Project 3:
Finding parameters of a function by Maximum Likelihood Estimation (MLE) method

Shahla Daneshmehr

April 2023

1 Abstract

Maximum likelihood estimation (MLE) is a statistical technique for estimating the parameters of a probability distribution that has been assumed given some observed data. The maximum likelihood estimate is the location in the parameter space where the likelihood function is maximized. Maximum likelihood is a popular approach for making statistical inferences since its rationale is clear and adaptable.

2 Introduction

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a model using a set of data. While MLE can be applied to many different types of models, this article will explain how MLE is used to fit the parameters of a probability distribution for a given set of failure and right censored data [1].

MLE works by calculating the probability of occurrence for each data point (we call this the likelihood) for a model with a given set of parameters. These probabilities are summed for all the data points. We then use an optimizer to change the parameters of the model in order to maximize the sum of the probabilities [2].

There are two major challenges with MLE. These are the need to use an optimizer (making hand calculations almost impossible for distributions with more than one parameter), and the need for a relatively accurate initial guess for the optimizer. The initial guess for MLE is typically provided using Least Squares Estimation. A variety of optimizers are suitable for MLE, though some may perform better than others so trying a few is sometimes the best approach [3].

There are several advantages of MLE which make it the standard method for fitting probability distributions in most software. MLE does not need the equation to be linearizable (which is needed in the least Squares Estimation) so any equation can be modeled. The other advantage of MLE is that unlike Least Squares Estimation which uses the plotting positions and does not directly use the right censored data, MLE uses the failure data and right censored data directly, making it more suitable for heavily censored datasets [4].

In this project, I describe the idea of maximum likelihood in so-called generative models. The idea is to describe the model in which my data is generated, rather than just fitting my data to some model, which means that I have to make some assumptions about the probabilistic nature of the data. Therefore, the concept of the maximum likelihood is a way of saying what is the most likely scenario for the data I have collected.

3 Hypotheses to Explain Maximum Likelihood Estimation

I suppose that you have in your hand a GPS that tells you where you are. However, you also are worried about this thing being not accurate so you bought another GPS so you have two GPSs now. One of them was purchased in the United States so uses u.s. satellites. The other one was purchased in Europe so it uses European satellites. Therefore, they are independent estimates of where you are. I want to set up the problem where is your location based on the readings that you have the data from those two GPSs [5].

4 Code and Experimental Simulation

A model exists first and it generates our samples. Let’s see our generative model:

```
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt
%matplotlib inline

# MLE of Gaussian distribution
# mu

m = 20
mu = 0
sigma = 5

x = np.random.normal(mu,sigma,[m,1])
xp = np.linspace(-20, 20, 100)
y0 = np.zeros([m, 1])

muhat = [-5, 0, 5, np.mean(x)]

plt.figure(figsize=(8, 8))

for i in range(4):
    yp = norm.pdf(xp, muhat[i], sigma)
    y = norm.pdf(x, muhat[i], sigma)
    logL = np.sum(np.log(y))

    plt.subplot(4, 1, i+1)
    plt.plot(xp, yp, ’r’)
    plt.plot(x, y, ’bo’)
    plt.plot(np.hstack([x, x]).T, np.hstack([y, y0]).T, ’k--’)

    plt.title(r’$\hat\mu$ = {0:.2f}’.format(muhat[i]), fontsize=15)
    plt.text(-15,0.06,np.round(logL,4),fontsize=15)
    plt.axis([-20, 20, 0, 0.11])

plt.tight_layout()
plt.show()
```

