Expectation-Maximization Algorithm

Module DES130: Computational Statistics

Dr. Erick A. Chacón Montalván (echacon@uni.edu.pe)

Escuela de Profesional de Ingeniería Estadística
Facultad de Ingeniería Económica, Estadística y Ciencias Sociales
Universidad Nacional de Ingeniería (UNI)
Lima – Perú

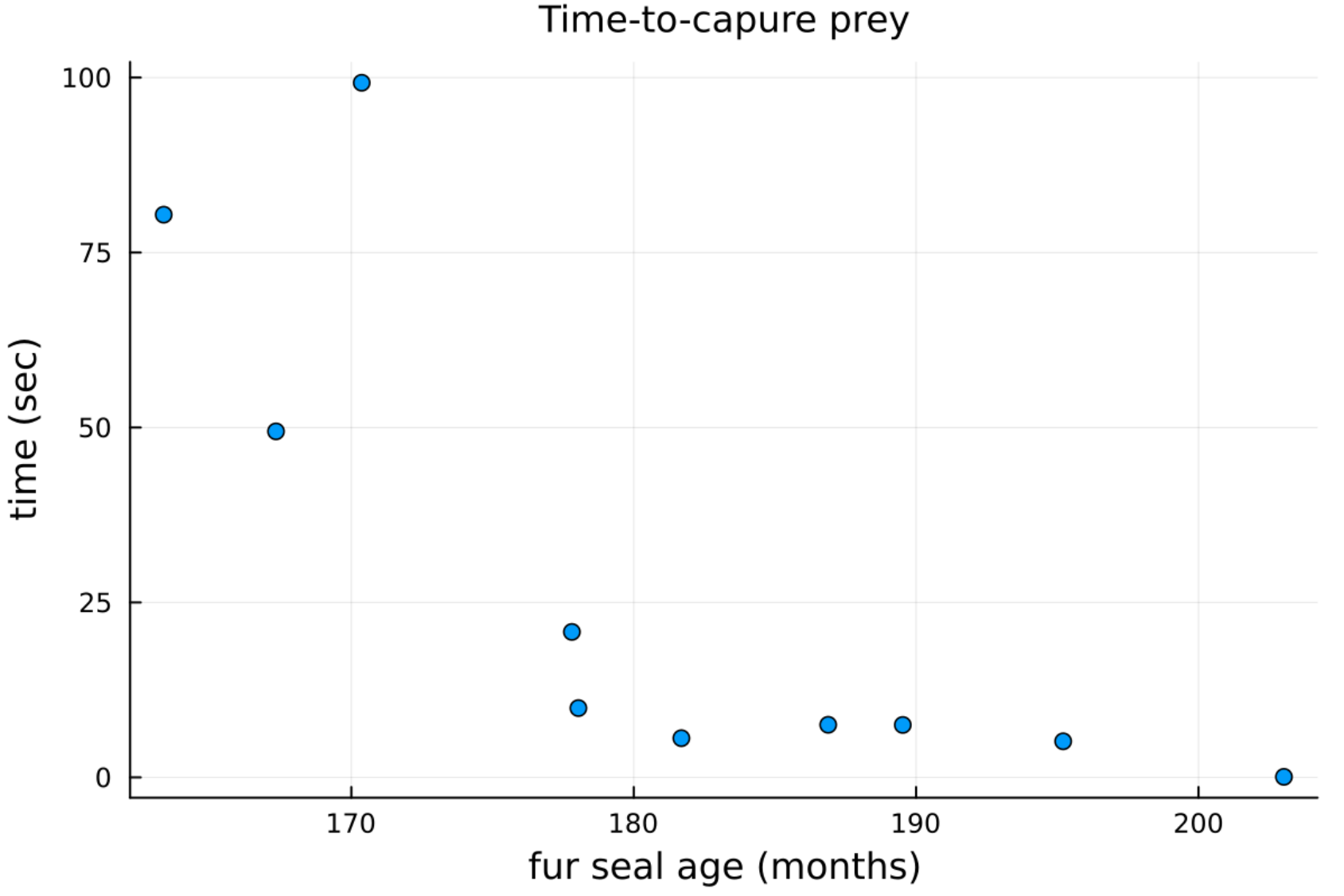
Likelihood function

Modelling time to capture with age

A biologist is studying the time that fur seals need to capture their first prey. The observed times (seconds) for 10 fur seals and their age (weeks) are below.

fur seal	age	time
1	177.811	20.7883
2	195.207	5.15748
3	167.333	49.4457
4	178.041	9.90466
5	203.024	0.0770514
6	163.358	80.4105
7	186.886	7.51927
8	170.368	99.2618
9	181.688	5.60274
10	189.528	7.49648

Modelling time to capture with age



Model-based inference

Given a collection of observations y_i for $i = 1, \ldots, n$. In a model-based inference framework

- We assume a known **stochastic model** for the data. For example,

$$Y_i \sim Exponential(\theta).$$

- The stochastic model have **unknown parameters**.

$$\theta : \text{Expected time-to-capture.}$$

- Likelihood inference is one way to obtain **estimates** about the parameters from the data. For example,

$$\hat{\theta} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Some stochastic models

Poisson model

$$Pr(Y_i = y; \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}$$

Exponential model

$$f_{Y_i}(y; \theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right)$$

Bernoulli model

$$Pr(Y_i = y; \pi) = y^{\pi}(1 - y)^{1-\pi}$$

Likelihood and log-likelihood functions

Likelihood function

Denoted as $L(\theta)$, is the chance (probability) of observing the data given a specific value of the parameter θ.

