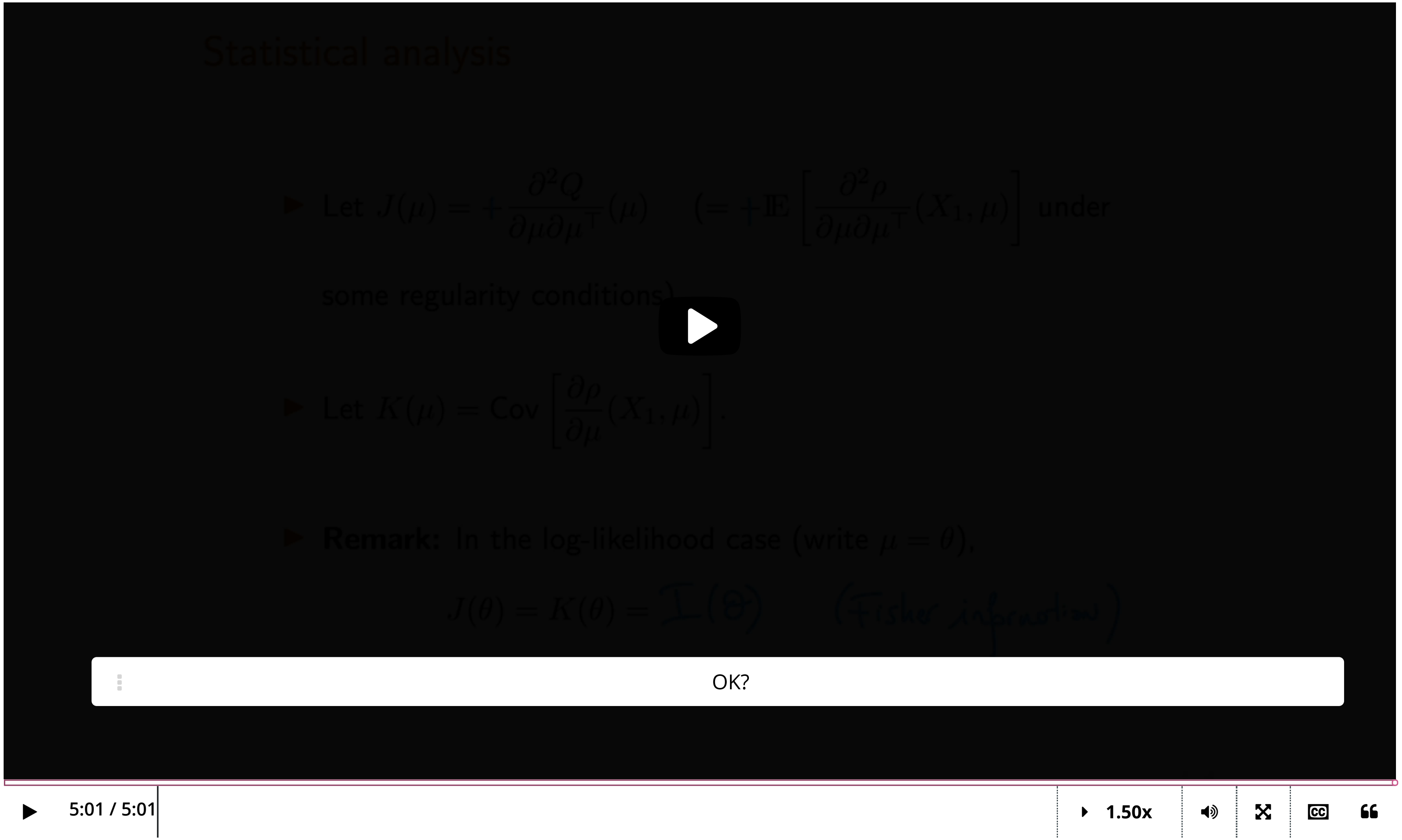
4. Preparations for the Asymptotic Normality of M-estimators

Video note:

In the video below and on the slides, we have used another common notation for the gradient and the Hessian. Below the video, we repeat the definitions of the matrices $\mathbf{J}$ and $\mathbf{K}$ in the familiar notation we had used earlier.

Also about 4:20, Prof Rigollet said "What is the expectation of the second derivative of the log likelihood minus the expectation of the second derivative of the log likelihood". What he meant was in fact "minus the expectation of the second derivative of the log liklihood", not the difference between this quantity and itself.

Preparation for Asymptotic Normality



Video

[Download video file]

Transcripts

[Download SubRip (.srt) file]

[Download Text (.txt) file]

The $\mathbf{J}$ and $\mathbf{K}$ matrices :

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random vector in $\mathbb{R}^k$ with some unknown distribution $\mathbf{P}$ with some associated parameter $\vec{\mu}^* \in \mathbb{R}^d$ on some sample space E. Let $\mathcal{Q}(\vec{\mu}) = \mathbb{E}[\rho(\mathbf{X}, \vec{\mu})]$ for some function $\rho : E \times \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the set of all possible values of the unknown true parameter $\vec{\mu}^*$.

