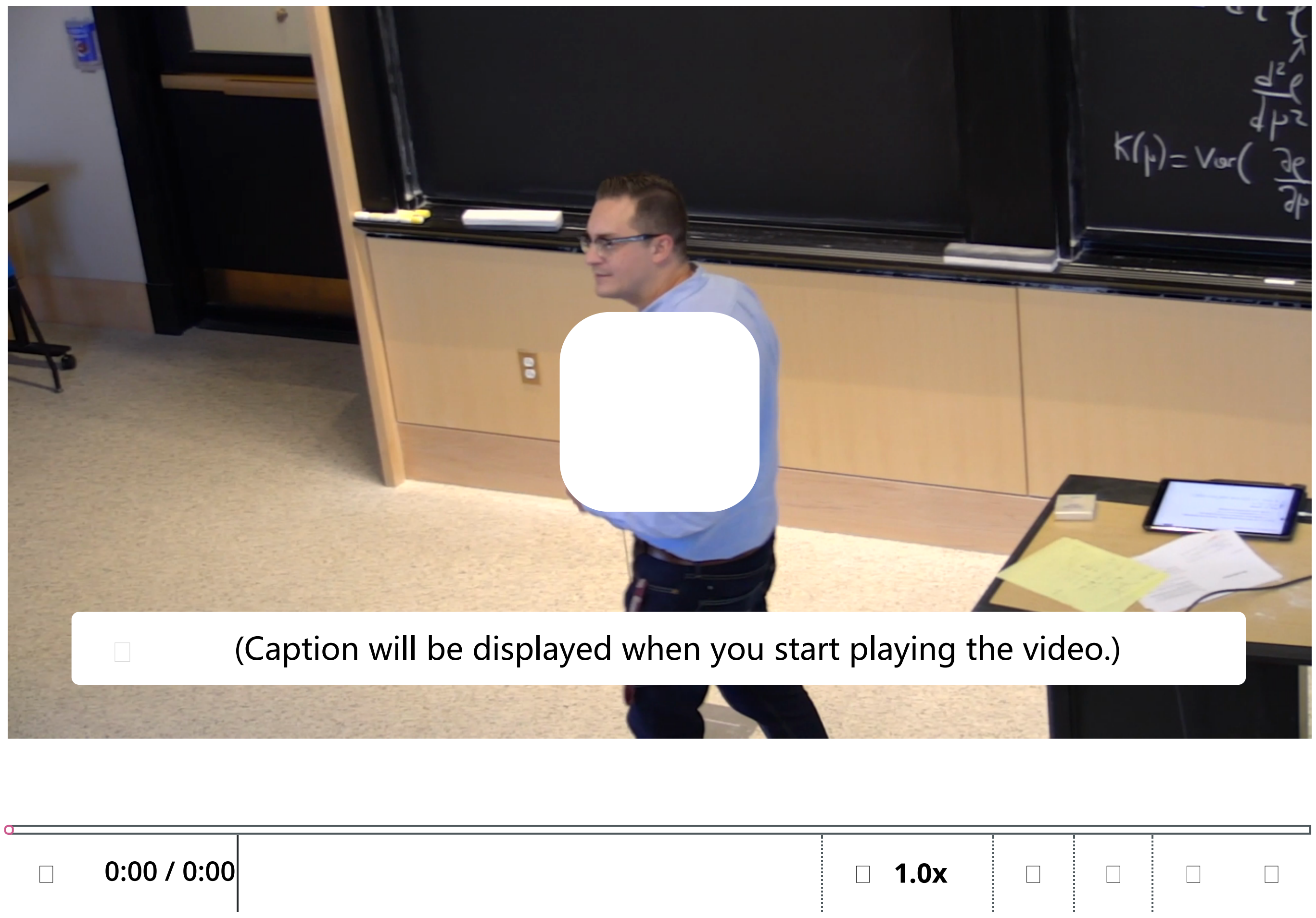
8. Applying Huber's loss to the Laplace distribution

Applying Huber's loss to the Laplace distribution



Start of transcript. Skip to the end.

If I was really in a great mood, and I

felt like integrating 1 plus 1 over x squared twice

and things like this, and do some nasty integration

by parts, I would actually show you

the example of what happens when you

apply this m estimator to the [? cowshed ?] distribution.