Notice that:

- $L(\theta) : \Theta \to \mathbb{R}^+$ is a function with respect to θ.
- $L(\theta)$ depends on the assumed stochastic model.

Log-likelihood function

Denoted as $l(\theta)$ is the logarithm of the likelihood function.

$$l(\theta) = \ln L(\theta).$$

Likelihood function

Discrete random variables

Given a set of observations $y_1, \ldots, y_n$, then the likelihood function is

$$L(\theta) = \Pr(Y_1 = y_1, \ldots, Y_n = y_n; \theta).$$

Continuous random variables

Given a set of observations $y_1, \ldots, y_n$, then the likelihood function is

$$\begin{aligned} L(\theta) &= \Pr(y_1 - \epsilon < Y_1 < y_1 + \epsilon, \ldots, y_n - \epsilon < Y_n < y_n + \epsilon; \theta) \\ &\approx (2 * \epsilon)^n f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) \\ &\propto f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta). \end{aligned}$$

Likelihood function a vector parameter

In real-world applications, we usually work with models than have more than one parameter. We consider it a vector parameter

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix}.$$

In those case, we will denote the likelihood function as

$$L(\boldsymbol{\theta})$$

and the log-likelihood function as

$$l(\boldsymbol{\theta}).$$

Time-to-capture model

Let consider that the time-to-capture observations y_i.

Assumed stochastic model

$$Y_i \overset{iid}{\sim} Exponential(\theta), \quad \text{for } \theta > 0 \text{ and } i = 1, \ldots, 10.$$

