LECTURE 8: MAXIMUM LIKELIHOOD ESTIMATION PART I

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
OCTOBER 29, 2025

1. Maximum Likelihood Estimators (MLE)

1.1. Likelihood function

Suppose $X_1, \ldots, X_n$ are i.i.d random sample from a population with pdf $f(x|\theta)$, where θ is an unknown vector of parameters that parameterize the pdf. The joint density of $X_1, \ldots, X_n$ is $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$.

Now suppose we observe a realization of the random sample which we denote as $X_1 = \tilde{x}_1, \ldots, X_n = \tilde{x}_n$. Can we say that the likelihood of observing the realization $\tilde{x}_1, \ldots, \tilde{x}_n$ is $\prod_{i=1}^{n} f(\tilde{x}_i|\theta)$? Yes, even though $\prod_{i=1}^{n} f(\tilde{x}_i|\theta)$ is strictly a density, it carries the connotation of a likelihood.[1]

The **likelihood function** of θ given that we observe the following data realized from the data-generating process, $X_1 = x_1, \ldots, X_n = x_n$, is defined as:

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1}$$

The **log-likelihood function** of θ given that we observe the realized data $X_1 = x_1, \ldots, X_n = x_n$, is defined as:

$$\mathcal{L}(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{2}$$

[1]When ϵ is small, we can approximate $P(x - \epsilon \leq X \leq x + \epsilon)$ as the area of a rectangle: the height $f(x)$ times the width 2ϵ. As such, $2\epsilon f(x) \approx P(x - \epsilon \leq X \leq x + \epsilon)$ for arbitrarily small ϵ. Moreover,

$$\frac{f(x_1)}{f(x_2)} \approx \frac{P(x_1 - \epsilon \leq X \leq x_1 + \epsilon)}{P(x_2 - \epsilon \leq X \leq x_2 + \epsilon)}$$

Hence if $f(x_1) > f(x_2)$, X is more likely to take values around x_1 than x_2.

Here, we treat $x_1, \ldots, x_n$ is the observed data and it is therefore fixed, while θ is the argument in the likelihood function that is varying.

1.2. Maximum likelihood estimators

The maximum likelihood estimate of θ given the data $x_1, \ldots, x_n$ is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta | x_1, \ldots, x_n) \tag{3}$$

Intuitively, we estimate θ by finding θ that maximizes the likelihood of observing the given data.

$\hat{\theta}$ depends on the dataset $(x_1, \ldots, x_n)$, which is a realization of the random sample $X_1, \ldots, X_n$. Hence the maximum likelihood estimator should be remembered as a random variable. The definition of MLE is the same regardless of whether X is a discrete or a continuous random variable.

1.3. Example: MLE of Normal parameters

$X_1, \ldots, X_n \sim$ i.i.d $\mathcal{N}(\mu, 1)$

Given the data $X_1 = x_1, \ldots, X_n = x_n$, the likelihood function is:

$$L(\mu | x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu)^2/2}$$

The log-likelihood function is:

$$\mathcal{L}(\mu | x_1, \ldots, x_n) = n \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\max_{\mu} L(\theta | x_1, \ldots, x_n) = \max_{\mu} \sum_{i=1}^{n} -(x_i - \mu)^2$$

Taking the first-order condition with respect to μ, we get $\sum_{i=1}^{n} (x_i - \mu) = 0$. Solving for μ, we obtain $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$. To verify that the solution $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$ is a local maximum, we check the second-order condition $\frac{\partial \mathcal{L}^2}{\partial \mu^2} = -n < 0$.

To verify that the solution is a global maximum, we check that $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$ is the only solution satisfying the first-order condition, and at the boundaries of the parameter space, the likelihood is strictly smaller.

1.4. Example: MLE of Bernoulli parameters

$X_1, \ldots, X_n \sim$ i.i.d Bernoulli(p).

