4. Consistency of Maximum Likelihood Estimator

Review: Definition of MLE

0/1 point (graded)

Let $\left\{E, (\mathbf{P}_\theta)_{\theta\in\Theta}\right\}$ be a statistical model associated with a sample of i.i.d. random variables $X_1, X_2, \ldots, X_n$. Assume that there exists $\theta^* \in \Theta$ such that $X_i \sim \mathbf{P}_{\theta^*}$.

Recall the **Kullback-Leibler (KL) divergence** between two distributions $\mathbf{P}_{\theta^*}$ and $\mathbf{P}_\theta$, with pdfs p_{θ^*} and p_θ respectively, is defined as

$$\mathrm{KL}\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right) = \mathbb{E}_{\theta^*}\left[\ln\left(\frac{p_{\theta^*}(X)}{p_\theta(X)}\right)\right],$$

and a consistent estimator of $\mathrm{KL}\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right)$ is

$$\widehat{\mathrm{KL}}_n\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right) = \text{a constant} - \frac{1}{n}\sum_{i=1}^{n}\ln p_\theta(X_i).$$

Which of the following represents the maximum likelihood estimator of θ^*? (Choose all that apply).

- [x] $\mathrm{argmin}_{\theta\in\Theta}\widehat{\mathrm{KL}}_n\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right)$
- [x] $\mathrm{argmax}_{\theta\in\Theta}\sum_{i=1}^{n}\ln p_\theta(X_i)$
- [x] $\mathrm{argmax}_{\theta\in\Theta}\ln\left(\prod_{i=1}^{n}p_\theta(X_i)\right)$
- [ ] $\mathrm{argmax}_{\theta\in\Theta}\ln\left(L_n(X_1, X_2, \ldots, X_n; \theta)\right)$

Solution:

Recall the **maximum likelihood estimator** can defined as the

$$\hat{\theta}_n^{MLE} = \mathrm{argmin}_{\theta\in\Theta}\widehat{\mathrm{KL}}_n\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right).$$

In other words, the maximum likelihood estimator is the (unique) θ that minimizes $\widehat{\mathrm{KL}}\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right)$ over the parameter space $\theta \in \Theta$. (The minimizer of the KL divergence is unique due to it being strictly convex in the space of distributions once $\mathbf{P}_{\theta^*}$ is fixed.)
All choices are equivalent to this definition:

$$\begin{aligned}
\hat{\theta}_n^{MLE} = \mathrm{argmin}_{\theta\in\Theta}\widehat{\mathrm{KL}}_n\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right) &= \mathrm{argmin}_{\theta\in\Theta}\left(\text{Constant} - \frac{1}{n}\sum_{i=1}^{n}\ln p_\theta(X_i)\right) \\
&= \mathrm{argmax}_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i=1}^{n}\ln p_\theta(X_i)\right) \qquad \text{(drop additive constant and negative sign))} \\
&= \mathrm{argmax}_{\theta\in\Theta}\left(\sum_{i=1}^{n}\ln p_\theta(X_i)\right) \qquad \text{(drop positive scaling factor)}
\end{aligned}$$

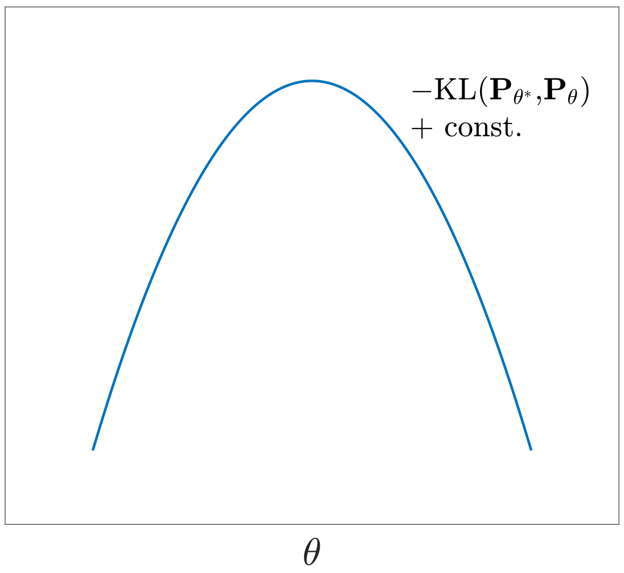
$$= \text{argmax}_{\theta\in\Theta} \left(\ln\left(\prod_{i=1}^{n} p_\theta\left(X_i\right)\right)\right) \qquad \text{(log property)}$$

