Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2020/2021

Review of maximum likelihood for complete data

Context

- ↪ We will study methods for estimation in the presence of missing data based on the principles of maximum likelihood, when it is reasonable to assume that the missing data mechanism is MAR.

- ↪ Before moving to maximum likelihood for missing/incomplete data, we review maximum likelihood inference for complete data.

Review of maximum likelihood for complete data

Likelihood function

- Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables with probability mass/density function $f(y; \boldsymbol{\theta})$ depending on a vector-valued parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$.

- The joint density of observations $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{y}). \tag{1}$$

- The expression in (1), when viewed as a function of the unknown parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$, is called the likelihood function.

Review of maximum likelihood for complete data

Likelihood function – example

- Let Y_1, Y_2, and Y_3 be iid from a Bernoulli distribution with parameter θ.
- The probability mass function is

$$f(y;\theta) = \theta^y(1-\theta)^{1-y}, \quad y \in \{0,1\}$$

and thus the likelihood is

$$L(\theta; y_1, y_2, y_3) = \prod_{i=1}^{3} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^{3} y_i}(1-\theta)^{3-\sum_{i=1}^{3} y_i}$$

- If $(y_1, y_2, y_3) = (0,0,0)$, then, for instance,

$$L(1/2, (0,0,0)) = \left(1-\frac{1}{2}\right)^{3-0} = \frac{1}{8} = 0.125, \quad L(1/3, (0,0,0)) = \left(1-\frac{1}{3}\right)^{3-0} = \frac{8}{27} \approx 0.296$$

- We say that $\theta = 1/3$ has a higher likelihood than $\theta = 1/2$ for these observed data.

Review of maximum likelihood for complete data

Maximum likelihood estimator

- The goal of statistical inference is to use the observed data $\mathbf{y}$ to estimate/infer $\boldsymbol{\theta}$.

- A sensible way to estimate the parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$ is to maximise the likelihood function, choosing the parameter value/vector that makes the data actually observed as likely as possible.

- Formally, we define the maximum likelihood estimator (mle) as that value $\widehat{\theta}_{\text{MLE}}$ such that

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta};\mathbf{y}),$$

that is, $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$ is the value that maximises the likelihood function.

- In other words,

$$L(\widehat{\boldsymbol{\theta}}_{\text{MLE}};\mathbf{y}) > L(\boldsymbol{\theta};\mathbf{y}), \quad \text{for all } \boldsymbol{\theta}.$$

Review of maximum likelihood for complete data

Maximum likelihood estimator–example

$\hookrightarrow$ For the previous example with Bernoulli data, with $(y_1, y_2, y_3) = (0, 0, 0)$ and where θ is either $1/2$ or $1/3$, then

$$\Theta = \{1/3, 1/2\},$$
$$\widehat{\theta}_{\text{MLE}} = 1/3,$$

because $L(1/3, (0, 0, 0)) > L(1/2, (0, 0, 0))$.

Review of maximum likelihood for complete data

Log likelihood

- It is often numerically convenient to use the log likelihood function, $\log L(\boldsymbol{\theta}; \mathbf{y})$ for computation of the mle.

- The logarithm is a strictly increasing function and therefore

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \log L(\boldsymbol{\theta}; \mathbf{y}).$$

- The first and second derivatives of the log likelihood are important and have their own names.

Review of maximum likelihood for complete data

Score function

- The first derivative of the log likelihood function is called score function

$$U(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}; \mathbf{y}).$$

- Note that the score function is a vector of first partial derivatives, one for each element of $\boldsymbol{\theta}$, i.e.,

$$U(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \log L(\boldsymbol{\theta}; \mathbf{y}) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \log L(\boldsymbol{\theta}; \mathbf{y}) \end{bmatrix}$$

- Computation of the mle is typically done by solving the system of equations

$$U(\boldsymbol{\theta}) = \mathbf{0}_p.$$

Review of maximum likelihood for complete data

Fisher information

- The expected Fisher information matrix is defined as

$$I(\boldsymbol{\theta}) = E\left[U(\boldsymbol{\theta})U(\boldsymbol{\theta})^T\right]$$

- Under general conditions, it simplifies to

$$I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta};\mathbf{Y})\right].$$

- The matrix of the negative observed second derivatives evaluated at the mle is called the observed Fisher information matrix

$$J(\widehat{\boldsymbol{\theta}}_{\text{MLE}}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta};\mathbf{y})\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{MLE}}}$$

Review of maximum likelihood for complete data

Asymptotic normality of the mle

- Additionally, and under certain regularity conditions, $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$ has approximately, in large samples, a multivariate normal distribution with mean equal to the true parameter and covariance matrix given by the inverse of the expected Fisher information matrix, so that

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} \sim \text{N}_p(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1}).$$

- It also holds and is of more convenience

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} \sim \text{N}_p(\boldsymbol{\theta}, J(\widehat{\boldsymbol{\theta}}_{\text{MLE}})^{-1}).$$

- The result above is used to derive approximate standard errors for $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$ and confidence intervals for $\boldsymbol{\theta}$.