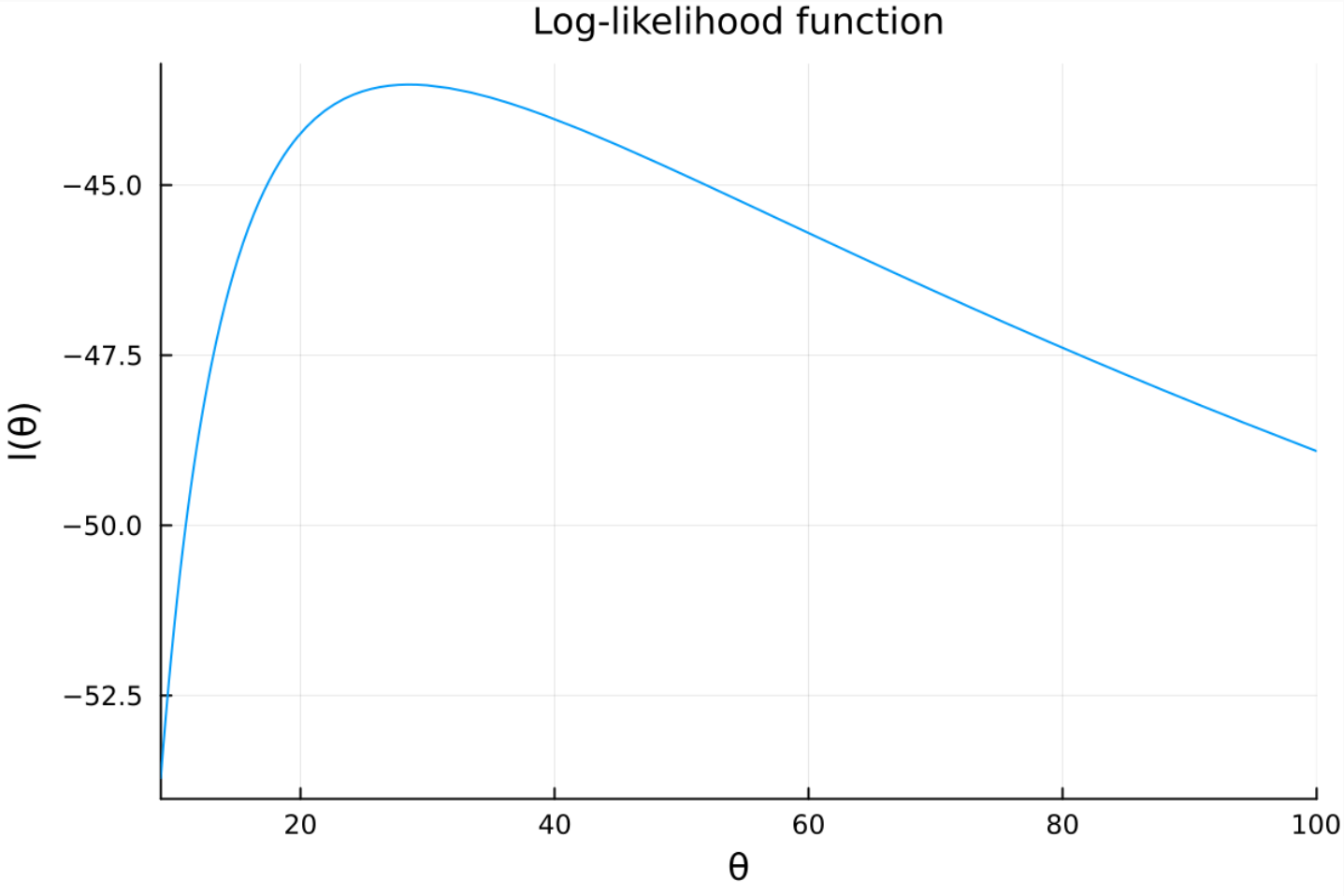
Likelihood function

$$\begin{aligned} L(\theta) &= f_{Y_1,\ldots,Y_{10}}(y_1, \ldots, y_n) \\ &= \prod_{i=1}^{10} f_{Y_i}(y_i) && \text{(due to independence)} \\ &= \prod_{i=1}^{10} \frac{1}{\theta} \exp\left(-\frac{y_i}{\theta}\right) \\ &= \theta^{-10} \exp\left(-\frac{\sum_{i=1}^{n} y_i}{\theta}\right). \end{aligned}$$

