

# IDA Assignment 2

Johnny Lee, s1687781

## Q1.

Suppose  $Y_1, \dots, Y_n$  are independent and identically distributed with cumulative distribution function given by

$$F(y; \theta) = 1 - e^{-y^2/(2\theta)}, \quad y \geq 0, \quad \theta > 0.$$

Further suppose that observations are (right) censored if  $Y_i > C$ , for some known  $C > 0$ , and let

$$X_i = \begin{cases} Y_i & \text{if } Y_i \leq C, \\ C & \text{if } Y_i > C, \end{cases} \quad R_i = \begin{cases} 1 & \text{if } Y_i \leq C \\ 0 & \text{if } Y_i > C \end{cases}$$

a)

Show that the maximum likelihood estimator based on the observed data  $\{(x_i, r_i)\}_{i=1}^n$  is given by

$$\hat{\theta} = \frac{\sum_{i=1}^n X_i^2}{2 \sum_{i=1}^n R_i}.$$

**Answer :**

We first define the Survival function (from **Workshop 3**) as

$$S(C; \theta) = \mathbb{P}(Y_i > C; \theta) = 1 - F(y_i; \theta)$$

which also represents the censored observations. For the uncensored observations, we have

$$f(y_i; \theta) = \frac{d}{dy_i} F(y_i; \theta) = \frac{ye^{-y^2/2\theta}}{\theta}$$

Given that  $Y_1, \dots, Y_n$  are independent and identically distributed, we have the likelihood function as,

$$\begin{aligned} L(\theta | \mathbf{y}, \mathbf{r}) &= \prod_{i=1}^n \left( [f(y_i; \theta)]^{r_i} [S(C; \theta)]^{1-r_i} \right) \\ &= \prod_{i=1}^n \left( \left[ \frac{ye^{-y^2/2\theta}}{\theta} \right]^{r_i} [e^{-C^2/2\theta}]^{1-r_i} \right) \\ &= \left( \frac{y_i}{2\theta} \right)^{\sum_{i=1}^n r_i} \exp \left( \frac{\sum_{i=1}^n (r_i y_i^2 + (1-r_i)C^2)}{2\theta} \right) \end{aligned} \tag{1}$$

Now we can rewrite the term in the exponential as  $X_i$  can we expressed as

$$x_i = r_i y_i + C(1 - r_i)$$

Then by taking square on both sides we have,

$$x_i^2 = r_i^2 y_i^2 + (1 - r_i)^2 C^2 + 2r_i y_i C(1 - r_i)$$

Noting that  $R_i$  is binary, we can then conclude with the expression as

$$x_i^2 = r_i y_i^2 + (1 - r_i) C^2 \quad (2)$$

Now we substitute (2) into (1) to have,

$$\begin{aligned} L(\theta|\mathbf{y}, \mathbf{r}) &= \left(\frac{y_i}{2\theta}\right)^{\sum_{i=1}^n r_i} \exp\left(\frac{\sum_{i=1}^n x_i}{2\theta}\right) \\ \implies \log(L(\theta|\mathbf{y}, \mathbf{r})) &= -\log \theta \sum_{i=1}^n r_i - \frac{\sum_{i=1}^n x_i^2}{2\theta} \\ \implies \frac{d}{d\theta} \log L(\theta|\mathbf{y}, \mathbf{r}) &= \frac{1}{\theta} \sum_{i=1}^n r_i + \frac{1}{2\theta^2} \sum_{i=1}^n x_i^2 \end{aligned} \quad (3)$$

By equating the derivative to 0, we can obtain the maximum likelihood estimate of  $\theta$  as below.

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^n x_i^2}{2 \sum_{i=1}^n r_i} \quad (\text{shown})$$

b)

Show that the expected Fisher Information for the observed data likelihood is

$$I(\theta) = \frac{n}{\theta^2} (1 - e^{-C^2/(2\theta)})$$

**Note:**  $\int_0^C y^2 f(y; \theta) dy = -C^2 e^{-C^2/(2\theta)} + 2\theta(1 - e^{-C^2/(2\theta)})$ , where  $f(y; \theta)$  is the density function corresponding to the cumulative distribution function  $F(y; \theta)$  defined above.

