# Visualizing the Fisher Information and the Variance of the MLE

Lukas Stein

## 1 Introduction

The aim of this work is to provide a graphical interpretation of the Variance of the MLE. Since the Variance of the MLE is inversely related to the Fisher Information (FI) of the respective distribution, I will mainly focus on deriving a visual intuition for the FI. In chapter 2, I will present two methods of calculating the FI and briefly explain their relevance to the following chapters. In 3.1 I will give some general information about the Cauchy Distribution and it's properties. The main section of this work deals with the visualization of the Cauchy distribution's log-likelihood and its derivatives in 3.2. In 4 I will discuss the implications of picking a false minimum to evaluate the log-likelihood at.

## 2 The Variance and the Fisher Information

The theorem of the asymptotic variance of the MLE states that for any extremum estimator $\hat{\theta}_N$ for the true parameter $\theta$, the asymptotic distribution is given by:

$$\hat{\theta} \xrightarrow{d} N(\theta, \frac{I(\theta)^{-1}}{N})$$

where $I(\theta)$ is the Fisher Information, which is defined as the variance of the score function[1]:

$$I(\theta) = Var(\frac{d}{d\theta}\log L(x|\theta)) = \int (\frac{d}{d\theta}\log f(x|\theta))^2 p_\theta(x)dx, \quad \text{with } p_\theta(x) = f(x|\theta) \tag{1}$$

An interpretation of this is that the Fisher Information as a function of $\theta$ describes the sensitivity of the log-likelihood function $\log L$ to changes in $\theta$ (given by the squared derivative of the log likelihood with respect to $\theta$), for all possible values of $x \in X$, weighted by their respective probability under $\theta$. It therefore describes the *expected* sensitivity of the log likelihood function to changes in $\theta$.

Under mild regularity conditions it can be shown that this is equivalent to (Ly et al. 2017):

$$I(\theta) = -E_\theta(\frac{d^2}{d\theta^2}\log f(x|\theta)) = -\int (\frac{d}{d\theta}\log f(x|\theta))^2 p_\theta(x)dx \tag{2}$$

Since we do not know the true parameter value $\theta$, we can use the aforementioned formula to *approximate* the Fisher Information using our empirical data[2]:

$$J_n(\theta) = \frac{d^2}{d\theta^2}\log L(x|\hat{\theta}) = \sum_{i=1}^{N} \frac{d^2}{d\theta^2}\log f(x_i|\hat{\theta}) \xrightarrow{p} \quad N \cdot \int (\frac{d}{d\theta}\log f(x|\theta))^2 p_\theta(x)dx = N \cdot I(\theta) \tag{3}$$

In the following chapters, I will explore how these expressions can be interpreted graphically.

[1] We define the score function as the first derivative of the likelihood function with respect to $\theta$

[2] In the multidimensional case, the observed Fisher Information is equal to the Hessian of the negative log likelihood.

# 3 Cauchy Distribution Likelihood

## 3.1 General Information about the Cauchy Distribution

The Cauchy Distribution not only makes for a relatively simple example of a (potentially) 'bumpy' likelihood function, but also has the added advantage of being relevant in real-life modeling, e.g., of stock returns on financial markets. It is often used in problems that generalize to most multimodal likelihood functions. Similarly to the normal distribution, the entire family of Cauchy distributions only differs in its two parameters: The location $\theta$ (also: $x_0$, analogous to $\mu$) and size $\gamma$ (analogous to $\sigma^2$):[3]

$$f(x, \gamma, \theta) = \frac{1}{\gamma\pi} \frac{1}{1 + (\frac{x-\theta}{\gamma})^2}$$

However, compared to the normal distribution, the Cauchy Distribution has relatively fat tails. This makes estimators like the mean, and the variance essentially useless when it comes to estimating the true parameters of the distribution from a sample. The log likelihood is:

$$\log L(x^n, \theta, \gamma) = n\log(\gamma) - n\log(\pi) - \sum_{n}^{i=1} \log[\gamma^2 + (x_i - \theta)^2]$$

Forming the derivative with respect to $\theta$ yields:

$$\frac{\delta \log L}{\delta\theta} = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{\gamma^2 + (x_i - \theta)^2} = 0$$

