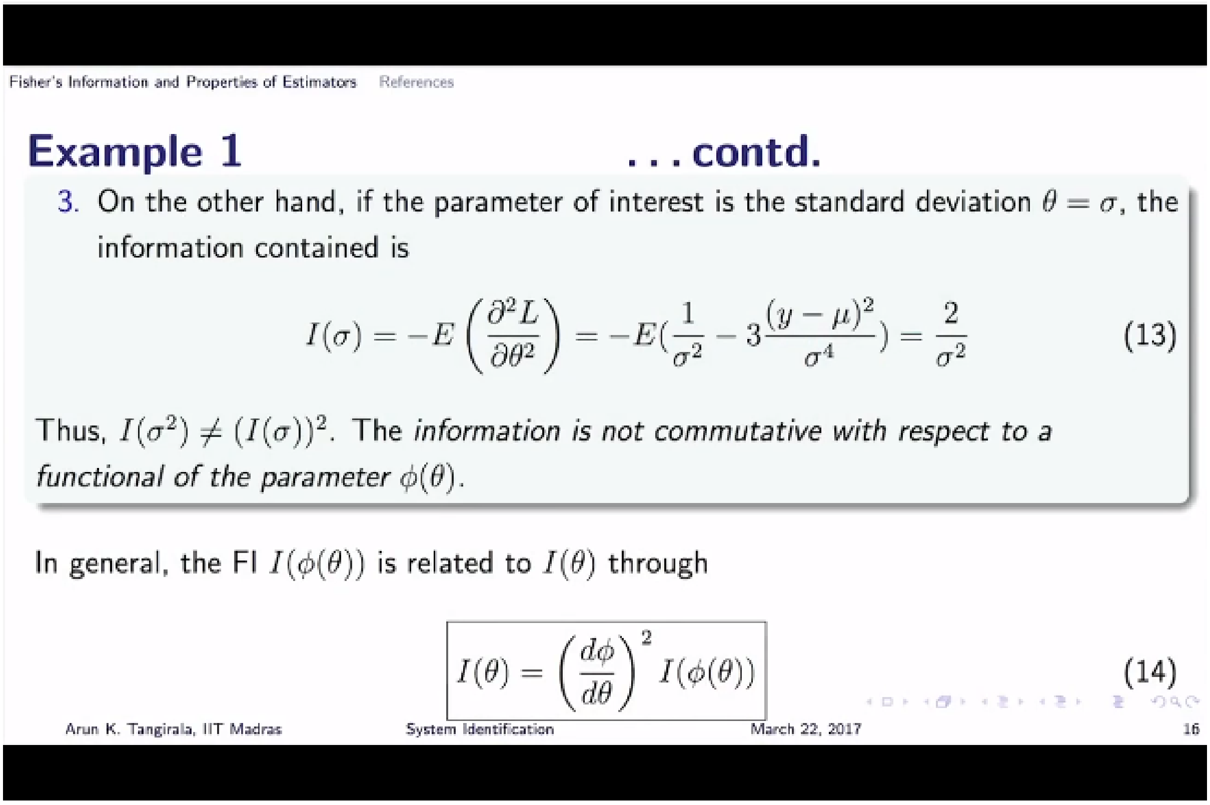
CH5230: System Identification

Fisher’s information and properties of estimators

Part 03

Very good to see all of you. So, if you recall, we were talking about Fisher’s information and this is in the context of estimation. Now before I proceed further, I have a proposition and this is already in connection with the email that one of the TAs must have sent you, right? So you’ve been given the links to the videos. How many of you have not watched the videos? All of you, they will not watch the videos at all? So my proposition is as follows. See, this part of the course actually overlaps with the time series almost 99% because estimation is common to both. In time series analysis also we develop models as you know we develop arma models, arima models, we estimate parameters and so on. And in System Identification you know why we are already discussing estimation. So the theory is almost the same. It is all about now placing the theoryin the context of either time series modelling or developing models is System Identification. So my suggestion is that, in the interest of System Identification, it is better that you watch these videos because I’m not going to teach anything much different. They are the same slides, they are perhaps the same perspectives. What I would rather like to do is, you watch these videos, you learn the, I mean, of course, note down any queries that you have. I would rather spend the classroom in showing certain examples in MATLAB. So for example, here in this goodness of estimators, for those of you who have watched, you must have already gone through in officials information bias, variance, consistency, and all those properties, right, up to confidence interval construction. I would rather probably show one or two examples on for example, how todo a Monte Carlo kind of simulation to come up with confidence interval construction, for example, I'm just giving an example.Or how do you go about computing the variants of an estimator using Monte Carlo simulations in MATLAB.What I teach you, sorry, in the videosare, how to theoretically arrive at the expressions for the variance of an estimator if possible, right. And of course, I'll quickly review.This is as far as the goodness of estimators is concerned. Now, when it comes to estimation methods, the methods that I discuss in time series as well as SysID, again are the same. I talk about method of Moments, least squares, maximum likelihood estimation, I've already explained to you the principle of likelihood,and base in estimation.The methods are the same, I would rather spend time in the classroom showing how to set up for example, the regressor matrix for the least squares estimation of FIR model or often ARX model. And so, and then doing so will actually enhance and complement what is already therein those videos. They are fairly generic, butby showing you a couple of demos in MATLAB, I would be placing all those methods in context here. So this week, I propose to do that, I suppose that you will go back now, and you will watch those videos.They are fairly sorry, comprehensive, I don't think I miss out on any concepts there. As I said, they're going to be the same set of notes that I will be using. And you also have the textbook in case you have a problem. But in the classroom, maybe the first 10, 15 minutes, I review the concepts and then I show you some examples in MATLAB, because then that actually allows you to supplement the theory with practice, because estimation theory is a lot of theory, in my opinion.And it is best understood when you start practicing.So therefore, if you're okay with this, I would suggest that you watch those videos, our TAs will send you a link to those videos that are already available on the web. But the specific links will be sent to you.Sit through those videos, and come back with any questions that you have.That will allow you also to watch these videos and learn these concepts at your leisure. You can say, this is kind of a semi flipped classroom model. It's not a complete flipped classroom model,because if it was a case, then I would have done it right from the beginning.But I just want to see how well you respond to this. I'm not experimenting here. But I think as far as SysID the way SysID has been taught over the last few years, yeah, this is the first time I'm doing this,primarily because the videos are now available. That's all I'm saying.Okay.Sowith that, what I'm going to do is I'm going to actually review these concepts of fisher’s information bias, and so on. But I'm not going to go so much in detail as I would have done in those videos. And if you have any questions, at any point in time, you should stop me and ask me, okay, these are concepts that aregoing to be important. And remember you have not been tested on this quiz. So your final exam will carry a lot of these concepts, you should remember that. Okay.So let's get going.

(Refer Slide Time 06:01)



Fisher's Information and Properties of Estimators   References

Example 1 ...contd.

3. On the other hand, if the parameter of interest is the standard deviation $\theta = \sigma$, the information contained is

$$I(\sigma) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E(\frac{1}{\sigma^2} - 3\frac{(y-\mu)^2}{\sigma^4}) = \frac{2}{\sigma^2} \qquad (13)$$

Thus, $I(\sigma^2) \neq (I(\sigma))^2$. The *information is not commutative with respect to a functional of the parameter* $\phi(\theta)$.