**Answer :**

From (3), we take another derivative of it and thus obtain as below

$$\frac{d^2}{d\theta^2} \log L(\theta) = \frac{1}{\theta^2} \sum_{i=1}^n r_i - \frac{x_i^2}{\theta^3}$$

Then, the Fisher Information for the observed data likelihood is,

$$\begin{aligned} I(\theta) &= -\mathbb{E} \left( \frac{\sum_{i=1}^n r_i}{\theta^2} - \frac{x_i^2}{\theta^3} \right) \\ &= -\frac{n\mathbb{E}(R)}{\theta^2} + \frac{n\mathbb{E}(X^2)}{\theta^3} \\ &= -\frac{n\mathbb{E}(R)}{\theta^2} + \frac{1}{\theta^3} \left( n\mathbb{E}(RY^2) + nC^2\mathbb{E}(1 - R) \right) \end{aligned} \quad (4)$$

Again, noting that  $R_i$  is binary,

$$\begin{aligned} \mathbb{E}(R) &= 1 \cdot \mathbb{P}(R = 1) + 0 \cdot \mathbb{P}(R = 0) \\ &= \mathbb{P}(R = 1) = \mathbb{P}(Y \leq C) \\ &= F(C; \theta) = 1 - e^{-C^2/2\theta} \end{aligned} \quad (5)$$

With the given equation,  $\mathbb{E}(RY^2) = \int_0^C y^2 f(y; \theta) dy = -C^2 e^{-C^2/(2\theta)} + 2\theta(1 - e^{-C^2/(2\theta)})$ , we can combine all the above equations as express the expected Fisher Information again,

$$\begin{aligned} I(\theta) &= \frac{n\mathbb{E}(R)}{\theta^2} + \frac{1}{\theta^3} \left( n\mathbb{E}(RY^2) + nC^2\mathbb{E}(1 - R) \right) \\ &= \frac{-n}{\theta^2} (1 - e^{-C^2/2\theta}) - \frac{n}{\theta^3} (C^2 e^{-C^2/2\theta}) + \frac{n}{\theta^3} (2\theta(1 - e^{-C^2/2\theta})) + \frac{n}{\theta^3} (C^2 e^{-C^2/2\theta}) \\ &= \frac{n}{\theta^2} (1 - e^{-C^2/2\theta}) \quad (\text{shown}) \end{aligned} \quad (6)$$

**c)**

Appealing to the asymptotic normality of the maximum likelihood estimator, provide a 95% confidence interval for  $\theta$ .

**Answer :**

Asymptotic normality of the maximum likelihood estimator is given as,

$$\hat{\theta} \sim N_p(0, I(\theta)^{-1})$$

Thus, with 0 and  $\frac{1}{I(\theta)}$  as mean and variance respectively, we can obtain the 95% confidence interval as below,

$$\hat{\theta} \pm \frac{1.96}{\sqrt{I(\theta)}}$$

## Q2.

Suppose that a dataset consists of 100 subjects and 10 variables. Each variable contains 10% of missing values. What is the largest possible subsample under a complete case analysis? What is the smallest? Justify.

Suppose that  $Y_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$  are iid for  $i = 1, \dots, n$ . Further suppose that now observations are (left) censored if  $Y_i < D$ , for some known  $D$  and let

$$X_i = \begin{cases} Y_i & \text{if } Y_i \geq D, \\ D & \text{if } Y_i < D, \end{cases} \quad R_i = \begin{cases} 1 & \text{if } Y_i \geq D \\ 0 & \text{if } Y_i < D \end{cases}$$

a)

Show that the log-likelihood of the observed data  $\{(x_i, r_i)\}_{i=1}^n$  is given by

$$\log L(\mu, \sigma^2 | \mathbf{x}, \mathbf{r}) = \sum_{i=1}^n \{r_i \log \phi(x_i; \mu, \sigma^2) + (1 - r_i) \log \Phi(x_i; \mu, \sigma^2)\}$$

where  $\phi(\cdot; \mu, \sigma^2)$  and  $\Phi(\cdot; \mu, \sigma^2)$  stands, respectively, for the density function and cumulative distribution function of the normal distribution with mean  $\mu$  and variance  $\sigma^2$ .