The equation above suggests the possibility of multiple local maxima. As Barnett (1966) explains, "[...] the fact that multiple relative maxima may occur is obvious from the form of $\delta l/\delta\theta$ [...]; $\delta l/\delta\theta$ tends to zero, from below when $\theta \to +\infty$ and from above when $\theta \to -\infty$, whilst any 'relatively isolated' observation $x_i$ must have the effect of making $\delta l/\delta\theta$ pass through zero from above, implying the presence of a relative maximum." (Barnett 1966)

The second derivative w.r.t. $\theta$:

$$\frac{\partial^2 \log L}{\partial\theta^2} = \sum_{i=1}^{n} \frac{2(\theta - \gamma - x_i)}{[(\theta - x_i)^2 + \gamma^2]^2}$$

## 3.2 Visualizing the Fisher Information on three different levels

To compare how the size parameter $\gamma$ and the sample size $n$ affect the Fisher Information contained in a sample, we will generate three different "sample types" that only differ in these two variables:

- Type 1: $x^{n=100} \overset{iid}{\sim} Ca(0, 100)$
- Type 2: $x^{n=7} \overset{iid}{\sim} Ca(0, 100)$
- Type 3: $x^{n=100} \overset{iid}{\sim} Ca(0, 50)$

Of each type, I will generate 200 samples. To be able to discuss specific aspects of the likelihood function, without relying too heavily on a single example, I will (only) graph the log-likelihoods of the first 15 samples generated for each type. For the remaining samples, I will only plot the density of values at the true parameter value. Assuming that this empirical density is an asymptotic approximation of the analytical density, also allows me to make visual arguments about the definition of the FI given by Equation (1).

The following analysis will follow the ideas and methods outlined in Zheng (n.d.) and Rich (2021). We will derive a visual intuition for the Fisher Information on three different levels by looking at the (1) log-likelihood, (2) it's first derivative w.r.t. $\theta$, the score function, and (3) it's second derivative.

[3] To express that a random variable X follows a Cauchy distribution with location $\theta$ and size $\gamma$, we write:

$$X \sim Ca(\theta, \gamma)$$

### 3.2.1 Level 1: The likelihood function

We are now plotting the log-likelihood functions for the first couple (~15) of samples (left, adjusted by their global minimum). Additionally, we can use a kernel estimator to estimate the distribution of values of the likelihood function, evaluated at the true value of $\theta$ for the different samples (right).

Each row corresponds to one of the three sample types defined above.