$$= \text{argmax}_{\theta\in\Theta} \ln\left(L_n\left(X_1, X_2, \ldots, X_n; \theta\right)\right) \qquad \text{(definition of likelihood)}\,.$$

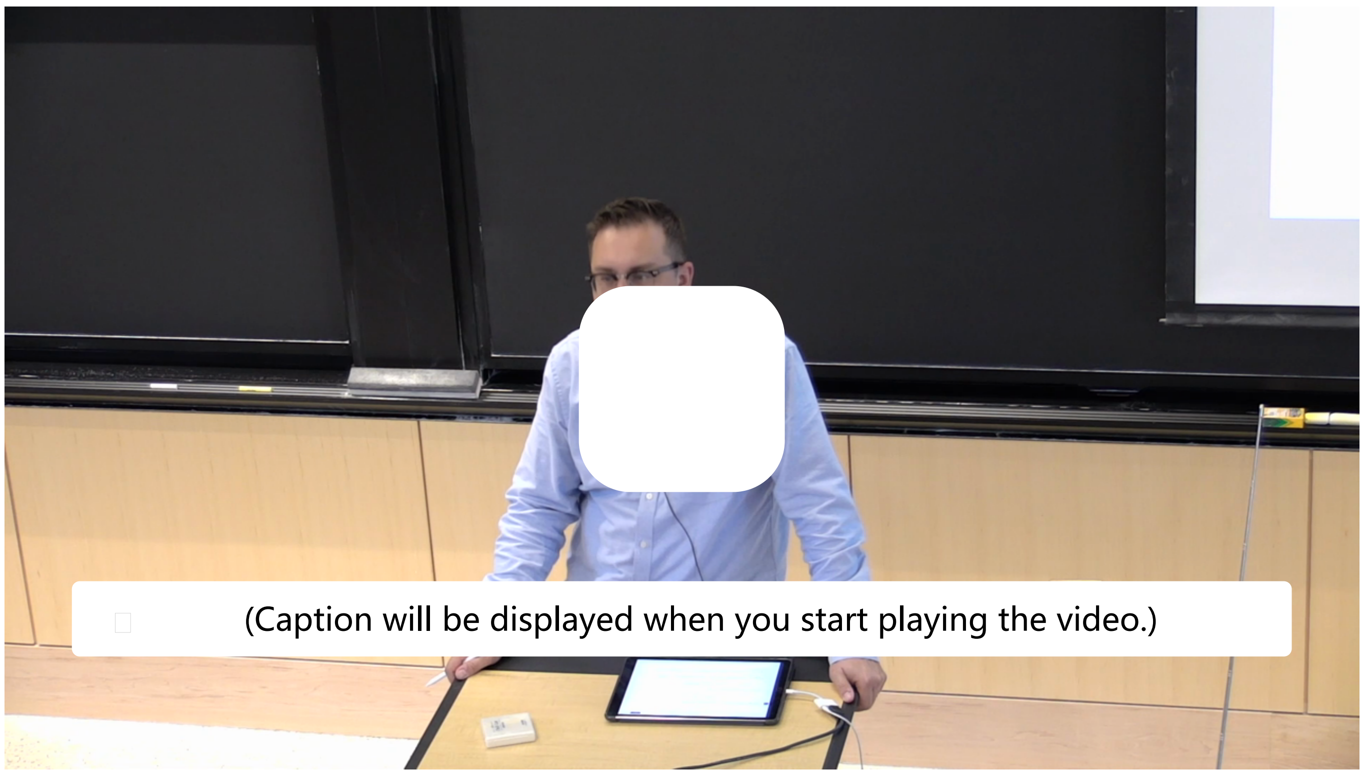
提交 你已经尝试了3次（总共可以尝试3次）

Answers are displayed within the problem

Note: In the following video, at around the 3:20 mark, the plot of $-\text{KL}\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right)$, with θ^* fixed and as a function of θ, is presented incorrectly as a convex curve while it should be concave. This error propagates until the end of the video and we request you to keep the following picture in mind instead:



Consistency of the Maximum Likelihood Estimator



Start of transcript. Skip to the end.

