1 Point estimation

1.1 Maximum Likelihood Estimators

The maximum likelihood method is the most popular way to estimate the parameter θ which specifies a probability model. The maximum likelihood estimate is the value $\hat{\theta}$ which maximize the likelihood function. That is, the maximum likelihood estimation chooses the model parameter $\hat{\theta}$ which is the **most likely to generate the observed data.**

Definition 1.1. *Let $X_{1:n} \sim f_\theta$, where $\theta \in \Theta$. Denote the likelihood function of θ by $L(\theta) = L(\theta \mid X_{1:n}) = f_\theta(X_{1:n})$, and log-likelihood function by $\ell(\theta) = \log L(\theta)$. The maximum likelihood estimator (MLE) of θ is*

$$\widehat{\theta} = \underset{\theta \in \Theta}{\arg\max} L(\theta) = \underset{\theta \in \Theta}{\arg\max} \ell(\theta)$$

Here we briefly review three good properties of the maximum likelihood estimation:

Theorem	Sketch of Proof
consistency (asymptotic correctness)	positivity of Kullback-Leibler divergence
asymptotic normality: $\hat{\theta} \sim \mathcal{N}\left(\bar{\theta}, \frac{1}{n} I^{-1}\right)$	central limit theorem for delta method
efficiency (minimum variance)	Cauchy-Schwarz inequality

Table 1: Theorems for maximum likelihood estimation.

To study the theoretical properties of MLEs, we need the notations below.

Definition 1.2. . *Let f and g be densities with support $\mathcal{X}$. The Kullback-Leibler (KL) distance between f and g is*

$$\mathrm{KL}(f \mid g) = \int_x f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x = \mathsf{E}\left(\log \frac{f(X)}{g(X)}\right)$$

if we assume that $X \sim f$

Remark 1.1. The definition is motivated by concepts from information theory and relative entropy.

- **relative entropy** measures the **inefficiency** or **additional information required** when using one distribution (denoted Q) to approximate or represent another distribution (denoted P).
- The KL distance is the difference of entropy between true model and another model.

We can see that

- $\mathrm{KL}(f \mid g) \geq 0$ since

$$\mathsf{E}\left(\log \frac{f(X)}{g(X)}\right) = \mathsf{E}\left(-\log \frac{g(X)}{f(X)}\right) \geq -\log \mathsf{E}\left(\frac{g(X)}{f(X)}\right) = -\log \int_{\mathcal{X}} \frac{g(x)}{f(x)} f(x) dx = 0$$

  and the equality holds iff $f = g$.
- $\mathrm{KL}(f \mid f) = 0$, and
- $\mathrm{KL}(f \mid g) \neq \mathrm{KL}(g \mid f)$ in general.

We consider well-specified and identifiable statistical models. Now define

$$\widehat{M}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_\star}(X_i)} \quad \text{and} \quad M(\theta) = \mathrm{E}_\star\left\{\log \frac{f_\theta(X_1)}{f_{\theta_\star}(X_1)}\right\}$$

Remark 1.2. Let $M(\theta)$ represent the negative KL divergence. Our objective is to find an estimator that minimizes this divergence, i.e.,

$$\theta_\star = \underset{\theta \in \Theta}{\arg\max}\, M(\theta).$$

However, since $M(\theta)$ cannot be computed directly, we use an empirical estimate $\widehat{M}(\theta)$ as a substitute. Thus, we seek

$$\widehat{\theta}=\arg\max_{\theta\in\Theta} \widehat{M}(\theta),$$

which is equivalent to the maximum likelihood estimator (MLE). Therefore, finding the MLE is equivalent to identifying the estimator that minimizes the estimated KL divergence.

Definition 1.3. *The statistical model $\mathscr{F}=\{f_\theta:\theta\in\Theta\}$ is said to be identifiable if*

$$\forall\theta\neq\theta',\quad \mathrm{KL}\left(f_{\theta'}\mid f_\theta\right)>0$$

Definition 1.4. *Let $f_\star$ be the data generating density. The model $\mathscr{F}=\{f_\theta:\theta\in\Theta\}$ is said to be well-specified if*

$$\forall\theta\neq\theta',\quad \mathrm{KL}\left(f_{\theta'}\mid f_\theta\right)>0$$

1.1.1 Consistency

The first property is that, as the number of observations n becomes large, the estimate $\hat{\theta}$ converges to the true value $\theta_\star$.