The likelihood function is

$$\begin{aligned} L(p|x_1, \ldots, x_n) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ \log L(p|x_1, \ldots, x_n) &= \sum_{i=1}^{n} x_i \log p + (1-x_i)\log(1-p) \\ &= (\sum_{i=1}^{n} x_i)\log p + \log(1-p)(n - \sum_{i=1}^{n} x_i) \end{aligned}$$

Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Taking the first-order-condition with respect to p,

$$\max_{0\leq p\leq 1} L(p|x_1, \ldots, x_n) \implies \frac{\partial \mathcal{L}}{\partial p} = \frac{n\bar{x}}{p} - \frac{n - n\bar{x}}{1-p} = 0 \implies \hat{p} = \bar{x}$$

Checking the second-order condition, we have $\frac{\partial^2 \mathcal{L}}{\partial p^2} = -\frac{n\bar{x}}{p^2} - \frac{n-n\bar{x}}{(1-p)^2} < 0$. Note that we maximize the likelihood within the range of the parameter space $0 \leq p \leq 1$, and that $\hat{p} = \bar{x}$ is the global maximum within this parameter space.

1.5. Example: uniform distribution

$X_1, \ldots, X_n \sim$ i.i.d $U[0, \theta]$

$$L(\theta|x_1, \ldots, x_n) = \begin{cases} (\frac{1}{\theta})^n & \text{if } \max(x_1, \ldots, x_n) \leq \theta \\ 0 & \text{if } \max(x_1, \ldots, x_n) > \theta \text{ or } \min(x_1, \ldots, x_n) < 0 \end{cases}$$

Assuming $\min(x_1, \ldots, x_n) \geq 0$, the MLE of θ is

$$\underset{0\leq\theta}{\operatorname{argmax}}\, L(\theta|x_1, \ldots, x_n) = \max(x_1, \ldots, x_n)$$

1.6. Example: MLE for more than 1 parameter

$X_1, \ldots, X_n \sim$ i.i.d $\mathcal{N}(\mu, \sigma^2)$ with μ and σ^2 unknown.

The likelihood functions is:

$$L(\mu, \sigma^2|x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\theta)^2/2\sigma^2}$$

The log-likelihood function is:

$$\mathcal{L}(\mu, \sigma^2|x_1, \ldots, x_n) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

The ML estimates $\hat{\mu}$ and $\hat{\sigma}^2$ satisfy the following first-order conditions:

$$\left.\frac{\partial \mathcal{L}(\mu, \sigma^2|\boldsymbol{x})}{\partial \mu}\right|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \tag{4}$$

$$\left.\frac{\partial \mathcal{L}(\mu, \sigma^2|\boldsymbol{x})}{\partial \sigma^2}\right|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = 0 \tag{5}$$

Therefore the ML solutions are $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. To further check that the solutions are local maximum, the second-order sufficient condition requires that the Hessian matrix be negative definite at the estimate, i.e. all eigenvalues are negative. When there are two parameters, this is equivalent to the determinant of the Hessian matrix being positive and the second-order derivative with respect to either one of the parameters be negative

1.7. Example: MLE for simple linear regressions

Let $Y_i \sim \mathcal{N}(\alpha + \beta x_i, 1)$ for $i = 1, \ldots, n$, where α, β are unknown parameters, and $x_i, i = 1, \ldots, n$ are fixed numbers. This model describes $Y_i = \alpha + \beta x_i + \epsilon_i$ for $i = 1, \ldots, n$ and $\epsilon_i \sim$ i.i.d $\mathcal{N}(0, 1)$.

The log-likelihood function is:

$$\mathcal{L}(\alpha, \beta|x_1, y_1, \ldots, x_n, y_n) = n\log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

The first-order conditions for maximizing the log-likelihood function with respect to α and β are:

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \tag{6}$$

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0 \tag{7}$$

Solving these two equations for α and β:

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The estimators recovered here are identical to the estimators obtained via Method of Moments (see last lecture).

Even though we obtain the same formulas for parameters of linear regressions, there is a distinct difference between the MLE approach and the GMM approach. In MLE, we have to make the assumption that ϵ is Normally distributed, whereas in GMM, we do not have to make the assumption that ϵ is Normally distributed. For this reason, GMM is known as a partial-information approach, while MLE is a full-information approach.