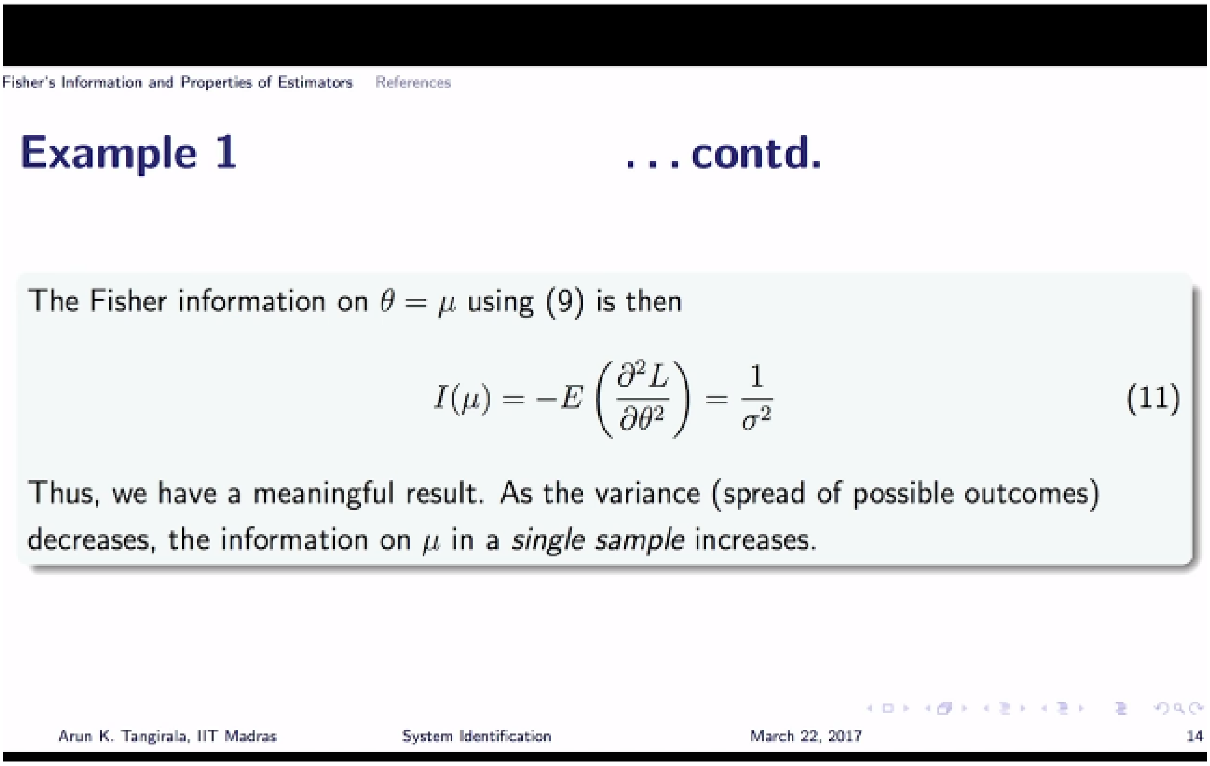
In general, the FI $I(\phi(\theta))$ is related to $I(\theta)$ through

$$\boxed{I(\theta) = \left(\frac{d\phi}{d\theta}\right)^2 I(\phi(\theta))} \qquad (14)$$

Arun K. Tangirala, IIT Madras   System Identification   March 22, 2017   16

I'll first briefly talk about the fisher’s information, couple of examples, which throw light on couple of aspects, officials information.And then we'll move on to reviewing the concepts of bias and so on. So,the example that we looked at in the last class was, what is information contain about mean in a single observation. And the single observation we assume it to be falling out of a Gaussian white noise process. And we showed that the, with the all the definitions, the singleobservation has one over sigma squared information about mean.

(Refer Slide Time 6:41)



Example 1 . . . contd.

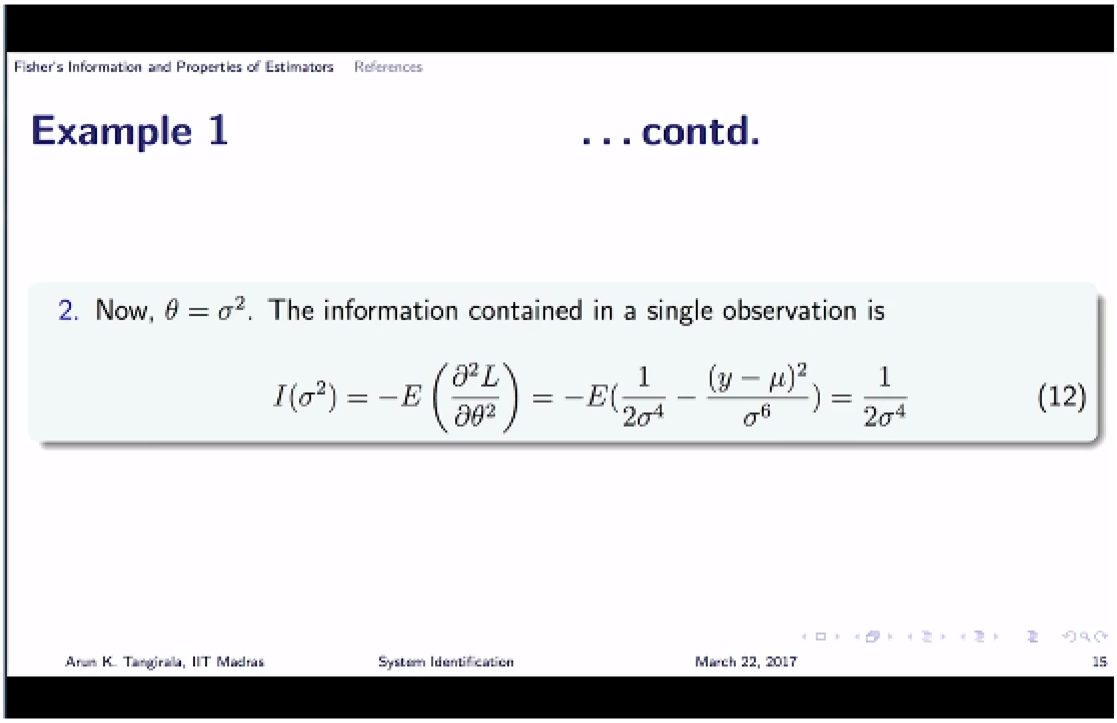
The Fisher information on $\theta = \mu$ using (9) is then

$$I(\mu) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = \frac{1}{\sigma^2} \tag{11}$$

Thus, we have a meaningful result. As the variance (spread of possible outcomes) decreases, the information on μ in a *single sample* increases.

If I asked the same question about variance,right, if I take a single observation and I ask you, from the single observation, can you say something about variability? What would be the information contained in that?

(Refer Slide Time 6:56)



Fisher's Information and Properties of Estimators References

Example 1 . . . contd.

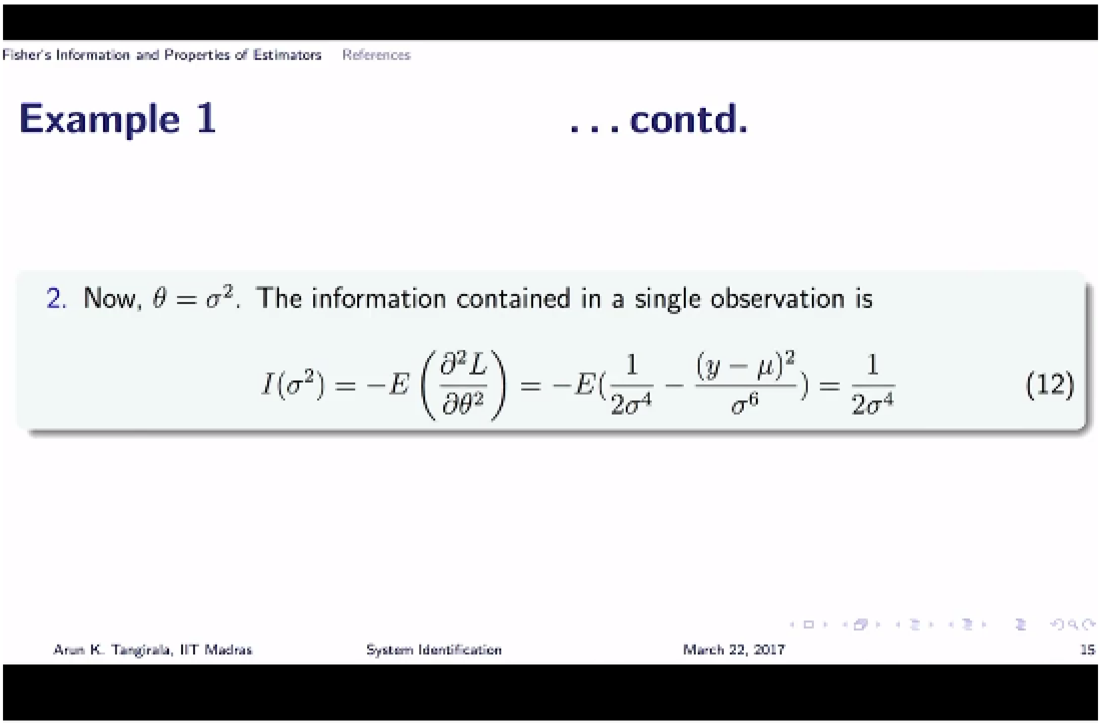
2. Now, $\theta = \sigma^2$. The information contained in a single observation is