Review of maximum likelihood for complete data

Example

- Let $Y_1, \ldots, Y_n$ form a random sample from a Bernoulli distribution with unknown parameter $0 \leq \theta \leq 1$. The goal is to find the mle of θ.

- The probability mass function is

$$f(y;\theta) = \theta^y (1-\theta)^{1-y}, \quad y \in \{0,1\}.$$

- The likelihood function is

$$\begin{aligned} L(\theta;\mathbf{y}) &= \prod_{i=1}^{n} \left\{ \theta^{y_i} (1-\theta)^{1-y_i} \right\} \\ &= \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i}. \end{aligned}$$

- The corresponding log likelihood is

$$\log L(\theta;\mathbf{y}) = \log\theta \sum_{i=1}^{n} y_i + \log(1-\theta) \left(n - \sum_{i=1}^{n} y_i \right).$$

Review of maximum likelihood for complete data

Example

- Taking the derivative and setting it to zero (i.e., equating the score function to zero)

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log L(\theta;\mathbf{y}) = 0 \Rightarrow \frac{1}{\theta}\sum_{i=1}^{n} y_i - \frac{1}{1-\theta}\left(n - \sum_{i=1}^{n} y_i\right) = 0,$$

  lead us to finally obtain

$$\widehat{\theta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}.$$

- **Remark**: formally, to be sure that we have obtained a maximum (the mle!), we would need to confirm that the derivative of the score function, evaluated at $\theta = \bar{y}$, is negative.

- We will now obtain the expected and observed Fisher information. For that, we need the second derivative:

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log L(\theta;\mathbf{y}) = -\frac{1}{\theta^2}\sum_{i=1}^{n} y_i - \frac{1}{(1-\theta)^2}\left(n - \sum_{i=1}^{n} y_i\right)$$

Review of maximum likelihood for complete data

Example

↪ Evaluating the second derivative at $\widehat{\theta}_{\text{MLE}} = \bar{Y}$, we obtain

$$J(\widehat{\theta}_{\text{MLE}}) = \frac{n}{\bar{Y}(1-\bar{Y})}.$$

↪ The expected Fisher information is (remembering that $E(Y) = \theta$)

$$\begin{aligned} I(\theta) = -E\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log L(\theta;\mathbf{Y})\right] &= \frac{1}{\theta^2}nE[Y] + \frac{1}{(1-\theta)^2}(n - nE[Y]) \\ &= \frac{n}{\theta(1-\theta)}. \end{aligned}$$

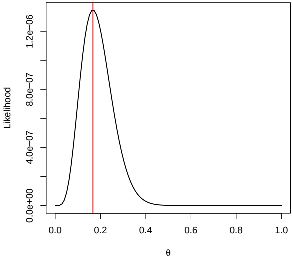
Review of maximum likelihood for complete data

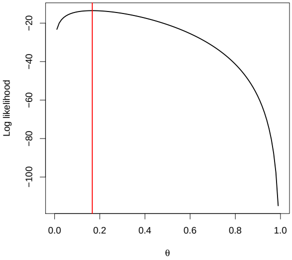
Example

- Suppose 5 people are infected in a sample size of 30, that is,

$$n = 30, \quad \sum_{i=1}^{30} y_i = 5.$$

- We know that $\widehat{\theta}_{\text{mle}} = 5/30 \approx 0.167$.





Review of maximum likelihood for complete data

Example

- Let $Y_1, \ldots, Y_n$ form a random sample from a Normal distribution with unknown parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. The goal is to find the mle of $\boldsymbol{\theta} = (\mu, \sigma^2)$.