In the meantime, let's just do it for a simpler

视频
下载视频文件

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

The Laplace distribution

3/3 points (graded)

The **Laplace distribution** (also known as the **double-exponential distribution**) is a continuous distribution with **location parameter** $m \in \mathbb{R}$ and density given by

$$f_m(x) = \frac{1}{2} e^{-|x-m|}.$$

Let X denote a Laplace random variable with location parameter set to be $m = 0$.

What is $\mathbb{E}[X]$?

0

Answer: 0

Does the variance $\sigma^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ exist?

● Yes

○ No

Which of the following are true about X? (Choose all that apply.)

Hint: The function $x^k e^{-x}$ is integrable, i.e. $\int_{-\infty}^{\infty} x^k e^{-x}\,dx$ is finite for all k.

☑ The distribution of X is symmetric in the sense that X and $-X$ have the same distribution.

☐ The function $\ln f_m(x)$ has a continuous first derivative.

☑ For any integer $k > 0$, the k-th moment $\mathbb{E}\left[X^k\right]$ exists.

Solution:

For the first question, observe that the function $xe^{-|x|}$ is odd and also integrable. Therefore,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \frac{1}{2} x e^{-|x|}\,dx = 0.$$

For the second question, the function $x^2 e^{-|x|}$ is integrable. Hence,

$$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} \frac{1}{2} x^2 e^{-|x|}\,dx$$

For the third question, we examine the choices in order.

- "The distribution of X is symmetric in the sense that X and $-X$ have the same distribution." is correct. This is because the density $\frac{1}{2}e^{-|x|}$ is an even function.
- "The function $\ln f_m(x)$ has a continuous first derivative." is incorrect. This is because $\ln f_m(x) = -|x - m|$, which is not differentiable at $x = m$.
- "For any integer $k > 0$, the k-th moment $\mathbb{E}\left[X^k\right]$ exists." is correct. In general, the function $x^k e^{-|x|}$ is integrable on $\mathbb{R}$, so the k-th moment $\mathbb{E}\left[X^k\right]$ exists for all $k > 0$.

提交 你已经尝试了3次（总共可以尝试3次）

Answers are displayed within the problem

The Sample Median

4/4 points (graded)

Let $S = x_1 < x_2 < \ldots < x_n$ denote a sorted list of numbers. We define the **elementary median** $\mathrm{med}_e(S)$ to be

$$\mathrm{med}_e(S) := \begin{cases} x_{\lceil n/2 \rceil} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{if } n \text{ is even} \end{cases}$$

In other words, when n is odd, the median is the middle number when the set S is sorted from smallest to largest. If n is even, we can just define the median to be the average of both middle numbers. This definition is likely familiar from prior math classes.

A more advanced definition, useful for statistical purposes, is to define the **sample median** $\mathrm{med}_s(S)$ of a sample $S := X_1, X_2, \ldots, X_n$ to be

$$\mathrm{med}(S) := \mathrm{argmin}_m \sum_{i=1}^{n} |X_i - m|\,.$$

While the elementary median is unique, this is not always the case for the **sample median**, as you will see in the next few questions.

Consider the data set $S = \{1, 2, 3\}$.

What is the **elementary median** of S?

2

☐ **Answer:** 2

What is the **sample median** of S?
Hint: Use computational software to graph the objective function.

2

☐ **Answer:** 2

Now consider the data set $T = \{1, 2, 3, 4\}$.

What is the **elementary median** of T?

2.5

☐ **Answer:** 2.5

Which of the following describes **all** of the sample medians of T? (*Hint:* Use computational software to graph the objective function)

○ 2.5

○ Any number in the open interval $(2, 3)$.

● Any number in the closed interval $[2, 3]$. ☐

○ None of the above.

Solution:

For the first question, the elementary median of S is 2. For the second question, we want to find m such that

$$F(m) = |1 - m| + |2 - m| + |3 - m|$$

is as small as possible. Note that $F(m)$ is a piecewise linear function with discontinuities in its first derivative precisely at the points $(1, 3), (2, 2)$, and $(3, 3)$ (one corresponding to each summand of the form $|x - m|$). Drawing the line segments connecting these three points, we see that the minimizer of $F(m)$ is at $m = 2$. Therefore, for the second question, the sample median is 2.