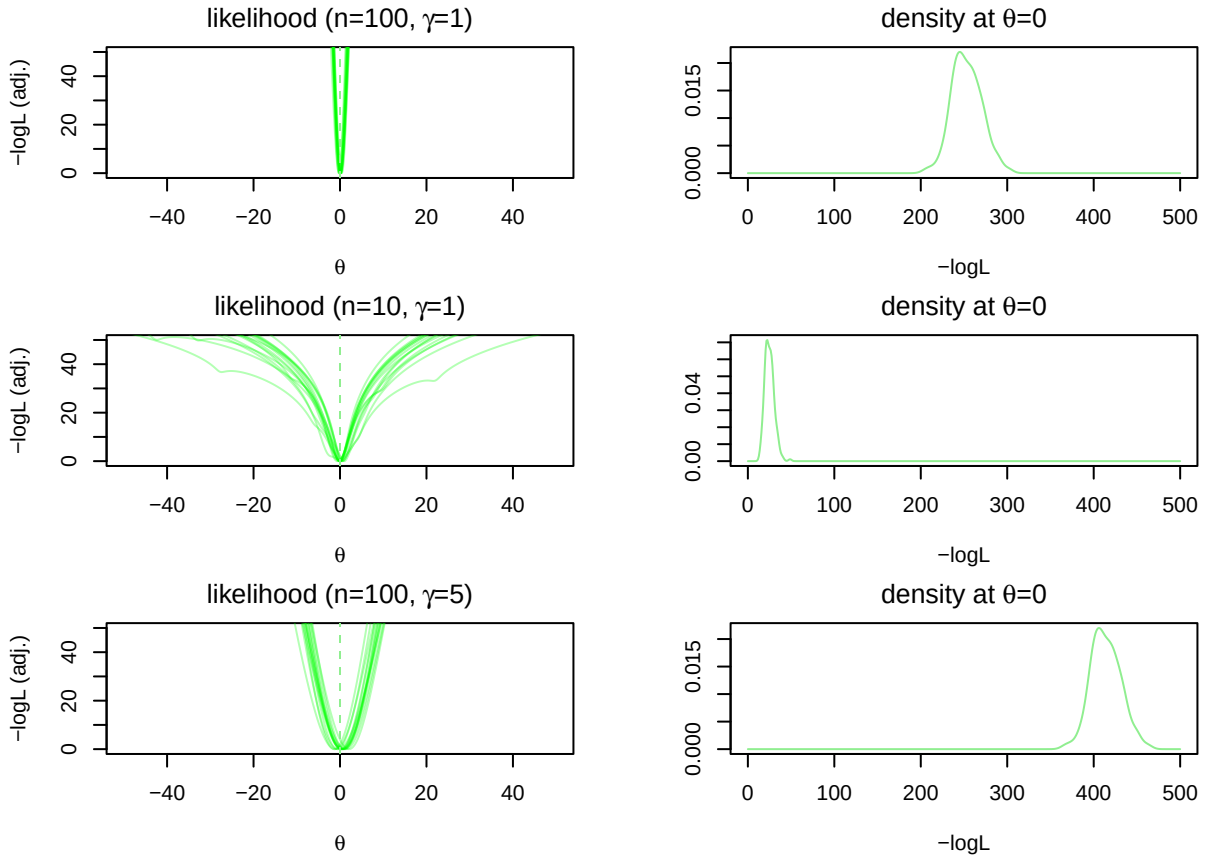


Figure 1: log-likelihood functions for the first 15 samples (left) and kernel density estimation over the likelihood functions of all samples evaluated at the true parameter value (right). Likelihood functions adjusted so their global minima lie on the x-axis.

As we can clearly see, in the low sample size case as well as the high variance case, the likelihood functions are less concentrated around the MLE (less "pointy"). It makes intuitive sense, that the resulting variance of the MLE must be lower. However, it would be difficult to try and spot the Fisher Information by just looking at the log-likelihood.

### 3.2.2 Level 2: The Score Function

Using the samples generated above, we can repeat the same process for the first derivative of the likelihood function. The resulting density estimations are asymptotic approximations of the analytical distribution of scores at the true parameter value.

The Fisher Information is defined as the variance of the aforementioned distribution and can therefore be read directly from the graph. In the following graphic, we calculate the variance of scores explicitly to estimate the FI. Note, that this is an approximation of the first definition of the FI as mentioned in Equation (1).