$$I(\sigma^2) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E(\frac{1}{2\sigma^4} - \frac{(y-\mu)^2}{\sigma^6}) = \frac{1}{2\sigma^4} \tag{12}$$

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 15

Then, of course, you know, I should also tell you the single observation, how would you estimate variance is a question that crosses your mind. But for now, assume that you’regoing to have the single observation alone, maybe you lose a square of the observation as a variance estimate.Remember, Fisher's information does not explicitly talk about how you're estimating that parameter, it just is concerned about the information contained. But still, it assumes that you're going to use some kind of a likelihood method.

(Refer Slide Time 7:26)



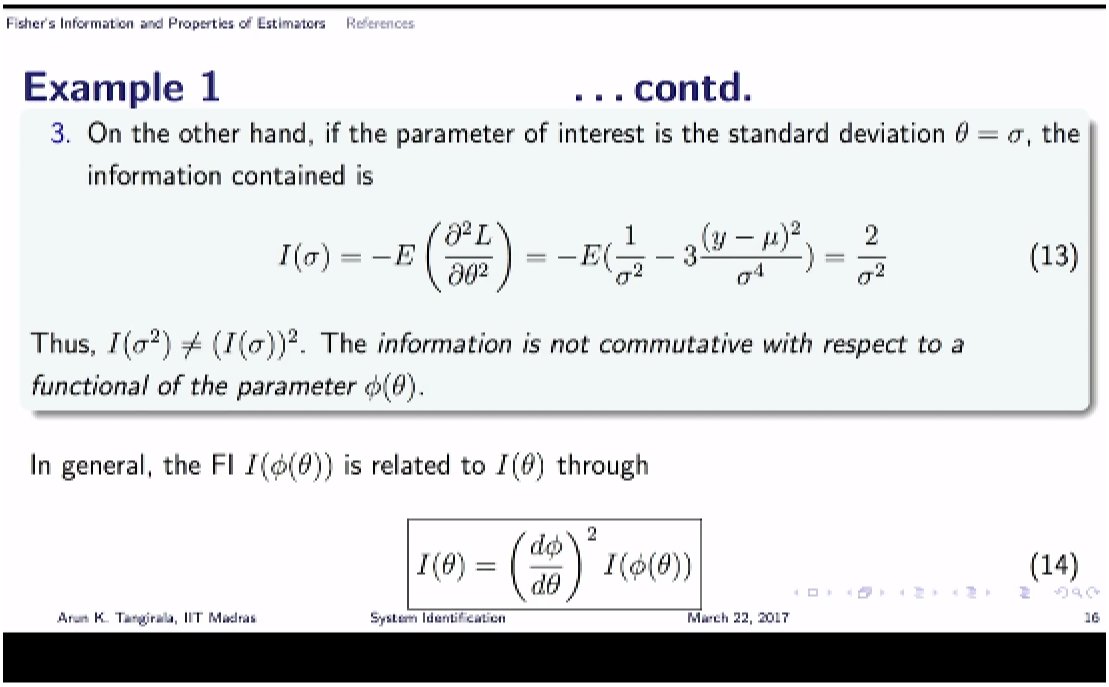
Example 1 . . . contd.

2. Now, $\theta = \sigma^2$. The information contained in a single observation is

$$I(\sigma^2) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E(\frac{1}{2\sigma^4} - \frac{(y-\mu)^2}{\sigma^6}) = \frac{1}{2\sigma^4} \qquad (12)$$

So if you work out the same example, the likelihood function doesn't change. The only difference is now the parameter of interest is sigma square, instead of mean. So when you work out the math, it turns out that you get this expression here for the variants of,sorry, information of contain in sigma square,it is one over two sigma power for four. Now, suppose I ask this question, I ask a slightly different question, which is that, what is information contain about sigma rather than sigma square?

(Refer Slide Time 8:05)



Example 1 . . . contd.

3. On the other hand, if the parameter of interest is the standard deviation $\theta = \sigma$, the information contained is

$$I(\sigma) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E(\frac{1}{\sigma^2} - 3\frac{(y-\mu)^2}{\sigma^4}) = \frac{2}{\sigma^2} \qquad (13)$$

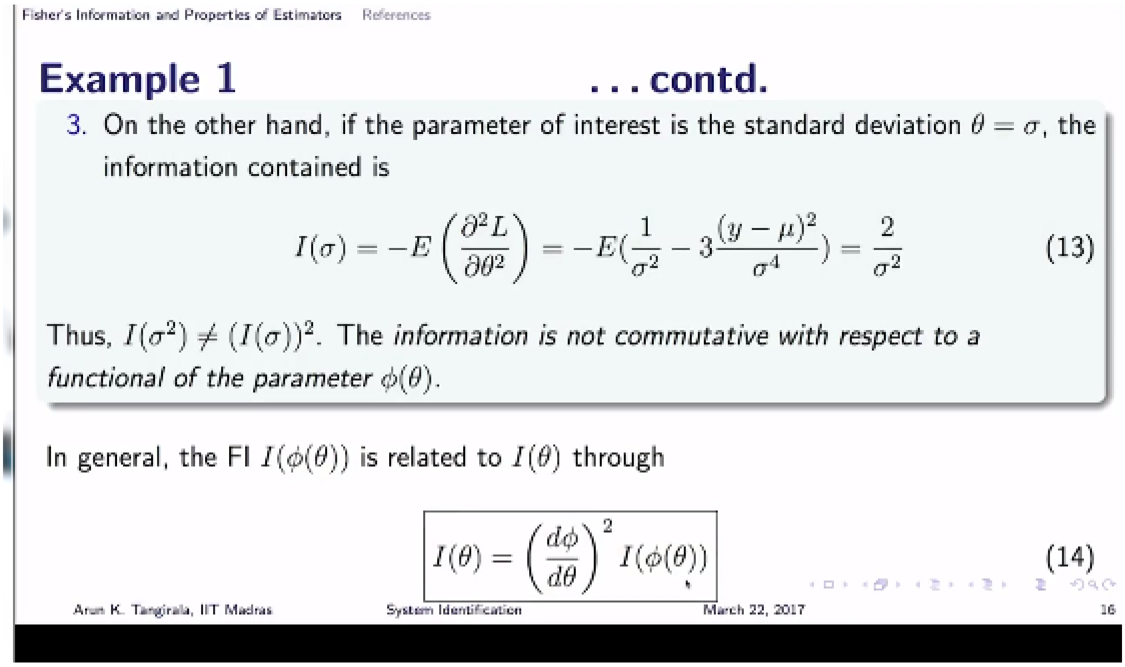
Thus, $I(\sigma^2) \neq (I(\sigma))^2$. The *information is not commutative with respect to a functional of the parameter* $\phi(\theta)$.

