4. Consistency of Maximum Likelihood Estimator

Exercises due Jun 29, 2021 19:59 EDT

Review: Definition of MLE

1/1 point (graded)

Let $\{E, (\mathbf{P}_\theta)_{\theta\in\Theta}\}$ be an identifiable statistical model for the sample of i.i.d. random variables $X_1, X_2, \ldots, X_n$. Assume that the model is well-specified, meaning that there exists $\theta^* \in \Theta$ such that $X_i \sim \mathbf{P}_{\theta^*}$.

Recall that 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, up to a constant, estimator of $\theta \mapsto \mathrm{KL}\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right)$ is

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

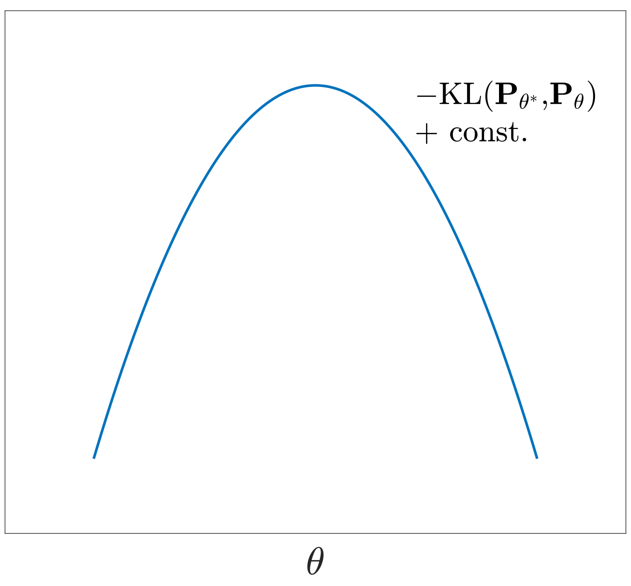
Which of the following represents the maximum likelihood estimator of θ^*? (Choose all that apply).

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

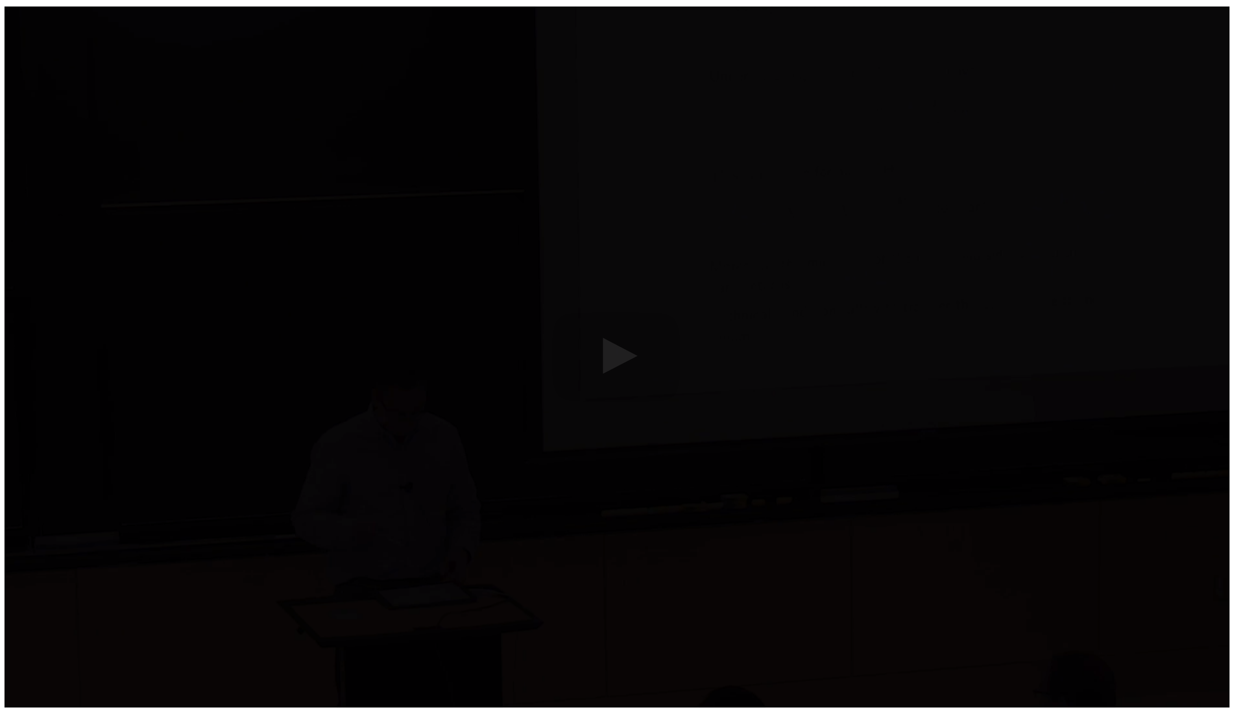
Submit You have used 1 of 3 attempts

✔ Correct (1/1 point)

Note: In the following video, at around the 3:20 mark, the plot of $-\mathrm{KL}\left(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta\right)$, 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



5:24 / 5:24 1.50x

this convergence on the y-axis of the functions

to a convergence of their minimizers.

And you know, things can go wrong when you have,

say, these discontinuous functions, when you

have weird things happening.

But in general, it's actually fine

when you have nice, smooth functions.

End of transcript. Skip to the start.

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download 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 $\hat{\theta}_n^{\mathrm{MLE}}$ of θ^* is a **consistent** estimator under mild regularity conditions (e.g. continuity in θ of the pdf p_θ almost everywhere), i.e.

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

Note that this is true even if the parameter θ is a vector in a higher dimensional parameter space Θ, and $\hat{\theta}_n^{\mathrm{MLE}}$ 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 $\mathbb{R}^d$:

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

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

$$F : \mathbb{R}^d \quad \to \quad [0, 1]$$

$$\mathbf{x} \quad \mapsto \quad \mathbf{P}\left(X^{(1)} \leq x^{(1)}, \ldots, X^{(d)} \leq x^{(d)}\right).$$

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

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

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 point possible (graded)

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

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.)

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

Submit You have used 0 of 2 attempts

Discussion

Hide Discussion

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

Add a Post

Show all posts by recent activity

?	Regarding the concave correction in the video:	3
	kl notation professor used a double bar to indicate asymmetry	1
?	Component Probabilities For the definition of multivariate convergence in probability, do we use the marginal probabilities of the joint distribution to show that each comp...	2

< Previous Next >