把不连续的点连起来，然后画图。

For the third question, the elementary median is 2.5. For the fourth question, we want to find m such that

$$G(m) = |1 - m| + |2 - m| + |3 - m| + |4 - m|$$

is as small as possible. As before, $G(m)$ is a piecewise linear function. Its discontinuities are at the points $(1, 6), (2, 4), (3, 4)$, and $(4, 6)$. Hence, $G(m)$ is a horizontal line segment on the interval $[2, 3]$, and $G(m)$ has positive slope otherwise. Hence, the correct response to the fourth question is "Any number in the closed interval $[2, 3]$.".

只知道minimizer在这两个点之间，有更多的数据就能发现更多信息。

Remark: In general, one can show that for an ordered sample $S = x_1 < \ldots < x_n$ that

$$\text{med}(S) = \begin{cases} x_{\lceil n/2 \rceil} & \text{if } n \text{ is odd} \\ \text{Any number in the interval } [x_{n/2}, x_{n/2+1}] & \text{if } n \text{ is even} \end{cases}$$

提交

你已经尝试了2次（总共可以尝试3次）

☐ Answers are displayed within the problem

Maximum Likelihood Estimator for the Laplace Distribution

2/2 points (graded)

Consider a Laplace statistical model $(\mathbb{R}, \{P_m\}_{m\in\mathbb{R}})$ where P_m denotes the Laplace distribution with location parameter m. Let $X_1, \ldots, X_n \overset{iid}{\sim} P_{m^*}$ denote a sample from a Laplace distribution with unknown parameter m^*. Recall that the density of P_m is given by

$$f_m(x) = \frac{1}{2}e^{-|x-m|}.$$

What is the log-likelihood $\ell_n(X_1, \ldots, X_n; m)$ for this statistical model?

● $-n\ln(2) - \sum_{i=1}^{n}|X_i - m|$ ☐

○ $-\sum_{i=1}^{n}|X_i - m|$

○ $-n\ln(2) - \sum_{i=1}^{n}(X_i - m)^2$

○ $-n\ln(2) + \sum_{i=1}^{n}|X_i - m|$

Recall that the maximum likelihood estimator $\widehat{m}_n^{\text{MLE}}$ is given by

$$\widehat{m}_n^{\text{MLE}} = \text{argmin}_{m\in\mathbb{R}} - \ell_n(X_1, \ldots, X_n; m).$$

Suppose you observe the sample

$$S = 0.5, 1.2, 0.6, -0.7, -0.2.$$

What is the value of the MLE for m^* for this data set? *Hint:* Use the previous question, in particular the remark at the end of the solution.

0.5

☐ **Answer:** 0.5

Solution:

For the first question, the likelihood for n observations is given by

$$\prod_{i=1}^{n} f_m(X_i) = \frac{1}{2^n}\prod_{i=1}^{n} e^{-|x-m|}.$$

Therefore,

$$\ell_n(X_1, \ldots, X_n; m) = -n\ln(2) - \sum_{i=1}^{n}|X_i - m|.$$

For the second question, we need to minimize the quantity

$$5\ln(2) + |m - 0.5| + |m - 1.2| + |m - 0.6| + |m + 0.2| + |m + 0.7|.$$

with respect to m. As stated in the previous quantity, any m that minimizes the above is a sample median of the data set S. Since S is odd, we have that $\widehat{m}_n^{\mathrm{MLE}} = 0.5$.

提交 你已经尝试了1次（总共可以尝试3次）

☐ Answers are displayed within the problem

Concept Question: Maximum Likelihood Estimator for the Laplace distribution

1/1 point (graded)

As in the previous problem, let $\widehat{m}_n^{\mathrm{MLE}}$ denote the MLE for an unknown parameter m^* of a Laplace distribution.

Can we apply the theorem for the asymptotic normality of the MLE to $\widehat{m}_n^{\mathrm{MLE}}$? (You must choose the correct answer that also has the correct explanation.)

○ No, because the log-likelihood is not concave.

● No, because the log-likelihood is not twice-differentiable, so the Fisher information does not exist. ☐

○ Yes, because the log-likelihood is concave.

○ Yes, because the other technical conditions required to apply the theorem are satisfied.

Solution:

We examine the choices in order.

- "No, because the log-likelihood is not concave." is incorrect. A sum of concave functions is concave, and for any constant c, the function $x \to -|x - c|$ is concave. Therefore, the log-likelihood

$$\ell_n\left(X_1, \ldots, X_n; m\right) = -n \ln\left(2\right) - \sum_{i=1}^{n} |X_i - m|$$

  is also concave. Hence, the reasoning for this response is incorrect.