In general, the FI $I(\phi(\theta))$ is related to $I(\theta)$ through

$$\boxed{I(\theta) = \left(\frac{d\phi}{d\theta}\right)^2 I(\phi(\theta))} \qquad (14)$$

Intuitively, you may be tempted to say,look, it may be the square root of, but it turns out the answer is different from the square root of what we had earlier. So the information contained in sigma, contained about sigma square is one over two sigma power four.Whereas,information contained about sigma is twoover sigma square.All right? So clearly the square of this is not what we had earlier.Now, is this true in general? The answer is yes. The information containedin theta and phi of theta are not necessarily related this way. So, you cannot say information contain about phi of theta is phi of information contained on theta. This is not true. All right. In fact, the expression here at the bottom tells you how the i of theta and i of phi of theta are connected.They are connected through the square of the derivative of phi.

(Refer Slide Time 9:20)



Fisher's Information and Properties of Estimators References

Example 1 ... contd.

3. On the other hand, if the parameter of interest is the standard deviation $\theta = \sigma$, the information contained is

$$I(\sigma) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E(\frac{1}{\sigma^2} - 3\frac{(y-\mu)^2}{\sigma^4}) = \frac{2}{\sigma^2} \qquad (13)$$

Thus, $I(\sigma^2) \neq (I(\sigma))^2$. The *information is not commutative with respect to a functional of the parameter* $\phi(\theta)$.

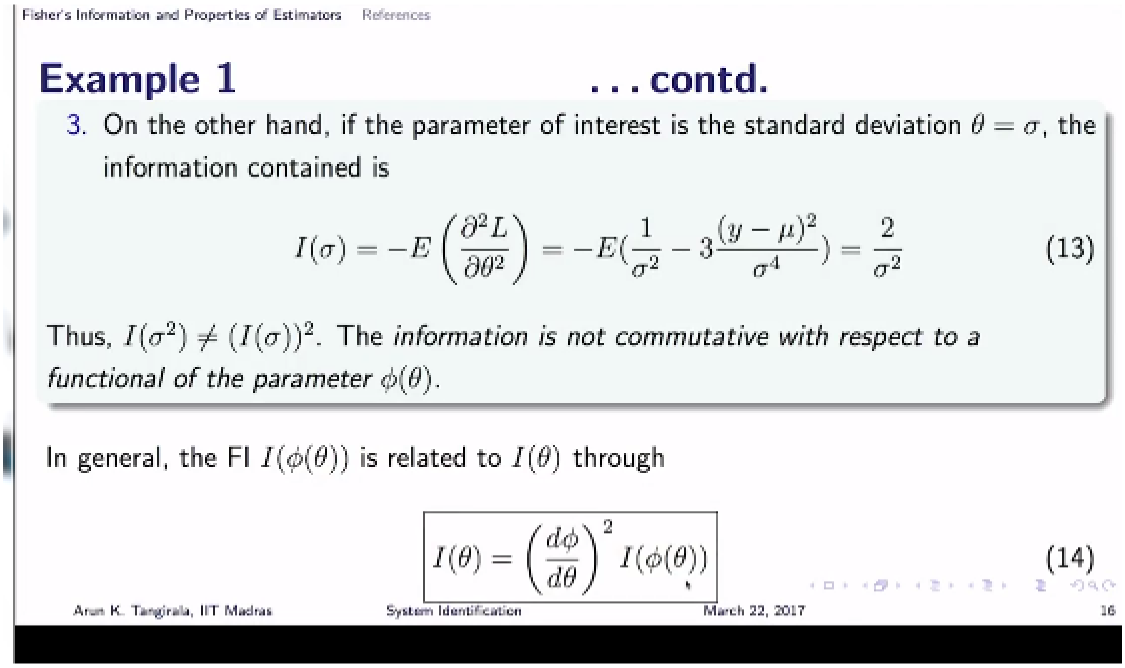
In general, the FI $I(\phi(\theta))$ is related to $I(\theta)$ through

$$\boxed{I(\theta) = \left(\frac{d\phi}{d\theta}\right)^2 I(\phi(\theta))} \qquad (14)$$

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 16

And you can cross check the answer this relation for this example. So, if you set theta as sigma and phi of theta sigma square,right? So,what would be information contained in sigma?That is two over sigma square, we already have that.And psiof theta is sigma square. So derivative of phi with respect to theta would be,what would it be?Right? So if you saytheta as, if you say, theta is sigma, then we have already shown this to be two over sigma square, this is your left hand side.The right hand side would be phi of theta for usis sigma square. Therefore, d phi or d theta would be,sorry,it has to be much, your volume has to be unfortunately, today much louder than the sound of the noise being producedthere. Two sigma.Still don't want to produce too much volume. Okay. Two sigma, that's correct. So, now, you can plug in into the relation there and verify quickly that you do have the right answer there.You already know information contained insigma square or you can cross check. So, this is in general the relation between i of theta and i of phi of theta. Now, while you can appreciate the beauty of this relation, the message that we get is, it may be different from estimating theta from phi of theta that is, suppose I want, what this tells us is that, suppose I have an estimator for sigmaright, I have estimated sigma, it may not necessarily mean that the optimal estimate or the most efficient estimate of sigma square will be simply the square of theestimate of sigma.Need not be, it may be, may not be. But in general, if you want to, therefore, estimate phi of theta, then estimate phi of theta directly. All right. Also, it says that suppose you do that. Suppose you estimate theta first and then take the transformation, then you have to account for this transformation also in the propagation of errors. Why I say that? So, I repeat. Suppose I have estimated theta, I have theta hat with me, it has some error. Now, I'm going to estimate phi of theta by simply applying the transformation to theta hat. When I do that, this relation kind of also tells me what accounts for how the error in theta hat propagates to phi of theta hat. This will also see in least squares later on. When we learn at least squares methods, we can ask if I have at least squares estimate of theta, what can I say about least squares estimate of phi of theta hat a phi of theta, if I just take a transformation. So the Jacobean of the function as you can see it is playing a role in the propagation of errors from theta hat to phi of theta hat, okay. And therefore, one has to be careful and also use this relation in calculating the variance of your phi of theta hat, okay. So in other words, if I'm estimating a model like this, suppose I'm estimating y equals, let's say alpha u, let's say this is the predictor and I have to estimate alpha. Suppose, I decided to fit this kind of a model beta square u, instead of fitting alpha u, I estimate beta square u and estimating this beta. I know the relation between beta square, beta and alpha. If I estimate beta and from where I can I try to estimate alpha, then I have to worry about this relation. When I can-- remember beta hat will have an error. How does beta hat error in beta hat propagate to alpha? This relation throws light on that? It says that the derivative of the Jacobean, sorry, the derivative of phi, of the Jacobean plays a role in the propagation of errors. The reason I say propagation of errors is as you must have seen in the videos, the most efficient estimator which is given by the Cramér Rao's inequality. Sorry, what is the bound on that? Inverse of Fisher's information, right. So the Cramér Rao's inequality tells me what is the bound that is what is a minimum error, I can expect in the estimate of a parameter. And that is related to Fisher's information. Therefore, this relation, in some sense throws light on how error propagates, that is something to remember.

(Refer Slide Time: 15:05)



Fisher's Information and Properties of Estimators References

Example 1 ...contd.

3. On the other hand, if the parameter of interest is the standard deviation $\theta = \sigma$, the information contained is

$$I(\sigma) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E(\frac{1}{\sigma^2} - 3\frac{(y-\mu)^2}{\sigma^4}) = \frac{2}{\sigma^2} \qquad (13)$$

Thus, $I(\sigma^2) \neq (I(\sigma))^2$. The *information is not commutative with respect to a functional of the parameter* $\phi(\theta)$.