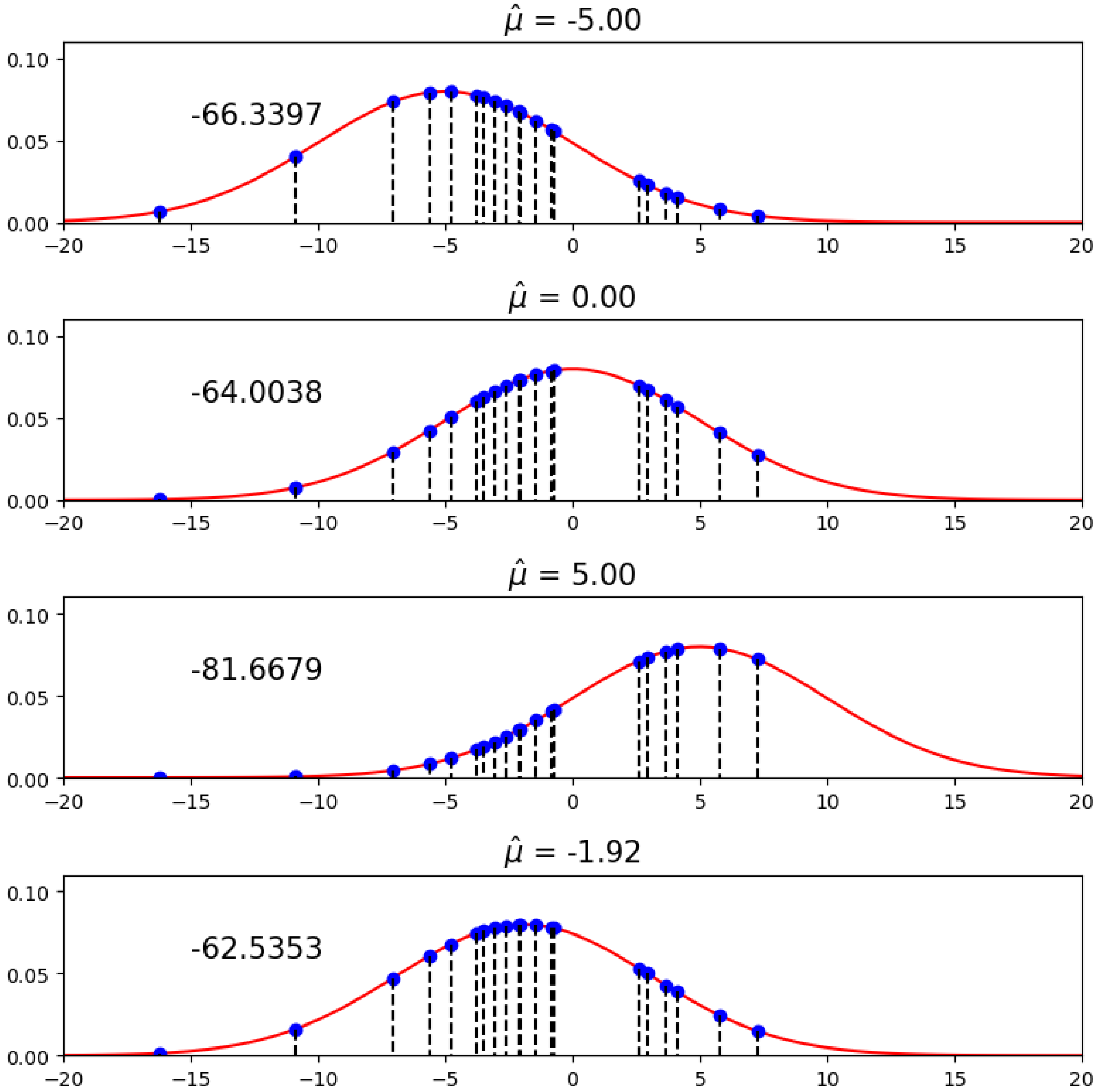


Figure 1: MLE of Gaussian distribution

5 Analysis

We are going to reconstruct the probability density function from a set of given data samples y_1, y_2,. . ., and m_y which are independent. We already know that the Gaussian distribution (also known as the normal distribution) is a bell-shaped curve, and it is assumed that during any measurement values will follow a normal distribution with an equal number of measurements above and below the mean value. In our assumption, we have the Gaussian distribution with unknown parameters μ and σ.

$$P\left(y_1, y_2, \cdots, y_m\,;\,\mu, \sigma^2\right) = \prod_{i=1}^{m} P\left(y_i\,;\,\mu, \sigma^2\right)$$

