Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2018/2019

Review of maximum likelihood for full data

- We will study methods for inference 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 full data.

Review of maximum likelihood for full data

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

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

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

- This expression, viewed as a function of the unknown parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$, is called the likelihood function.

- Under the assumption of identically distributed random variables, the likelihood function simplifies to

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

Review of maximum likelihood for full data

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

- A sensible way to estimate the parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$ is to maximise the likelihood function, choosing the parameter value 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

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

- In other words,

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

- The following (invariance) property of mles can be important in several problems: let $g(\boldsymbol{\theta})$ be a function of the parameter $\boldsymbol{\theta}$. If $\widehat{\boldsymbol{\theta}}$ is the mle of $\boldsymbol{\theta}$, then $g(\widehat{\boldsymbol{\theta}})$ is the mle of $g(\boldsymbol{\theta})$.

Review of maximum likelihood for full data

- 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 monotone function and therefore

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

- The log likelihood function has much larger importance besides simplifying the computation of the mle.

- Its first and second derivatives are important and have their own names.

Review of maximum likelihood for full data

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

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}).$$

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

$$S(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \log L(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \log L(\boldsymbol{\theta}) \end{bmatrix}$$

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

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

Review of maximum likelihood for full data

- The expected Fisher information matrix is defined as

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

- Under general conditions

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

with $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

- Note that under the assumption that $Y_1, \ldots, Y_n$ are iid, we have

$$I(\boldsymbol{\theta}) = -nE\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta}; Y_1)\right]$$

- The matrix of negative observed second derivatives is sometimes called the observed Fisher information matrix

$$I(\boldsymbol{\theta}; \mathbf{Y}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta}; \mathbf{Y})$$

Review of maximum likelihood for full data

↪ Under certain regularity conditions, the mle converges to the true parameter, i.e., the mle is a consistent estimator

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} \xrightarrow{p} \boldsymbol{\theta}.$$

↪ Additionally, and again 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 information matrix, so that

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

↪ If $\boldsymbol{\theta}$ is unknown, then so is $I(\boldsymbol{\theta})$ and $I(\boldsymbol{\theta}; \mathbf{Y})$. It also holds and is of more convenience

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

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

Review of maximum likelihood for full data

$\hookrightarrow$ The use of $I(\widehat{\theta}_{\text{MLE}}; \mathbf{Y})$ over $I(\widehat{\theta}_{\text{MLE}})$ to compute the variance of $\widehat{\theta}_{\text{MLE}}$ was advocated by Efron and Hinkley (1978) in their article "Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information".

$\hookrightarrow$ Also, $I(\widehat{\theta}_{\text{MLE}}; \mathbf{Y}) = I(\widehat{\theta}_{\text{MLE}})$ for distributions that belong to the exponential family.

Review of maximum likelihood for full data

- Let $Y_1, \ldots, Y_n$ form a random sample from a Bernoulli distribution with unknown parameter $0 < \theta < 1$. We need to find the mle for θ.

- The probability mass function is

$$f(y; \theta) = \theta^y (1 - \theta)^{1-y}.$$

- The likelihood function is

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

- Taking the log, we get

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

Review of maximum likelihood for full data

$\hookrightarrow$ Taking the derivative and setting it 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

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

$\hookrightarrow$ 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 full data

↪ The observed Fisher information is then

$$I(\theta; \mathbf{Y}) = \frac{1}{\theta^2} \sum_{i=1}^{n} Y_i + \frac{1}{(1-\theta)^2} \left(n - \sum_{i=1}^{n} Y_i \right),$$

which evaluated at $\widehat{\theta}_{\mathrm{MLE}} = \bar{Y}$, becomes

$$I(\widehat{\theta}_{\mathrm{MLE}}; \mathbf{Y}) = \frac{n}{\bar{Y}(1-\bar{Y})}.$$

↪ The expected Fisher information is

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

Review of maximum likelihood for full data

- Let $Y_1, \ldots, Y_n$ form a random sample from an Exponential distribution with unknown parameter $\theta > 0$. We need to find the mle for θ.

