Lecture 8 —Parameter Estimation

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

Overview. We distinguish between a parameter of interest and nuisance parameters, and define what it means for a parametric model to be identifiable. We then discuss two systematic ways to estimate a parameter of interest: 1) the plug-in method, and 2) maximum likelihood estimation.

1 Parameters and identifiability

Given a statistical model with several parameters, we may only be interested in some of them, or in a function of the parameters. Parameters we care about are "parameters of interest". Parameters we don't care about are "nuisance" parameters.

Example.

> Consider the statistical model $\{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma \geq 0\}$. The following are examples of parameters of interest vs nuisance parameters:
>
> 1. $\theta = (\mu, \sigma^2)$ is the (two-dimensional) parameter of interest
> 2. $\theta = \mu$ is the parameter of interest, and σ^2 is the nuisance parameter
> 3. μ is the nuisance parameter, and $\theta = \sigma^2$ is the parameter of interest.
> 4. $\theta = \mu/\sigma$ is the parameter of interest, and μ, σ^2 are nuisance parameters.

1.1 Identifiability

Note that we observe data $X_1, \ldots, X_n$ from the *distribution* $\mathbb{P}_\theta$, which means that we are only indirectly collecting information about the *parameter* θ. We can only hope to recover θ if the model is *identifiable*:

Definition 1.1: Identifiability

The full parameter θ is identifiable from the statistical model $\{\mathbb{P}_\theta \mid \theta \in \Theta\}$ if

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \implies \theta = \theta'.$$

Equivalently, distinct parameters θ, θ' correspond to distinct probability distributions $\mathbb{P}_\theta, \mathbb{P}_{\theta'}$.
A *parameter of interest* $f(\theta)$ is identifiable from the statistical model $\{\mathbb{P}_\theta \mid \theta \in \Theta\}$ if

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \implies f(\theta) = f(\theta').$$

Example.

> If the statistical model is $\{\mathcal{N}(0, \sigma^2) \mid \sigma \in \mathbb{R}\}$, then σ^2 is identifiable but σ is *not* identifiable (it could be positive or negative). If the statistical model is $\{\mathcal{N}(0, \sigma^2) \mid \sigma \geq 0\}$ then σ *is* identifiable.

2 Methods to estimate parameters

2.1 The plug-in method

Informally, the "plug-in" method can be represented as $\mathbb{E} \rightsquigarrow \frac{1}{n}\sum_{i=1}^n$ (recall $\rightsquigarrow$ means "estimate by"). In other words, if a parameter can be written in terms of expectations, replace each expectation you see by the corresponding sample average.

Example.

> $X_1, \ldots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.
>
> - μ is the parameter of interest. $\mu = \mathbb{E}[X_1]$, so we take
>
> $$\mu \rightsquigarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$
>
> - σ^2 is the parameter of interest. $\sigma^2 = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2$, so we take
>
> $$\sigma^2 \rightsquigarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^2.$$
>
> - μ/σ is the parameter of interest. $\mu/\sigma = \mathbb{E}[X_1]/(\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2)^{1/2}$, so
>
> $$\mu/\sigma \rightsquigarrow \frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right)^{1/2}}.$$

2.2 Maximum Likelihood

Consider a statistical model $\{\mathbb{P}_\theta \mid \theta \in \Theta\}$. We observe data $X_1, \ldots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{\theta^*}$. Here, θ^* is the true parameter, which we should think of as fixed. In contrast, $\theta \in \Theta$ will be allowed to vary.

Definition 2.1: (Log) likelihood and maximum likelihood estimator

Let $f_\theta(x)$ be the pdf corresponding to $\mathbb{P}_\theta$. The *likelihood* function is

$$L_n(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

The *log likelihood* function is

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

The *maximum likelihood estimator* $\hat{\theta}_n$ (MLE) is the point which maximizes the function ℓ_n, i.e.

$$\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} \ell_n(\theta) = \operatorname{argmax}_{\theta \in \Theta} L_n(\theta).$$

(Note that the maximum value of $\ell_n = \log L_n$ is different from the maximum value of L_n, but the point at which the maximum is achieved is the same for both functions.)

The short and sweet interpretation of maximum likelihood: $L_n(\theta)$ is the probability to observe i.i.d. samples $X_1, \ldots, X_n$ under the distribution with parameter θ. Out of all possible θ's, we find the one for which this probability is greatest. This θ — which is the MLE $\hat{\theta}_n$ — *is most likely to have generated the data* $X_1, \ldots, X_n$, hence the term maximum likelihood.

