Statistical Inference

Lecture 06a

ANU - RSFAS

Last Updated: Wed Mar 28 11:39:40 2018

MLE Computation: Expectation - Maximization (EM) Algorithm

- Presentation adapted from CB & *Computational Statistics*.
- The EM algorithm is a general algorithm to find MLEs when some of the data are missing (or the problem can be set in a manner that there are missing data).
- Suppose we observe all of the data $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, then all we do to find the MLE is maximize:

$$\ell(\boldsymbol{\theta}; \boldsymbol{y})$$

- Suppose we don't observe all the $\boldsymbol{y}$s then based on the notation by Donald Rubin we have $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss})$.

$$\begin{aligned} f(\boldsymbol{y};\boldsymbol{\theta}) &= f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss};\boldsymbol{\theta}) \\ &= k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs},\boldsymbol{\theta})g(\boldsymbol{y}_{obs};\boldsymbol{\theta}) \end{aligned}$$

- This leads to: $g(\boldsymbol{y}_{obs};\boldsymbol{\theta}) = \frac{f(\boldsymbol{y};\boldsymbol{\theta})}{k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs},\boldsymbol{\theta})}$

$$\begin{aligned} log\,[g(\boldsymbol{y}_{obs};\boldsymbol{\theta})] &= log\,[f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss};\boldsymbol{\theta})] - log\,[k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs},\boldsymbol{\theta})] \\ \ell_{obs}(\boldsymbol{\theta};\boldsymbol{y}_{obs}) &= \ell_{comp}(\boldsymbol{\theta};\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}) - log\,[k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs},\boldsymbol{\theta})] \end{aligned}$$

- As $\boldsymbol{y}_{miss}$ is missing, we replace the right side of the equation with its expectation:

$$\begin{aligned}
\ell_{obs}(\boldsymbol{\theta}; \boldsymbol{y}_{obs}) &= E\left\{\ell_{comp}(\boldsymbol{\theta}; \boldsymbol{y}_{obs}, \boldsymbol{y}_{miss})\middle|\boldsymbol{\theta}', \boldsymbol{y}_{obs}\right\} \\
&\quad -E\left\{log\left[k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs}, \boldsymbol{\theta})\right]\middle|\boldsymbol{\theta}', \boldsymbol{y}_{obs}\right\}
\end{aligned}$$

- The EM algorithm seeks to maximize $\ell(\boldsymbol{\theta}; \boldsymbol{y}_{obs})$ with respect to $\boldsymbol{\theta}$ through the following process:

1. **E step**: Calculate the expectation of the complete likelihood conditional on the observed data and the current value of $\boldsymbol{\theta}$:

$$
\begin{aligned}
Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right) &= E\left\{\ell_{comp}(\boldsymbol{\theta}; \boldsymbol{y}_{obs}, \boldsymbol{y}_{miss})\middle|\boldsymbol{\theta}^{(r)}, \boldsymbol{y}_{obs}\right\} \\
&= \int \left[\ell_{comp}(\boldsymbol{\theta}; \boldsymbol{y}_{obs}, \boldsymbol{y}_{miss})\right] k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs}, \boldsymbol{\theta}) d\boldsymbol{y}_{miss}
\end{aligned}
$$

2. **M step**: Maximize $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$ with respect to $\boldsymbol{\theta}$. Set $\boldsymbol{\theta}^{(r+1)}$ equal to the maximizer of Q.

3. Return to the E step unless a stopping criterion has been reached.

- I will not present a proof that the EM algorithm maximizes $\ell(\boldsymbol{\theta}; \boldsymbol{y}_{obs})$. If you are interested please see either Casella and Berger Exercise 7.31 or *Computational Statistics* Section 4.2.1.