Our generated model:

$$P\left(y = y_i\,;\,\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right)$$

$$\begin{aligned} \mathcal{L} = P\left(y_1, y_2, \cdots, y_m\,;\,\mu, \sigma^2\right) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\,(y_i - \mu)^2\right) \\ &= \frac{1}{(2\pi)^{\frac{m}{2}}\sigma^m} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mu)^2\right) \end{aligned}$$

$$\ell = \log \mathcal{L} = -\frac{m}{2}\text{log}2\pi - m\text{log}\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mu)^2$$

To maximize:

$$\frac{\partial \ell}{\partial \mu} = 0, \frac{\partial \ell}{\partial \sigma} = 0$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{m}(y_i - \mu) = 0 \qquad \Longrightarrow \qquad \mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} y_i \quad \text{: sample mean}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{m}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{m}(y_i - \mu)^2 = 0 \quad \Longrightarrow \quad \sigma^2_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \mu)^2 \quad \text{: sample variance}$$

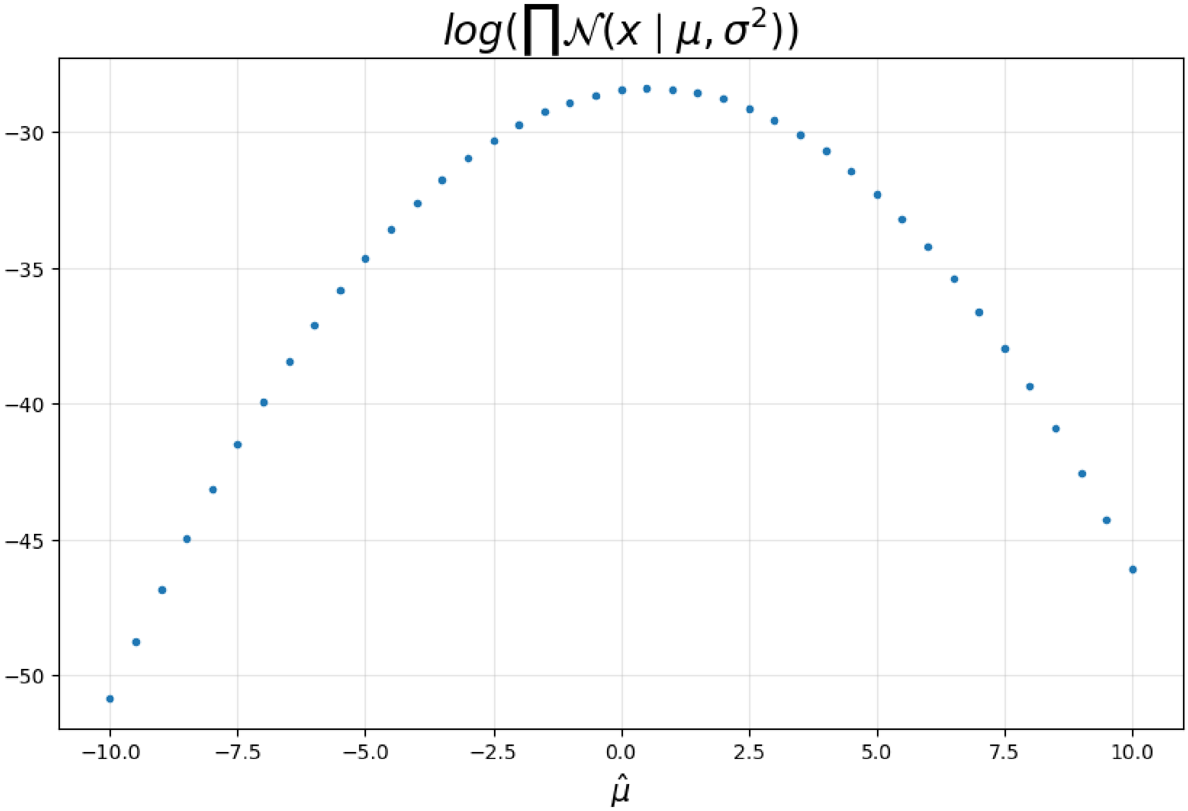
In order to represent data statistics, we frequently compute a mean and variance. Our sample mean is Gaussian distributed by the central limit theorem. The likelihood function is the probability density at a particular outcome D when the true value of the parameter is θ.