- The probability density function is

$$f(y; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}, \quad y \in \mathbb{R}.$$

- The likelihood is

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{y}) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\}\right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}, \end{aligned}$$

and then log likelihood is

$$\log L(\boldsymbol{\theta}; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Review of maximum likelihood for complete data

Example

↪ The score function is given by

$$U(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial\mu} \log L(\boldsymbol{\theta}; \mathbf{y}) \\ \frac{\partial}{\partial\sigma^2} \log L(\boldsymbol{\theta}; \mathbf{y}) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2 \end{bmatrix}.$$

↪ We then have

$$U(\boldsymbol{\theta}) = \mathbf{0} \Rightarrow \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) = 0 \quad \wedge \quad -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2 = 0,$$

leading to

$$\widehat{\mu} = \bar{Y}, \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Review of maximum likelihood for complete data

Example

$\hookrightarrow$ The matrix of second derivatives:

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta};\mathbf{y}) = \begin{bmatrix} \frac{\partial^2}{\partial\mu^2}\log L(\boldsymbol{\theta};\mathbf{y}) & \frac{\partial^2}{\partial\mu\partial\,\sigma^2}\log L(\boldsymbol{\theta};\mathbf{y}) \\ \frac{\partial^2}{\partial\mu\partial\,\sigma^2}\log L(\boldsymbol{\theta};\mathbf{y}) & \frac{\partial^2}{\partial(\sigma^2)^2}\log L(\boldsymbol{\theta};\mathbf{y}) \end{bmatrix}$$
$$= \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^n(y_i-\mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^n(y_i-\mu) & \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^n(y_i-\mu)^2 \end{bmatrix}.$$

$\hookrightarrow$ The observed Fisher information is then

$$J(\widehat{\boldsymbol{\theta}} = (\mu,\widehat{\sigma}^2)) = \begin{bmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\widehat{\sigma}^4} \end{bmatrix}.$$

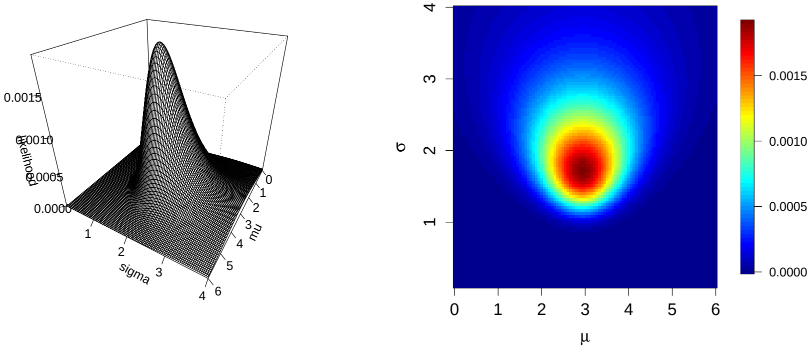
$\hookrightarrow$ The expected Fisher information matrix is

$$I(\mu,\sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

Review of maximum likelihood for complete data

Example

↪ Suppose that we observe $\mathbf{y} = (1.747, 3.367, 1.329, 6.191, 3.659, 1.359)$. We have $\widehat{\mu} = 2.942$ and $\widehat{\sigma}^2 = 2.964$.



Review of maximum likelihood for complete data

Right censored observations

- Let $Y_1, \ldots, Y_n$ be a random sample from an Exponential distribution with parameter $\theta > 0$. Suppose that some of the Ys are right censored and let

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

  be the observations and the censoring indicator, where C is a known censoring point.
- Note that we can write

$$X_i = Y_i I(Y_i \leq C) + CI(Y_i > C) = Y_i R_i + C(1 - R_i).$$

- Our observed data is of the form $\{(x_i, r_i)\}_{i=1}^n$.
- The contribution of a non censored observation to the likelihood is $f(y; \theta)$ and the contribution of a censored observation to the likelihood is $\Pr(Y > C; \theta) = S(C; \theta)$, where S here denotes the survival function.

Review of maximum likelihood for complete data

Right censored observations

- The likelihood is thus of the form

$$L(\theta) = \prod_{i=1}^{n} \left\{ f(y_i;\theta)^{r_i} S(C;\theta)^{1-r_i} \right\}.$$

- For the exponential distribution we have $f(y;\theta) = \theta e^{-\theta y}$ and $S(y;\theta) = e^{-\theta y}$ and therefore

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \left\{ [\theta e^{-\theta y_i}]^{r_i} [e^{-\theta C}]^{1-r_i} \right\} \\ &= \theta^{\sum_{i=1}^{n} r_i} e^{-\theta \sum_{i=1}^{n} y_i r_i} e^{-\theta \sum_{i=1}^{n} C(1-r_i)} \\ &= \theta^{\sum_{i=1}^{n} r_i} e^{-\theta \sum_{i=1}^{n} [y_i r_i + C(1-r_i)]} \\ &= \theta^{\sum_{i=1}^{n} r_i} e^{-\theta \sum_{i=1}^{n} x_i}, \end{aligned}$$

which leads to

$$\widehat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^{n} R_i}{\sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n} I(Y_i \leq C)}{\sum_{i=1}^{n} Y_i I(Y_i \leq C) + C I(Y_i > C)},$$

where $\sum_{i=1}^{n} R_i$ is the number of uncensored observations.