1.8. Restricting the parameter space

When we maximize the likelihood function $L(\theta|x)$ with respect to θ, we need to specify the range of θ in which to search for the maximum. For example, when estimating a scale parameter like σ^2, we maximize the objective function over positive values.

As another example, to estimate the parameters of the Binomial distribution from the dataset $x_1, x_2, \ldots, x_n$, we find an integer k that maximizes:

$$L(k, p|x) = \Pi_{i=1}^{n}\binom{k}{x_i}p^{x_i}(1-p)^{k-x_i}$$

We numerically search over $k = \{\max x_i, 1 + \max x_i, 2 + \max x_i, \ldots, N\}$. We cannot take the first-order conditions here.

1.9. Probit regression

Suppose Y_i is a binary random variable. For example, $Y_i = 1$ if a user i clicks on the advertisement, and $Y_i = 0$ otherwise. We want to model Y_i (so as to be able to predict it). Let Y_i be a Bernoulli random variable, such that $Y_i = 1$ with probability p, and $Y_i = 0$ with probability $1 - p$. In addition, we parameterize the probability $p = g(\alpha + \beta x_i)$, for some unknown parameters α and β. Here, x_i is a feature/covariate/explanatory variable of the user i, say the age of the user i.

Because the probability $p = g(\alpha + \beta x_i)$ must be between 0 and 1, we need to pick a function g such that $g : \mathbb{R} \to [0, 1]$. One such function is the cdf of $\mathcal{N}(0, 1)$, which we denote by $\Phi(\cdot)$. In another words, the model for the random variable Y_i is:

$$Y_i = \begin{cases} 1 & \text{with probability } \Phi(\alpha + \beta x_i) \\ 0 & \text{with probability } 1 - \Phi(\alpha + \beta x_i) \end{cases}$$

Suppose we observe i.i.d realizations from this model: $(y_1, x_1, y_2, x_2, \ldots, y_n, x_n)$, how do we estimate a and b?

The pdf of Y_i is just the Bernoulli pdf:

$$f(y_i) = \Phi(\alpha + \beta x_i)^{y_i}(1 - \Phi(\alpha + \beta x_i))^{1-y_i}, \quad y_i \in \{0, 1\}$$

Because Y_i are independent, the joint pdf is just:

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} \Phi(\alpha + \beta x_i)^{y_i}(1 - \Phi(\alpha + \beta x_i))^{1-y_i}$$

Therefore the likelihood function is:

$$L(\alpha, \beta | \boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} \Phi(\alpha + \beta x_i)^{y_i}(1 - \Phi(\alpha + \beta x_i))^{1-y_i}$$

$$\mathcal{L}(\alpha, \beta | \boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} y_i \log(\Phi(\alpha + \beta x_i)) + \sum_{i=1}^{n} (1 - y_i) \log(1 - \Phi(\alpha + \beta x_i))$$

The Maximum Likelihood estimates of α and β are $\text{argmax}_{\alpha,\beta} \mathcal{L}(\alpha, \beta | \boldsymbol{x}, \boldsymbol{y})$. There is no analytical closed-form solution, however it can be shown that the likelihood function is concave in α and β. Therefore standard numerical algorithms for optimization (such as Newton-Raphson method or stochastic gradient ascent) will

converge rapidly to the unique maximum. Moreover, the gradient can be evaluated easily as:

$$\nabla \mathcal{L}(\alpha, \beta) = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial \alpha} \\ \frac{\partial \mathcal{L}}{\partial \beta} \end{bmatrix} = \sum_{i=1}^{n} \left(\frac{y_i}{\Phi(\alpha + \beta x_i)} - \frac{1 - y_i}{1 - \Phi(\alpha + \beta x_i)} \right) \cdot \phi(\alpha + \beta x_i) \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

A different model called the Logit model uses the logistic function $\frac{1}{1+e^{-a-bx_i}}$ to model the probabilities.

1.10. Invariance property of MLE*

(Optional reading)

Suppose instead of maximizing $L(\theta|x)$ with respect to θ, we are interested in obtaining an estimate of $\eta = \tau(\theta)$ from the likelihood function $L(\theta|x)$. For example, $\eta = \frac{3\theta-2}{5}$ or $\eta = \log(\theta)$.