Theorem 1.1. *Assume:*

$$\begin{aligned}
\text{(Maximizer:)}\qquad & \widehat{\theta}=\arg\max_{\theta\in\Theta}\widehat{M}(\theta),\\
\text{(Uniform LLN):}\qquad & \sup_{\theta\in\Theta}|\widehat{M}(\theta)-M(\theta)|\xrightarrow{\text{pr}}0,\\
\text{(Well-separation):}\qquad & \forall\varepsilon>0,\quad \sup_{\theta:|\theta-\theta_\star|>\varepsilon} M(\theta)<M\left(\theta_\star\right).
\end{aligned}$$

Then $\widehat{\theta}\xrightarrow{\text{pr}}\theta_\star$.
Intuition*: Since two functions $\widehat{M}(\theta)$ and $M(\theta)$ are getting closer, the points of maximum should also get closer which exactly means that $\hat{\theta}\to\theta_0$.*

Proof. First, we examine the difference in the true Kullback-Leibler (KL) divergence between the true parameter $\theta_\star$ and the maximum likelihood estimator (MLE) $\hat{\theta}$:

$$\begin{aligned}
0\le M(\theta_\star)-M(\hat{\theta}) &= \left\{M(\theta_\star)-\widehat{M}(\hat{\theta})\right\}+\left\{\widehat{M}(\hat{\theta})-M(\hat{\theta})\right\} && \text{Add and subtract } \widehat{M}(\hat{\theta})\\
&\le \left\{M(\theta_\star)-\widehat{M}(\theta_\star)\right\}+\left\{\widehat{M}(\hat{\theta})-M(\hat{\theta})\right\} && \hat{\theta}\text{ maximizes } \widehat{M}(\theta)\\
&\le 2\sup_{\theta\in\Theta}\left|\widehat{M}(\theta)-M(\theta)\right|\xrightarrow{\text{pr}}0. && \text{Uniform convergence}
\end{aligned}$$

indicating that the difference converges to zero in probability. Given the strong identifiability of $\theta_\star$, there exists a $\delta>0$ such that:

$$|\theta-\theta_\star|\ge\varepsilon\quad\Rightarrow\quad M(\theta)<M\left(\theta_\star\right)-\delta$$

Note that if $A\Rightarrow B$, then $\mathrm{P}(A)\le\mathrm{P}(B)$. So, putting $\theta=\widehat{\theta}$, we have

$$\mathrm{P}\left(\left|\widehat{\theta}-\theta_\star\right|\ge\varepsilon\right)\le\mathrm{P}\left\{M(\widehat{\theta})<M\left(\theta_\star\right)-\delta\right\}\to0$$

□

There are many different versions of conditions for proving consistency of MLE. Many of them **rely heavily on the uniform LLN.** We provide an alternative set of conditions.

Theorem 1.2. *Let $\{f_\theta : \theta \in \Theta\}$ be the model, and $X_1, X_2, \ldots \overset{IID}{\sim} f_{\theta_\star}$. Assume*

1. *(Compactness.) Θ is compact;*
2. *(Uniqueness of maximizer.) $\theta^\star$ is the unique maximizer of $\theta \mapsto M(\theta)$;*
3. *(Continuous.) $M(\theta)$ is continuous in θ;*
4. *(Uniform LLN.) $\widehat{M}(\theta)$ converges uniformly in probability.*

Then $\widehat{\theta} \xrightarrow{\text{pr}} \theta_\star$.

Proof. For $\epsilon > 0$, define the ϵ-neighborhood around $\theta_\star$ as:

$$\Theta(\epsilon) = \{\theta : \|\theta - \theta_\star\| < \epsilon\}$$

We aim to show that

$$P_{\theta_0}\left[\hat{\theta} \in \Theta(\epsilon)\right] \to 1$$

Since $\Theta(\epsilon)$ is an open set, we know that $\Theta \cap \Theta(\epsilon)^C$ is a compact set. Since $M(\theta)$ is a continuous function, then $\sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} \{M(\theta)\}$ is achieved for a θ in this compact set. Denote this value by θ_0. Since $\theta_\star$ is the unique max, let $M(\theta_\star) - M(\theta_0) = \delta > 0$. For any θ, we distinguish between two cases.

- $\theta \in \Theta \cap \Theta(\epsilon)^C$.