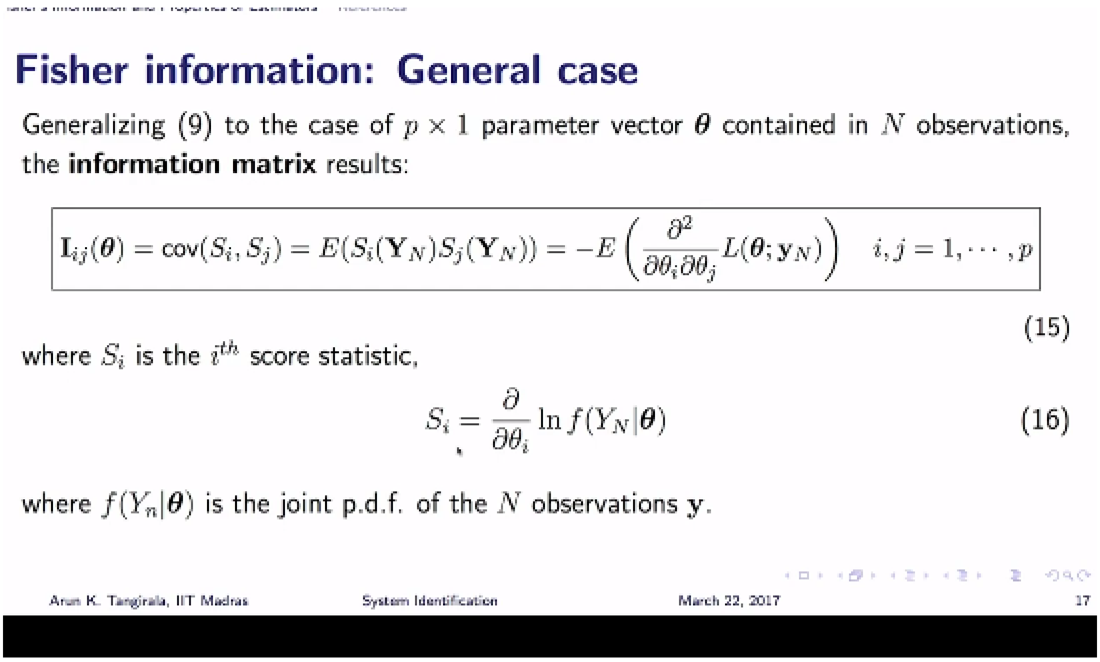
In general, the FI $I(\phi(\theta))$ is related to $I(\theta)$ through

$$\boxed{I(\theta) = \left(\frac{d\phi}{d\theta}\right)^2 I(\phi(\theta))} \qquad (14)$$

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 16

Now let's move on and look at the general case. This is the second example that will talk about where I have more than one parameter that I'm estimating, typically there is a case, right? And that I have more than one observation, there is also the general scenario. So when that is the case,you will have to set up the joint likelihood that is, or the joint p.d.f of those n observations and also now your Fisher's information is a matrix, it's going to be a matrix, and the I, jth element of this matrix will be governed by the partial derivative of the score of the likelihood function.

(Refer Slide Time: 15 50)



Fisher information: General case

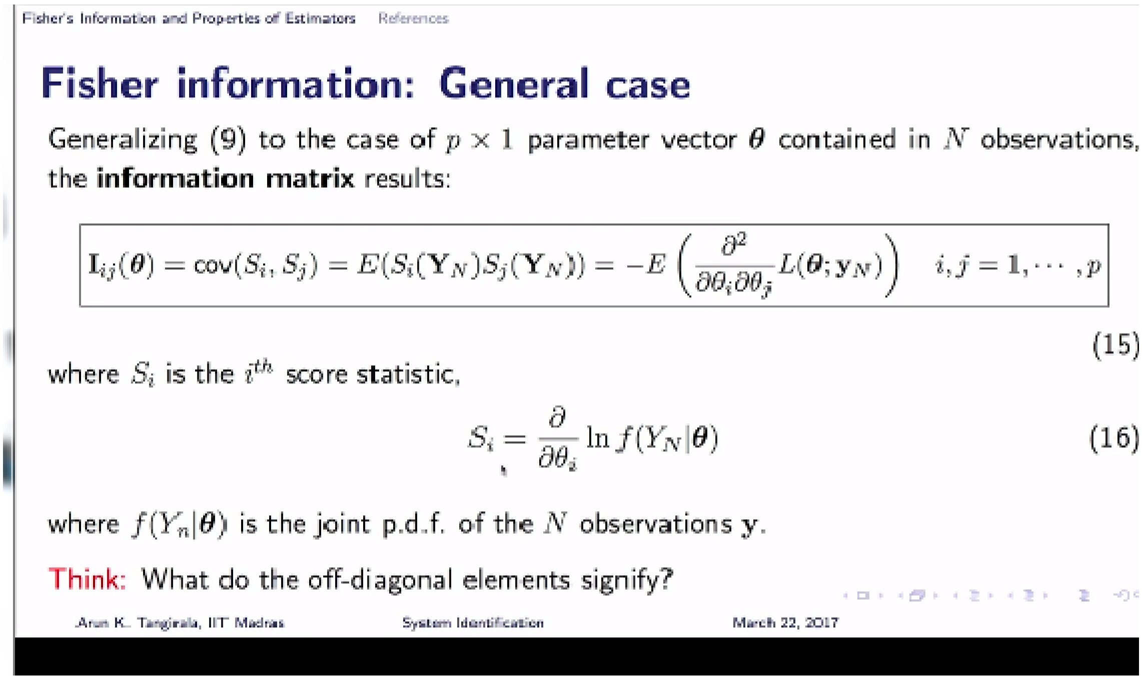
Generalizing (9) to the case of $p \times 1$ parameter vector $\boldsymbol{\theta}$ contained in N observations, the **information matrix** results:

$$\boxed{\mathbf{I}_{ij}(\boldsymbol{\theta}) = \text{cov}(S_i, S_j) = E(S_i(\mathbf{Y}_N)S_j(\mathbf{Y}_N)) = -E\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}L(\boldsymbol{\theta};\mathbf{y}_N)\right) \quad i,j = 1,\cdots,p} \tag{15}$$

where S_i is the i^{th} score statistic,

$$S_i = \frac{\partial}{\partial\theta_i}\ln f(Y_N|\boldsymbol{\theta}) \tag{16}$$

where $f(Y_n|\boldsymbol{\theta})$ is the joint p.d.f. of the N observations $\mathbf{y}$.



Fisher's Information and Properties of Estimators References

Fisher information: General case

Generalizing (9) to the case of $p \times 1$ parameter vector $\boldsymbol{\theta}$ contained in N observations, the **information matrix** results:

$$\boxed{\mathbf{I}_{ij}(\boldsymbol{\theta}) = \text{cov}(S_i, S_j) = E(S_i(\mathbf{Y}_N)S_j(\mathbf{Y}_N)) = -E\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}L(\boldsymbol{\theta};\mathbf{y}_N)\right) \quad i,j = 1,\cdots,p} \tag{15}$$

