Computational Statistics
Project 2

The code for this project is available under
https://github.com/max607/computational-statistics-em.

By
Maximilian Schneider

01.02.2023

Contents

1 Maximum Likelihood estimation of θ

1.1 Notation

In the following random variables (RV) are denoted with capital letters, e.g. Y, their realizations with lowercase letters, e.g., y_i, where always $i = 1, ..., n$. Loglikelihoods are written, e.g., as $\ell(\theta)$. For parameters, Greek letters are used and their estimators are distinguished with a hat, e.g., $\hat{\theta}$. In the context of this project, derivatives are only taken when functions are viewed as functions of one variable, denoted, e.g., as $\ell'(\theta)$.

1.2 Likelihood

Given are 150 one dimensional data. Figure 1.1 shows a histogram of them. They exhibit a positive skew.

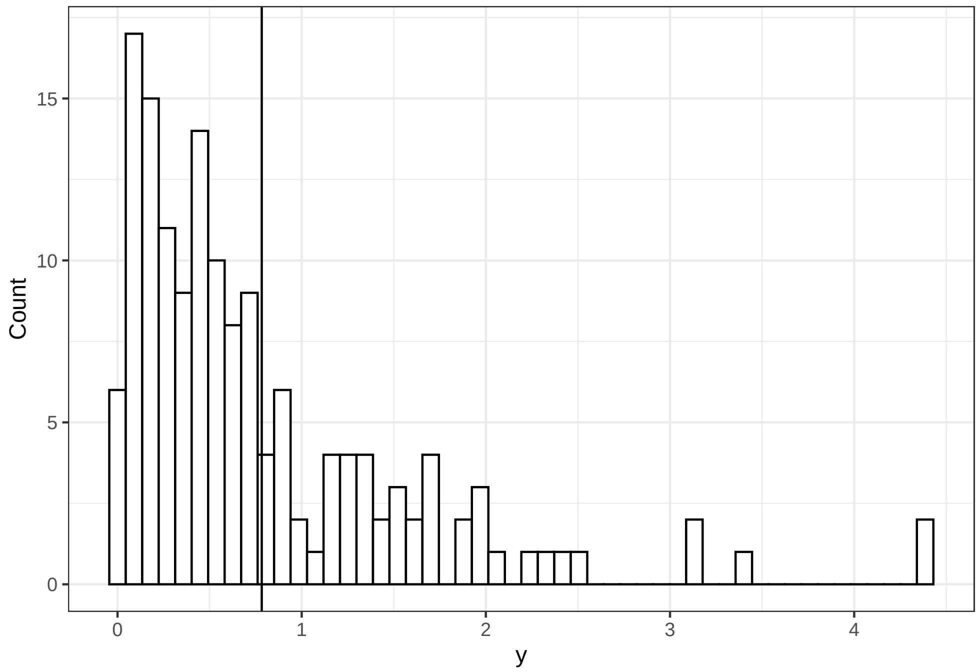

Figure 1.1: Histogram of provided data. The black line indicates the sample mean.

It is assumed they are independently identically distributed (i.i.d.) realizations y_i of a RV Y with probability density function (PDF)

$$f_\theta(y_i) = \frac{\theta^2}{\theta + 1}(1 + y_i)\exp(-\theta y_i), \quad y_i, \theta > 0. \tag{1.1}$$

The goal is to estimate θ via maximum likelihood (ML). Its loglikelihood

$$
\begin{aligned}
\ell(\theta) &= \log\left(\prod_{i=1}^{n} f_\theta(y_i)\right) \\
&= 2n\log(\theta) - n\log(\theta+1) - \theta\sum_{i=1}^{n} y_i + c \\
&\propto 2\log(\theta) - \log(\theta+1) - \theta\bar{y} + c, \qquad (1.2)
\end{aligned}
$$

where c is a constant, which does not depend on θ, and thus is irrelevant for the following calculations. The last term (1.2) results by dividing the loglikelihood by the number of observations n, so $\bar{y}$ denotes the sample mean.
The first derivative of (1.2)

$$
\ell'(\theta) = \frac{2}{\theta} - \frac{1}{\theta+1} - \bar{y}. \qquad (1.3)
$$

There is no analytical solution available for equating (1.3) to zero and solving for θ.

1.3 Newton-Raphson

The maximum of $\ell(\theta)$ has to be found numerically, here via Newton-Raphson, i.e., iteratively applying

$$
\theta^* = \theta - \frac{\ell'(\theta)}{\ell''(\theta)}, \qquad (1.4)
$$

where $\ell''(\theta) = -\frac{2}{\theta^2} + \frac{1}{(\theta+1)^2}$, θ is the value at iteration t and θ^* is the updated value at iteration $t+1$, until the update gets very close to zero. The final value is taken as the ML estimate.

For the given data $\hat{\theta} = 1.7424899$.

2 Bootstrapping for standard error of $\hat{\theta}$

2.1 Sampling from $f_{\hat{\theta}}(y_i)$

The goal is to quantify the uncertainty of $\hat{\theta}$. For this parametric bootstrap is applied, for which sampling from $f_{\hat{\theta}}(y_i)$ is necessary. The starting point is restating (1.1) as

$$f_\theta(y_i) = \frac{\theta}{\theta+1}\theta\exp(-\theta y_i) + \frac{1}{\theta+1}\theta^2 y_i \exp(-\theta y_i) \tag{2.1}$$

and recognizing this as a mixture of two gamma distributions in shape and rate parameterization, where θ is the rate and the shapes are equal to one and two, respectively.

