Lecture 4: Decision theory and Cramer Rao efficiency

Sébastien Gadat

TSE

November 5, 2023

Loss function for the estimation problem

Let $(\Omega, \mathcal{P})$ be a parametric model with

$$\mathcal{P} = \{\mathbb{P}_\theta;\, \theta \in \Theta\}.$$

- **Objective :** guess the truth about the DGP (*i.e.* estimate θ_0) using the available observed data.
- Among **set of possible decisions $\mathcal{D}$,** what is the best achievable one ?
- Point estimation problem : $\mathcal{D} = \Theta$ and $r(x) = \hat{\theta}$.
- Below, we will focus on the Mean Square Error (M.S.E. for short) :

$$R(\theta_0, \hat{\theta}_n) = \mathbb{E}_{\theta_0}[\|\theta_0 - \hat{\theta}_n\|^2]$$

Loss decomposition

Loss function : map $L : \Theta \times \mathcal{D} \mapsto \mathbb{R}^+$, which assigns a non negative real number to each pair (θ, d) where $\theta \in \Theta$ and $d \in \mathcal{D}$ is a decision :

$$L(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|^2.$$

Important Result : Bias Variance decomposition :

$$\begin{aligned} R(\theta_0, \hat{\theta}_n) &= \mathrm{Bias}^2 + \mathrm{Var}(\hat{\theta}_n) \\ &= \left[\mathbb{E}_{\theta_0}[\hat{\theta}_n] - \theta_0\right]^2 + \mathrm{Var}(\hat{\theta}_n). \end{aligned}$$

Unbiased estimation : natural restriction ?

We sometimes restrict the class of estimators to unbiased ones :

$$\mathbb{E}_{\theta_0}[\hat{\theta}_n] = \theta_0.$$

Example : Consider $X_1, \ldots, X_n$ i.i.d. $\mathcal{U}([0, \theta])$

- $\hat{\theta}_n^{(1)} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- $\hat{\theta}_n^{(2)} = X_{n:n}$
- $\hat{\theta}_n^{(3)} = \lambda_n X_{n:n}$ where λ_n is computed to ensure that :

$$\mathbb{E}[\hat{\theta}_n^{(3)}] = \theta_0$$

Among the three estimators, what is the best one in terms of MSE ?

Regular models : definition of the Likelihood

Definition

We consider $\Theta \subset \mathbb{R}^d$ and a statistical model $\mathcal{P} = \{X, \mathcal{X}, \mathbb{P}_\theta;\ \theta \in \Theta\}$.
We assume that all the distributions $\mathbb{P}_\theta$ are a.c. w.r.t. a reference measure μ, with a density p_θ :

$$\mathbb{P}_\theta = p_\theta . \mu$$

For any $x \in \mathcal{X}$, we define the likelihood / log likelihood of x as :

$$L(\theta, x) = p_\theta(x) \qquad \text{and} \qquad \ell(\theta, x) = \log p_\theta(x)$$

The (log)-likelihood quantifies the plausibility to observe x when assuming the value θ of the parameter.

Likelihood and regular models

- (Log)-likelihood : key tool for our statistical / machine learning purpose these two years.
- Powerful for estimation, test, classification ...

Assume that we observe $(X_1, \dots, X_n)$, we denote by L_n/ℓ_n :

$$L_n(\theta) = p_\theta(X_1, \dots, X_n) \qquad \text{and} \qquad \ell_n(\theta) = \log L_n(\theta).$$

- L_n/ℓ_n is the (log)-likelihood computed at $\theta \in \Theta$
- L_n is a random function as it depends on the sample $(X_1, \dots, X_n)$.
- When $n = 1$, we simply denote the (log)-Likelihood by $L_\theta(x)$ and $\ell_\theta(x)$.
- When the sample is i.i.d., $\ell_n(\theta)$ is a sum of individual log-likelihood :

$$\ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(X_i).$$

Likelihood : examples

Several easy computations : imagine we observe $X_1, \ldots, X_n$ i.i.d. Give the reference measure μ and compute the log-likelihood of the next models.

- Gaussian model $\mathcal{N}(\mu, \sigma^2)$, and $\mathcal{N}(\mu, \Sigma^2)$
- Exponential model $\mathcal{E}(\theta)$
- Uniform model $\mathcal{U}([0, \theta])$
- Poisson $\mathcal{P}(\lambda)$
- Bernoulli $\mathcal{B}(p)$

Likelihood and regular models

We consider Θ an open set of $\mathbb{R}^d$ and a parametric model $\mathcal{P} = \{X, \mathcal{X}, \mathbb{P}_\theta;\, \theta \in \Theta\}$.