Assuming τ is an invertible one-to-one function, then we can easily rewrite and transform the likelihood function.

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} f(x_i|\tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|x)$$

Suppose $\hat{\theta} = \text{argmax}\, L(\theta|x)$, and let $\hat{\eta} = \tau(\hat{\theta})$, then $\hat{\eta}$ also maximizes $L(\tau^{-1}(\eta)|x)$.

The invariance property of MLE holds more generally for any function τ. If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.[2]

For example, if $\hat{p}$ is the MLE of the Binomial distribution with unknown parameter p, then $n\hat{p}(1-\hat{p})$ is the MLE of the the variance.

This invariance property is nice, but there is a somewhat undesirable consequence: MLEs are generally NOT unbiased. Both of the exercises above demonstrate this. For a simpler example, consider $X \sim \mathcal{N}(\theta, 1)$. The MLE of θ is $\hat{\theta} = \bar{X}$ and, therefore the MLE of θ^2 is $\bar{X}^2$.

However we know from Jensen's inequality that $\mathbb{E}[\bar{X}^2] \geq \mathbb{E}[\bar{X}]^2 = \theta^2$, therefore $\bar{X}^2$ is a biased estimator of θ^2. Despite being generally biased, MLE enjoys desirable large sample properties, such as being consistent and efficient, as we will see later.

[2]For a general function τ, if we are interested in estimating the parameter $\eta = \tau(\theta)$ from the likelihood function $L(\theta|x)$, then we transform the likelihood as follows: $L(\eta|x) = \max_{\theta:\tau(\theta)=\eta} L(\theta|x)$

1.11. Numerical implementation of MLE

In practice, MLE typically lacks a closed-form solution and requires numerical optimization. This process involves several challenges:

(i) **Non-linearity of the Likelihood Function:** Likelihood functions are often highly non-linear, which complicates finding the global maximum. Numerical solvers (such as fmincon and fminsearch in MATLAB, or Python's scipy.optimize.minimize) are designed to locate local optima, meaning they may converge only to a local maximum rather than the global one.

(ii) **Multiple Parameters and Local Maxima:** With a high-dimensional parameter space, the likelihood surface may have multiple local maxima. To improve the chances of finding the global maximum, we should start the optimization from several different initial values.

(iii) **Implementation of the Likelihood Function and its Derivatives:** In many cases, the likelihood function, as well as its gradient and Hessian, must be explicitly coded. The MLE is then obtained using iterative methods, such as gradient ascent or stochastic gradient descent.

(iv) **Accurate Gradient Calculation:** Finite-difference methods can approximate gradients but often lack precision. Instead, exact gradients can be obtained through Automatic Differentiation, a technique commonly used in backpropagation within tools like PyTorch and TensorFlow, providing more accurate and efficient gradient computations.

GMM shares a lot of these difficulties.

2. Methods of evaluating estimators

In this section, we will introduce a general framework for evaluating how good an estimator is.

If θ is the ground truth of the parameters, and your proposed estimate is a, you incur a loss of $L(a, \theta)$. The function L is called the Loss Function and it is the utility function that is specific to an individual researcher. The notion of loss functions is central to statistical decision theory.

Common function of L includes the squared error loss function, $L(a, \theta) = (a - \theta)^2$ and the absolute error loss function $L(a, \theta) = |a - \theta|$. These loss functions penalize under- and over-estimate symmetrically and equally. A loss function that penalizes overestimation more than underestimation is:

$$L(\theta, a) = \begin{cases} (a-\theta)^2 \text{ if } a < \theta \\ 4(a-\theta)^2 \text{ if } a \geq \theta \end{cases}$$