where S_i is the i^{th} score statistic,

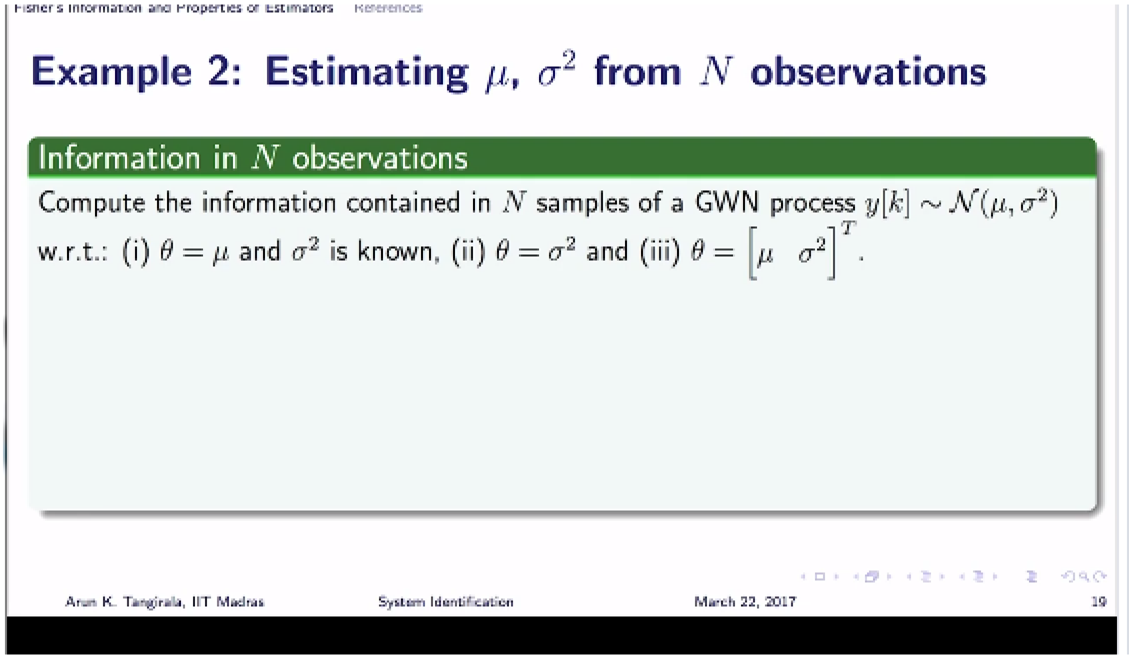
$$S_i = \frac{\partial}{\partial\theta_i}\ln f(Y_N|\boldsymbol{\theta}) \tag{16}$$

where $f(Y_n|\boldsymbol{\theta})$ is the joint p.d.f. of the N observations $\mathbf{y}$.

Think: What do the off-diagonal elements signify?

So, your Fisher's information is a matrix where you have a diagonal and then an off diagonal. The diagonal, kind of signifies the information contained jointly in about two parameters. And again, the way you define I, jth element of the matrix is simply the covariance between Si, Sj. Earlier, we defined for a single parameter, the Fisher's information to be simply variance of the score. Now, we are generalizing to covariance, that's the only difference and Si is nothing but the i partial derivative of the likelihood function with respect to the, no that's the theta i that you're looking at. So let's look at an example we have n observations, we went through one example already in the previous class.

(Refer Slide Time: 16:50)



Fisher's Information and Properties of Estimators References

Example 2: Estimating μ, σ^2 from N observations

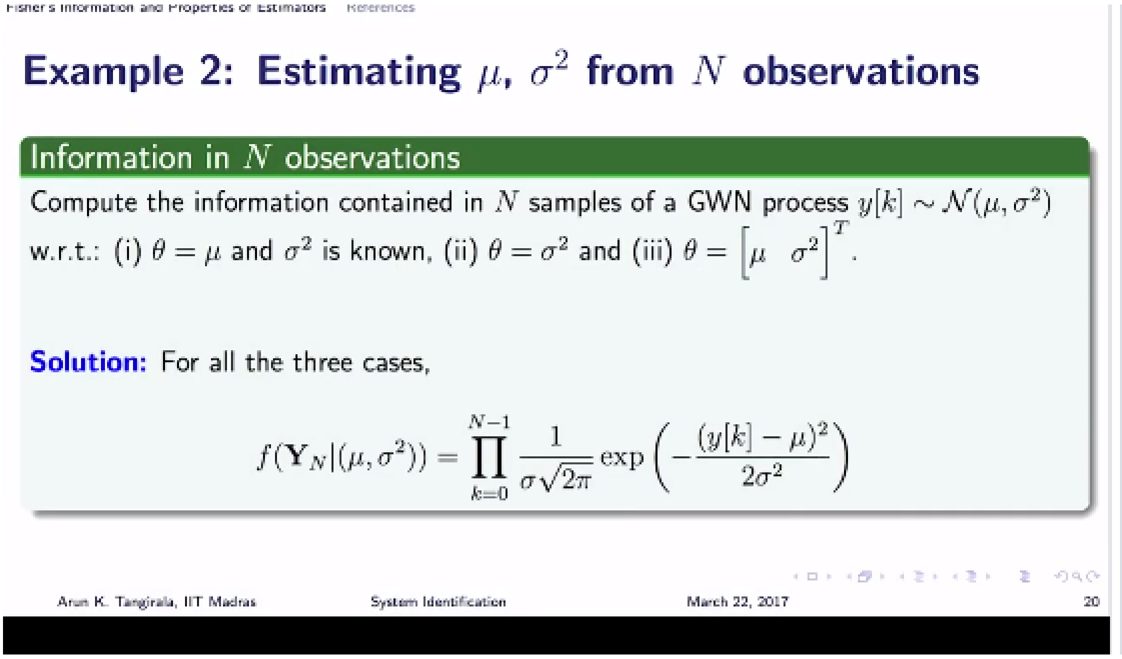
Information in N observations

Compute the information contained in N samples of a GWN process $y[k] \sim \mathcal{N}(\mu, \sigma^2)$ w.r.t.: (i) $\theta = \mu$ and σ^2 is known, (ii) $\theta = \sigma^2$ and (iii) $\theta = \begin{bmatrix} \mu & \sigma^2 \end{bmatrix}^T$.

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 19

So there are three scenarios here. I'm given n observations, I'm also given that these observations fall out of Gaussian white noise process. In the first scenario, I would like to know how much information is contained in this n observations about mean, assuming sigma square is known. In the second scenario, I would like to know what is information contained about sigma square with me being fixed.And the third one is where I want to know how much information is contained in this and observations when both are unknown. We have already gone through this example in the last class. And we kind of intuitively said that the Fisher's information in n observations about the mean is simply going to be n times the information contained in the single observation. But why when does it hold? Yeah. When they are uncorrelated or independent rather, right? When the observations are independent, here we have a Gaussian and uncorrelated, but we know already Gaussian and uncorrelated means independent. So each observation, when we say we have n independent observations what that means is, each observation is getting something unique, giving me some information that is not contained in the other observation, right? Therefore, I should expect the information to be proportionally increased by a factor of n.

(Refer Slide Time: 18:24)



Fisher's Information and Properties of Estimators References

Example 2: Estimating μ, σ^2 from N observations

Information in N observations

Compute the information contained in N samples of a GWN process $y[k] \sim \mathcal{N}(\mu, \sigma^2)$ w.r.t.: (i) $\theta = \mu$ and σ^2 is known, (ii) $\theta = \sigma^2$ and (iii) $\theta = \begin{bmatrix} \mu & \sigma^2 \end{bmatrix}^T$.