Definition (Regular model)

- For μ a.s. z, the function $\theta \mapsto p_\theta(z)$ is cont. differentiable on Θ
- We can switch ∇_θ and $\mathbb{E}_\theta$:

$$\nabla_\theta \int p_\theta(z) d\mu(z) = \int \nabla_\theta p_\theta(z) d\mu(z) = 0$$

- $$\int \|\nabla_\theta \ell(\theta, z)\|^2 p_\theta(z) d\mu(z) < +\infty$$

Fisher score

Assume that we have a regular model, we define the Fisher score as :

Definition (Fisher score)

For any r.v. Z and a parametric model $\mathcal{P} = \{X, \mathcal{X}, \mathbb{P}_\theta;\ \theta \in \Theta\}$, we define the score as :

$$S(\theta, Z) = \nabla_\theta[\ell_\theta(Z)].$$

For a regular model, the score is a centered random variable.

Likelihood and regular models : examples

Verify wether the three conditions for the following models hold or not.

- Uniform model $\mathcal{U}([0,\theta])$ (is not regular)
- Exponential model $\mathcal{E}(\theta)$ (is regular)
- Gaussian model $\mathcal{N}(\mu,\sigma^2), \theta = (\mu,\sigma^2)$ (is regular)
- Bernoulli model $\mathcal{B}(p)$ (is regular)
- Poisson model $\mathcal{P}(\lambda)$ (is regular)
- Geometric model $\mathcal{G}(p)$ (is regular)

Regular estimator

We assume that the statistical model $\mathcal{P} = \{X, \mathcal{X}, \mathbb{P}_\theta;\, \theta \in \Theta\}$ is regular.

Definition (Regular estimator)

An estimator T is a regular estimator of $g(\theta)$ is

- $T(Z)$ has a second order moment for any θ :

$$\mathbb{E}_\theta[T(Z)^2] < \infty.$$

- The function $\theta \longmapsto \mathbb{E}_\theta[T(Z)]$ is differentiable over Θ and

$$\forall \theta \in \Theta \qquad \nabla_\theta \mathbb{E}_\theta[T(Z)] = \int T(z) \nabla_\theta [p_\theta(z)] d\mu(z)$$

Fisher information

We assume that the statistical model $\mathcal{P} = \{X, \mathcal{X}, \mathbb{P}_\theta;\ \theta \in \Theta\}$ is regular. The final fundamental definition is as follows.

Definition (Fisher information)

The Fisher information of the model $\mathcal{P}$ is defined as :

$$\mathbb{I} : \theta \longmapsto \mathbb{E}_\theta \left[S(\theta, Z) S(\theta, Z)^T \right].$$

- $\mathbb{I}(\theta)$ is a $d \times d$ symetric and positive matrix.
- Since the score is a centered random variable :

$$\mathbb{I}(\theta) = \mathrm{Cov}\left(S(\theta, Z)\right).$$

Fisher information : examples

Compute the Fisher information in the following examples.

- Bernoulli model $\mathcal{B}(p)$

$$\mathbb{I}(p) = \frac{1}{p(1-p)}$$

- Binomial model $\mathcal{B}(n, p)$

$$\mathbb{I}(p) = \frac{n}{p(1-p)}$$

- Gaussian model $\mathcal{N}(\mu, 1)$

$$\mathbb{I}(\mu) = 1$$

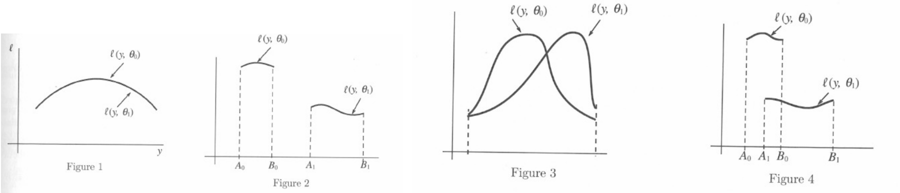
- Gaussian model $\mathcal{N}(\mu, \sigma^2)$

$$\mathbb{I}(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^2} \end{pmatrix}$$

Why Fisher information ?