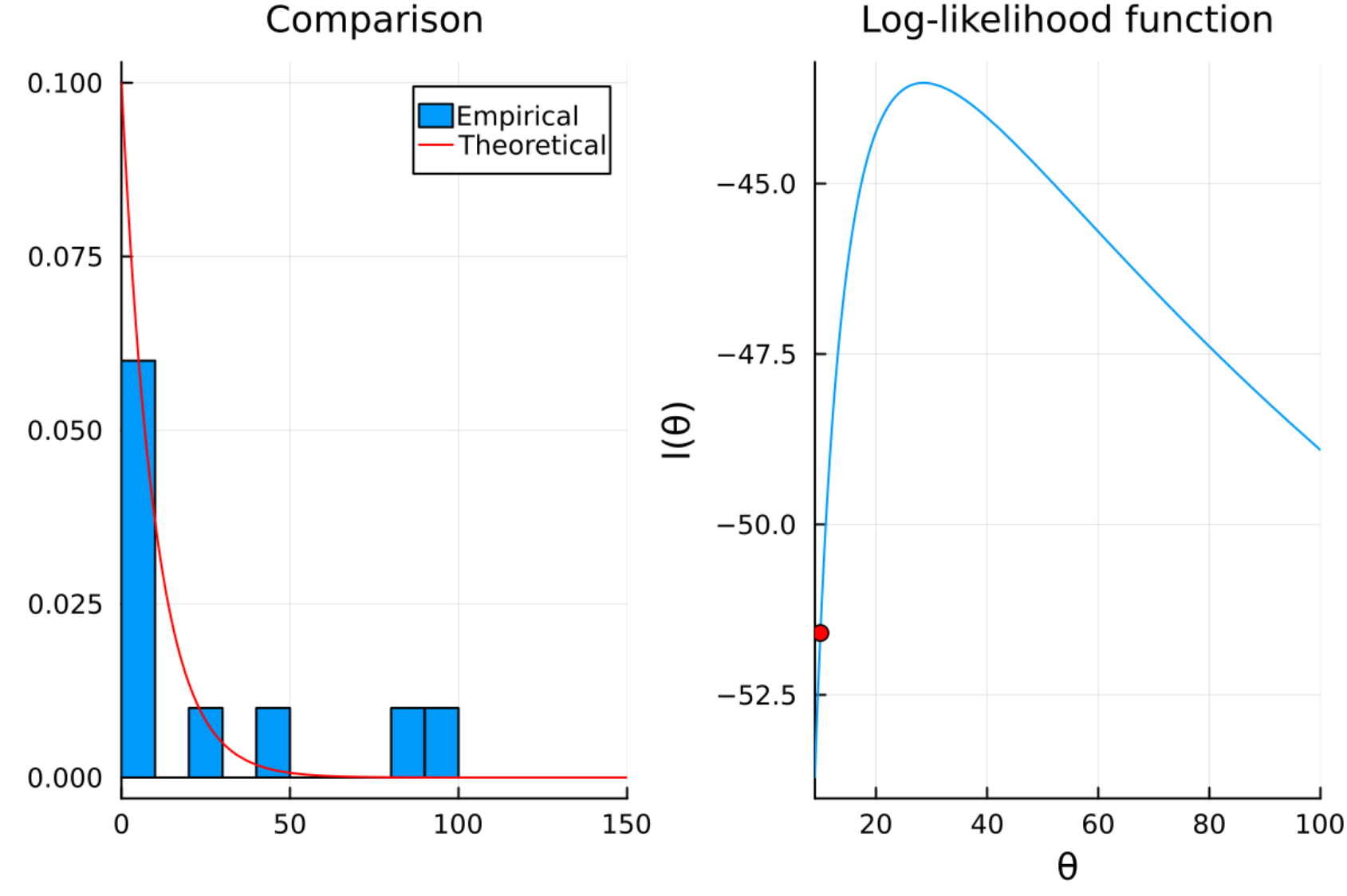
Time-to-capture model

Log-likelihood function

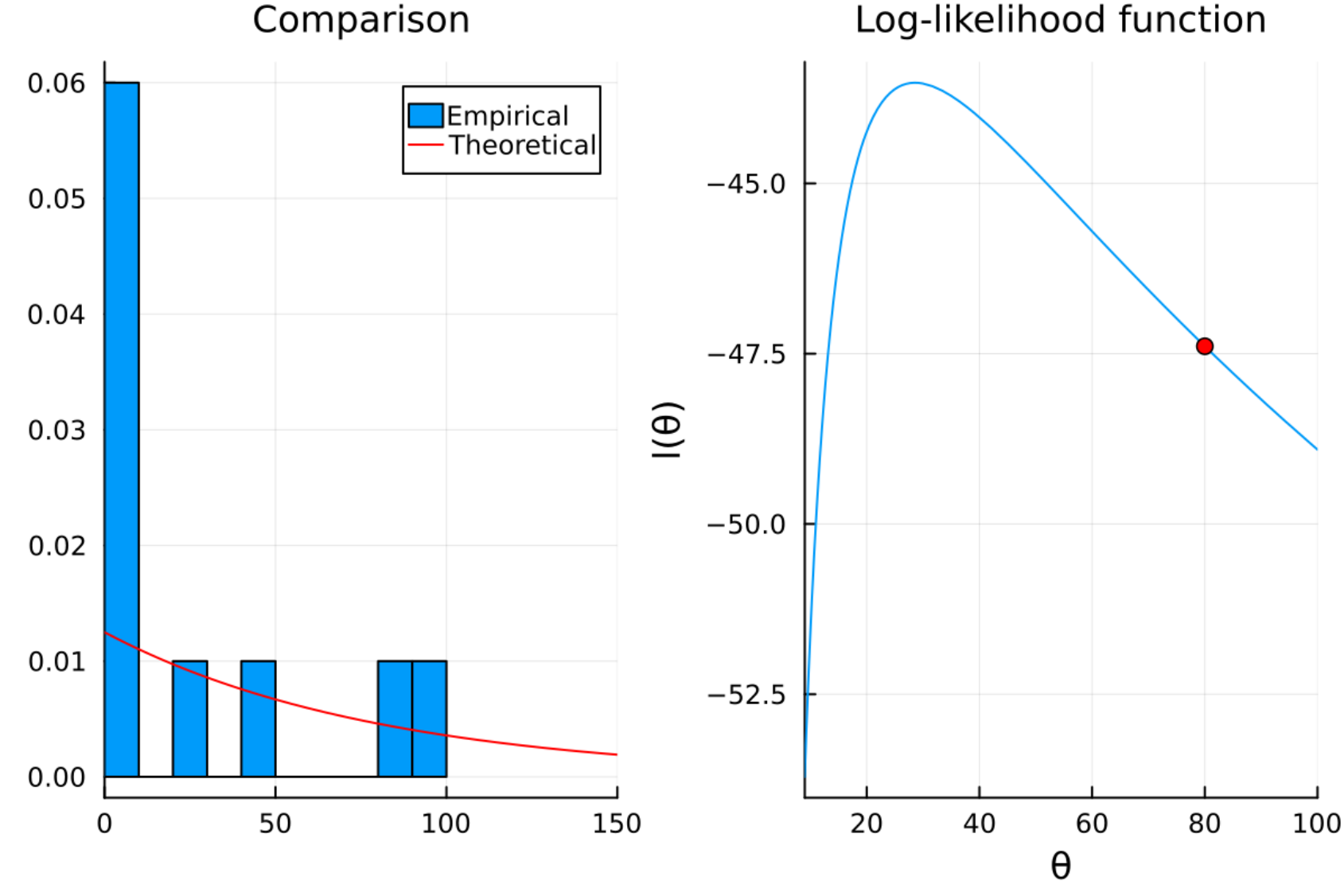
$$l(\theta) = -10 \log(\theta) - \frac{\sum_{i=1}^{n} y_i}{\theta}$$