Since the estimator a is a random variable that depends on the random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, we are interested in the expected loss that will be incurred if the estimator $a(\boldsymbol{X})$ is used:

$$\mathbb{E}_{\boldsymbol{X}}[L(a(\boldsymbol{X}), \theta)]$$

2.1. Mean Squared Error

The mean squared error (MSE) is the expected loss of an estimator $a(\boldsymbol{X})$ under the square loss function. That is,

$$\mathbb{E}[(a(\boldsymbol{X}) - \theta)^2]$$

An important property of the MSE is:

$$\begin{aligned} \mathbb{E}[(a(\boldsymbol{X}) - \theta)^2] &= \mathbb{E}[(a(\boldsymbol{X})^2] - 2\,\mathbb{E}[a(\boldsymbol{X})]\theta + \theta^2 \\ &= \mathbb{E}[(a(\boldsymbol{X})^2] - \mathbb{E}[a(\boldsymbol{X})]^2 + \mathbb{E}[a(\boldsymbol{X})]^2 - 2\,\mathbb{E}[a(\boldsymbol{X})]\theta + \theta^2 \\ &= \text{Var}(a(\boldsymbol{X})) + (\mathbb{E}[a(\boldsymbol{X})] - \theta)^2 \\ &= \text{Var}(a(\boldsymbol{X})) + \text{bias}^2 \end{aligned}$$

Therefore the MSE measures both the variability of an estimator (precision), as well as its bias (accuracy). A good estimator according to the mean squared error (in the sense of having a low MSE) is both precise and accurate.

2.2. Example

Let $X_1, \ldots, X_n \sim$ i.i.d $\mathcal{N}(\mu, \sigma^2)$.

The sample mean and sample variance estimators $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are both unbiased estimators of μ and σ^2. Because the bias is zero for both estimators, the mean squared errors are given by:

$$\text{MSE}(\bar{X}) = \mathbb{E}[(\bar{X} - \mu)^2] = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

$$\text{MSE}(S^2) = \mathbb{E}[(S^2 - \sigma^2)^2] = \text{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

Note that the MSE will usually be a function of the true underlying parameters of the DGP. As a result, the MSE can be larger or smaller depending on the values of these parameters.

When one estimator has a smaller MSE than another, it is a better estimator (according to the mean square error criterion).

Unbiased estimators may not be optimal in terms of MSE. There is usually a trade-off between bias and variance so that a small increase in bias can be traded for a larger decrease in variance, resulting in a better MSE. This is the case for the sample variance versus the Maximum-Likelihood estimator of σ^2 which is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$. Although $\hat{\sigma}^2$ is biased, it has a much lower variance, such that the overall MSE is smaller.

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$$

.

$$\text{Var}(\hat{\sigma}^2) = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}$$

.

$\text{MSE}(\hat{\sigma}^2)$ is therefore given by:

$$\mathbb{E}[(\hat{\sigma}^2 - \sigma^2)^2] = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \frac{2n-1}{n^2}\sigma^4$$

.

Now, $\mathbb{E}[(S^2 - \sigma^2)^2] = \frac{2}{n-1}\sigma^4 = \frac{2n}{n(n-1)}\sigma^4 > \frac{2n-1}{n(n-1)}\sigma^4 > \frac{2n-1}{n^2}\sigma^4 = \mathbb{E}[(\hat{\sigma}^2 - \sigma^2)^2]$.

This shows that $\text{MSE}(S^2) > \text{MSE}(\hat{\sigma}^2)$. The MLE of σ^2 has a lower MSE than the sample variance.

However the MSE here is calculated assuming the DGP is $\mathcal{N}(\mu, \sigma^2)$. We can use Monte Carlo simulation to calculate the MSE for different DGPs, as demonstrated in the accompanying Python Notebook.

Regardless of the data generating process, for large n, we have $S^2 \approx \mathcal{N}(\sigma^2, \frac{1}{n}(\mu_4 - \sigma^4))$, as well as $\hat{\sigma}^2 \approx \mathcal{N}(\sigma^2, \frac{1}{n}(\mu_4 - \sigma^4))$, where $\mu_4 = \mathbb{E}[(X-\mu)^4]$. As such, S^2 and $\hat{\sigma}^2$ are asymptotically equivalent, and they have approximately the same mean-squared error for large n.

If the mean-squared error converges to zero, then the estimator also converges in probability to the true parameter. In general, convergence in L^p is sufficient for consistency, but it is not necessary. We will see this in the problem set.