MATH3091 Statistical Modelling II
Lecture 3: Score Function and Information Matrix

Dr. Chao Zheng

4th Feb 2022

Recap

Previously, we

- reintroduced the likelihood: "the probability that we would have seen the data we actually did, for each value of the parameter".
- reviewed the "usual" recipe for finding maximum likelihood estimates: find a stationary point of the log-likelihood (and check it is a maximum).

Let's look at one more example.

Example: gamma with known rate parameter

Suppose $Y_i \sim \text{gamma}(\theta, 1)$ i.i.d, with

$$f(y;\theta) = \frac{1}{\Gamma(\theta)} y^{\theta-1} e^{-y}, \quad y > 0.$$

Example: gamma with known rate parameter

Suppose $Y_i \sim \text{gamma}(\theta, 1)$ i.i.d, with

$$f(y;\theta) = \frac{1}{\Gamma(\theta)} y^{\theta-1} e^{-y}, \quad y > 0.$$

The likelihood is

$$L(\theta; y_1, \dots, y_n) = \prod_{i=1}^{n} \frac{1}{\Gamma(\theta)} y_i^{\theta-1} e^{-y_i},$$

The log-likelihood is

$$\begin{aligned} \log L(\theta; y_1, \dots, y_n) &= \sum_{i=1}^{n} \log\left(\frac{1}{\Gamma(\theta)} y_i^{\theta-1} e^{-y_i}\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{\Gamma(\theta)}\right) + (\theta - 1) \sum_{i=1}^{n} \log y_i - \sum_{i=1}^{n} y_i. \end{aligned}$$

Differentiating the log-likelihood

Differentiating, we have

$$\frac{d}{d\theta} \log L(\theta; y_1, \ldots, y_n) = -n \frac{d}{d\theta} \log \Gamma(\theta) + \sum_{i=1}^{n} \log y_i,$$

where $\psi(\theta) = \frac{d}{d\theta} \log \Gamma(\theta)$ is called the *digamma function*.

So $\hat{\theta}$ satisfies

$$-n\psi(\hat{\theta}) + \sum_{i=1}^{n} \log y_i = 0,$$

or

$$\psi(\hat{\theta}) = \frac{\sum_{i=1}^{n} \log y_i}{n}$$

which has no closed-form solution for $\hat{\theta}$.

Differentiating the log-likelihood

Differentiating, we have

$$\frac{d}{d\theta} \log L(\theta; y_1, \ldots, y_n) = -n \frac{d}{d\theta} \log \Gamma(\theta) + \sum_{i=1}^{n} \log y_i,$$

where $\psi(\theta) = \frac{d}{d\theta} \log \Gamma(\theta)$ is called the *digamma function*.

So $\hat{\theta}$ satisfies

$$-n\psi(\hat{\theta}) + \sum_{i=1}^{n} \log y_i = 0,$$

or

$$\psi(\hat{\theta}) = \frac{\sum_{i=1}^{n} \log y_i}{n}$$

which has no closed-form solution for $\hat{\theta}$.

It is very common in practice that it is not possible to write down a closed-form expression for the MLE, so we must use numerical methods to maximise the log-likelihood.

2.2.1 Score Function

Score

Let

$$u_i(\boldsymbol{\theta}) \equiv \frac{\partial}{\partial \theta_i} \ell(\theta), \quad i = 1, \ldots, p$$

and $\boldsymbol{u}(\boldsymbol{\theta}) \equiv [u_1(\boldsymbol{\theta}), \ldots, u_p(\boldsymbol{\theta})]^T$. Then we call $\boldsymbol{u}(\boldsymbol{\theta})$ the *vector of scores* or *score vector*.

Where $p = 1$ and $\boldsymbol{\theta} = (\theta)$, the *score* is the scalar defined as

$$u(\theta) \equiv \frac{\partial}{\partial \theta} \ell(\theta).$$

Score quiz 1

Suppose we have observations $y_1, \dots, y_n$ of $Y_1, \dots, Y_n$, i.i.d. Bernoulli(p) random variables, and we want to estimate p.