Starting from first principals, it is assumed only RVs $U \overset{iid}{\sim} U(0,1)$ are available. This is not to much of a hassle, as Gamma RVs with shape j are the sum of j Exponential RVs, which in turn can be easily obtained via inversion

$$f^{-1}(u;\theta) = -\frac{\log(u)}{\theta}, \tag{2.2}$$

where θ already is the desired rate.

This is implemented in the following steps:

1) Draw n u_i.
2) Calculate n temporary $y_i = f^{-1}(u_i;\hat{\theta})$ with shape one.
3) Draw the number of shape two gammas n_2.
    1) Draw another n u_j.
    2) $n_2 = \#\{u_j | u_j < \frac{1}{\hat{\theta}+1}\}$.
4) Draw n_2 u_k.
5) Calculate n_2 temporary $y_k = f^{-1}(u_k;\hat{\theta})$ with shape one.
6) Add y_ks component-wise to the fist n_2 y_i.
7) Return y_is.

Returned is a sample of size n, which can be seen as a realization of Y_i with PDF $f_\theta(y_i)$. n_2 observations are realizations of Gamma RVs with shape two and $n - n2$ observation of are realizations of Gamma RVs with shape one.

For the purpose of estimating θ with the estimator of section 1.3 it is of no importance that the sample is sorted by shape.

2.2 Bootstrap standard error

Given B samples of size n from the sampler of the previous section, and corresponding bootstrap estimates $\hat{\theta}_b^*,\ b = 1, ..., B$, the bootstrap standard error

$$\widehat{\mathrm{se}}_B(\hat{\theta}) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*)^2}, \tag{2.3}$$

according to the lecture slides, where $\hat{\theta}^*$ is the sample mean of all $\hat{\theta}_b^*$.

For this simulation $B = 10^4$ and the resulting $\widehat{\mathrm{se}}_B(\hat{\theta}) = 0.1142654$.

3 EM

3.1 Augmented data

Consider a more complex PDF for the given data

$$f_\pi(y_i) = \pi f_\theta(y_i) + (1-\pi) f_\lambda(y_i), \quad y_i > 0, \pi \in [0,1], \tag{3.1}$$

where $f_\theta(y_i)$ is the PDF from before and $f_\lambda(y_i) = \lambda \exp(-\lambda y_i)$, i.e., $f_\pi(y_i)$ is a mixture of two PDFs.
The goal is to estimate θ, λ and π using an EM algorithm.

Start by introducing augmented data x_i and assume they are realizations from $X_i \sim Ber(\pi)$. An equivalent formulation of (3.1) thus is

$$f(\theta, \lambda | x_i, y_i) = f_\theta(y_i)^{x_i} f_\lambda(y_i)^{1-x_i}, \tag{3.2}$$

where the likelihood of θ and λ is computed given x_i, y_i. I.e., simulation from $f_\pi(y_i)$ given θ, λ and π is possible, by first drawing x_i and then drawing from $f_\theta(y_i)$ if $x_i = 1$ or $f_\lambda(y_i)$ if $x_i = 0$.

3.2 Expectation

For the first part of the EM algorithm an expression for $\mathbb{E}(X_i | \pi, \theta, \lambda, y_i)$ is needed. This is obtained by applying Bayes' theorem

$$\begin{aligned} f(x_i | \theta, \lambda, \pi, y_i) &= \frac{f(\theta, \lambda | x_i, y_i) f(x_i | \pi)}{f(\theta, \lambda | y_i)} \\ &= \frac{f_\theta(y_i)^{x_i} f_\lambda(y_i)^{1-x_i} \ \pi^{x_i} (1-\pi)^{1-x_i}}{(1-\pi) f_\lambda(y_i) + \pi f_\theta(y_i)}. \end{aligned} \tag{3.3}$$

The second line is implied by the distribution of X_i and by the fact that $f(\theta, \lambda | y_i) = f(x_i = 0 | \pi) f(\theta, \lambda | x_i = 0, y_i) + f(x_i = 1 | \pi) f(\theta, \lambda | x_i = 1, y_i)$.
One can immediately see

$$\mathbb{E}(X_i | \pi, \theta, \lambda, y_i) = \frac{f_\theta(y_i) \pi}{(1-\pi) f_\lambda(y_i) + \pi f_\theta(y_i)}. \tag{3.4}$$

3.3 Maximization

The second step is simple maximum likelihood estimation given the augmented data. The loglikelihoods of the respective parameters are

$$\ell(\pi) = \log(\pi) \sum_{i=1}^{n} x_i + \log(1-\pi) \sum_{i=1}^{n} (1-x_i), \tag{3.5}$$

$$\ell(\lambda) = n \log(\lambda) - \lambda \sum_{i=1}^{n} y_i - \log(\lambda) \sum_{i=1}^{n} x_i + \lambda \sum_{i=1}^{n} x_i y_i + c, \tag{3.6}$$

$$\begin{aligned} \ell(\theta) &= 2\log(\theta) \sum_{i=1}^{n} x_i - \log(\theta+1) \sum_{i=1}^{n} x_i - \theta \sum_{i=1}^{n} x_i y_i + c \\ &\propto 2\log(\theta) - \log(\theta+1) - \theta\tilde{y} + c, \end{aligned} \tag{3.7}$$

where $\tilde{y} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i$.

The maximizer of (3.5) is $\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
It is also possible to maximize (3.6) analytically, via $\hat{\lambda} = (n - \sum_{i=1}^{n} x_i)/(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i)$.
This can be interpreted as the inverse of the sample mean of the y_i, where $x_i = 0$.
(3.7) has no close form and is maximized using Newton-Raphson (see section 1.3), substituting $\bar{y}$ with $\tilde{y}$, which can be interpreted as the sample mean of the y_i, where $x_i = 1$.

3.4 Application

Terminating condition