OK, so now I have this MLEs.

And I have two ways of computing MLEs,

either setting derivative equal to 0

or just looking at the plot.

And once I do this, which is really just one way, which

is taking the maximum of the likelihood, once I have this,

I would like you--

you might question whether this estimator is

视频
下载视频文件

字幕
下载 SubRip (.srt) file
下载 Text (.txt) file

Consistency of MLE

Given i.i.d samples $X_1, \ldots, X_n \sim \mathbf{P}_{\theta^*}$ and an associated statistical model $\left(E, \{\mathbf{P}_\theta\}_{\theta\in\Theta}\right)$, the maximum likelihood estimator $\hat{\theta}_n^{\text{MLE}}$ of θ^* is a **consistent** estimator under mild regularity conditions (e.g. continuity in θ of the pdf p_θ almost everywhere), i.e.

$$\hat{\theta}_n^{\text{MLE}} \xrightarrow[(p)]{n\to\infty} \theta^*.$$

Note that this is true even if the parameter θ is a vector in a higher dimensional parameter space Θ, and $\hat{\theta}_n^{\text{MLE}}$ is a multivariate random variable, e.g. if $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \in \mathbb{R}^2$ for a Gaussian statistical model.

Multivariate Random Variables

A **multivariate random variable** , or a **random vector** , is a vector-valued function whose components are (scalar) random variables on the same underlying probability space. More specifically, a random vector $\mathbf{X} = \left(X^{(1)}, \ldots, X^{(d)}\right)^T$ of dimension $d \times 1$ is a vector-valued function from a probability space Ω to $\mathbb{R}^d$:

$$\begin{aligned} \mathbf{X} : \Omega &\to \mathbb{R}^d \\ \omega &\mapsto \begin{pmatrix} X^{(1)}(\omega) \\ X^{(2)}(\omega) \\ \vdots \\ X^{(d)}(\omega) \end{pmatrix} \end{aligned}$$

where each $X^{(k)}$ is a (scalar) random variable on Ω. We will often (but not always) use the bracketed superscript (k) to denote the k-th component of a random vector, especially when the subscript is already used to index the samples.

The **probability distribution** of a random vector $\mathbf{X}$ is the **joint distribution** of its components $X^{(1)}, \ldots, X^{(d)}$.

The **cumulative distribution function (cdf)** of a random vector $\mathbf{X}$ is defined as

$$\begin{aligned} F : \mathbb{R}^d &\to [0,1] \\ \mathbf{x} &\mapsto \mathbf{P}\left(X^{(1)} \le x^{(1)}, \ldots, X^{(d)} \le x^{(d)}\right). \end{aligned}$$

Convergence in Probability in Higher Dimension

To make sense of the consistency statement $\hat{\theta}_n^{\text{MLE}} \xrightarrow[(p)]{n\to\infty} \theta^*$ where the MLE $\hat{\theta}_n^{\text{MLE}}$ is a random vector, we need to know what convergence in probability means in higher dimensions. But this is no more than convergence in probability for **each component**.

Let $\mathbf{X}_1, \mathbf{X}_2 \ldots$ be a sequence of random vectors of size $d \times 1$, i.e. $\mathbf{X}_i = \begin{pmatrix} X_i^{(1)} \\ \vdots \\ X_i^{(d)} \end{pmatrix}$.

Let $\mathbf{X} = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(d)} \end{pmatrix}$ be another vector of size $d \times 1$.