As a first guess at p, I guess that $p = 0.5$, and calculate the score there, and find $u(0.5) = -2$

Is the MLE:

- greater than 0.5?
- less than 0.5?
- equal to 0.5?
- Not possible to tell.

Score quiz 2

Suppose we have observations $y_1, \ldots, y_n$ of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables, and we want to estimate p.

After finding $u(0.5) = -2$, I now guess that $p = 0.2$, calculate the score there, and find $u(0.2) = 1$

Is the MLE:

- greater than or equal to 0.5?
- greater than 0.2 but less than 0.5?
- equal to 0.2?
- less than 0.2?
- Not possible to tell.

Notes on the score

- The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ satisfies

  $$u(\hat{\boldsymbol{\theta}}) = \mathbf{0},$$

  that is,

  $$u_i(\hat{\boldsymbol{\theta}}) = 0, \quad i = 1, \ldots, p.$$

- $u(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ for fixed (observed) $\boldsymbol{y}$. However, if we replace $y_1, \ldots, y_n$ in $u(\boldsymbol{\theta})$, by the corresponding random variables $Y_1, \ldots, Y_n$ then we obtain a vector of random variables $U(\boldsymbol{\theta}) \equiv [U_1(\boldsymbol{\theta}), \ldots, U_p(\boldsymbol{\theta})]^T$.

An important result about the score

Theorem: The expected score at the true (but unknown) value of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}_0$, is zero:

$$E[U(\boldsymbol{\theta}_0)] = \mathbf{0}$$

i.e.

$$E[U_i(\boldsymbol{\theta}_0)] = 0, \quad i = 1, \ldots, p,$$

provided that

1. The expectation exists.
2. The sample space for $\boldsymbol{Y}$ does not depend on $\boldsymbol{\theta}$.

Proof (continuous $\boldsymbol{y}$ – in discrete case replace $\int$ by $\sum$)

Ignore the subsript, for each $i = 1, \dots, p$

$$\begin{aligned}
E[U_i(\boldsymbol{\theta})] &= \int U_i(\boldsymbol{\theta}) f_{\boldsymbol{Y}}(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \int \frac{\partial}{\partial \theta_i} \ell(\theta) f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \int \frac{\partial}{\partial \theta_i} \log f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \int \frac{\frac{\partial}{\partial \theta_i} f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})} f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \int \frac{\partial}{\partial \theta_i} f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \frac{\partial}{\partial \theta_i} \int f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \frac{\partial}{\partial \theta_i} 1 = 0.
\end{aligned}$$

Example (Bernoulli)

$y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Here $\boldsymbol{\theta} = (p)$ and

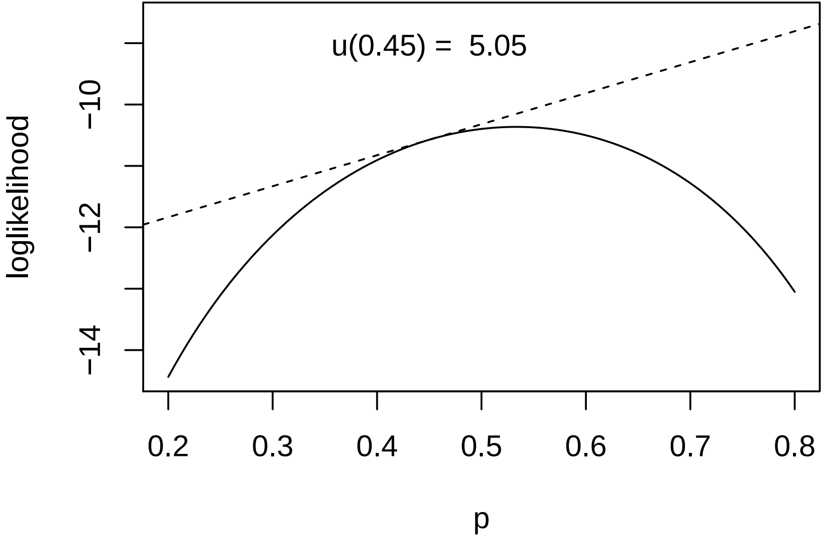
$$u(p) = n\bar{y}/p - n(1 - \bar{y})/(1 - p).$$

Since $E[U(p)] = 0$, we must have $E[\bar{Y}] = p$ (which we already know is correct).