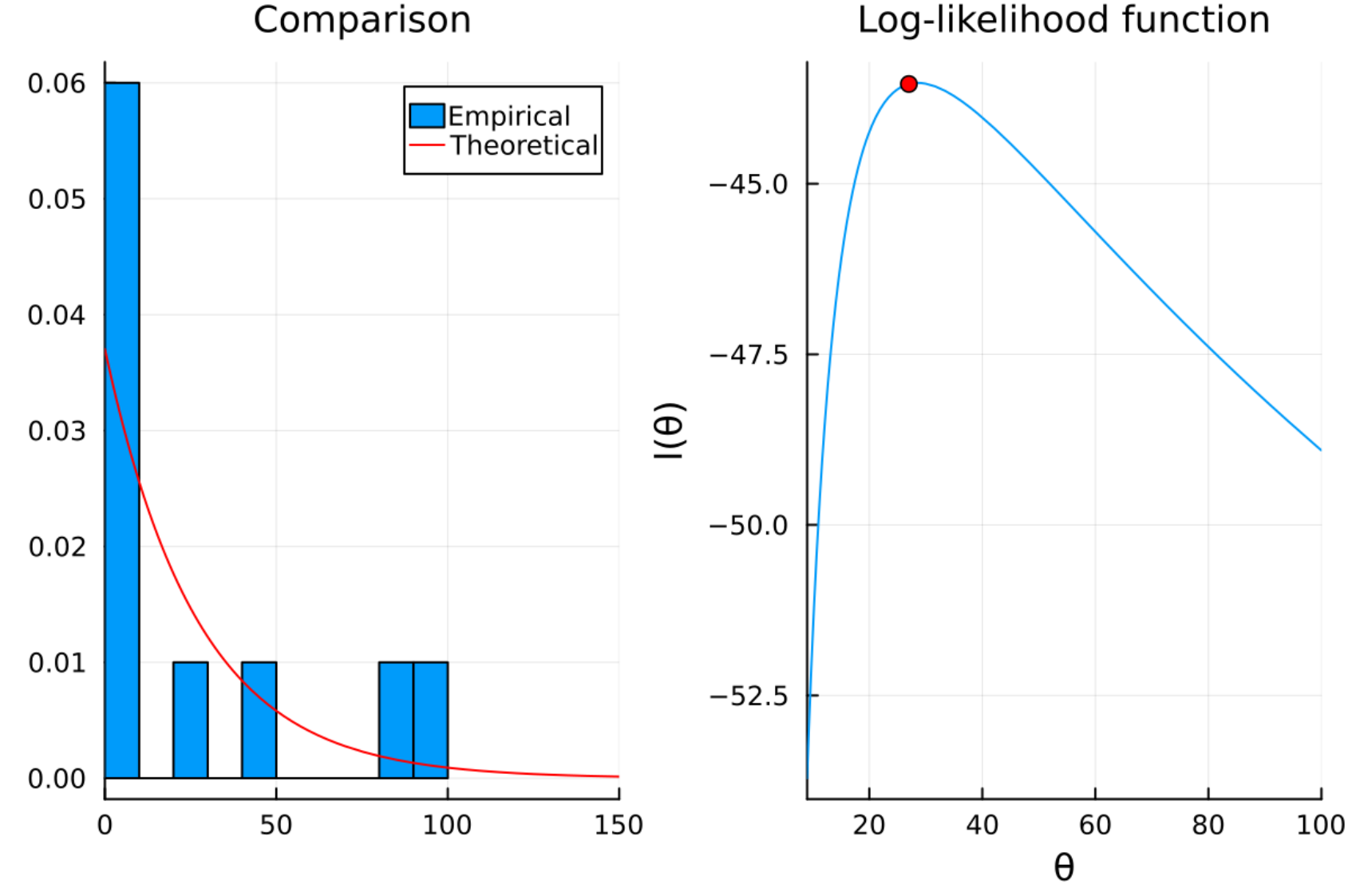
Time-to-capture model: Interpret θ = 10



Time-to-capture model: Interpret θ = 80



Time-to-capture model: Interpret θ = 27



Time-to-capture model

Let consider that the time-to-capture observations y_i.