$$\mathcal{L} = P(D \mid \theta) = P(D\,;\,\theta)$$

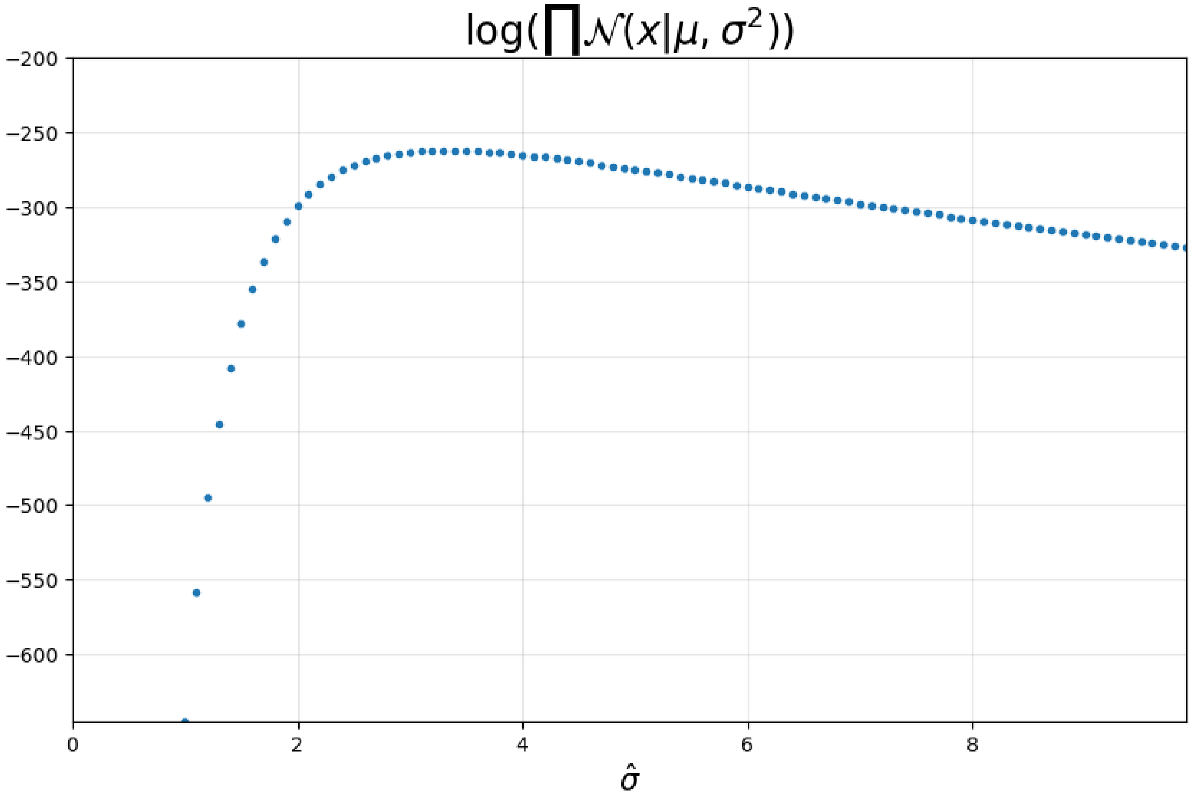
$$\theta_{MLE} = \underset{\theta}{\text{argmax}}\;\; P(D\,;\,\theta)$$

Now, we are going to compare our data with data from formula:

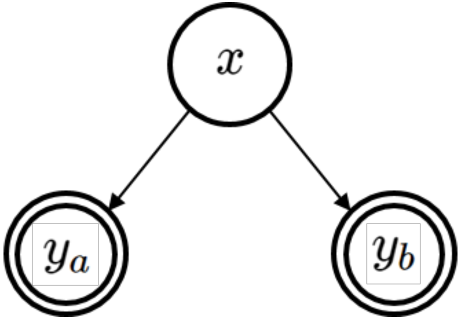
$$\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x_i$$


$log(\prod \mathcal{N}(x \mid \mu, \sigma^2))$
−30
−35
−40
−45
−50
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
$\hat{\mu}$

$$\sigma^2_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \mu)^2$$



Now, I show all the above equations in our real example (GPSs). Let’s assume there are two GPSs.



