Maximum likelihood estimation

Olga Klopp

- A **fundamental** principle in statistics. Particular cases known since the 18th century. General definition: Fisher (1922).
- Provides a first **systematic method** to build a M-estimator (often a GMM-estimator, often also a simple substitution estimator).
- **Optimal** procedure (in what way?) under assumptions of **regularity** of the family $\{\mathbb{P}_\theta, \theta \in \Theta\}$.
- Sometimes difficult to implement in practice → **optimization problem**.

Example of application: Logistic regression is a technique used often in machine learning to classify data points. For example, logistic regression can be used to classify whether an email is spam or is not spam. The logistic regression model says that the probability a data point X_i is in class 1 is as follows: $h_\theta(x_i) = \mathbb{P}[y_i = 1] = \dfrac{1}{1 + e^{-\theta^T x_i}}$ The parameter vector θ is typically estimated using MLE.

Another example: estimation of Neural Network weights.

1 Likelihood function

Definition 1. *In the sampling model (on $\mathbb{R}$) with densities $f(\theta, x)$ the* **likelihood function** *of a random sample of size n $(X_1, \ldots, X_n)$ associated to $\{f(\theta, \cdot), \theta \in \Theta\}$ is:*

$$\boxed{\theta \in \Theta \mapsto \mathcal{L}_n(\theta, X_1, \ldots, X_n) = \prod_{i=1}^n f(\theta, X_i)}$$

- It's a random function
- It's the density of the observations

1.1 Examples

- Example 1: Poisson model. We observe

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Poisson}(\theta),$$

$\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$.

- The density of $\mathbb{P}_\theta$

$$f(\theta, x) = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 0, 1, 2, \ldots.$$

- The associated likelihood function

$$\begin{aligned}\theta \mapsto \mathcal{L}_n(\theta, X_1, \ldots, X_n) &= \prod_{i=1}^n e^{-\theta} \frac{\theta^{X_i}}{X_i!} \\ &= \frac{1}{\prod_{i=1}^n X_i!} e^{-n\theta} \theta^{\sum_{i=1}^n X_i}\end{aligned}$$

- Example 2 Cauchy model. We observe

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} \text{Cauchy centered in } \theta,$$

  $\theta \in \Theta = \mathbb{R}$.

- The density

$$\frac{d\,\mathbb{P}_\theta}{d\lambda}(x) = f(\theta, x) = \frac{1}{\pi\big(1 + (x-\theta)^2\big)}$$

- The associated **likelihood function**:

$$\theta \mapsto \mathcal{L}_n(\theta, X_1, \dots, X_n) = \frac{1}{\pi^n} \prod_{i=1}^{n} \frac{1}{\big(1 + (X_i - \theta)^2\big)}$$

2 Principle of maximum likelihood

- Consider a model **restricted to two points** with only two distributions: $\{\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_1}\}$

$$\Theta = \{\theta_1, \theta_2\} \subset \mathbb{R},$$

  where $\mathbb{P}_{\theta_i}$ discreet on $\mathbb{N}$.

- For any $(x_1, \dots, x_n) \in \mathbb{N}^n$, and $\theta \in \{\theta_1, \theta_2\}$,

$$\mathbb{P}_\theta\left[X_1 = x_1, \dots, X_n = x_n\right] = \prod_{i=1}^{n} \mathbb{P}_\theta\left[X_i = x_i\right] = \prod_{i=1}^{n} f(\theta, x_i).$$

  This is the probability to observe $(x_1, \dots, x_n)$ if the true distribution is $\mathbb{P}_\theta$.

For the observations $X_1, \dots, X_n$, the likelihood

$$\theta \in \Theta \mapsto \mathcal{L}_n(\theta, X_1, \dots, X_n) = \prod_{i=1}^{n} f(\theta, X_i)$$

is the **probability to have observed $X_1, \dots, X_n$** if the true distribution is given by $\mathbb{P}_\theta$.