Assumed stochastic model

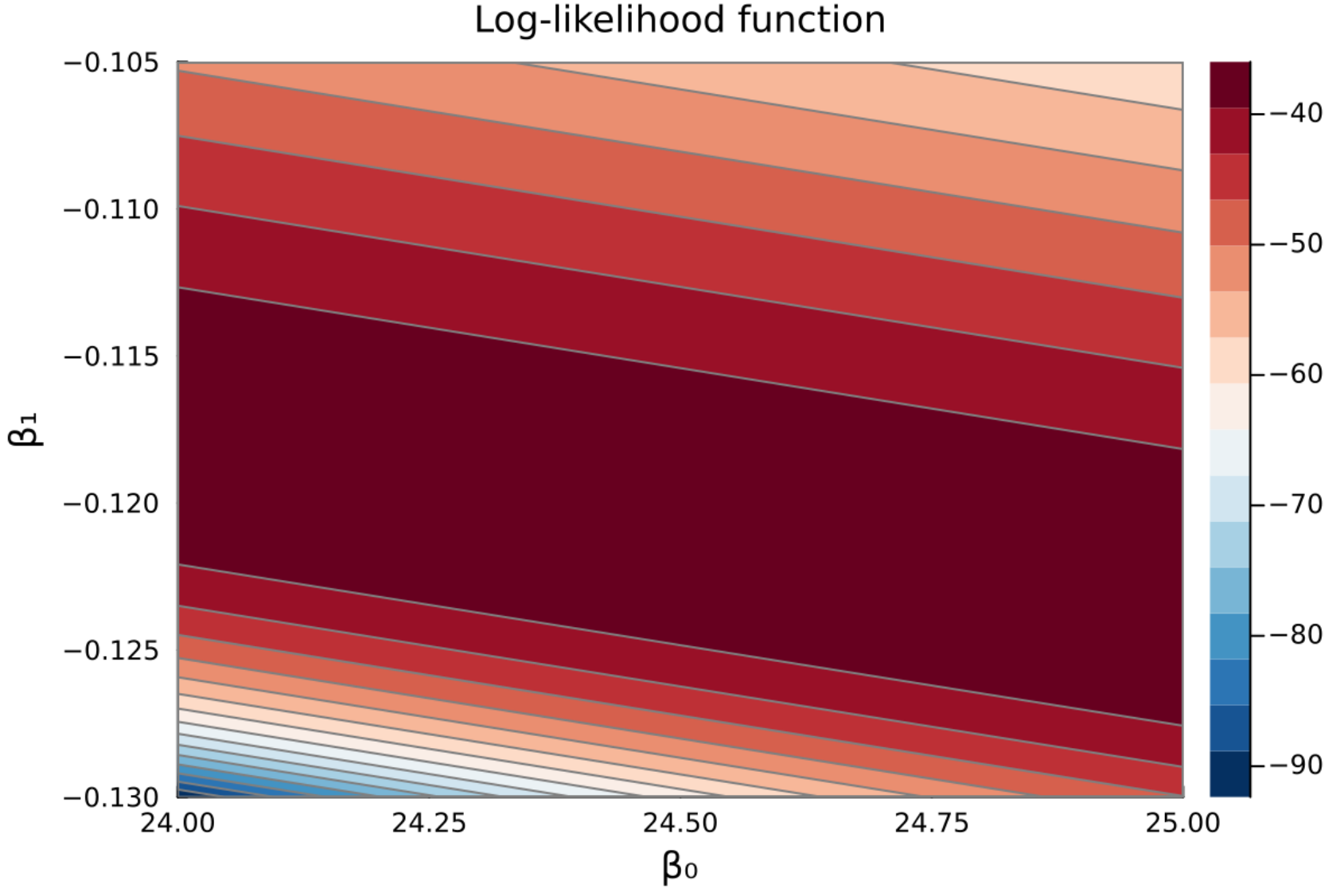
$$Y_i \overset{iid}{\sim} Exponential(\mu_i), \quad \text{for } \mu_i > 0 \text{ and } i = 1, \ldots, 10.$$

$$log(\mu_i) = \beta_0 + \beta_1 age_i$$

Likelihood function

$$\begin{aligned} L(\beta_0, \beta_1) &= f_{Y_1,\ldots,Y_{10}}(y_1, \ldots, y_n) \\ &= \prod_{i=1}^{10} f_{Y_i}(y_i) \qquad \text{(due to independence)} \\ &= \prod_{i=1}^{10} \frac{1}{\exp(\beta_0 + \beta_1 age_i)} \exp\left(-\frac{y_i}{\exp(\beta_0 + \beta_1 age_i)}\right). \end{aligned}$$

Time-to-capture model with predictor



Score function and information

Score function

$$S_i(\theta) = \frac{\partial}{\partial \theta_i} \log L(\theta)$$

Observed information

$$I_{Oij}(\theta) = -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L(\theta)$$

Fisher information

$$I_{Eij}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L(\theta)\right]$$

Maximum likelihood estimation

Maximum likelihood estimate

Definition

Given a set of observations $y_1, \ldots, y_n$, then the maximum likelihood estimate is obtained by getting the value of θ that maximizes the likelihood function,

$$\hat{\theta}_{ml} = \arg\max_{\theta} L(\theta).$$

It is equivalent to the value that maximizes the log-likelihood function,

$$\hat{\theta}_{ml} = \arg\max_{\theta} l(\theta).$$

The idea it to obtain the value of θ under which it is more likely to have observed the data $y_1, \ldots, y_n$.

Maximum likelihood estimation

Score function

It is the gradient vector of the log-likelihood function with respect to the parameter,

$$U(\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} l(\theta) \\ \frac{\partial}{\partial \theta_2} l(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_p} l(\theta). \end{bmatrix}$$

Due to the definition of the MLE, then

$$U(\hat{\theta}_{ml}) = \mathbf{o}.$$

Solving the previous equation system allow us to obtain the maximum and minimum local points of $l(\theta)$.

Maximum likelihood estimator

Maximum likelihood estimator (MLE)

It is a random variable used to make inference about a parameter. It is asymptotically normally distributed under regular conditions:

$$\hat{\theta}_{ml} \sim MVN(\theta_0, I_E(\hat{\theta})^{-1});$$

where θ_0 is the true parameter value.

In case $\boldsymbol{I}_E(\theta_0)$ is not available, it is usually replaced by another aymptotically equivalent term. For example, $\boldsymbol{I}_E(\hat{\theta}), \boldsymbol{I}_O(\theta_0), \boldsymbol{I}_O(\hat{\theta})$. This result is used to obtain confidence intervals and perform hypothesis testing.

Examples

- Generalized linear models
- Generalized additive models
- Time series models
- Survival analysis
- Point processes
- Spatial models

Expectation-Maximization algorithm

Species modelling

In an study of 4 animal species, it is know that the probability to belong to each category es $\frac{1}{2} + \frac{\theta}{4}$, $\frac{1-\theta}{4}$, $\frac{1-\theta}{4}$ and $\theta/4$. Estimate the value of θ given that, from a total of 197, we observed $125, 18, 20, 34$ animales respectively.