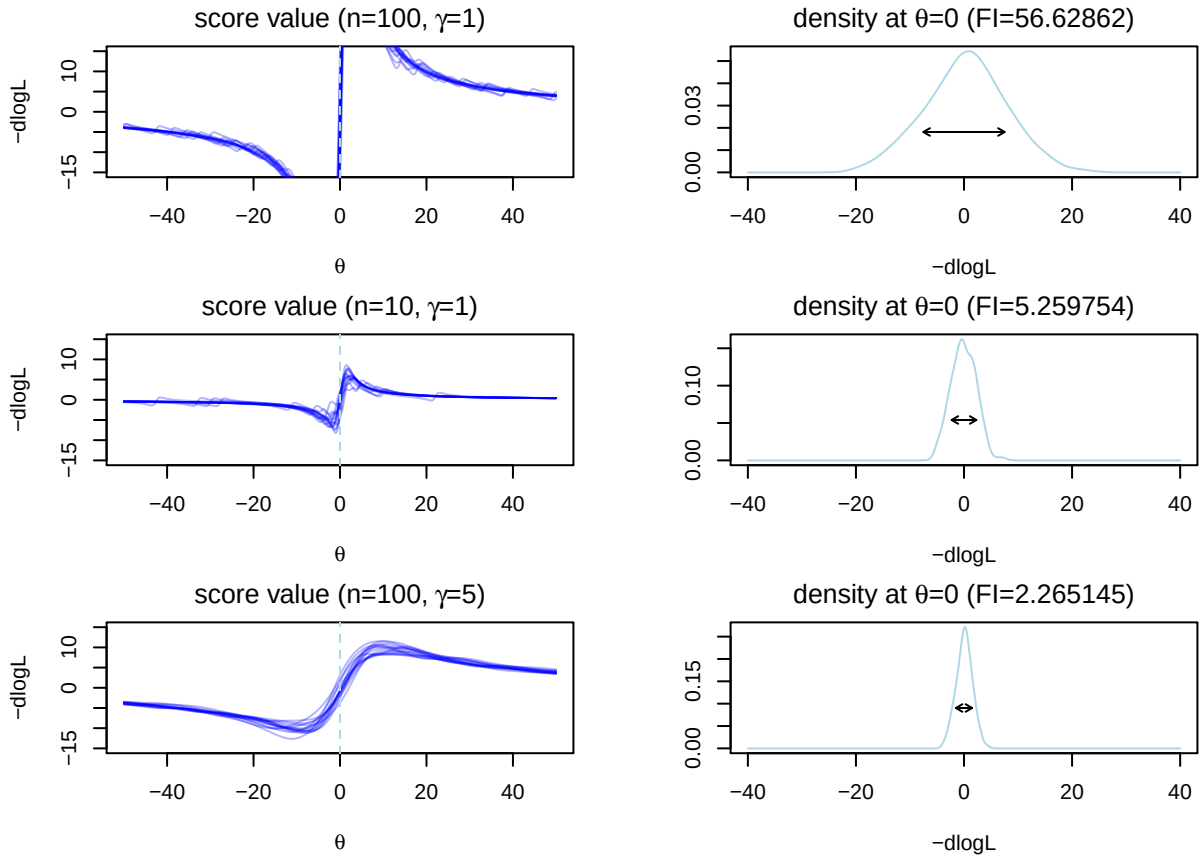


Figure 2: score functions for the first 15 samples (left) and kernel density estimation over the score functions of all samples evaluated at the true parameter value (right). Arrow with length of 2*sqrt(FI).

The fact that the distribution of scores is centered around zero confirms that $\theta = 0$ is a local minimum of the analytical likelihood.

A wider variance in scores, implies a greater Fisher Information and therefore a smaller Variance of the MLE. In the small $n$ case as well as the large $\gamma$ case, the scores are more closely concentrated around 0. This means that we are more likely to encounter a score value close to zero when evaluating the score function at the true parameter value.

It is a tempting interpretation of a lower variance in scores to assume that it means, that we are more likely to pick the true parameter value as our MLE, and therefore the variance should in fact be lower. However, this is not true, since this observation holds for all possible parameter values in (close) proximity to the true $\theta$. Instead, if scores are often close to zero, regardless of the data that produced the likelihood function, it implies that the likelihood-function around the true parameter value is (on average) relatively insensitive to the parameter (i.e., more flat). A higher variance in scores on the other hand means that the function around the true $\theta$ is relatively sensitive to different inputs (i.e., sharper).

### 3.2.3 Level 3: The Second derivative

However, when working with real data, we obviously do not have the necessary information to estimate the density function of score values. Instead, we use the second derivative of the log-likelihood.

The average (mean) FI implied by the second derivative log-likelihoods evaluated at $\theta$ across all different samples is displayed in brackets. Note, that this relates to the second definition of the FI as mentioned in Equation (2) and (3).