Then the matrices $\mathbf{J}$ and $\mathbf{K}$ are defined as

$$\mathbf{J} = \mathbb{E}[\mathbf{H}\rho] = \mathbb{E}\left[\begin{pmatrix} \frac{\partial^2 \rho}{\partial \mu_1 \partial \mu_1}(\mathbf{X}_1, \vec{\mu}) & \ldots & \frac{\partial^2 \rho}{\partial \mu_1 \partial \mu_d}(\mathbf{X}_1, \vec{\mu}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \rho}{\partial \mu_d \partial \mu_1}(\mathbf{X}_1, \vec{\mu}) & \ldots & \frac{\partial^2 \rho}{\partial \mu_d \partial \mu_d}(\mathbf{X}_1, \vec{\mu}) \end{pmatrix}\right] \qquad (d \times d)$$

$$\mathbf{K} = \mathsf{Cov}[\nabla \rho(\mathbf{X}_1, \vec{\mu})] = \mathsf{Cov}\left[\begin{pmatrix} \frac{\partial \rho}{\partial \mu_1}(\mathbf{X}_1, \vec{\mu}) \\ \vdots \\ \frac{\partial \rho}{\partial \mu_d}(\mathbf{X}_1, \vec{\mu}) \end{pmatrix}\right] \qquad (d \times d).$$

In one dimension, i.e. $d = 1$, the matrices reduce to the following:

$$J(\mu) = \mathbb{E}\left[\frac{\partial^2 \rho}{\partial \mu^2}(X_1, \mu)\right]$$

$$K(\mu) = \mathrm{Var}\left[\frac{\partial \rho}{\partial \mu}(X_1, \mu)\right]$$

Concept Check: M-estimators vs. Maximum Likelihood Estimation

1/1 point (graded)

Let ρ denote a loss function, and let $X_1, \ldots, X_n \stackrel{iid}{\sim} \mathbf{P}$. Let $\widehat{\mu}$ denote the M-estimator for some unknown parameter $\mu^* = \mathrm{argmin}_{\mu \in \mathbb{R}} \mathbb{E}[\rho(X_1, \mu)] \in \mathbb{R}$ associated with $\mathbf{P}$. (Here we are assuming that μ^* is a one-dimensional parameter.)

Consider the following functions

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

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

- [ ] It is always true that $J(\mu) = K(\mu)$.
- [x] $J(\mu) = K(\mu)$ when ρ is the negative log-likelihood– in this case, both of these functions are equal to the Fisher information.
- [x] Under some technical conditions, the functions $J(\mu)$ and $K(\mu)$ determine the asymptotic variance of the M-estimator $\widehat{\mu}$.

✔

Solution:

- The response "It is always true that $J(\mu) = K(\mu)$." is incorrect. In general, the functions $J(\mu)$ and $K(\mu)$ will not be equal to each other. For example, if the loss function is given in terms of Huber's loss (as we will see later in this lecture), $J(\mu) \neq K(\mu)$.
- The choice "$J(\mu) = K(\mu)$ when ρ is the negative log-likelihood– in this case, both of these functions are equal to the Fisher information." is correct. In the special case where $\rho(x, \mu)$ is defined to be the negative log-likelihood of the statistical model, then it is true that $J(\mu) = K(\mu)$. This was derived in Lecture 11.
- The choice "Under some technical conditions, the functions $J(\mu)$ and $K(\mu)$ determine the asymptotic variance of the M-estimator $\widehat{\mu}$." is correct. This is content of the theorem on the slide "Asymptotic Normality," which shows that the asymptotic variance of $\widehat{\mu}_n$, assuming some hypotheses, is given by $J(\mu^*)^{-1} K(\mu^*) J(\mu^*)^{-1}$.

Remark on signs:
Let us match the signs in the definition of $\mathbf{J}$ and $\mathbf{K}$ with those in the definition of Fisher information. For maximum likelihood estimation,

$$\rho_n(\theta) := \rho(\mathbf{X}_1, \ldots, \mathbf{X}_n, \theta) = -\ell_n(\theta) \qquad \text{where } \ell_n(\theta) = \ln L_n(\mathbf{X}_1, \ldots, \mathbf{X}_n, \theta).$$

For this particular loss function ρ, the $\mathbf{J}$ and $\mathbf{K}$ matrices are

$$\begin{aligned}\mathbf{J} &= \mathbb{E}\left[\mathbf{H}\rho_1(\theta)\right] = -\mathbb{E}\left[\mathbf{H}\ell_1(\theta)\right]\\ \mathbf{K} &= \mathsf{Cov}\left[\nabla\rho_1(\theta)\right] = \mathsf{Cov}\left[-\nabla\ell_1(\theta)\right] = \mathsf{Cov}\left[\nabla\ell_1(\theta)\right] \qquad (\mathsf{Cov}\left[\mathbf{Y}\right] = \mathsf{Cov}\left[-\mathbf{Y}\right] \text{ for any random vector } \mathbf{Y}.\end{aligned}$$

Both of these matrices equals the Fisher information matrix.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 3 Methods of Estimation:Lecture 12: M-Estimation / 4. Preparations for the Asymptotic Normality of M-estimators

Add a Post

Show all posts ▼ by recent activity ▼

There are no posts in this topic yet.

✖

Learn About Verified Certificates