Likelihood function

$$L(\theta; \mathbf{x}) = \frac{197!}{125!18!20!34!}\left(\frac{1}{2} + \frac{\theta}{4}\right)^{125}\left(\frac{1-\theta}{4}\right)^{18+20}\left(\frac{\theta}{4}\right)^{34}$$
$$\propto (2+\theta)^{125}(1-\theta)^{38}\theta^{34},$$
$$L(\theta; \mathbf{x}, y) = \frac{197!}{y!(125-y)!18!20!34!}\left(\frac{\theta}{4}\right)^{y}\left(\frac{1}{2}\right)^{125-y}\left(\frac{1-\theta}{4}\right)^{38}\left(\frac{\theta}{4}\right)^{34}$$
$$\propto (1-\theta)^{38}\theta^{y+34}.$$

Definition

Definition

The EM algorithm is an **iterative** algorithm to obtain **maximum likelihood estimates** under the presence of **missing data**.

Notes

- It is iterative.
- It converges to a local maximum, being a maximum likelihood estimate when the likelihood function is unimodal.
- It is used in incomplete data problems; for instance, mixture models.

The algorithm

The name EM is used because it performs two steps at each iteration.

- E-step: Expectation
- M-step: Maximization

Instead of using $l(\theta; \boldsymbol{x})$ such as

$$\hat{\theta}_{ml} = \arg\max l(\theta; \boldsymbol{x}).$$

We can think of an additional data that has not been observed to obtain

$$l(\theta; \boldsymbol{x}, \boldsymbol{y})$$

to perform the estimation. These augmented log-likelihood should be easier to work with but unfortunately $\boldsymbol{y}$ is not available, so we will take the expected value.

EM algorithm

Let θ^* be the current estimate of θ, and consider

$$Q(\theta, \theta^*) = \int l(\theta; x, y) f(y \mid x; \theta^*) dy, \tag{1}$$

which is the expected value of $l(\theta; x, y)$ with respect to $f(y \mid x; \theta^*)$.

EM algorithm

1. Set-up an initial estimate θ^*
2. Estimate $l(\theta; x, y)$ by taking the expectation with respect Y and conditioning in $X = x, \theta = \theta^*$. $Q(\theta, \theta^*)$.
3. Maximize the estimated log-likelihood $Q(\theta, \theta^*)$ with respect to θ to update it.
4. Repeats steps until converge.

EM algorithm

Then the steps at each iteration are

- **E-step**: Compute $Q(\theta, \theta^*)$ as a function of θ only.
- **M-step**: Maximize $Q(\theta, \theta^*)$ with respect to θ to get θ^{**}.

Species modelling

In an study of 4 animal species, it is know that the probability to belong to each category es $\frac{1}{2} + \frac{\theta}{4}$, $\frac{1-\theta}{4}$, $\frac{1-\theta}{4}$ and $\theta/4$. Estimate the value of θ given that, from a total of 197, we observed $125, 18, 20, 34$ animales respectively.

$$L(\theta; x) = \frac{197!}{125!18!20!34!}\left(\frac{1}{2} + \frac{\theta}{4}\right)^{125}\left(\frac{1-\theta}{4}\right)^{18+20}\left(\frac{\theta}{4}\right)^{34}$$
$$\propto (2+\theta)^{125}(1-\theta)^{38}\theta^{34}.$$

$$L(\theta; x, y) = \frac{197!}{y!(125-y)!18!20!34!}\left(\frac{\theta}{4}\right)^{y}\left(\frac{1}{2}\right)^{125-y}\left(\frac{1-\theta}{4}\right)^{38}\left(\frac{\theta}{4}\right)^{34}$$
$$\propto (1-\theta)^{38}\theta^{y+34}.$$

Example

$$l(\theta; x, y) = c + 83 \log(1 - \theta) + (y + 34) \log(\theta).$$

Expectation-Step:

$$\mathbb{E}\left[l(\theta; x, y) \mid x, \theta^*\right] = c + 83 \log(1 - \theta) + (\mathbb{E}\left[y \mid x, \theta^*\right] + 34) \log(\theta)$$

$$\Pr(z \in A \mid z \in G1, \theta^*) = \frac{\theta^*/4}{1/2 + \theta^*/4} = \frac{\theta^*}{2 + \theta^*}$$

$$y \mid x, \theta^* = \mathit{Binomial}\left(125, \frac{\theta^*}{2 + \theta^*}\right)$$

$$\mathbb{E}\left[l(\theta; x, y) \mid x, \theta^*\right] = c + 83 \log(1 - \theta) + \left(\frac{125\theta^*}{2 + \theta^*} + 34\right) \log(\theta)$$

$$Q(\theta, \theta^*) = c + 83 \log(1 - \theta) + \left(\frac{125\theta^*}{2 + \theta^*} + 34\right) \log(\theta).$$

Example

Maximization-Step:

$$\frac{dQ(\theta, \theta^*)}{d\theta} = -\frac{83}{1-\theta} + \left(\frac{125\theta^*}{2+\theta^*} + 34\right)\frac{1}{\theta}.$$

$$0 = -\frac{83}{1-\theta^{**}} + \left(\frac{125\theta^*}{2+\theta^*} + 34\right)\frac{1}{\theta^{**}}.$$

$$\theta^{**} = \frac{125\theta^*/(2+\theta^*) + 34}{125\theta^*/(2+\theta^*) + 34 + 38}.$$

Iterating the previous equation until converge perform the Expectation and Maximization steps implicitely.