Then

$$\mathbf{X}_n \xrightarrow[n\to\infty]{(p)} \mathbf{X} \iff X_n^{(k)} \xrightarrow[n\to\infty]{(p)} X^{(k)} \text{ for all } 1 \le k \le d.$$

In other words, the sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots$ **converges in probability** to $\mathbf{X}$ if and only if each component sequence $X_1^{(k)}, X_2^{(k)}, \ldots$ converges in probability to $X^{(k)}$.

Hence, for example, in the Gaussian model $\left((-\infty, \infty), \{\mathcal{N}(\mu, \sigma^2)\}_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_{>0}}\right)$, consistency of the MLE $\hat{\theta}_n^{\text{MLE}} = \begin{pmatrix} \widehat{\mu} \\ \widehat{\sigma^2} \end{pmatrix}$ means that $\widehat{\mu}$ and $\widehat{\sigma^2}$ are consistent estimators of μ^* and $(\sigma^2)^*$, respectively.

Remark: You can check that this condition is equivalent to the following definition of convergence in probability, which is a straightforward generalization of the 1-dimensional case:

$$P\left(\{\omega \in \Omega : |\mathbf{X}_n(\omega) - \mathbf{X}(\omega)| < \epsilon\}\right) \xrightarrow{n\to\infty} 1 \qquad \text{for any } \epsilon > 0.$$

Consistency of the MLE of a Uniform Model

1/1 point (graded)

Let $X_1, \ldots, X_n \stackrel{iid}{\sim} \mathrm{Unif}[0, \theta^*]$ where θ^* is an unknown parameter. We construct the associated statistical model $(\mathbb{R}_{\geq 0}, \{\mathrm{Unif}[0, \theta]\}_{\theta > 0})$

Consider the maximum likelihood estimator $\hat{\theta}_n^{\text{MLE}} = \max_{i=1,\ldots,n} X_i$.

Which of the following are true about $\hat{\theta}_n^{\text{MLE}}$. (Choose all that apply.)

- [x] $\max_{i=1,\ldots,n} X_i$ is a consistent estimator
- [x] For any $0 < \epsilon \leq \theta^*$, $\mathbf{P}\left(|\max_{i=1,\ldots,n} X_i - \theta^*| \geq \epsilon\right) \to 0$ as $n \to \infty$
- [ ] For any $0 < \epsilon \leq \theta^*$, $\mathbf{P}\left(|\max_{i=1,\ldots,n} X_i - \theta^*| \geq \epsilon\right) \to c$ as $n \to \infty$, where $c > 0$ is a constant
- [x] For any $0 < \epsilon \leq \theta^*$, $\mathbf{P}\left(|\max_{i=1,\ldots,n} X_i - \theta^*| \geq \epsilon\right) = \left(\frac{\theta^* - \epsilon}{\theta^*}\right)^n$

Solution:

Choices 1, 2, and 4 are true because of the following proof for consistency of this ML estimator. Let $0 < \epsilon \leq \theta^*$:

$$\begin{aligned}
\mathbf{P}\left(\left|\max_{i=1,\ldots,n} X_i - \theta^*\right| \geq \epsilon\right) &= \mathbf{P}\left(\theta^* - \max_{i=1,\ldots,n} X_i \geq \epsilon\right) \\
&= \mathbf{P}\left(\max_{i=1,\ldots,n} X_i \leq \theta^* - \epsilon\right) \\
&= \left(\frac{\theta^* - \epsilon}{\theta^*}\right)^n \longrightarrow 0 \text{ as } n \to \infty.
\end{aligned}$$

Choice 3 is not true because if a sequence (the relevant sequence here is $\left(\frac{\theta^*-\epsilon}{\theta^*}\right)^n$) converges to a limit, then the limit is unique.

提交 你已经尝试了1次（总共可以尝试2次）

Answers are displayed within the problem

讨论

显示讨论

主题： Unit 3 Methods of Estimation:Lecture 10: Consistency of MLE, Covariance Matrices, and Multivariate Statistics / 4. Consistency of Maximum Likelihood Estimator