EM Example

bimodal

two component mixture model

whole density

must integrate to 1

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(y;\theta)$.

prob

$$f(y;\theta) = \underset{\uparrow\,\text{prob}}{p}\ \mathrm{normal}(\mu_0, \sigma_0^2) + (1-p)\ \mathrm{normal}(\mu_1, \sigma_1^2)$$

- For this problem generally we have $\theta = (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, p)$.
- For our example let's simplify the problem and assume $p = \frac{1}{2}, \sigma_0^2 = \sigma_1^2 = 1$
- We have the following likelihood:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{1}{2}\ \mathrm{normal}(\mu_0, 1) + \frac{1}{2}\ \mathrm{normal}(\mu_1, 1) \right]$$

- Directly optimizing this is hard.

- To make things easier, we can introduce latent (missing) variables $Z_1, \ldots, Z_n$.
  - Where $Z_i = 0$ if Y_i is from $\text{normal}(\mu_0, 1)$.
  - Where $Z_i = 1$ if Y_i is from $\text{normal}(\mu_1, 1)$.
- Why is this easier? If we know what normal distribution each Y comes from then it is easy to determine the MLEs for μ_0 and μ_1.
- We can use the EM algorithm, where $\boldsymbol{y}_{miss} = \boldsymbol{z}$.
- Note: Z_i is a Bernoulli random variable. $P(Z_i = 1) = \frac{1}{2}$.

$$L(\theta)_{comp} = \prod_{i=1}^{n} \text{normal}(y_i; \mu_0, 1)^{1-z_i} \text{normal}(y_i; \mu_1, 1)^{z_i}$$

$$
\begin{aligned}
\ell_{comp}(\theta) &= \sum_{i=1}^{n}(1-z_i)log\left(\frac{1}{\sqrt{2\pi}}exp\left(-\frac{1}{2}(y_i-\mu_0)^2\right)\right)\\
&\quad+\sum_{i=1}^{n}z_i log\left(\frac{1}{\sqrt{2\pi}}exp\left(-\frac{1}{2}(y_i-\mu_1)^2\right)\right)\\
&\quad+constants\\
&= -\frac{1}{2}\sum_{i=1}^{n}(1-z_i)(y_i-\mu_0)^2+-\frac{1}{2}\sum_{i=1}^{n}z_i(y_i-\mu_1)^2+constants
\end{aligned}
$$

1. Let's determine $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$. Note that z is linear in the log likelihood! Makes our job much easier!

$$
\begin{aligned}
E\left[\ell_{comp}(\theta)\right] &= -\frac{1}{2}\sum_{i=1}^{n}(1 - E[Z_i|\boldsymbol{y}_{obs}, \theta^r])(y_i - \mu_0)^2 \\
&\quad -\frac{1}{2}\sum_{i=1}^{n} E[Z_i|\boldsymbol{y}_{obs}, \theta^r](y_i - \mu_1)^2 + constants
\end{aligned}
$$

- We need to determine: $E[Z_i|\boldsymbol{y}_{obs}, \theta^r]$.
- Note: $E[Z_i|\boldsymbol{y}_{obs}; \theta^r] = Pr(Z_i = 1|\boldsymbol{y}_{obs}; \theta^r)$

$Z = 0, 1$

$E(Z) = P(Z=1)\cdot 1 + 0 P(Z=0)$
$= P(Z=1)$

- We will use Bayes' rule. $P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)} = \frac{P(A\cap B)}{P(B)}$

$$
\begin{aligned}
Pr(Z_i = 1|\boldsymbol{y}_{obs}; \theta^r) &= \frac{f(\boldsymbol{y}_{obs}|Z_i = 1; \theta^r)Pr(Z_i = 1)}{f(\boldsymbol{y}_{obs}|Z_i = 1; \theta^r)Pr(Z_i = 1) + f(\boldsymbol{y}_{obs}|Z_i = 0; \theta^r)Pr(Z_i = 0)} \\
&= \frac{\mathrm{normal}(y_i; \mu_1, 1)\frac{1}{2}}{\mathrm{normal}(y_i; \mu_1, 1)\frac{1}{2} + \mathrm{normal}(y_i; \mu_0, 1)\frac{1}{2}} \\
&= \frac{\mathrm{normal}(y_i; \mu_1, 1)}{\mathrm{normal}(y_i; \mu_1, 1) + \mathrm{normal}(y_i; \mu_0, 1)} \\
&= p_i
\end{aligned}
$$