- The probability density function is

$$f(y; \theta) = \theta e^{-\theta y},$$

implying that the likelihood is

$$L(\theta; \mathbf{y}) = \prod_{i=1}^{n} \{\theta e^{-\theta y_i}\} = \theta^n e^{-\theta \sum_{i=1}^{n} y_i}.$$

- The log likelihood is then

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

Review of maximum likelihood for full data

- Taking the derivative and setting it to zero

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

  lead us to finally obtain

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

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

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} L(\theta;\mathbf{y}) = -\frac{n}{\theta^2}.$$

- In this case

$$I(\theta) = I(\theta;\mathbf{Y}) = \frac{n}{\theta^2}.$$

- Suppose further that we are interested in finding the mle for θ^3. Instead of re-doing all the calculations, we can simply make use of the invariance property of the mle, and so $\widehat{\theta}^3_{\mathrm{MLE}} = \frac{1}{\bar{Y}^3}$.

Review of maximum likelihood for full data

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

- The probability density function is

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

- 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 full data

↪ The score function is given by

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

↪ We then have

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

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 full data

$\hookrightarrow$ We will now obtain the expected and observed Fisher information matrices. For that, we need matrix of second derivatives:

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

$\hookrightarrow$ Evaluating at $(\widehat{\mu}, \widehat{\sigma}^2)$, the observed Fisher information becomes

$$\begin{bmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\widehat{\sigma}^4} \end{bmatrix}.$$

$\hookrightarrow$ The expected Fisher information matrix is

$$\begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

Review of maximum likelihood for full data

- Let $Y_1, \ldots, Y_n$ form a random sample from a Uniform$(0, \theta)$ distribution, for some $\theta > 0$. We need to find the mle for θ.

- Note that the density function is $1/\theta$ only for y values in the interval $[0, \theta]$ and is zero otherwise, i.e.,

$$f(y; \theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq y \leq \theta \\ 0, & \text{otherwise.} \end{cases}$$

- Thus, the likelihood function for the n observations is

$$L(\theta; \mathbf{y}) = \begin{cases} \frac{1}{\theta^n}, & 0 \leq y_i \leq \theta \quad \text{for all } i = 1, \ldots, n, \\ 0, & \text{otherwise.} \end{cases}$$

- For parameter θ to be greater or equal to all y_i, it is equivalent to θ being greater or equal to the maximum of $(y_1, \ldots, y_n)$, $y_{\max}$.

Review of maximum likelihood for full data

↪ That is

$$L(\theta; \mathbf{y}) = \begin{cases} \frac{1}{\theta^n}, & \theta \geq y_{\text{max}} \\ 0, & \text{otherwise.} \end{cases}$$

↪ The likelihood function is zero for $\theta < y_{\text{max}}$, and is monotone decreasing for $\theta \geq y_{\text{max}}$, and hence it must be that $\widehat{\theta}_{\text{MLE}} = \max(Y_1, \ldots, Y_n)$.

Review of maximum likelihood for full data

$\hookrightarrow$ Let $T_1, \ldots, T_n$ be a random sample from an Exponential distribution with parameter $\theta > 0$. Suppose that some of the times are right censored and let

$$Y_i = \begin{cases} T_i & \text{if } T_i \leq c, \\ c & \text{if } T_i > c, \end{cases}$$

be the observed times, where c is a known censoring time.

$\hookrightarrow$ We thus have data of the form $\{(y_i, I(t_i \leq c))\}_{i=1}^{n}$. Note further that we can write

$$y_i = t_i I(t_i \leq c) + cI(t_i > c), \quad \text{and} \quad I(t_i \leq c) + I(t_i > c) = 1.$$

Review of maximum likelihood for full data

$\hookrightarrow$ The likelihood function is then

$$L(\theta; \mathbf{y}, I(t_1 \leq c), \ldots, I(t_n \leq c)) = \prod_{i=1}^{n} \left\{ f(t_i; \theta)^{I(t_i \leq c)} S(c; \theta)^{I(t_i > c)} \right\}$$
$$= \theta^{\sum_{i=1}^{n} I(t_i \leq c)} e^{-\theta[\sum_{i=1}^{n} \{t_i I(t_i \leq c) + cI(t_i > c)\}]}$$

$\hookrightarrow$ We can easily derive that

$$\widehat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^{n} I(T_i \leq c)}{\sum_{i=1}^{n} Y_i}.$$

Review of maximum likelihood for full data

- In some cases, it will not be possible to obtain a closed form expression for the mle.
- In such cases, we need to resort to numerical iterative procedures.
- There are several numerical procedures that one can employ in order to calculate mle estimates and, luckily, `R` has very good optimisation tools.
- Among these, chief are the Newton–Raphson/Fisher-Scoring method, the method of bisection, the method of gradient descent and the EM algorithm.
- Which one is more appropriate depends on the specific example.
- What it is common to all of them is that they are iterative: they start at a given input value and iterate some operation until the convergence criterion is attained.