The MLE (Maximum Likelihood Estimator) chooses the most likely θ: that is the parameter $\theta \in \Theta$ which **maximizes the probability to have observed $X_1, \dots, X_n$**

1. Case 1 : "θ_1 is more likely than θ_2" when

$$\prod_{i=1}^{n} f(\theta_1, X_i) \geq \prod_{i=1}^{n} f(\theta_2, X_i)$$

2. Case 2 : "θ_2 is more likely than θ_1" when

$$\prod_{i=1}^{n} f(\theta_2, X_i) > \prod_{i=1}^{n} f(\theta_1, X_i)$$

Principe of maximum likelihood:

$$\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{ml}} = \begin{cases} \theta_1 & \text{when } \theta_1 \text{ is more likely} \\ \theta_2 & \text{when } \theta_2 \text{ is more likely} \end{cases}$$

- We can generalize this principle for a family of distributions and any set of parameters .
- $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta$, $\{\mathbb{P}_\theta, \theta \in \Theta\}$ having density, $\Theta \subset \mathbb{R}^d$, $\theta \mapsto \mathcal{L}_n(\theta, X_1, \dots, X_n)$ associated likelihood.

Definition 2. *We call* **maximum likelihood estimator** *any estimator $\widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}}$ satisfying*

$$\mathcal{L}_n(\widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}}, X_1, \dots, X_n) = \max_{\theta \in \Theta} \mathcal{L}_n(\theta, X_1, \dots, X_n).$$

- Outline: **Existence, uniqueness, statistical properties**

Remarks:

- Log-likelihood:

$$\begin{aligned}\theta \mapsto \ell_n(\theta, X_1, \dots, X_n) &= \log \mathcal{L}_n(\theta, X_1, \dots, X_n) \\ &= \sum_{i=1}^n \log f(\theta, X_i).\end{aligned}$$

  Well defined if $f(\theta, \cdot) > 0$.

$$\text{Max. likelihood} = \text{max. log-likelihood.}$$

  (log-likelihood is sometimes easier to maximize)

- **Likelihood equation** :

$$\nabla_\theta \ell_n(\theta, X_1, \dots, X_n) = 0$$

Example 1: normal model We have a random sample of size n from the normal distribution $\mathcal{N}(\mu, \sigma^2)$, the parameter is $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$.

- **Likelihood**

$$\mathcal{L}_n((\mu, \sigma^2), X_1, \dots, X_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\big(-\tfrac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\big).$$

- **Log-likelihood**

$$\ell_n\big((\mu, \sigma^2), X_1, \dots, X_n\big) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2.$$

  Likelihood Equation(s) : $\nabla_\theta \ell_n(\theta, X_1, \dots, X_n) = 0$,

$$\begin{cases} \partial_\mu \ell_n\big((\mu, \sigma^2), X_1, \dots, X_n\big) & = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^n (X_i - \mu) \\[2em] \partial_{\sigma^2} \ell_n\big((\mu, \sigma^2), X_1, \dots, X_n\big) & = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^n (X_i - \mu)^2 \end{cases}$$

  Solution (for $n \geq 2$):

$$\boxed{\left(\overline{X}_n, \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right) = (\bar{X}_n, \hat{\sigma}_n)}$$

  and we check that it is a global maximum then $\widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}} = (\bar{X}_n, \hat{\sigma}_n)$.

Example 2: Poisson model

- **Likelihood**

$$\mathcal{L}_n(\theta, X_1, \dots, X_n) = \frac{1}{\prod_{i=1}^n X_i!} e^{-n\theta} \theta^{\sum_{i=1}^n X_i}$$

- **Log-likelihood**

$$\ell_n(\theta, X_1, \dots, X_n) = c(X_1, \dots, X_n) - n\theta + \sum_{i=1}^n X_i \log \theta$$

- **Equation of likelihood**

$$-n + \sum_{i=1}^n X_i \frac{1}{\theta} = 0, \ \text{ that is } \ \boxed{\widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}} = \frac{1}{n} \sum_{i=1}^n X_i = \overline{X}_n}$$

Example 3: Laplace model $X_1, \dots, X_n \overset{i.i.d.}{\sim}$ Laplace of parameter $\theta \in \Theta = \mathbb{R}:$ density

$$f(\theta, x) = \frac{1}{2\sigma} \exp\big(-\frac{|x-\theta|}{\sigma}\big),$$

where $\sigma > 0$ is **known**.

- **Likelihood**

$$\mathcal{L}_n(\theta, X_1, \dots, X_n) = (2\sigma)^{-n} \exp\big(-\frac{1}{\sigma} \sum_{i=1}^n \big|X_i - \theta\big|\big)$$

- **Log-likelihood**

$$\ell_n(\theta, X_1, \dots, X_n) = -n \log(2\sigma) - \frac{1}{\sigma} \sum_{i=1}^n \big|X_i - \theta\big|$$

Maximize $\mathcal{L}_n(\theta, X_1, \dots, X_n)$ is equivalent to minimize the function $\theta \mapsto \sum_{i=1}^n \big|X_i - \theta\big|$, which is differentiable almost everywhere with a derivate that is piecewise constant. **Equation of likelihood:**

$$\sum_{i=1}^n \mathrm{sign}(X_i - \theta) = 0.$$

let $X_{(1)} \leq \ldots \leq X_{(n)}$ order statistics.