**Answer :**

We first define the Survival function (from **Workshop 3**) as

$$S(D; \mu, \sigma^2) = \mathbb{P}(Y_i < D; \mu, \sigma^2) = \Phi(x_i; \mu, \sigma^2)$$

which also represents the censored observations. For the uncensored observation, we have

$$\phi(x_i; \mu, \sigma^2)$$

Given that  $X_1, \dots, X_n$  are independent and identically distributed, we have the likelihood function as,

$$\begin{aligned} L(\mu, \sigma^2 | \mathbf{x}, \mathbf{r}) &= \prod_{i=1}^n \left( [\phi(x_i; \mu, \sigma^2)]^{r_i} [\Phi(x_i; \mu, \sigma^2)]^{1-r_i} \right) \\ \Rightarrow l(\mu, \sigma^2 | \mathbf{x}, \mathbf{r}) &= \log \prod_{i=1}^n \left( [\phi(x_i; \mu, \sigma^2)]^{r_i} [\Phi(x_i; \mu, \sigma^2)]^{1-r_i} \right) \\ &= \log \left( [\phi(x_i; \mu, \sigma^2)]^{\sum_{i=1}^n r_i} [\Phi(x_i; \mu, \sigma^2)]^{\sum_{i=1}^n (1-r_i)} \right) \\ &= \sum_{i=1}^n \left( r_i \log \phi(x_i; \mu, \sigma^2) + (1 - r_i) \log \Phi(x_i; \mu, \sigma^2) \right) \end{aligned} \tag{7}$$

b)

Determine the maximum likelihood estimate of  $\mu$  based on the data available in the file `dataex2.Rdata`. Consider  $\sigma^2$  known and equal to  $1.5^2$ . **Note:** You can use a built in function such as `optim` or the `maxLik` package in your implementation.

Answer :

```
#defining a function to simulate the log likelihood
log.lik <- function(mu, data){
  x <- data[, 1]
  r <- data[, 2]
  sum((r*dnorm(x, mu, 1.5, log = TRUE) +
      (1-r)*pnorm(x, mu, 1.5, log = TRUE)))
}

#computing the maximum likelihood estimate of mu
mle <- maxLik(logLik = log.lik, data = dataex2, start = c(mu = 5))
summary(mle)

## -----
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 3 iterations
## Return code 1: gradient close to zero (gradtol)
## Log-Likelihood: -336.3821
## 1 free parameters
## Estimates:
##      Estimate Std. error t value Pr(> t)
## mu    5.5328      0.1075   51.48 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## -----
```

We built a function `log.lik()` that produces the log likelihood and then used `maxLik()` to simulate  $\mu$  based on the data. With Newton-Raphson method, we estimated  $\hat{\mu} = 5.5328$  and standard error of 0.1075

### Q3.

Consider a bivariate normal sample  $(Y_1, Y_2)$  with parameters  $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_{12}, \sigma_2^2)$ . The variable  $Y_1$  is fully observed, while some values of  $Y_2$  are missing. Let  $R$  be the missingness indicator, taking the value 1 for observed values and 0 for missing values. For the following missing data mechanisms state, justifying, whether they are ignorable for likelihood-based estimation.

a)

$$\text{logit}\{\mathbb{P}(R = 0|y_1, y_2, \theta, \psi)\} = \psi_0 + \psi_1 y_1, \quad \psi = (\psi_0, \psi_1) \text{ distinct from } \theta.$$

**Answer :**

Referring to the ignorability assumption (from **Lecture 6.1**), the missing in  $Y_2$  is either **MAR** or **MCAR** and its model parameters,  $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_{12}, \sigma_2^2)$  and missing mechanism parameter,  $\psi$ .

First, the missing mechanism is **MAR**. This is because the missingness is only dependent on  $Y_1$  which is a fully observed variable. The parameters,  $\{\theta, \psi\}$  are also distinct. Therefore, the ignorability assumption holds here and (a) is ignorable for likelihood-based estimation.

b)

$$\text{logit}\{\mathbb{P}(R = 0|y_1, y_2, \theta, \psi)\} = \psi_0 + \psi_1 y_2, \quad \psi = (\psi_0, \psi_1) \text{ distinct from } \theta.$$