  Let A_n be the event that $\sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} \left|M(\theta) - \widehat{M}(\theta)\right| < \delta/2$. Then

$$\begin{aligned} A_n \Rightarrow \widehat{M}(\theta) &< M(\theta) + \delta/2 \\ &\leq M(\theta_0) + \delta/2 \\ &= M(\theta_\star) - \delta + \delta/2 \\ &= M(\theta_\star) - \delta/2 \end{aligned}$$

- $\theta \in \Theta(\epsilon)$.

  Let B_n be the event that $\sup_{\Theta(\epsilon)} \left|M(\theta) - \widehat{M}(\theta)\right| < \delta/2$. Then

$$\begin{aligned} B_n &\Rightarrow \widehat{M}(\theta) > M(\theta) - \delta/2 \text{ for all } \theta \\ &\Rightarrow \widehat{M}(\theta) > M(\theta_\star) - \delta/2 \end{aligned}$$

We conclude that if both A_n and B_n hold then $\hat{\theta} \in \Theta(\epsilon)$. By the proof of theorem 1.1, we know that as long as $\widehat{M}(\theta)$ converges uniformly in probability, $M(\theta_\star) - M(\widehat{\theta}) \xrightarrow{\text{pr}} M(\theta)$. Comparing the two cases above we have $\hat{\theta} \in \Theta(\epsilon)$. □

A key element of above two proof is that converges uniformly in probability. But it is difficult to prove.

Lemma 1.3. *Let $\{f_\theta : \theta \in \Theta\}$ be the model, and $X_1, X_2, \ldots \overset{IID}{\sim} f_{\theta_\star}$. Θ is compact, $\log f(x;\theta)$ is continuous in θ for all $\theta \in \Theta$ and all $x \in \mathscr{X}$, and if there exists a function $d(x)$ such that $|\log f(x;\theta)| \leq d(x)$ for all $\theta \in \Theta$ and $x \in \mathscr{X}$, and $E_{\theta_0}[d(X)] < \infty$, then*

- *$M(\theta)$ is continuous in θ;*
- $\sup_{\theta \in \Theta} \left|\widehat{M}(\theta) - M(\theta)\right| \xrightarrow{\text{pr}} 0$

Proof. To establish continuity, we need to demonstrate that if $\theta_k \to \theta$, then $M(\theta_k) \to M(\theta)$. Specifically, we need to show that

$$M(\theta_k) = \mathsf{E}\left(\log\left(\frac{f_{\theta_k}}{f_{\theta_\star}}\right)\right) \to M(\theta) = \mathsf{E}\left(\log\left(\frac{f_\theta}{f_{\theta_\star}}\right)\right)$$

By the continuity of f_θ , we know that $\log f_{\theta_k} \to \log f_\theta$. Furthermore, since $\log f_\theta \leq d(X)$ and $\mathsf{E}d(X) < \infty$, we can apply the dominated convergence theorem, which ensures the desired convergence holds. Since Θ is compact, we also know that $M(\theta)$ is **uniformly continuous**.

Regarding uniform convergence, we need to establish that

$$\sup_{\theta\in\Theta}\left|\widehat{M}(\theta)-M(\theta)\right| \overset{\text{pr}}{\to} 0$$

We have already shown that $M(\theta)$ is uniformly continuous. Now, we consider the properties of $\widehat{M}(\theta)$. Since $\widehat{M}(\theta)$ is the average of the log-likelihood, its behavior is closely related to the properties of the likelihood function $f_\theta(x)$. As $f_\theta(x)$ is continuous in Θ and Θ is compact, it follows that $\log f_\theta(x)$ is also uniformly continuous. Thus, for any $\epsilon > 0$, there exists $\delta(\epsilon)$ such that if $\|\theta_1-\theta_2\| < \delta$, we have

$$\sup_{\|\theta_1-\theta_2\|<\delta} |\log f_{\theta_1}(x) - \log f_{\theta_2}(x)| < \epsilon$$

which implies that

$$\Delta(x,\delta) = \sup_{\|\theta_1-\theta_2\|<\delta} |\log f_{\theta_1}(x) - \log f_{\theta_2}(x)| \to 0$$