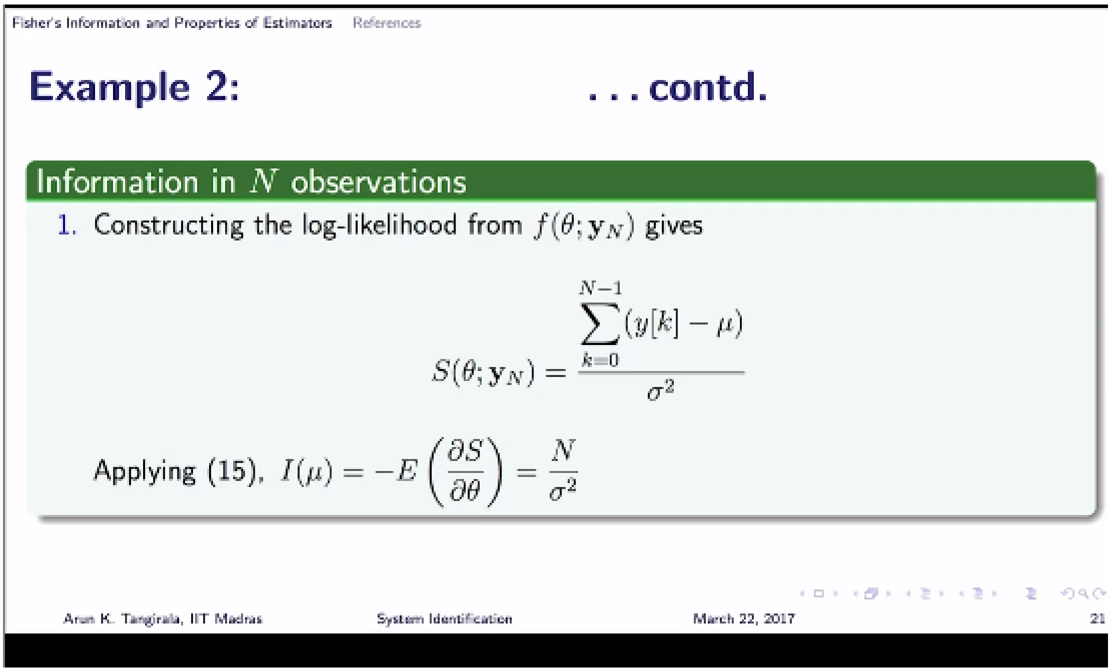
Solution: For all the three cases,

$$f(\mathbf{Y}_N | (\mu, \sigma^2)) = \prod_{k=0}^{N-1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y[k]-\mu)^2}{2\sigma^2} \right)$$

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 20

And that is what turns out when you work out the math, you begin with a joint p.d.f. and essentially the likelihood function, work out the score, essentially take the derivative of the p.d.f. with respect to mean and then take the variance, I mean, apply the Fisher's information definition minus expectation of the database of the score, and you get n over sigma squared, right?

(Refer Slide Time: 18:55)



Fisher's Information and Properties of Estimators References

Example 2: ...contd.

Information in N observations

1. Constructing the log-likelihood from $f(\theta; \mathbf{y}_N)$ gives

$$S(\theta; \mathbf{y}_N) = \frac{\sum_{k=0}^{N-1} (y[k]-\mu)}{\sigma^2}$$

Applying (15), $I(\mu) = -E\left(\frac{\partial S}{\partial \theta}\right) = \frac{N}{\sigma^2}$

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 21

This is exactly n times what we had for the single observation, no surprise. From this, you can only so kind of infer how many degrees of freedom do you have. This degrees of freedom is a very important concept in statistics. What are degrees of freedom in linear algebra? When you run, when we are solving a set of equations, we run into degrees of freedom concept there as well as right? What is the degrees of freedom concept in linear algebra, when you're solving a set of equations? In choosing the unknowns. So if I have n unknown than any equation, what is the degrees of freedom? Zero, right.

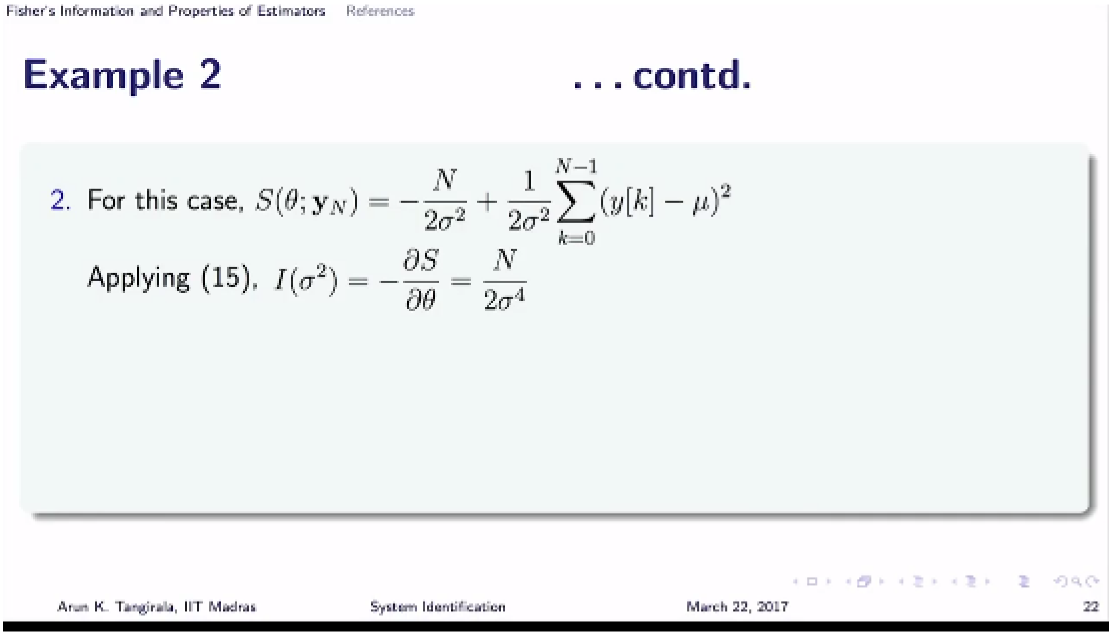
Correct. Here, the degrees of freedom concept in statistics is different.The degrees of freedom concept in statistics is the number of independent sources of variability with respect to some random variable. So the random variable that I'm looking at is the estimate of something mean here. Some theta hat. So if I were to ask in how many different ways is theta hat affected when I have any observations here, I say there are no degrees of freedom.But this is true only because I'm looking at n observations that are independent. If I had a correlated process, the degrees of freedom will come down.Why is this degrees of freedom important? This degrees of freedom is important in estimation because it tells the designer or the analyst how in what way you can actually affect the variance or the information. Ultimately, I've been saying that right, right? Cramér Rao's inequality will make use of this Fisher's information and show us that the lowest variability that I can achieve is the inverse of Fisher's information. So when we talk official's information, we might as well treat it as some variants of the parameter estimate. So this degrees of freedom tells an analyst, data analyst in statistics as to what is the controlling factor, whether there is a controlling factor? And if there is, what is the controlling factor by which you can affect the variability of the estimate. So here, I have this example tells me as I, the degrees of freedom is n. That means if I and we know that the inverse of this is going to play a role in the error, minimum error that I can get. So, which means, if I turn to Cramér's Rao's inequality, and ask what is the minimum variance, I can expect in the estimation of mean. So if I ask the Cramér Rao's inequality will tell me that the minimum variance is, I'm going to drop this way b, but this is what they see a Cramér Rao's inequality is going to tell me. What is the sigma square on the right hand side? Of the data generating process, right. So, you should not confuse, you should be very clear. On the left hand side, what is a variance that we're talking about of the estimate, right? So, you should always learn to distinguish between sigma square of theta hat and sigma square of the process. The fact that there are n degrees of freedom is very encouraging now. Why? Don't want to say anything? Correct. So, which means, and the degrees of freedom is, is there anything that I can control?

Both the observations compare those.