Exercise: make sure you can compute the MLE for the following models:

$$\begin{aligned} \{\text{Bernoulli}(p) \mid p \in [0,1]\} \\ \{\mathcal{N}(\mu, 1) \mid \mu \in \mathbb{R}\} \\ \{\text{Unif}([0,\theta]) \mid \theta \geq 0\}. \end{aligned}$$

2.3 Where the MLE comes from, more formally

We will see in this section that the MLE stems from the following procedure: for each θ, we compute an approximation $\widehat{\text{dist}}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ to the exact distance $\text{dist}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ between $\mathbb{P}_{\theta^*}$ and $\mathbb{P}_\theta$. We then find the θ for which $\widehat{\text{dist}}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ is the smallest. In other words, we find the minimizer of the function $\theta \mapsto \widehat{\text{dist}}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$. This minimizer will be our estimate for θ.

Properties that $\text{dist}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ **should satisfy**:

1. **Computable from samples.** We can't compute the exact distance $\text{dist}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ because we don't know θ^*. But we *do* have access to samples $X_1, \ldots, X_n$ from $\mathbb{P}_{\theta^*}$. So we should choose a distance metric which can be approximated using the samples.

2. **Minimized *only* at θ^*.** Consider the ideal case in which we *could* compute $\text{dist}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ for each θ. This distance should have the property that

$$\text{dist}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) \quad \begin{matrix} = 0 & \text{if} & \theta = \theta^* \\ > 0 & \text{if} & \theta \neq \theta^* \end{matrix} \tag{1}$$

Definition 2.2: Kullback-Leibler (KL) divergence

Let f_{θ^*}, f_θ be the pdfs associated to $\mathbb{P}_{\theta^*}$ and $\mathbb{P}_\theta$, respectively. The KL divergence between $\mathbb{P}_{\theta^*}$ and $\mathbb{P}_\theta$ is defined as

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \,||\, \mathbb{P}_\theta) &= \int f_{\theta^*}(x) \log\left(\frac{f_{\theta^*}(x)}{f_\theta(x)}\right) dx \\ &= \int f_{\theta^*}(x) \log f_{\theta^*}(x) dx - \int f_{\theta^*}(x) \log f_\theta(x) dx. \end{aligned} \tag{2}$$

We make the following important observations.

- The KL divergence is actually a "divergence", not a distance. It doesn't satisfy the triangle inequalities and it is not symmetric: $D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \,||\, \mathbb{P}_\theta) \neq D_{\mathrm{KL}}(\mathbb{P}_\theta \,||\, \mathbb{P}_{\theta^*})$.

- Nevertheless, the KL divergence satisfies the property (1): $D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \,||\, \mathbb{P}_\theta) \geq 0$ always (this can be proved using Jensen's inequality), but $D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \,||\, \mathbb{P}_\theta) = 0$ *only* if $\mathbb{P}_\theta = \mathbb{P}_{\theta^*}$. If the parameter is *identifiable* (recall Definition 1.1), this implies θ^* is the unique minimizer of the function $\theta \mapsto D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \,||\, \mathbb{P}_\theta)$.

- Consider the second line in the equation (1). Note that the term $\int f_{\theta^*}(x) \log f_{\theta^*}(x) dx$ does not depend on the variable θ, only on the fixed point θ^*. Therefore, we can drop it when minimizing the KL divergence:

$$\begin{aligned} \mathrm{argmin}_{\theta \in \Theta} D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \,||\, \mathbb{P}_\theta) &= \mathrm{argmin}_{\theta \in \Theta} \left[- \int f_{\theta^*}(x) \log f_\theta(x) dx \right] \\ &= \mathrm{argmax}_{\theta \in \Theta} \int f_{\theta^*}(x) \log f_\theta(x) dx \\ &= \mathrm{argmax}_{\theta \in \Theta} \mathbb{E}\left[\log f_\theta(X)\right], \quad X \sim \mathbb{P}_{\theta^*}. \end{aligned} \tag{3}$$

  To get the third line, we used that $\mathbb{E}[g(X)] = \int f(x) g(x) dx$ for a random variable X that has pdf $f(x)$.

- Recall the plug-in approach: expectations can be replaced by sample averages! We can finally use our samples $X_1, \ldots, X_n$. Continuing from the third line of (3), we get

$$\mathrm{argmax}_{\theta \in \Theta} \mathbb{E}\left[\log f_\theta(X)\right] \approx \mathrm{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log f_\theta(X_i) =: \hat{\theta}_n.$$

  The function being maximized on the right is *precisely* the log likelihood from Definition 2.1. Thus, we have recovered the original definition of the MLE.