- if n is even: $\widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}}$ is not unique; any point of the interval $\big[X_{\left(\frac{n}{2}\right)}, X_{\left(\frac{n}{2}+1\right)}\big]$ is a MLE.
- if n is odd: $\widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}} = X_{\left(\frac{n+1}{2}\right)}$, the MLE is unique.
- **for any** n, the empirical median is a MLE.

Example 4: Cauchy model Density

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

- **Likelihood**

$$\mathcal{L}_n(\theta, X_1, \dots, X_n) = \pi^{-n} \prod_{i=1}^n \frac{1}{1 + (X_i - \theta)^2}$$

- **Log-likelihood**

$$\ell_n(\theta, X_1, \dots, X_n) = -n \log \pi - \sum_{i=1}^n \log\big(1 + (X_i - \theta)^2\big)$$

- **Equation of likelihood**

$$\boxed{\sum_{i=1}^n \frac{X_i - \theta}{1 + (X_i - \theta)^2} = 0}$$

no explicit solution and generally admits several solutions.

2.1 Choice of the statistical model

- The data analyst has the choice of the family $\{\mathbb{P}_\theta, \theta \in \Theta\}$. The MLE depends on this choice.
- Example: we have the following sample ($n = 10$):

  $$0.92, -0.20, -1.80, 0.02, 0.49, 1.41, -1.59, -1.29, 0.34, 100$$

  We choose a shift model $f(x - \theta)$ for two different f:

  1. f density of normal distribution $\Rightarrow \widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}} = \overline{X}_n = 9.83$.
  2. f density of Laplace distribution $\Rightarrow$ any point of the interval $[0.02, 0.34]$ is a $\widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}}$, in particular, the median:
     $$\widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}} = Med(\hat{F}_n) = \widehat{q}_{n,1/2} = 0.02$$

- **Other choice of model ...**

3 Asymptotic properties of estimators

Sampling model: $(X_n)_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta \in \{\mathbb{P}_\theta : \theta \in \Theta\}$. Let $(\hat{\theta}_n)_n$ be an estimator. We say that:

1. $\widehat{\theta}_n$ is **consistent** if for any $\theta \in \Theta$,
   $$\boxed{\widehat{\theta}_n \xrightarrow{\mathbb{P}_\theta} \theta}$$
   ($\widehat{\theta}_n$ est **strongly consistent** if the cv is a.s.)
2. $\widehat{\theta}_n$ is **asymptotically normal** if, for any $\theta \in \Theta$, there exists a growing sequence of positives real numbers $(a_n) \uparrow \infty$ and V_θ, a random variable, such that:
   $$\boxed{a_n\big(\widehat{\theta}_n - \theta\big) \xrightarrow{d} V_\theta}$$
   When $V_\theta \sim \mathcal{N}(0, v(\theta))$, $v(\theta)$ is called the **asymptotic variance**; $1/a_n$ **is the asymptotic rate of convergence** (generally, $a_n = \sqrt{n}$)

Asymptotic properties of GMM, M and ML estimators is, in general, a delicate problem.

3.1 Asymptotic normality of MLE

- Setting: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta$ for $\theta \in \Theta \subset \mathbb{R}$
- $\{\mathbb{P}_\theta : \theta \in \Theta\}$ and the associated likelihood:
  $$\theta \mapsto \mathcal{L}_n(\theta, X_1, \ldots, X_n) = \prod_{i=1}^{n} f(\theta, X_i)$$
- The MLE is given by
  $$\widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}} \in \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta, X_1, \ldots, X_n)$$

Under certain conditions, we have
$$\sqrt{n}\big(\widehat{\theta}_{\mathrm{n}}^{\mathtt{ml}} - \theta\big) \xrightarrow{d} \mathcal{N}\big(0, v(\theta)\big)$$
where the asymptotic variance is given by
$$\boxed{v(\theta) = \frac{1}{\boldsymbol{I}(\theta)} \text{ where } \boldsymbol{I}(\theta) = \mathbb{E}_\theta\left[(\partial_\theta \log f(\theta, X))^2\right]}$$

In the sampling model $\{\mathbb{P}_\theta : \theta \in \Theta\}$ with density $f(\theta, x)$ we define

$$\ell(\theta, x) = \log f(\theta, x)$$

with the convention$(\log 0 = -\infty)$ and if $\ell(\cdot, x)$ is differentiable, we call :

1. **score function**: for a fixed $x \in \mathbb{R}$,

   $$\theta \mapsto \partial_\theta \ell(\theta, x)$$

   It indicates how sensitive a likelihood function is to its parameter θ.