The Bernoulli loglikelihood and score

e.g. take 15 samples from a Bernoulli(0.45) distribution.

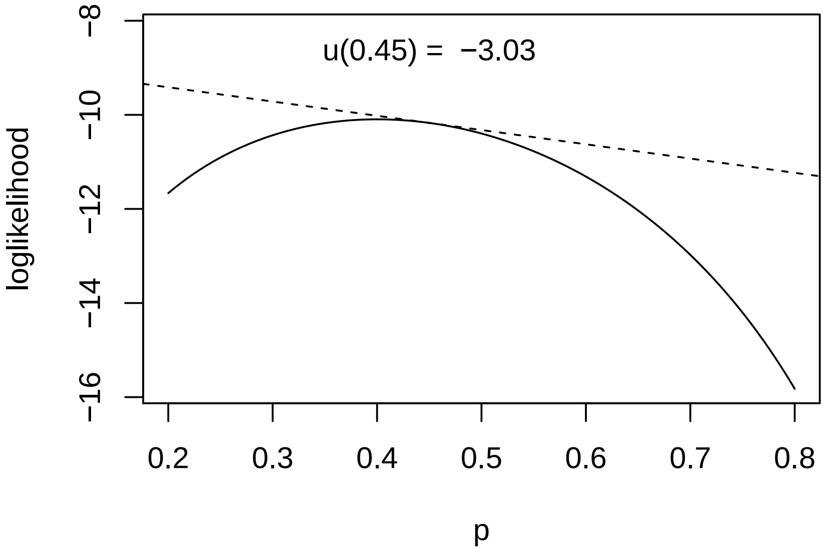
```
##  [1] 0 0 1 1 0 1 1 1 1 0 0 0 1 0 1
```



The Bernoulli loglikelihood and score

e.g. take 15 samples from a Bernoulli(0.45) distribution.

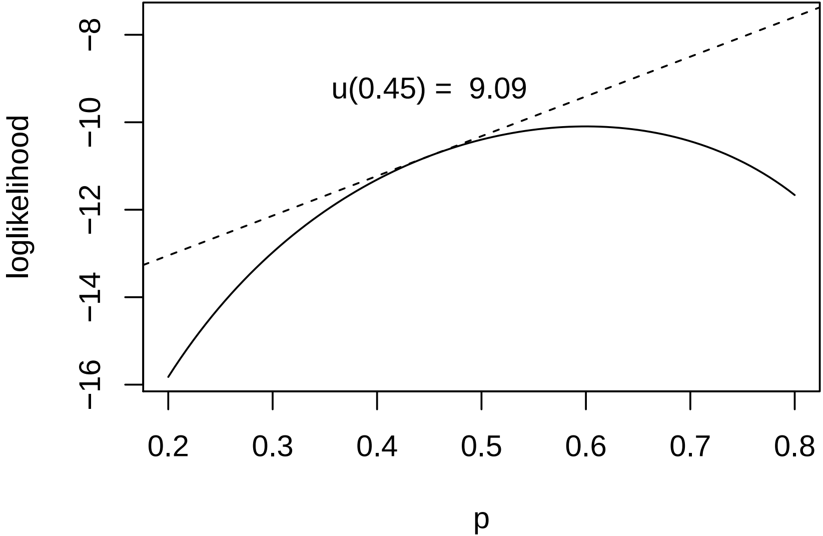
```
##  [1] 0 1 1 0 1 1 0 1 0 0 0 0 0 1 0
```



The Bernoulli loglikelihood and score

e.g. take 15 samples from a Bernoulli(0.45) distribution.