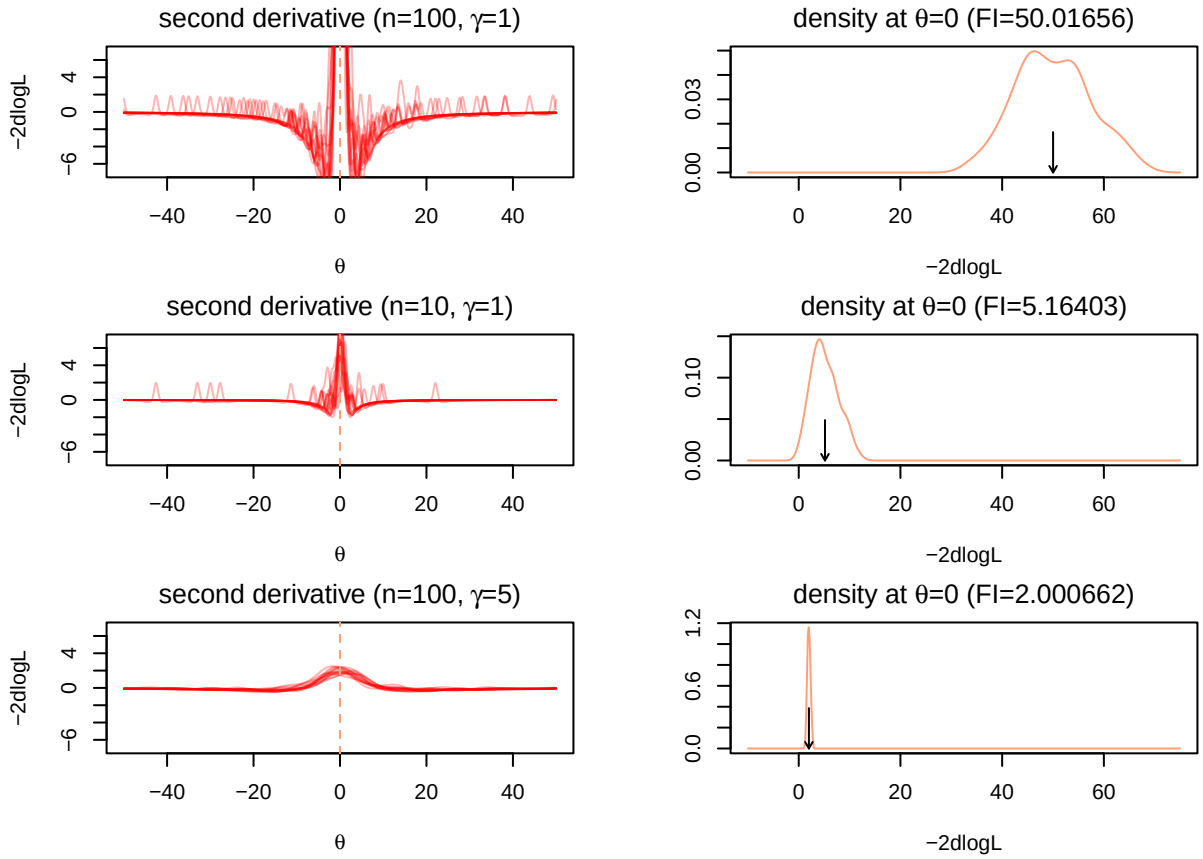


Figure 3: second derivative log-likelihood functions for the first 15 samples (left) and kernel density estimation over the second derivative log-likelihood of all samples evaluated at the true parameter value (right). Arrow points at FI.

The second derivative measures the curvature of the likelihood function at the true parameter value and therefore more closely aligns with the idea behind our initial intuition about how the variance should behave according to the degree to which the likelihood is concentrated around the MLE. A high curvature implies that the score function is highly sensitive to changes in $\theta$ and therefore its root is relatively insensitive.

As we can see here, each sample provides a different estimation for the Fisher Information but the results are asymptotically normally distributed around the true FI.

# 4 Implications

In the simulations above, both methods provide similar results when it comes to estimating the Fisher Information from the generated samples:[4]

Table 1: average Fisher Information estimates for both methods

| method | var(1d) | mean(2d) | diff | diff% |
|---|---|---|---|---|
| $n = 100$ , $\gamma = 1$ | 56.62861899 | 50.01655991 | -6.61205908 | -0.11676179 |
| $n = 10$ , $\gamma = 1$ | 5.25975401 | 5.16402998 | -0.09572403 | -0.01819934 |
| $n = 100$ , $\gamma = 5$ | 2.26514476 | 2.0006624 | -0.26448236 | -0.11676179 |

So far we have worked under the assumption that the true parameter value $\theta$ is already known. This gave us a nice visual representation of the Fisher Information, and consequently the variance of the MLE. We now have two graphical interpretations of the Fisher Information:

1. The variance of the distribution of first derivatives at the *true* $\theta$.
2. The Fisher Information measures the sensitivity of the score-function (/the curvature of the likelihood-function) around the *true* $\theta$.