Discuss a little about the term *information*, at least informally.

$$\log \frac{p_{\theta_0}(Z)}{p_{\theta_1}(Z)}.$$



1. Fig1 : Z does not permit to distinguish between θ_0 and θ_1. $\mathbb{I}$ is zero.
2. Fig2 : $Z \in [A_0, B_0]$: S is infinite. We can perfectly distinguish between θ_0 and θ_1
3. Fig3 : We cannot distinguish between θ_0 and θ_1 except with a real number. The log is positive when $p_{\theta_0}(Z) > p_{\theta_1}(Z)$.
4. Fig4 : Mix between 2 and 3.

Why Fisher information ?

Discuss a little about the term *information*, at least informally.

$$\log \frac{p_{\theta_0}(Z)}{p_{\theta_1}(Z)}.$$

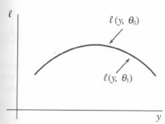


Figure 1

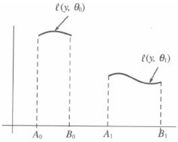


Figure 2

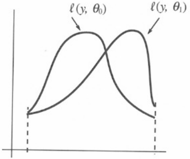


Figure 3

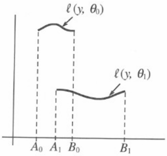


Figure 4

Implicitely, $\mathbb{I}$ is infinite when it is possible to perfectly identify θ without any mistake. It appears to be possible in Fig. 2 and Fig. 4.
Oppositely, $\mathbb{I}$ is 0 when it is *impossible*.
Reasonnably : situations like Fig 3. stand for the general case.

Exhaustivity

Imagine that you have at your disposal $Y = (X_1, \ldots, X_n)$, an i.i.d. sample of a Bernoulli model $\mathcal{B}(p)$.

- Instead of giving you Y, we only give you

$$S_n = \sum_{i=1}^{n} X_i$$

- Is there a loss of information ?

In our example, identify $\mathcal{L}(Y|S_n)$:

$$\mathbb{P}\left[Y = x | S_n = s\right] = \frac{\mathbb{P}\left[Y = x \& S_n = s\right]}{\mathbb{P}\left[S_n = s\right]}$$

$$= \begin{cases} 0 \quad \text{if} \quad \sum_{i=1}^{n} x_i \neq s \\ \frac{p^s(1-p)^{n-s}}{C_n^s p^s (1-p)^{n-s}} \quad \text{if} \quad \sum_{i=1}^{n} x_i = s \end{cases}$$

The conditional distribution is independent from p : means that once S_n is known the whole dependency of Y through p is determined.

Exhaustivity : definition

Definition (Exhaustive statistics)

We consider a statistical model $\mathcal{P} = \{X, \mathcal{X}, \mathbb{P}_\theta;\ \theta \in \Theta\}$. A statistics S is exhaustive if and only if

$$\forall \theta \in \Theta \qquad \mathcal{L}(X|S) \qquad \text{is independent from} \qquad \theta.$$

We can state a powerful criterion for exhaustivity.

Theorem (Factorization criterion for exhausitivity)

Consider a statistical model for which $\mathbb{P}_\theta$ is a.c. w.r.t. μ of density p_θ. S is exhaustive if and only if we can find g and ψ such that :

$$p_\theta(x) = g(x)\psi_\theta(S(x))$$

Exhaustivity : examples

Consider the Gaussian model of n i.i.d. samples $X = (X_1, \ldots, X_n)$ of $\mathcal{N}(\mu, 1)$. We verify that :

$$
\begin{aligned}
p_\theta(x) &= \prod_{i=1}^{n} \frac{\exp(-(X_i-\mu)^2/2)}{\sqrt{2\pi}} \\
&= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^2\right) \\
&= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}X_i^2\right) \exp\left(\mu\sum_{i=1}^{n}X_i - n\mu^2/2\right).
\end{aligned}
$$

We observe that S_n defined below is exhaustive :

$$S_n = \sum_{i=1}^{n} X_i$$

We just have to use the factorization criterion !

Exhaustivity and Information

We consider X a random variable and S a statistics. We denote by $\mathbb{I}_S(\theta)$ the Fisher information on θ brought by S in the image model.

Theorem

- $\mathbb{I}_S(\theta) \leq \mathbb{I}_X(\theta)$
- *If S is exhaustive, then :* $\mathbb{I}_S(\theta) = \mathbb{I}_X(\theta)$.
- *If S and T are independent, then* $\mathbb{I}_{(S,T)}(\theta) = \mathbb{I}_S(\theta) + \mathbb{I}_T(\theta)$.

Cramer-Rao lower bound

To be continued in Semester 2.