- "No, because the log-likelihood is not twice-differentiable, so the Fisher information does not exist." is correct. This is because $\ell_n\left(X_1, \ldots, X_n; m\right)$ has discontinuities in its first derivative with respect to m at $m = X_i$ for $i = 1, \ldots, n$.
- "Yes, because the log-likelihood is concave." is incorrect. Although the log-likelihood is concave, the Fisher information does not exist, as discussed in the analysis of the previous two responses.
- "Yes, because the other technical conditions required to apply the theorem are satisfied." is incorrect. The remaining technical conditions are not enough to guarantee asymptotic normality since the Fisher information does not exist, as discussed above.

提交 你已经尝试了1次（总共可以尝试2次）

☐ Answers are displayed within the problem

Applying Huber's loss to a Laplace distribution I

2/2 points (graded)

As above, let m^* denote an unknown parameter for a Laplace distribution. In this problem, we will use the principles of M-estimation and the smoothness of Huber's loss function to construct an asymptotically normal estimator for m^*. Let P_m denote the Laplace distribution with location parameter m.

Recall Huber's loss is defined as

$$h_\delta(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| < \delta \\ \delta(|x| - \delta/2) & \text{if } |x| > \delta \end{cases}.$$

As computed in lecture, the derivative of Huber's loss is the **clip function** :

$$\text{clip}_\delta(x) := \frac{d}{dx} h_\delta(x) = \begin{cases} \delta & \text{if } x > \delta \\ x & \text{if } -\delta \leq x \leq \delta \\ -\delta & \text{if } x < -\delta \end{cases}$$

Find the value of

$$\frac{\partial}{\partial m} \mathbb{E}_{X \sim P_{m^*}} \left[h_\delta(X - m) \right] \Big|_{m = m^*}.$$

Hint: You are allowed to switch the derivative and expectation.

0

✓ **Answer:** 0

In the framework of M-estimation, our loss function is not Huber's loss itself, but rather

$$\rho(x, m) := h_\delta(x - m)$$

Recall the functions

$$\begin{aligned} J(m) &= \mathbb{E}\left[\frac{\partial^2 \rho}{\partial m^2}(X_1, m)\right] \\ K(m) &= \text{Var}\left[\frac{\partial \rho}{\partial m}(X_1, m)\right] \end{aligned}$$

Do the functions K and J exist for a Laplace statistical model?

○ No, because the log-likelihood is not twice-differentiable.

○ No, because $J(m)$ exists but $K(m)$ does not.

○ Yes, because the Fisher information is well-defined for a Laplace statistical model.

● Yes, because the function $\rho(x, m)$ as defined above is twice-differentiable. ✓

Solution:

The answer to the first question is 0. To see this, observe that

$$\begin{aligned} \frac{\partial}{\partial m} \mathbb{E}_{X \sim P_m^*} \left[h_\delta(X) \right] &= \mathbb{E}_{X \sim P_m^*} \left[\frac{\partial}{\partial m} h_\delta(X) \right] \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \text{clip}_\delta(x - m) \, e^{-|x - m^*|} \, dx \\ &= \frac{1}{2} \left(\delta \int_{m+\delta}^{\infty} e^{-|x - m^*|} \, dx - \delta \int_{-\infty}^{-\delta + m} e^{-|x - m^*|} \, dx + \int_{-\delta + m}^{\delta + m} x e^{-|x - m^*|} dx \right). \end{aligned}$$

Applying the change of variables $y = x - m$, we have

$$= \frac{1}{2}\left(\delta \int_{\delta}^{\infty} e^{-|y+m-m^*|}\,dy - \delta \int_{-\infty}^{-\delta} e^{-|y+m-m^*|}\,dy + \int_{-\delta}^{\delta} y e^{-|y+m-m^*|}\,dy\right).$$

Setting $m = m^*$, we have

$$\frac{\partial}{\partial m}\mathbb{E}_{X\sim P_m^*}\left[h_\delta(X)\right]\Big|_{m=m^*} = \frac{1}{2}\left(\delta \int_{\delta}^{\infty} e^{-|y|}\,dy - \delta \int_{-\infty}^{-\delta} e^{-|y|}\,dy + \int_{-\delta}^{\delta} y e^{-|y|}\,dy\right) = 0.$$

Remark: The function $m \mapsto \mathbb{E}_{X\sim P_m^*}\left[h_\delta(X)\right]$ is strictly convex, so this means the loss function has a unique critical point, and this is where the minimum is attained. The above calculation guarantees that the minimum is at $m = m^*$, the value of the true parameter.