However, when working with real data, we are observing the function at the *estimated* parameter $\hat{\theta}$. Especially when working with a bumpy likelihood function, the MLE can be off. The Cauchy Distribution is one example of a pdf that can have multiple roots when solving for the MLE.[5] A consequence of this is that the MLE significantly differs from the true parameter value.

This has an important implication: The MSE derived by calculating the Fisher information explicitly assumes that we will evaluate the function at or close to the true parameter. It does not take into account the possibility of a "false" local extremum. However, depending on the complexity of the likelihood function, and how well we are able to solve for it's global minimum, it is very likely, that our method will not return the "true" global minimum. The confidence intervals therefore may never cover the true parameter value, even if the true Fisher Information is used to estimate the MSE.

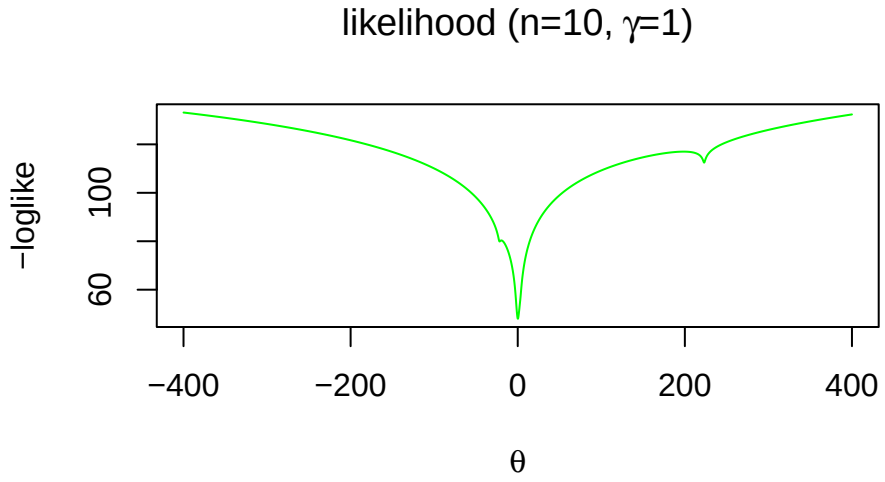


Figure 4: Example of a likelihood function generated from a Cauchy distribution with a second local minimum at 223.07

[4]For reference: The analytical Fisher Information fro the Cauchy Distribution with $\gamma = 1$ is given by $n/2$, which is close to our first two results (Pati 2016)

[5]Note that, while there are several ways to solve for the global maximum of the Cauchy Distribution (Barnett 1966), the following argument generalizes to other distributions that produce more complex multimodal likelihood functions.

An example of such a function with multiple minima is referenced in Figure (4). Note, that the Type-2 samples ($\gamma = 50,\ n = 7$, second row) generated in our effort to visualize the FI stem from the same data generating process (dgp) that produced the log-likelihood in Figure (4). We will now examine the implications mentioned above, using the data from Figure (4) as an example.