- So we have:

$$
Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right) = -\frac{1}{2}\sum_{i=1}^{n}(1 - p_i)(y_i - \mu_0)^2 - \frac{1}{2}\sum_{i=1}^{n} p_i(y_i - \mu_1)^2 + constants
$$

2. Let's maximize $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$ with respect to μ_0, μ_1. We find:

$$\hat{\mu}_0^{(r+1)} = \frac{\sum_{i=1}^{n}(1-p_i)y_i}{\sum_{i=1}^{n}(1-p_i)}$$

$$\hat{\mu}_1^{(r+1)} = \frac{\sum_{i=1}^{n} p_i y_i}{\sum_{i=1}^{n} p_i}$$

- Recompute p_i with $\hat{\mu}_0^{(r+1)}$ and $\hat{\mu}_1^{(r+1)}$. Thus iterate between the E and M steps till convergence.

```r
## generate data from a bivariate normal:
set.seed(1001)
n <- 1000
z <- rbinom(n, 1, 1/2)

y <- rep(NA, n)
y[z==0] <- rnorm(length(z[z==0]), -2, 1)
y[z==1] <- rnorm(length(z[z==1]), 2, 1)
```

- Beacuse we generated the data, we know z, so we know the MLEs:

$$\hat{\mu}_0 = \bar{y}|_{z=0}$$

```r
mean(y[z==0])
```

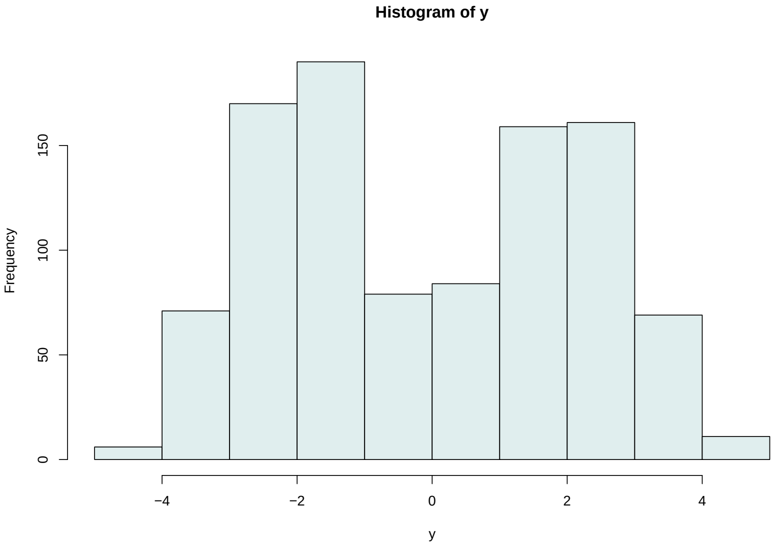
```
## [1] -1.957452
```

$$\hat{\mu}_1 = \bar{y}|_{z=1}$$

```r
mean(y[z==1])
```

```
## [1] 1.997553
```

```r
hist(y, col="azure2")
```



E-M

```r
## starting values
mu.0 <- -0.5
mu.1 <- 0.5

##
check <- 10
eps <- 1e-10

##
while(check > eps){

  # vector E[z|y] - E step
  rho <- dnorm(y, mu.1, 1)/(dnorm(y, mu.1, 1) + dnorm(y, mu.0, 1))

  # M step
  mu.0.new <- sum((1-rho)*y)/(sum((1-rho)))
  mu.1.new <- sum((rho)*y)/(sum((rho)))

  check <- sum( c(abs(mu.0.new - mu.0), abs(mu.1.new - mu.1)))
  mu.0 <- mu.0.new
  mu.1 <- mu.1.new
}

mu.0.hat <- mu.0
mu.1.hat <- mu.1
```

MLEs based on E-M algorithm

```
mu.0.hat
```

```
## [1] -1.942764
```

```
mu.1.hat
```

```
## [1] 2.007483
```

Invariance Property of MLEs

Lemma 3.2: Suppose that $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ represent two alternative parameterizations for some probability distirbution and that $\boldsymbol{\eta}$ is a (1-1) function of $\boldsymbol{\theta}$, so that we can write $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta}), \boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\eta})$ for appropriate functions $\boldsymbol{g}(\cdot), \boldsymbol{h}(\cdot)$.