Review of maximum likelihood for full data

- Newtons's method for finding the solution $\widehat{\boldsymbol{\theta}}$ to $S(\boldsymbol{\theta}) = \mathbf{0}$ can be described as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \left[I\left(\boldsymbol{\theta}^{(t)}; \mathbf{Y}\right)\right]^{-1} S\left(\boldsymbol{\theta}^{(t)}\right).$$

- The behaviour of $I\left(\boldsymbol{\theta}^{(t)}; \mathbf{Y}\right)$ can be problematic if $\boldsymbol{\theta}^{(t)}$ is far from the mle $\widehat{\boldsymbol{\theta}}$.

- Thus, instead of using the observed Fisher information $I(\boldsymbol{\theta}; \mathbf{Y})$, we can use the expected Fisher information to get

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \left[I\left(\boldsymbol{\theta}^{(t)}\right)\right]^{-1} S\left(\boldsymbol{\theta}^{(t)}\right).$$

- This algorithm is called the Fisher scoring method.

Review of maximum likelihood for full data

- We will consider applying the above methods to the Cauchy and Weibull distributions.

- Let us suppose that $Y_1, \ldots, Y_n$ are iid random variables following the Cauchy distribution with unknown location θ and scale equal to one, whose density function is given by

$$f(y;\theta) = \frac{1}{\pi(1+(y-\theta)^2)}, \qquad y \in \mathbb{R}.$$

- The log likelihood function in this case is

$$\log L(\theta;\mathbf{y}) = -n\log\pi - \sum_{i=1}^{n}\log\left\{1+(y_i-\theta)^2\right\}.$$

- The root of the score equation has no closed form expression, as one can appreciate below

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log L(\theta;\mathbf{y}) = 0 \Rightarrow 2\sum_{i=1}^{n}\frac{y_i-\theta}{1+(y_i-\theta)^2} = 0.$$

Review of maximum likelihood for full data

- To implement the Newton–Raphson and Fisher-Scoring methods we need to calculate the second derivative of the log likelihood, which is given by

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

- After some calculations, the expected Fisher information is given by $I(\theta) = n/2$.

- See the supplementary file on Learn for details on the implementation of the numerical methods.

Review of maximum likelihood for full data

- The Weibull distribution is widely used in survival analysis. Its density function is given by

$$f(y;\theta) = \frac{\alpha}{\beta}\left(\frac{y}{\beta}\right)^{\alpha-1} \exp\left\{-\left(\frac{y}{\beta}\right)^{\alpha}\right\}, \quad y > 0, \quad \alpha > 0, \quad \beta > 0.$$

- Here α is a shape parameter and β a scale parameter, and $\theta = (\alpha, \beta)$.

- The log likelihood function is given by

$$\log L(\boldsymbol{\theta}; \mathbf{y}) = n\log(\alpha) - n\alpha\log(\beta) + (\alpha - 1)\sum_{i=1}^{n}\log(y_i) - \frac{1}{\beta^{\alpha}}\sum_{i=1}^{n} y_i^{\alpha}.$$

- Differentiation leads to the system of equations:

$$\begin{cases} -\frac{n\alpha}{\beta} + \frac{\alpha}{\beta}\sum_{i=1}^{n}\left(\frac{y_i}{\beta}\right)^{\alpha} = 0 \\ \frac{n}{\alpha} + \sum_{i=1}^{n}\log\left(\frac{y_i}{\beta}\right) - \sum_{i=1}^{n}(\frac{y_i}{\beta})^{\alpha}\log\left(\frac{y_i}{\beta}\right) = 0 \end{cases}$$

- To compute the solutions of this system we resort to iterative numerical methods.