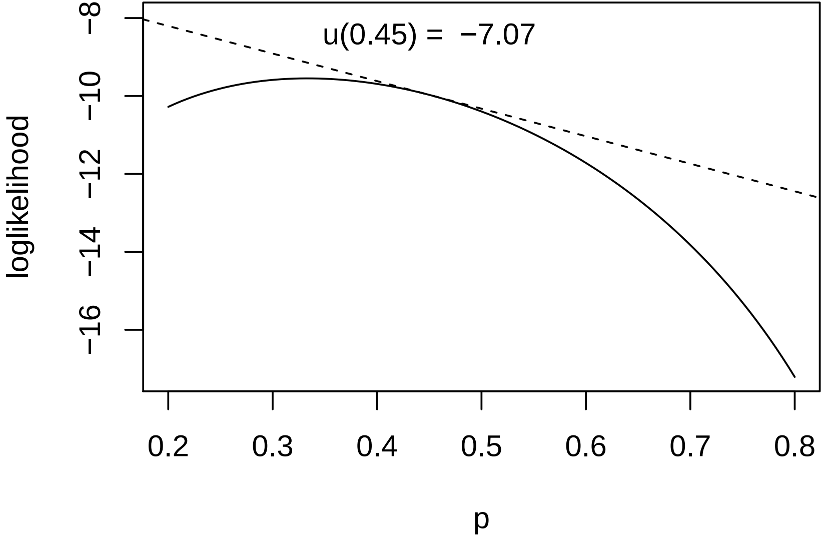
```
##  [1] 0 1 0 0 1 1 1 0 1 0 1 1 1 1 0
```



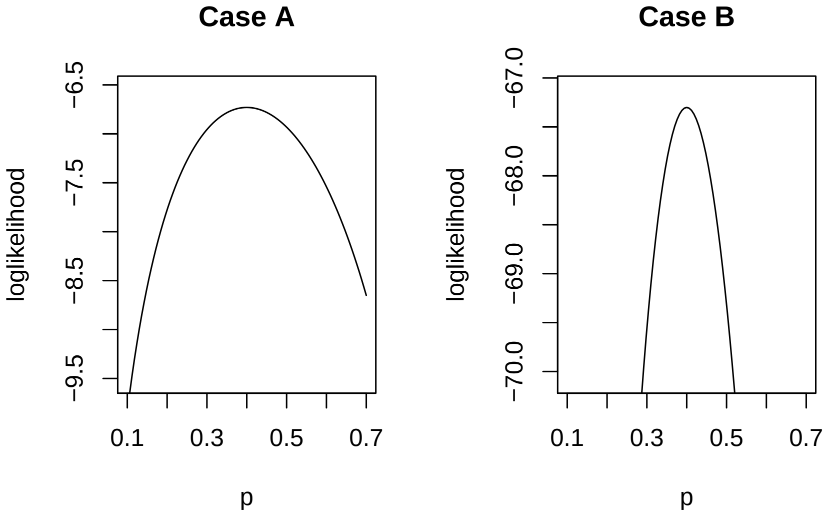
The Bernoulli loglikelihood and score

e.g. take 15 samples from a Bernoulli(0.45) distribution.

```
##  [1] 1 0 0 1 1 0 1 0 0 0 0 0 0 1 0
```



Information: Bernoulli loglikelihoods



Bernoulli loglikelihoods quiz

In both cases, $\hat{p} = 0.4$. In case A, a 95% confidence interval for p is $[0.15, 0.70]$. Which of the following could be a 95% confidence interval for p in case B?

- $[0.01, 0.79]$
- $[0.31, 0.50]$
- Neither of the above

Bernoulli loglikelihoods quiz

In both cases, $\hat{p} = 0.4$. In case A, a 95% confidence interval for p is $[0.15, 0.70]$. Which of the following could be a 95% confidence interval for p in case B?

- $[0.01, 0.79]$
- $[0.31, 0.50]$
- Neither of the above

The data provide more *information* about the parameter in case B than in case A, so we have a higher confidence in case B that the true parameter value is close to 0.4, leading to a smaller confidence interval.

We call minus the second derivative of the log-likelihood the *observed information*, which is 41.7 in case A, and 417 in case B, both at 0.4.

Summary of Socre Function

For now, we

- Defined the score as the gradient of the log-likelihood.
- show that score is useful in finding the MLE.