"No, because the log-likelihood is not twice-differentiable.","No, because $J(m)$ exists but $K(m)$ does not.","Yes, because the Fisher information is well-defined for a Laplace statistical model.","Yes, because the function $\rho(x,m)$ as defined above is twice-differentiable."

For the second question, we consider the responses in order.

- "No, because the log-likelihood is not twice-differentiable." is incorrect. In the problem "Huber's loss" on the page "Robust Statistics and Huber's Loss", we showed that $\rho(x,m) = h_\delta(x-m)$ is twice-differentiable with respect to m.
- "No, because $J(m)$ exists but $K(m)$ does not." is also incorrect. Both $K(m)$ and $J(m)$ exist because ρ is twice-differentiable, and its derivatives are integrable.
- "Yes, because the Fisher information is well-defined for a Laplace statistical model." is incorrect. The Fisher information does not exist for a Laplace statistical model, as was shown in the problem "Concept Question: Maximum Likelihood Estimator for the Laplace Distribution" on the page "Applying Huber's loss to the Laplace distribution."
- "Yes, because the function $\rho(x,m)$ as defined above is twice-differentiable." is the correct answer. In a previous problem, we showed that $\rho(x,m) = h_\delta(x-m)$ is twice-differentiable with respect to m.

提交 你已经尝试了2次（总共可以尝试3次）

☐ Answers are displayed within the problem

Applying Huber's loss to a Laplace distribution II

1/1 point (graded)
We use the same set-up as in the previous problem. Recall that m^* is an unknown location parameter for a Laplace distribution.

The M-estimator $\widehat{m}$ for m^* associated to the loss function $\rho(x,m) = h_\delta(x-m)$ is given by

$$\widehat{m} = \textbf{argmin}_{m\in\mathbb{R}}\frac{1}{n}\sum_{i=1}^{n} h_\delta(X_i - m).$$

Consider the slide *Asymptotic normality* in Unit 3. Suppose that the technical conditions in 3. of slide are satisfied. Also, note that for any fixed x, the function $m \mapsto h_\delta(x-m)$ is strictly convex. (You can see this by observing $h_\delta(x-m) = h_\delta(m-x)$, so the graph of $h_\delta(x-m)$ as a function of m for a fixed x is the same as the graph of $h_\delta(x-m)$ as a function of x for a fixed m.) Finally, consider the calculation of $J(m)$ from lecture.

Can we apply the theorem on the slide *Asymptotic normality* to conclude that $\widehat{m}$ is asymptotically normal?

(Choose the correct answer, ‘Yes' or 'No', that also has a correct explanation.)

○ No, because m^* is not the unique minimizer of the function $m \mapsto \mathbb{E}_{X\sim P_{m^*}}\left[\rho(X,m)\right]$.

○ No, because $J(m)$ is not invertible.

○ Yes, because m^* is the unique minimizer of the function $m \mapsto \frac{1}{n}\sum_{i=1}^{n} h_\delta(X_i - m)$ and $J(m)$ is invertible.

● Yes, because m^* is the unique minimizer of the function $m \mapsto \mathbb{E}_{X\sim P_{m^*}}[\rho(X,m)]$ and $J(m)$ is invertible.

Solution:

"Yes, because m^* is the unique minimizer of the function $m \mapsto \mathbb{E}_{X\sim P_{m^*}}[\rho(X,m)]$ and $J(m)$ is invertible." is the correct answer. As shown in the Remark in the solution to the previous problem, m^* is the unique minimizer of the loss function $m \mapsto \mathbb{E}_{X\sim P_{m^*}}[\rho(X,m)]$. Moreover,