- If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ then $\hat{\boldsymbol{\eta}} = \boldsymbol{g}(\hat{\boldsymbol{\theta}})$ is the MLE for $\boldsymbol{\eta}$
- If the mapping is (1-1) we simply note:

$$\eta = \tau(\theta) \to \tau^{-1}(\eta) = \theta$$

Invariance Property of MLEs

- Define our likelihood based on the reparameterization ($\theta = \tau^{-1}(\eta)$):

$$L^*(\eta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i; \tau^{-1}(\eta)) = L(\tau^{-1}(\eta); \boldsymbol{x}) = L(\theta; \boldsymbol{x})$$

- We find the supremumum of likelihood

$$\overset{max}{\underset{\eta}{sup}} L^*(\eta; \boldsymbol{x}) = \overset{max}{\underset{\eta}{sup}} L(\tau^{-1}(\eta); \boldsymbol{x}) = \overset{max}{\underset{\theta}{sup}} L(\theta; \boldsymbol{x})$$

to see that the maximum of $L^*(\eta; \boldsymbol{x})$ is when $\eta = \tau(\theta) = \tau(\hat{\theta})$.

Invariance Property of MLEs

$\hat{\theta}$ is ~~the~~ MLE for θ

$\tau = (\theta)^2 \Rightarrow \hat{\tau} = (\hat{\theta})^2$

- However, many functions of interest are not one-to-one: $\theta \to \theta^2$.
- We proceed by defining the induced likelihood function of L^* for $\tau(\theta)$

$$L^*(\eta; \boldsymbol{x}) = \overset{\max}{\cancel{\sup_{\theta:\tau(\theta)=\eta}}} L(\theta; \boldsymbol{x}) \qquad \eta = \tau(\theta)$$

- The value $\hat{\eta}$ that maximizes $L^*(\eta; \boldsymbol{x})$ will be called the MLE of η.

not 1-1.

$\tau(-5) = \tau(5) = 10$

Invariance Property of MLEs

Proof: Let $\hat{\eta}$ denote the value that maximizes $L^*(\eta; \boldsymbol{x})$. Let's show for all values of η that

$$L^*\left(\hat{\eta} = \tau(\hat{\theta}); \boldsymbol{x}\right) \geq L^*(\eta; \boldsymbol{x})$$

$L(\hat{\theta}) \geq L(\theta)$

· Steve Stern's Note

· Casella & Berger

$$\begin{aligned} \underbrace{L^*(\eta; \boldsymbol{x})} &= \sup_{\theta:\tau(\theta)=\eta} L(\theta; \boldsymbol{x}) \quad \text{constrained} \\ &\leq \sup_{\theta} L(\theta; \boldsymbol{x}) \quad \text{un constrained} \\ &= L(\hat{\theta}; \boldsymbol{x}) \\ &= \sup_{\theta:\tau(\theta)=\tau(\hat{\theta})} L(\theta; \boldsymbol{x}) \\ &= L^*(\tau(\hat{\theta}); \boldsymbol{x}) \end{aligned}$$

$\hat{\eta}$ $\shortparallel$

Invariance Property of MLEs

Eg. Normal: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(\mu, \sigma^2)$.

- If we want the MLE of μ^2 it is $\widehat{\mu^2} = (\hat{\mu})^2$.
- If we want the MLE of σ it is $\widehat{\sigma} = \sqrt{\hat{\sigma}^2}$.
- This tends to be helpful in a computational sense as well (we can remove bounds on parameters):

$$\sigma^2 = exp(\theta) \qquad -\infty < \theta < \infty$$