$$y_a = x + \varepsilon_a, \ \varepsilon_a \sim \mathcal{N}\left(0, \sigma_a^2\right)$$
$$y_b = x + \varepsilon_b, \ \varepsilon_b \sim \mathcal{N}\left(0, \sigma_b^2\right)$$

In a matrix form:

$$y = \begin{bmatrix} y_a \\ y_b \end{bmatrix} = Cx + \varepsilon = \begin{bmatrix} 1 \\ 1 \end{bmatrix} x + \begin{bmatrix} \varepsilon_a \\ \varepsilon_b \end{bmatrix} \qquad \varepsilon \sim \mathcal{N}\left(0, R\right), \quad R = \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}$$

$$P\left(y \mid x\right) \sim \mathcal{N}\left(Cx, R\right) \& = \frac{1}{\sqrt{\left(2\pi\right)^2 |R|}} \exp\left(-\frac{1}{2}\left(y - Cx\right)^T R^{-1} \left(y - Cx\right)\right)$$

Let’s find

$$\hat{x}$$

$$\ell = -\log 2\pi - \frac{1}{2} \log |R| - \frac{1}{2} \underbrace{\left(y - Cx\right)^T R^{-1} \left(y - Cx\right)}$$
$$\left(y - Cx\right)^T R^{-1} \left(y - Cx\right) = y^T R^{-1} y - y^T R^{-1} Cx - x^T C^T R^{-1} y + x^T C^T R^{-1} Cx$$

$$\Longrightarrow \ \frac{d\ell}{dx} = 0 = -2C^T R^{-1} y + 2C^T R^{-1} Cx$$
$$\hat{x} = \left(C^T R^{-1} C\right)^{-1} C^T R^{-1} y$$
$$\left(C^T R^{-1} C\right) = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_a^2} & 0 \\ 0 & \frac{1}{\sigma_b^2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}$$
$$C^T R^{-1} = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_a^2} & 0 \\ 0 & \frac{1}{\sigma_b^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_a^2} & \frac{1}{\sigma_b^2} \end{bmatrix}$$

$$
\begin{aligned}
\hat{x} &= \left(C^T R^{-1} C\right)^{-1} C^T R^{-1} y = \left(\frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}\right)^{-1} \begin{bmatrix} \frac{1}{\sigma_a^2} & \frac{1}{\sigma_b^2} \end{bmatrix} \begin{bmatrix} y_a \\ y_b \end{bmatrix} \\
&= \frac{\frac{1}{\sigma_a^2} y_a + \frac{1}{\sigma_b^2} y_b}{\frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}}
\end{aligned}
$$

$$
\begin{aligned}
\text{var}\left(\hat{x}\right) &= \left(\left(C^T R^{-1} C\right)^{-1} C^T R^{-1}\right) \cdot \text{var}(y) \cdot \left(\left(C^T R^{-1} C)^{-1} C^T R^{-1}\right)\right)^T \\
&= \left(\left(C^T R^{-1} C\right)^{-1} C^T R^{-1}\right) \cdot R \cdot \left(\left(C^T R^{-1} C)^{-1} C^T R^{-1}\right)\right)^T \\
&= \left(C^T R^{-1} C\right)^{-1} C^T \cdot \left(R^{-1}\right)^T C \left(\left(C^T R^{-1} C\right)^{-1}\right)^T \\
&= \underbrace{\left(C^T R^{-1} C\right)^{-1}} \underbrace{C^T R^{-1} C} \left(\left(C^T R^{-1} C\right)^{-1}\right)^T = \left(C^T R^{-1} C\right)^{-1} \\
&= \frac{1}{\frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}} \leq \ \sigma_a^2,\ \sigma_b^2
\end{aligned}
$$

Our outline:

$$
\hat{x} = \frac{\frac{1}{\sigma_a^2} y_a + \frac{1}{\sigma_b^2} y_b}{\frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}} = \frac{\frac{1}{\sigma_a^2}}{\frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}} y_a + \frac{\frac{1}{\sigma_b^2}}{\frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}} y_b
$$

$$
\text{var}(\hat{x}) = \frac{1}{\frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}} \quad < \quad \sigma_a^2,\ \sigma_b^2 \qquad \Longrightarrow \text{ more accurate}
$$