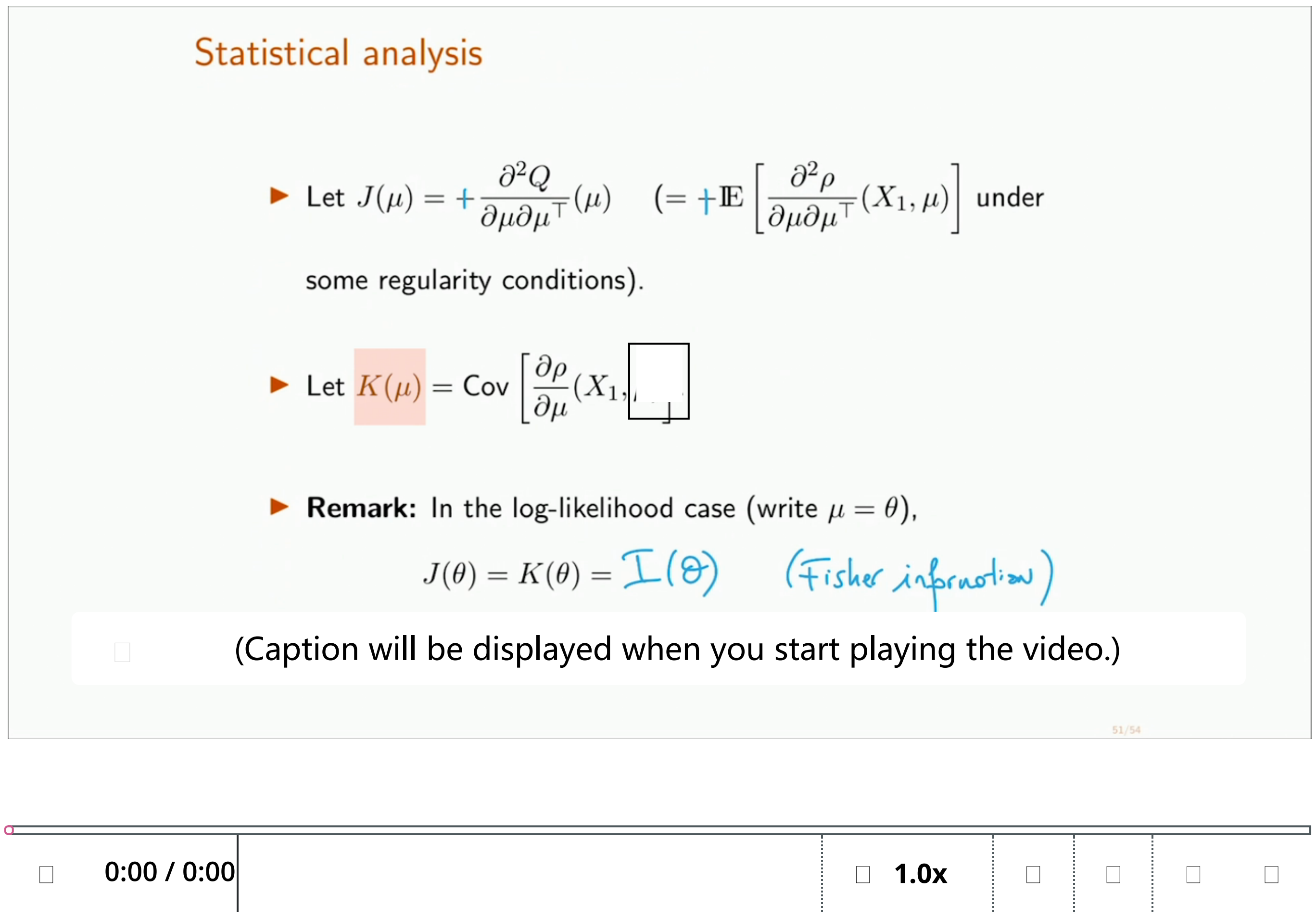
$$J(m) = 1 - e^{-\delta}$$

as was shown in the lecture. Hence $J(m)$ is invertible. Assuming that the required technical conditions are satisfied, we conclude that the estimator $\widehat{m}$ is asymptotically normal.

提交 你已经尝试了1次（总共可以尝试2次）

Answers are displayed within the problem

Applying Huber's loss to the Laplace distribution: Computation



Statistical analysis

- Let $J(\mu) = +\frac{\partial^2 Q}{\partial\mu\partial\mu^\top}(\mu)$ $\left(= +\mathbb{E}\left[\frac{\partial^2 \rho}{\partial\mu\partial\mu^\top}(X_1,\mu)\right]\right.$ under some regularity conditions).
- Let $K(\mu) = \text{Cov}\left[\frac{\partial\rho}{\partial\mu}(X_1,\right.$
- **Remark:** In the log-likelihood case (write $\mu = \theta$), $J(\theta) = K(\theta) = I(\theta)$ (Fisher information)

(Caption will be displayed when you start playing the video.)

So it's just delta squared, f0 of x dx.

[INAUDIBLE] with me?

All right, so now I have to do limit of competition.

So let's just write what this thing actually is.