**Answer :**

The missing mechanism is **MNAR** as the mechanism is only dependent on  $Y_2$ . Therefore, the missing value is depending on itself and possibly other factors. Therefore, by referring to the ignorability assumption (from **Lecture 6.1**), we conclude that (b) is not ignorable for likelihood-based estimation.

c)

$$\text{logit}\{\mathbb{P}(R = 0|y_1, y_2, \theta, \psi)\} = 0.5(\mu_1 + \psi y_1), \text{ scalar } \psi \text{ distinct from } \theta.$$

**Answer :**

The missing mechanism here is dependent on both  $\mu_1$  and  $Y_1$ . We can observe similarity to (a). Distinctness of the parameters means that the parameter space of  $(\theta, \psi)$  is equal to the Cartesian product of their individual product spaces. However, the  $\mu_1$  also exists in the parameter space. This violates the ignorability assumption. Hence, (c) is not ignorable for likelihood-based estimation.

Q4.

$$Y_i \stackrel{\text{ind.}}{\sim} \text{Bernoulli}(p_i(\boldsymbol{\beta}))$$

$$p_i(\boldsymbol{\beta}) = \frac{\exp(\beta_0 + x_i\beta_1)}{1 + \exp(\beta_0 + x_i\beta_1)},$$

for  $i = 1, \dots, n$  and  $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ . Although the covariate  $x$  is fully observed, the response variable  $Y$  has missing values. Assuming ignorability, derive and implement the EM algorithm to compute the MLE of  $\boldsymbol{\beta}$  based on the data available in `dataex4.Rdata`. **Note:** 1) For simplicity, and without loss of generality because we have a univariate pattern of missingness, when writing down your expressions, you can assume that the first  $m$  values of  $Y$  are observed and the remaining  $n - m$  are missing. 2) You can use a built in function such as `optim` or the `maxLik` package for the M-step.

**Answer :**

```
head(dataex4)
```

```
##           X   Y
## 1 -0.4689827  1
## 2 -0.2557522  1
## 3  0.1457067  1
## 4  0.8164156 NA
## 5 -0.5966361  1
## 6  0.7967794 NA
```

```
cat("Number of missing values in Y:", sum(is.na(dataex4)))
```

```
## Number of missing values in Y: 95
```

Scrutinising on the dataset, we can observe that the missing value only occurs in  $Y$  and there are 95 missing values occurring in a univariate pattern

We first derive the likelihood function to implement the EM algorithm given that  $y_{obs} = y_1, \dots, y_m$  and  $y_{mis} = y_{m+1}, \dots, y_n$ .

$$\begin{aligned} L(\beta_0, \beta_1 | \mathbf{x}, \mathbf{y}_{obs}, \mathbf{y}_{mis}) &= \prod_{i=1}^n \left( [p_i(\beta_0, \beta_1)]^{y_i} [1 - p(\beta_0, \beta_1)]^{1-y_i} \right) \\ \Rightarrow L(\beta_0, \beta_1 | \mathbf{x}, \mathbf{y}_{obs}, \mathbf{y}_{mis}) &= \prod_{i=1}^n \left( \frac{e^{\beta_0 + x_i\beta_1}}{1 + e^{\beta_0 + x_i\beta_1}} \right)^{y_i} \left( \frac{1}{1 + e^{\beta_0 + x_i\beta_1}} \right)^{1-y_i} \\ \Rightarrow \log L(\beta_0, \beta_1 | \mathbf{x}, \mathbf{y}_{obs}, \mathbf{y}_{mis}) &= \sum_{i=1}^n \left( y_i \log \left( \frac{e^{\beta_0 + x_i\beta_1}}{1 + e^{\beta_0 + x_i\beta_1}} \right) + (1 - y_i) \log \left( \frac{1}{1 + e^{\beta_0 + x_i\beta_1}} \right) \right) \\ \Rightarrow \log L(\beta_0, \beta_1 | \mathbf{x}, \mathbf{y}_{obs}, \mathbf{y}_{mis}) &= \sum_{i=1}^n \left( y_i \log(e^{\beta_0 + x_i\beta_1}) - \log(1 + e^{\beta_0 + x_i\beta_1}) - y_i \log(1 + e^{\beta_0 + x_i\beta_1}) + y_i \log(1 + e^{\beta_0 + x_i\beta_1}) \right) \\ \Rightarrow \log L(\beta_0, \beta_1 | \mathbf{x}, \mathbf{y}_{obs}, \mathbf{y}_{mis}) &= \sum_{i=1}^n \left( y_i(\beta_0 + x_i\beta_1) - \log(1 + e^{\beta_0 + x_i\beta_1}) \right) \\ &= l(\boldsymbol{\beta} | \mathbf{x}, \mathbf{y}_{obs}, \mathbf{y}_{mis}) \end{aligned} \tag{8}$$