2. we call Fisher information at $\theta \in \Theta$,

   $$\boldsymbol{I}(\theta) = \mathbb{E}_\theta\left[(\partial_\theta \ell(\theta, X))^2\right]$$

   It is a way of measuring the amount of information that each observation X_i carries about an unknown parameter θ of a distribution that models X_i.

So we have "under certain assumptions" that:

$$\sqrt{n}\big(\widehat{\theta}_{\text{n}}^{\,\mathtt{ml}} - \theta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{\boldsymbol{I}(\theta)}\Big)$$

Sufficient conditions to ensure the asymptotic **normality of MLE with the asymptotic variance $\boldsymbol{I}(\theta)^{-1}$** are at the origin of the definition of a regular model.

Conclusion: The study of the asymptotic normality of MLE leads us to introduce the following notions:

1. the score
2. Fisher information
3. a regular model

In the sampling model we observe

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta, \theta \in \Theta \subset \mathbb{R}$$

with the density $f(\theta, x)$ that models X_i. Then the density of the sample $\forall z = (x_1, \dots, x_n) \in \mathbb{R}^n$,

$$\boxed{f_n(\theta, z) = \prod_{i=1}^{n} f(\theta, x_i)}$$

Fisher's information contained in an observation $Z = (X_1, \dots, X_n)$ is

$$\boxed{\boldsymbol{I}_n(\theta) = \boldsymbol{I}(\theta|\mathcal{E}^n) = \mathbb{E}_\theta\left[(\partial_\theta \log f_n(\theta, Z))^2\right]}$$

Theorem 1.

$$\mathbb{I}(\theta \,|\, \mathcal{E}^n) = n\mathbb{I}(\theta \,|\, \mathcal{E})$$

Proposition 1. *In a regular model:*

$$\begin{aligned}\boldsymbol{I}(\theta) &= \mathbb{E}_\theta\left[(\partial_\theta \log f(\theta, X))^2\right] \\ &= -\,\mathbb{E}_\theta\, \partial_\theta^2 \log f(\theta, X) \\ &= -\partial_a^2 \mathbb{D}(a, \theta)\big|_{a=\theta}\end{aligned}$$

where $\mathbb{D}(a, \theta) = \mathbb{E}_\theta\left[\log f(a, X)\right]$.

- Numerical computations of a MLE can be difficult to achieve.

- If we have an estimator $\widehat{\theta}_n$ that is asymptoticly normal and if we can easily compute

$$\ell_n'(\theta) = \frac{1}{n}\sum_{i=1}^{n} \partial_\theta \log f(\theta, X_i), \quad \ell_n''(\theta) = \frac{1}{n}\sum_{i=1}^{n} \partial_\theta^2 \log f(\theta, X_i)$$

then we can correct $\widehat{\theta}_n$ so that we have the same asymptotic behavior as the MLE:

$$\boxed{\widetilde{\theta}_n = \widehat{\theta}_n - \frac{\ell_n'(\widehat{\theta}_n)}{\ell_n''(\widehat{\theta}_n)}} \quad \text{(Newton's algorithme)}$$

satisfies

$$\boxed{\sqrt{n}\big(\widetilde{\theta}_n - \theta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\theta)}\Big)}$$

Exercise: Distribution of genotypes in a population When the gene frequencies are in equilibrium, the genotypes AA, Aa and aa are manifested in a population with probabilities $(1-\theta)^2, 2\theta(1-\theta)$ and θ^2 respectively, where θ is an unknown parameter. Plato et al. (1964) published the following data on the type of haptoglobin (the protein that in humans is encoded by the HP gene) in a sample of 190 people:

Type of haptoglobin	Hp-AA	Hp-Aa	Hp-aa
Number of people	10	68	112

1. How to interpret the parameter θ? Propose a statistical model for this problem.
2. Calculate the maximum likelihood estimator $\hat{\theta}_n$ of θ
3. Give the asymptotic distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$
4. Construct 95% asymptotic confidence interval for θ.