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

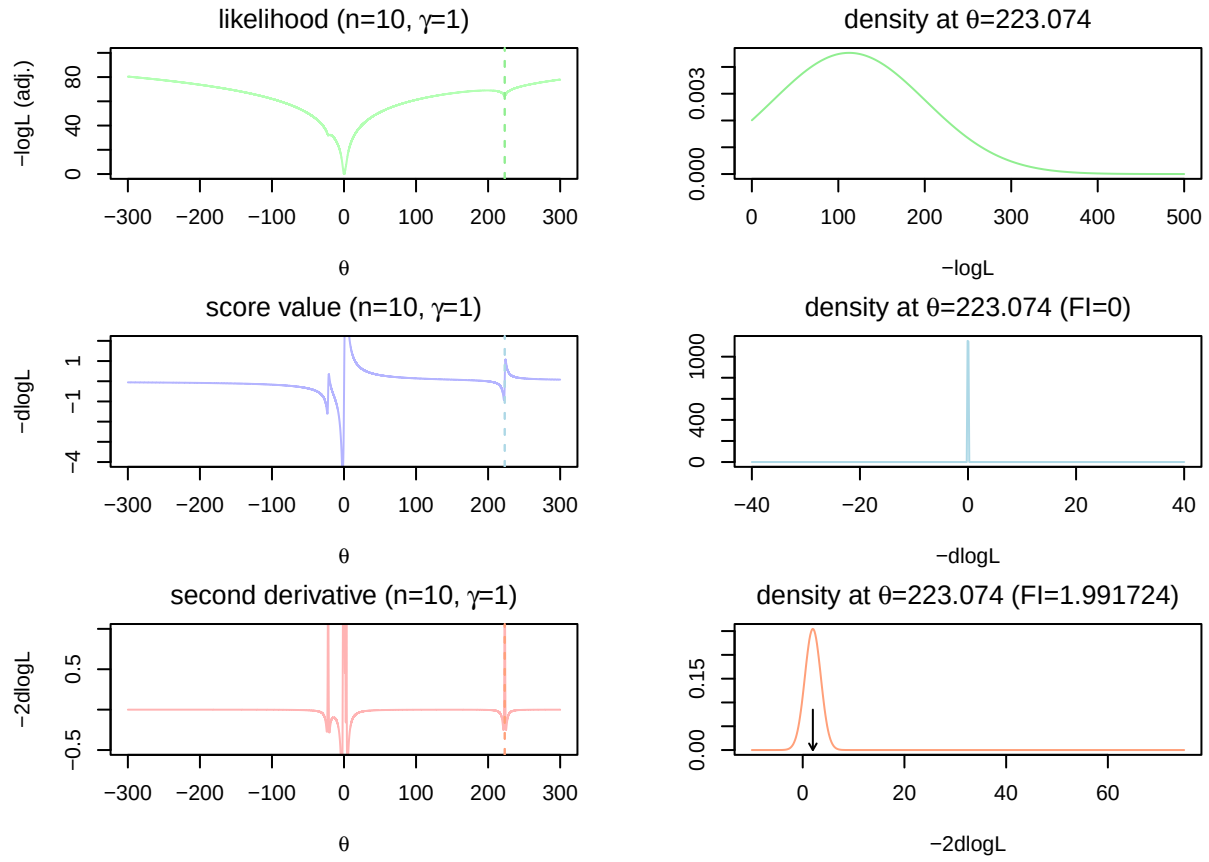


Figure 5: The log-likelihood and its derivatives when the log-likelihood has multiple minima.

As Figure (5) shows, the second minimum is represented in the score function as a second root, that closely resembles the structure of the first one, making it difficult to discern the local minimum from a global one. Simple methods of gradient descent like the BFGS will in fact return the false local minimum when using a starting value >224.

Unsurprisingly, the same holds true for the second derivative. When evaluating the second derivative at the false minimum, the estimated FI is 1.99, implying a MSE of:

$$MSE = \sqrt{\frac{1}{1.99}} = 0.7085$$

The true FI for the displayed likelihood is $n/2 = 5$, which yields a true variance of:

$$MSE = \sqrt{\frac{1}{5}} = 0.4472$$

In either case, the resulting 99%-confidence interval would not cover the true parameter value, when we naively assume that we have found the global minimum.

## References


Barnett, V. D. 1966. "Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots." *Biometrika* 53 (1-2): 151–65. https://doi.org/10.1093/biomet/53.1-2.151.

Ly, Alexander, Maarten Marsman, Josine Verhagen, Raoul Grasman, and Eric-Jan Wagenmakers. 2017. "A Tutorial on Fisher Information." https://arxiv.org/abs/1705.01064.

Pati, Debdeep. 2016. "Fisher Information." https://ani.stat.fsu.edu/~debdeep/Fisher.pdf.

Rich, Duane. 2021. "The Fisher Information." Youtube. 2021. https://www.youtube.com/watch?v=pneluWj-U-o.

Zheng, Songfeng. n.d. "Fisher Information and Cramér-Rao Bound." https://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/Fisher_info.pdf.