So it's 2 times the integral between 0

and delta of x squared times e to the minus--

sorry, that's delta-- e to the minus x dx divided by 2.

So those two cancel.

That's just my Laplace.

I just plugged in f0 for what it is.

And then I have plus-- well, those two

are going to cancel again.

So I have delta squared integral between delta and infinity of e

to the minus lambda dx.

Agreed?

OK, let's split work.

视频
下载视频文件

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

xuetangX.com 学堂在线

讨论

隐藏讨论

主题： Unit 3 Methods of Estimation:Lecture 12: M-Estimation / 8. Applying Huber's loss to the Laplace distribution

Add a Post

All Posts

Applying Huber's loss to a Laplace distribution I; what does it mean to be differentiable?

question posted 3 days ago by SergK (Community TA)

We have seen in the forums, thanks mrBB, example of a function that is differentiable but not continuously differentiable; but this is tricky.

I thought that being differentiable means to have derivative at each point, so derivative of a differentiable function can't have simple jumps, because having different left and right derivatives means derivative does not exist, and so the function is not differentiable.

Now after reading the solution I feel like an idiot. What does it mean to be differentiable, in context of the solution ???

此帖对所有人可见。

添加回复

1 response

sudarsanvsr_mit (Staff)
3 days ago

This is a great question. I feel like we've been a bit casual when dealing with the second derivative in the Huber loss case.

The second derivative h_δ'' does not exist at $m-\delta$ and $m+\delta$ but that does not matter because we are interested in the expected value of h_δ'' with respect to a continuous density.

So is it OK to understand that "twice differentiable" really means twice differentiable *almost surely*?

kentaro_nakaishi 在3 days ago前发表

"*Assume that ℓ is a.s. twice differentiable*" Slide 35/54, lecture 3.

Mark_B2 在3 days ago前发表

@kentaro_nakaishi: That is correct.

sudarsanvsr_mit (Staff) 在3 days ago前发表

Thank both of you, @Mark_B2 and @sudarsanvsr_mit , for clarification!

kentaro_nakaishi 在3 days ago前发表

That still makes me wonder how that differs from the absolute value function, that also is twice differentiable almost everywhere, except at a single point. Is there an essential difference in having also a single discontinuity in the first derivative? I'd say, but didn't carefully think it over, that that also doesn't stop us from calculating an expected value (or variance) w.r.t. a continuous density function.

Edit: gave it a bit more thought (but perhaps not enough yet). Is the problem with using the Maximum Likelihood Estimator for the Laplace distribution not so much the (twice-)differentiability of ρ, but rather that we find a $J(m)$ that is zero and therefore not invertible?

mrBB (Community TA) 在about 11 hours ago前发表

This topic belongs to measure theory. If $F(x)$ is differentiable just almost everywhere, we cannot expect the fundamental theorem of calculus $\int_a^b F'(x)\,dx = F(b) - F(a)$ in general. It actually holds if and only If $F(x)$ is absolutely continuous (a stronger notion of continuity). So if $F(x)$ has a jump like the derivative of $|x|$, we cannot use various techniques of calculus.

Anyway this will explain why the staff won't go into the details here:)

kentaro_nakaishi 在about 5 hours ago前发表

Thanks for that note, kentaro_nakaishi. I will chew on it a bit. It doesn't immediately answer the question to me why we bother about a discontinuity in ρ' but not in ρ'' when calculating $E\left[(\rho')^2\right]$ and $E\left[\rho''\right]$ respectively. In both instances piecewise integration seems to get us a long way. But guess you're right that the subject is probably more technical than we can expect the course material to carefully cover.

mrBB (Community TA) 在about 2 hours ago前发表

Discontinuity in ρ' prevents us from finding its extremum at the points of discontinuity whereas discontinuity in ρ'' doesn't prevent us from finding its expectation (or variance of ρ').

Mark_B2 在about an hour ago前发表

Sure, but these questions, and my pondering, are related to calculating the Fisher information and not so much to finding the extremum of the log likelihood. To me it isn't immediately obvious that we *can't* calculate the Fisher information in the absence of a continuous derivative, but that the absence of a continuous second derivative is no impediment for doing that. Somehow the concept questions take this as something obvious, but it still isn't obvious to me.

mrBB (Community TA) 在18 minutes ago前发表

Does Fisher information has the meaning if extrema of likelihood couldn't be found?

Mark_B2 在about a minute ago前发表

添加评论

显示所有的回复

添加一条回复：

Preview

提交