Now we proceed to implement the EM algorithm by calculating  $Q(\beta|\beta^{(t)})$

$$\begin{aligned}
Q(\beta|\beta^{(t)}) &= \mathbb{E}_{\mathbf{y}_{mis}}[l(\beta|\mathbf{x}, \mathbf{y}_{obs}, \mathbf{y}_{mis})|\mathbf{y}_{obs}, \mathbf{x}, \beta^{(t)}] \\
&= \sum_{i=1}^m \left( y_i(\beta_0 + x_i\beta_1) \right) - \sum_{i=1}^n \left( \log(1 + e^{\beta_0 + x_i\beta_1}) \right) + \sum_{i=m+1}^n \left( (\beta_0 + x_i\beta_1) \mathbb{E}_{\mathbf{y}_{mis}}[y_i|\mathbf{x}, \mathbf{y}_{obs}, \beta^{(t)}] \right) \\
&= \sum_{i=1}^m \left( y_i(\beta_0 + x_i\beta_1) \right) - \sum_{i=1}^n \left( \log(1 + e^{\beta_0 + x_i\beta_1}) \right) + \sum_{i=m+1}^n \left( (\beta_0 + x_i\beta_1) p_i(\beta) \right) \\
&(\mathbb{E}(Y_i) = p_i(\beta) \text{ as } Y_i \sim \text{Bernoulli}(p_i(\beta)))
\end{aligned} \tag{9}$$

Next step follow with M-step by computing the partial derivatives.

In the code, we have used for the stopping criterion as below

$$|p^{(t+1)} - p^{(t)}| + |\beta_0^{(t+1)} - \beta_0^{(t)}| + |\beta_1^{(t+1)} - \beta_1^{(t)}| < \varepsilon$$

## Q5

Consider a random sample  $Y_1, \dots, Y_n$  from the mixture distribution with density

$$f(y) = pf_{\text{LogNormal}}(y; \mu, \sigma^2) + (1-p)f_{\text{Exp}}(y; \lambda),$$

with

$$\begin{aligned}
f_{\text{LogNormal}}(y; \mu, \sigma^2) &= \frac{1}{y\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (\log y - \mu)^2 \right\}, \quad y > 0, \quad \mu \in \mathbb{R}, \quad \sigma > 0 \\
f_{\text{Exp}}(y; \lambda) &= \lambda e^{-\lambda y}, \quad y \geq 0, \quad \lambda > 0
\end{aligned}$$

and  $\theta = (p, \mu, \sigma^2, \lambda)$

### a) Derive the

EM algorithm to find the updating equations for  $\theta^{(t+1)} = (p^{(t+1)}, \mu^{(t+1)}, (\sigma^{(t+1)})^2, \lambda^{(t+1)})$ .

**Answer :**

Let us consider a mixture model of Log-Normal and Exponential distributions.

$$\mathbb{P}(Y \leq y) = p \cdot \frac{1}{y\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (\log y - \mu)^2 \right\} + (1-p) \cdot \lambda e^{-\lambda y}$$

Let  $z_i$  be binary latent variables indicating component membership, i.e.

$$z_i = \begin{cases} 1 & \text{if } y_i \text{ belong to } f_{\text{LogNormal}}(y; \mu, \sigma^2) \\ 0 & \text{if } y_i \text{ belong to } f_{\text{Exp}}(y; \lambda) \end{cases}$$

### b)

Using the dataset `datasetex5.Rdata` implement the EM algorithm and find the MLEs for each component of  $\theta$ . As starting values, you might want to consider  $\theta^{(0)} = (p^{(0)}, \mu^{(0)}, (\sigma^{(0)})^2, \lambda^{(0)}) = (0.1, 1, 0.5^2, 2)$ . Draw the histogram of the data with the estimated density superimposed.

**Answer :**