Theory

Why does EM work?

We can proof the at each iteration the log-likelihood function is increasing.

$$Q(\theta^{**}, \theta^{*}) - Q(\theta^{*}, \theta^{*}) \geq 0,$$

$$\int \log(f(x, y; \theta^{**}))f(y \mid x; \theta^{*})dy - \int \log(f(x, y; \theta^{*}))f(y \mid x; \theta^{*})dy \geq 0,$$

$$\int \log\left(\frac{f(x, y; \theta^{**})}{f(x, y; \theta^{*})}\right) f(y \mid x; \theta^{*})dy \geq 0.$$

Considering that $f(x, y; \theta) = f(y \mid x; \theta)f(x; \theta)$,

$$\int \log\left(\frac{f(y \mid x; \theta^{**})}{f(y \mid x; \theta^{*})}\right) f(y \mid x; \theta^{*})dy + \int \log\left(\frac{f(x; \theta^{**})}{f(x; \theta^{*})}\right) f(y \mid x; \theta^{*})dy \geq 0.$$

Given that $\log(x) \leq x - 1$,

$$\int \log\left(\frac{f(x; \theta^{**})}{f(x; \theta^{*})}\right) f(y \mid x; \theta^{*})dy \geq 0$$

$$\log(f(x; \theta^{**})) - \log(f(x; \theta^{*})) \geq 0.$$

Estandard error

The covariance matrix of the estimator is approximated by the inverse of the observed information evaluated at $\hat{\theta}$:

$$\left[-\frac{\partial^2 \log f(x;\theta)}{\partial \theta_i \theta_j} \right]^{-1}.$$

$$f(x;\theta) = f(x,y;\theta)/f(y \mid x;\theta)$$

$$-\log f(x;\theta) = -\log f(x,y;\theta) - \{-\log f(y \mid x;\theta)\}$$

$$-\frac{d^2}{d\theta^2} \log f(x;\theta) = -\frac{d^2}{d\theta^2} \log f(x,y;\theta) - \{-\frac{d^2}{d\theta^2} \log f(y \mid x;\theta)\}$$

Estandard error

Multiplying by $f(y \mid x, \phi)$ and integrate with respect to y.

$$-\frac{d^2}{d\theta^2}\log f(x;\theta) = -\int \frac{d^2}{d\theta^2} f(y \mid x, \phi)\log f(x, y;\theta)dy - \int \{-\frac{d^2}{d\theta^2}\log f(y \mid x;\theta)\} f(y \mid x, \phi)dy$$

$$-\frac{d^2}{d\theta^2}\log f(x;\theta) = -\frac{d^2}{d\theta^2}Q(\theta, \phi) - -\frac{d^2}{d\theta^2}H(\theta, \phi).$$

Which can be computed considering that

$$\begin{aligned}
-\frac{d^2}{d\theta^2}H(\theta, \phi)|_{\hat{\theta}} &= \mathbb{V}\left[\frac{d\log f(y \mid x, \theta)}{d\theta} \mid x, \hat{\theta}\right] \\
&= \mathbb{V}\left[\frac{d\log f(y, x \mid \theta)}{d\theta} - \frac{d\log f(x;\theta)}{d\theta} \mid x, \hat{\theta}\right] \\
&= \mathbb{V}\left[\frac{d\log f(y, x \mid \theta)}{d\theta} \mid x, \hat{\theta}\right].
\end{aligned}$$

Example

For the species modelling example:

$$-\frac{d^2}{d\theta^2} Q(\theta, \phi)|_{\hat{\theta}} = \frac{\mathbb{E}\left[Y \mid x_1; \hat{\theta}\right] + x_4}{\hat{\theta}^2} + \frac{x_2 + x_3}{(1 - \hat{\theta}^2)}.$$

$$\begin{aligned}
\mathbb{V}\left[\frac{d \log f(y \mid x, \theta)}{d\theta} \mid x, \hat{\theta}\right] &= \mathbb{V}\left[(Y + x_4)/\hat{\theta} - (x_2 + x_3)/(1 - \theta) \mid x_1, \hat{\theta}\right] \\
&= \frac{\mathbb{V}\left[Y \mid x_1, \hat{\theta}\right]}{\hat{\theta}^2}, \\
&= \frac{125}{\hat{\theta}^2}\left(\frac{\hat{\theta}}{2 + \hat{\theta}}\right)\left(\frac{2}{2 + \hat{\theta}}\right).
\end{aligned}$$

Examples

- Augmented data.
- Censored data.
- Mixtures.
- Random effect models.
- Missing data.