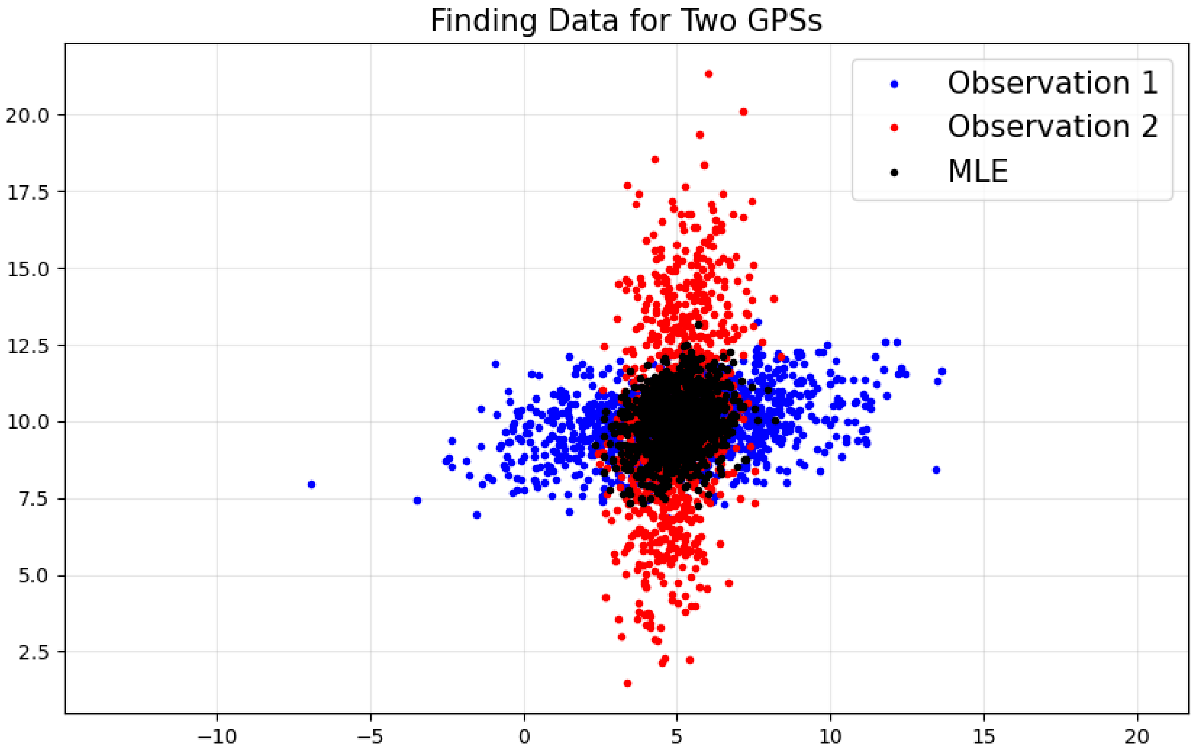


Figure 2: 2D data for two GPSs, MLE of Gaussian distribution

6 Conclusion

In conclusion, based on our data, we could find out that the two GPSs are better than one GPS because of the less uncertainty. Therefore, we see that accuracy or uncertainty information is also important in GPSs or in general words in sensors. This simple example shows us the importance of maximum likelihood estimation in real life.

7 References

1. Rossi, Richard J. (2018). Mathematical Statistics: An Introduction to Likelihood Based Inference. New York: John Wiley Sons. p. 227.

2. Chambers, Raymond L.; Steel, David G.; Wang, Suojin; Welsh, Alan (2012). Maximum Likelihood Estimation for Sample Surveys. Boca Raton: CRC Press.

3. Ward, Michael Don; Ahlquist, John S. (2018). Maximum Likelihood for Social Science: Strategies for Analysis. New York: Cambridge University Press.

4. Hendry, David F.; Nielsen, Bent (2007). Econometric Modeling: A Likelihood Approach. Princeton: Princeton University Press.

5. http://courses.shadmehrlab.org/learningtheory.html.