Therefore, if $\|\theta_1-\theta_2\| < \delta$, for all $x \in \mathscr{X}$, we obtain

$$\begin{aligned}\left|\widehat{M}(\theta_1)-\widehat{M}(\theta_2)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}(\log f_{\theta_1}(x_i) - \log f_{\theta_2}(x_i))\right| \\ &\le \frac{1}{n}\sum_{i=1}^{n}|\log f_{\theta_1}(x) - \log f_{\theta_2}(x)| \\ &\le \frac{1}{n}\sum_{i=1}^{n}\Delta(x,\delta) \to 0\end{aligned}$$

So $\widehat{M}(\theta)$ is also uniform continuous.
Now we can maginify the target inequality and use the properties of uniformly continuous.
To begin, we first cut off our set Θ by considering an open ball of radius δ around each $\theta \in \Theta$, i.e., $B(\theta,\delta) = \{\tilde{\theta} : \|\tilde{\theta}-\theta\| < \delta\}$. The union of these balls forms an open cover of Θ. By compactness, we can find a finite subcover, denoted as $\{B(\theta_j,\delta), j = 1,\dots,J\}$. For each $\theta \in \Theta$, there exists a θ_j such that $\theta \in B(\theta_j,\delta)$. Therefore, for any $\theta \in \Theta$, we have

$$|\widehat{M}(\theta)-M(\theta)| \le |\widehat{M}(\theta)-\widehat{M}(\theta_j)| + |\widehat{M}(\theta_j)-M(\theta_j)| + |M(\theta_j)-M(\theta)|$$

Since $\widehat{M}(\theta)$ and $M(\theta)$ are uniformly continuous, we can choose δ small enough such that both $|\widehat{M}(\theta)-\widehat{M}(\theta_j)|$ and $|M(\theta_j)-M(\theta)|$ can be bounded by $\epsilon/3$.
So

$$|\widehat{M}(\theta)-M(\theta)| \le \epsilon/3 + \max_{j=1,2,\cdots,J}|\widehat{M}(\theta_j)-M(\theta_j)| + \epsilon/3$$

which implies that

$$\sup_{\theta\in\Theta}|\widehat{M}(\theta)-M(\theta)| \le 2\epsilon/3 + \max_{j=1,2,\cdots,J}|\widehat{M}(\theta_j)-M(\theta_j)|$$

Further, we have

$$\sup_{\theta\in\Theta}|\widehat{M}(\theta)-M(\theta)| > \epsilon \implies \max_{j=1,2,\cdots,J}|\widehat{M}(\theta_j)-M(\theta_j)| > \epsilon/3$$

We now show that

$$\mathsf{P}\left(\max_{j=1,2,\cdots,J}|\widehat{M}(\theta_j)-M(\theta_j)| > \epsilon/3\right) \overset{\text{pr}}{\to} 0.$$

It is easy since

$$\begin{aligned}\mathsf{P}\left(\max_{j=1,2,\cdots,J}|\widehat{M}(\theta_j)-M(\theta_j)| > \epsilon/3\right) &= \mathsf{P}\left(\bigcup_{j=1,2,\cdots,J}\left\{|\widehat{M}(\theta_j)-M(\theta_j)| > \epsilon/3\right\}\right) \\ &\le \sum_{j=1,2,\cdots,J}\mathsf{P}\left(\left\{|\widehat{M}(\theta_j)-M(\theta_j)| > \epsilon/3\right\}\right)\end{aligned}$$

then by WLLN, we know that for each θ_j and for any $\epsilon > 0, \eta > 0$ there exists $N(\epsilon, \eta)$ so that for all $n > N_j(\epsilon, \eta)$

$$\mathsf{P}\left(\left\{|\widehat{M}(\theta_j) - M(\theta_j)| > \epsilon/3\right\}\right) < \eta/J$$

Let $N = \max N_j$ we have that

$$\sum_{j=1,2,\cdots,J} \mathsf{P}\left(\left\{|\widehat{M}(\theta_j) - M(\theta_j)| > \epsilon/3\right\}\right) < \eta$$

Finally,

$$\mathsf{P}\left(\sup_{\theta\in\Theta}\left|\widehat{M}(\theta) - M(\theta)\right| > \epsilon\right) \leq \mathsf{P}\left(\max_{j=1,2,\cdots,J}|\widehat{M}(\theta_j) - M(\theta_j)| > \epsilon/3\right) \stackrel{\text{pr}}{\to} 0$$

which completes the proof. □