We also mentioned that the amount of *information* provided by the data is largest when the curvature of the log-likelihood is greatest.

We will formally define *information* to reflect this intuition.

2.2.2 Information Matrix

Observed Information

Suppose that $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, whose joint p.d.f. $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$ is completely specified except for the values of p unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$.

Previously, we defined the Hessian matrix $H(\boldsymbol{\theta})$ to be the matrix with components

$$[H(\boldsymbol{\theta})]_{ij} \equiv \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\boldsymbol{\theta}) \qquad i = 1, \ldots, p;\ j = 1, \ldots, p.$$

We call $-H(\boldsymbol{\theta})$ the *observed information (matrix)*. Where $p = 1$ and $\boldsymbol{\theta} = (\theta)$, the *observed information* is a scalar defined as

$$-H(\theta) \equiv -\frac{\partial^2}{\partial\theta^2} \ell(\theta).$$

Fisher information

As with the score, if we replace $y_1, \ldots, y_n$ in $H(\boldsymbol{\theta})$, by the corresponding random variables $Y_1, \ldots, Y_n$, we obtain a matrix of random variables. Then, we define the *expected information (matrix)* or *Fisher information (matrix)* as

$$[\mathcal{I}(\boldsymbol{\theta})]_{ij} = E_{\boldsymbol{\theta}}(-[H(\boldsymbol{\theta})]_{ij}) \qquad i = 1, \ldots, p;\ j = 1, \ldots, p.$$

Note that here the expectation is with respect to $\boldsymbol{\theta}$, not the true value of the parameter.

Example (Bernoulli)

$y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Here $\boldsymbol{\theta} = (p)$ and

$$\begin{aligned} u(p) &= \frac{n\bar{y}}{p} - \frac{n(1-\bar{y})}{(1-p)} \\ -H(p) &= \frac{n\bar{y}}{p^2} + \frac{n(1-\bar{y})}{(1-p)^2} \\ \mathcal{I}(p) &= \frac{n}{p} + \frac{n}{(1-p)} = \frac{n}{p(1-p)}. \end{aligned}$$

Example (Normal)

$y_1, \dots, y_n$ are observations of $Y_1, \dots, Y_n$, i.i.d. $N(\mu, \sigma^2)$ random variables. Here $\boldsymbol{\theta} = (\mu, \sigma^2)$ and

$$\begin{aligned} u_1(\mu, \sigma^2) &= \frac{n(\bar{y} - \mu)}{\sigma^2} \\ u_2(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum (y_i - \mu)^2. \end{aligned}$$

Therefore

$$-\boldsymbol{H}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{n(\bar{y}-\mu)}{(\sigma^2)^2} \\ \frac{n(\bar{y}-\mu)}{(\sigma^2)^2} & \frac{1}{(\sigma^2)^3} \sum (y_i - \mu)^2 - \frac{n}{2(\sigma^2)^2} \end{pmatrix}$$

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

A link between the score and expected information

Theorem: The variance-covariance matrix of the score vector is equal to the expected information matrix *i.e.*

$$\mathsf{Var}_{\boldsymbol{\theta}}[U(\boldsymbol{\theta})] = \mathcal{I}(\boldsymbol{\theta})$$

or

$$\mathsf{Var}_{\boldsymbol{\theta}}[U(\boldsymbol{\theta})]_{ij} = [\mathcal{I}(\boldsymbol{\theta})]_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, p$$

provided that

1. The variance exists.
2. The sample space for $\boldsymbol{Y}$ does not depend on $\boldsymbol{\theta}$.

This result holds uniformly for all $\boldsymbol{\theta}$, not just at the $\boldsymbol{\theta}_0$.

