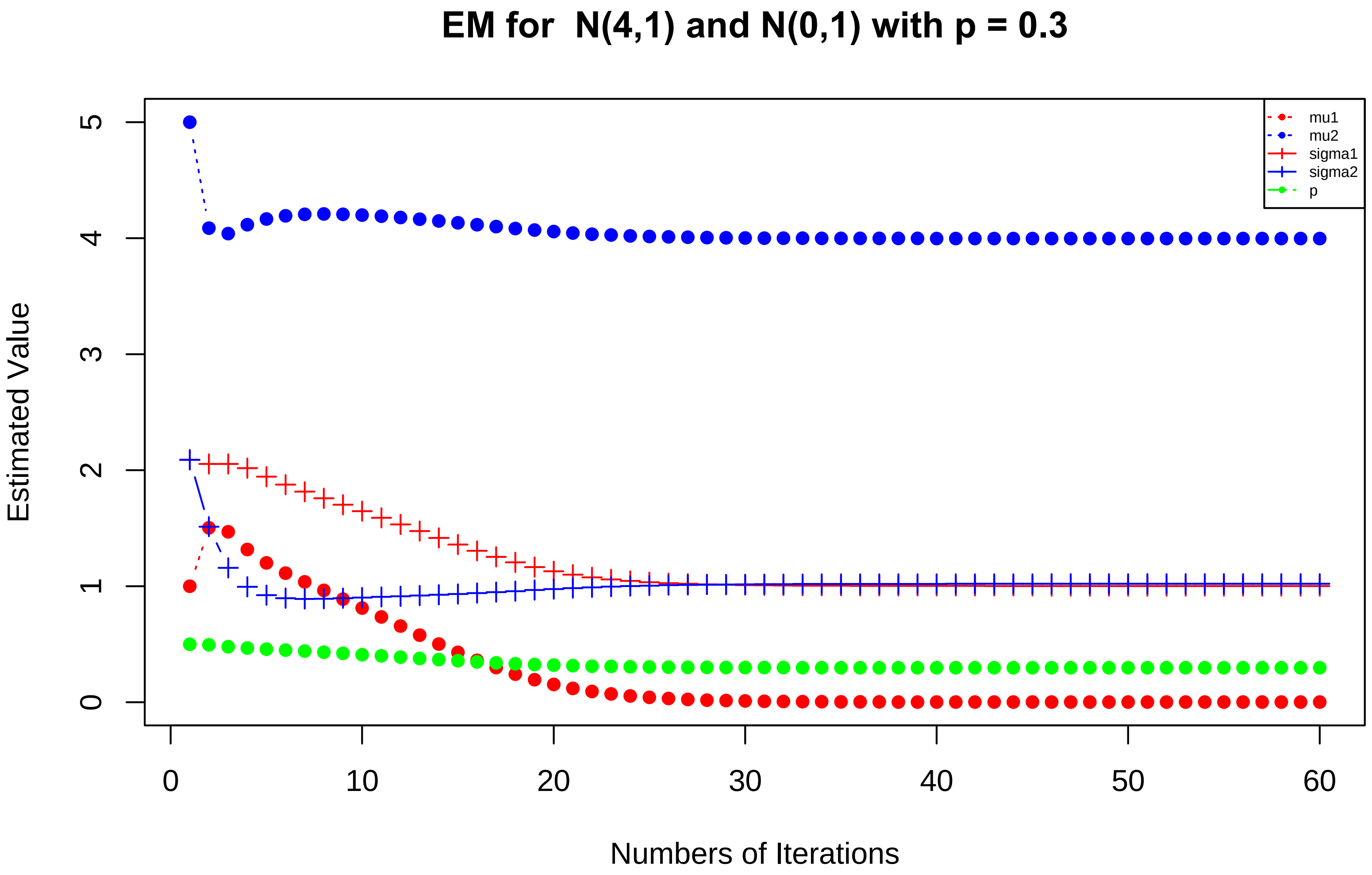
p=0.2972996

$\mu_1 = 0.0010892$

$\mu_2 = 3.996653$

$\sigma_1 = 1.001920$

$\sigma_2 = 1.0203203$



Cons and suggested solutions

<u>Local vs. global max</u>

- There may be multiple modes
- EM may converge to a saddle point

<u>Solution</u>: Multiple starting points

<u>Starting points</u>

- Bad starting points may hurt

<u>Solution:</u>
- Based on the information of the data
- Use the method of moment

<u>Slow Convergence</u>

- EM can be painfully slow

to converge near the maximum

<u>Solution</u>:
Find quicker convergence methods

References

http://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/MLE.pdf

http://web1.sph.emory.edu/users/hwu30/teaching/statcomp/Notes/Lecture3_EM.pdf

https://people.eecs.berkeley.edu/~pabbeel/cs287-fa13/slides/Likelihood_EM_HMM_Kalman.pdf

https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Meilijson, Isaac. “A Fast Improvement to the EM Algorithm on Its Own Terms.” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 51, no. 1, 1989, pp. 127–138. *JSTOR*, www.jstor.org/stable/2345847.

Dempster, Laird & Rubin (1977, JRSSB,39:1-38

Questions