Proof (continuous $\boldsymbol{y}$ – in discrete case replace $\int$ by $\sum$)}

For each $i = 1, \dots, p$ and $j = 1, \dots, p$,

$$\begin{aligned}
\text{Var}_{\boldsymbol{\theta}}[U(\boldsymbol{\theta})]_{ij} &= E_{\boldsymbol{\theta}}[U_i(\boldsymbol{\theta})U_j(\boldsymbol{\theta})] \\
&= \int \frac{\partial}{\partial\theta_i}\ell(\theta)\frac{\partial}{\partial\theta_j}\ell(\theta) f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= \int \frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta}) f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= \int \frac{\frac{\partial}{\partial\theta_i}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}\frac{\frac{\partial}{\partial\theta_j}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})} f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y}
\end{aligned}$$

Proof

Now

$$
\begin{aligned}
[\mathcal{I}(\boldsymbol{\theta})]_{ij} &= E_{\boldsymbol{\theta}}\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta)\right] \\
&= \int\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\right] f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= \int -\frac{\partial}{\partial\theta_i}\left[\frac{\frac{\partial}{\partial\theta_j}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}\right] f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= \int\left[-\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}+\frac{\frac{\partial}{\partial\theta_i}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})^2}\right] f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\int f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y}+\int\frac{1}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}\frac{\partial f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{\partial\theta_j}d\boldsymbol{y} \\
&= \mathrm{Var}[U(\boldsymbol{\theta})]_{ij}
\end{aligned}
$$

Maximum likelihood estimation

Maximum likelihood estimation is an attractive method of estimation:

- It is intuitively sensible (choose $\boldsymbol{\theta}$ which makes the observed data most probable).
- It is general: you just need a probability model and some data!
- Computing the the MLE is often fairly simple. Even when the simultaneous equations we obtain by differentiating the loglikelihood function are impossible to solve directly, solution by numerical methods is usually feasible.
- Other good properties: efficiency, consistency and asymptotic normality, . . .

Numerical-Method to find MLE

We can use score to determine if our guess $\theta^{(1)}$ is to the left or right of the MLE $\hat{\theta}$.

Question: After knowing the direction, by how much should we adjust our guess?

Newton-Raphson Method

Note that by Taylor expansion:
$u(\hat{\theta}) = u(\theta^{(1)}) + (\hat{\theta} - \theta^{(1)})H(\theta^{(1)}) + ...$, note that $u(\hat{\theta}) = 0$ and ignore the remainder terms:

$$0 \approx u(\theta^{(1)}) + (\hat{\theta} - \theta^{(1)})H(\theta^{(1)})$$

Therefore

$$\hat{\theta} \approx \theta^{(1)} - u(\theta^{(1)})/H(\theta^{(1)})$$

This suggests the following *Newton-Raphson algorithm:*

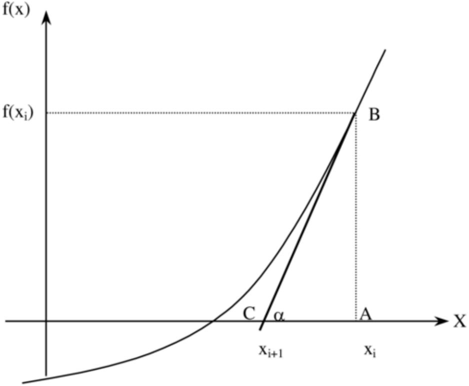
1. *Let $\theta^{(1)}$ denote your initial guess of the MLE.*
2. *For $i = 2, 3, \ldots$, update $\theta^{(i)}$ as:*

$$\theta^{(i)} = \theta^{(i-1)} - \frac{u(\theta^{(i-1)})}{H(\theta^{(i-1)})}$$

3. *Terminate the iteration until $|\theta_i - \theta_{i-1}| < \epsilon$, where ϵ is a small constant denote our convergence tolerance.*

Newton-Raphson method

To find the root of a function $f(x) = 0$



$$\tan(\alpha) = \frac{AB}{AC}$$

$$f'(x_i) = \frac{f(x_i)}{x_i - x_{i+1}}$$

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Conclusion

- We have defined the score, the Hessian matrix, the observed information and the expected information.
- The score plays an important role in numerical methods to maximise the log-likelihood.
- We showed that that the expected value of the score vector at the true value of θ is zero.
- We proved that "variance of the score equals the expected information"
- We show the Newton-Raphson method to find the MLE numerically