Yes. The number of observation something that mostly is within my control. Yeah, in some cases may not. But by and large, if I put in more effort, I can obtain more observations which is good. So there is a controlling factor on the error. Because its variability is directly going to affect the standard error in estimate, square root of this is a standard error. The fact that there is a degrees of freedom that is within the within my control is good news. Now, by increasing n, I can decrease the error in the estimate. And if n goes to infinity, the error goes to zero. Which is what is the hallmark of a consistent estimator, right? Consistency is all about. So the, on the other hand with the single observation, there was no control. With a single observation, the fisher's information was one over sigma square. There's nothing you could do. But it is telling now that if I increase the number of observations, if I collect more observations, then the error will come down. He may say, what's the big deal I? That's should be kind of intuitive, that as a collect more data, my the estimate should improve correct, it seems very intuitive. However, it is not necessarily the case all the time. It depends on how you're estimating. So the classic example ETFE, we have talked about ETFE, right? Empirical time for function estimator, although we have not talked about estimator but we've talked about ETF, if you estimate the ETF, the empirical transfer function from n observations, unfortunately, the situation that is error in the estimate of ETF doesn't improve at all. It is an inconsistent estimate. Or the other example is, spectral density. If I'm using periodogram to estimate the spectral density, the same storywith how do you compute the period of time we have gone through that already right, you take the DFT square and divide by N. And if you use that to estimate the spectral density of random signal, unfortunately, even though you supply billion points, billion observations, the quality of the periodogram doesn't improve at all. So there are glaring examples in estimation theory that give rise to inconsistent estimators.

Therefore, it is important before you choose to work with any estimator you should be assured that by increasing the number of observations the quality of the estimate will improve. And for this, you have to either turn to theory if it is an existing estimator or if it is a new estimate that you have come up with, then you have to either prove it theoretically, that yes, the number of observations, you know, increase in n will give rise to decrease in error or you prove it through Monte Carlo simulations. One of this has to happen, otherwise, you should not be straight away jumping and using any estimator that you want.

(Refer Slide Time: 26:20)



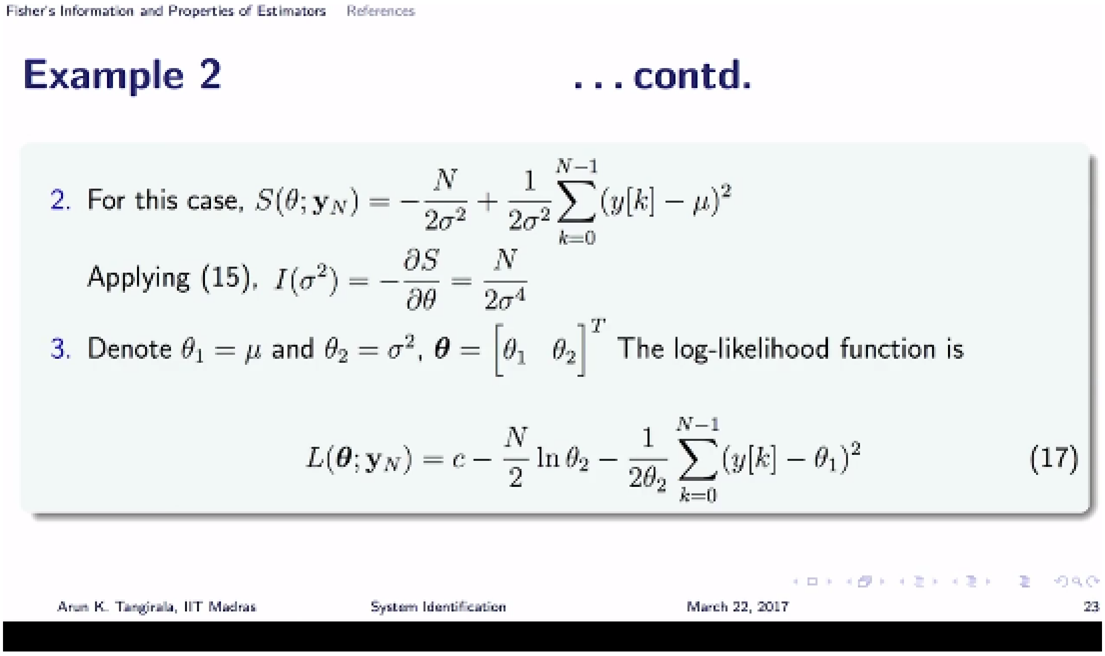
Fisher's Information and Properties of Estimators References

Example 2 . . . contd.

2. For this case, $S(\theta; \mathbf{y}_N) = -\dfrac{N}{2\sigma^2} + \dfrac{1}{2\sigma^2}\sum_{k=0}^{N-1}(y[k]-\mu)^2$

   Applying (15), $I(\sigma^2) = -\dfrac{\partial S}{\partial \theta} = \dfrac{N}{2\sigma^4}$

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 22



Fisher's Information and Properties of Estimators References

Example 2 . . . contd.

2. For this case, $S(\theta; \mathbf{y}_N) = -\dfrac{N}{2\sigma^2} + \dfrac{1}{2\sigma^2}\sum_{k=0}^{N-1}(y[k]-\mu)^2$

   Applying (15), $I(\sigma^2) = -\dfrac{\partial S}{\partial \theta} = \dfrac{N}{2\sigma^4}$

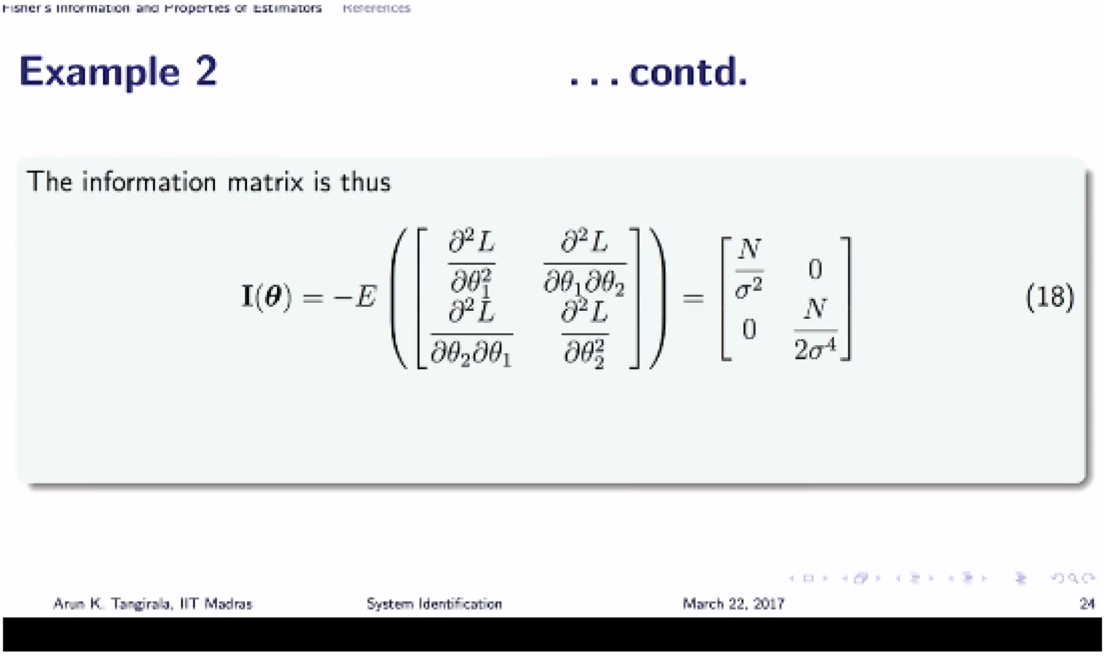
3. Denote $\theta_1 = \mu$ and $\theta_2 = \sigma^2$, $\boldsymbol{\theta} = \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}^T$ The log-likelihood function is

$$L(\boldsymbol{\theta}; \mathbf{y}_N) = c - \frac{N}{2}\ln\theta_2 - \frac{1}{2\theta_2}\sum_{k=0}^{N-1}(y[k]-\theta_1)^2 \qquad (17)$$

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 23

Okay. So, these are the two things that I wanted to talk about of course,the you see the same increase in the number of, in the degrees of freedom when it comes to estimating sigma square. Earlier, we had one over two sigma to the power of four for IO sigma square. Now, we have n over two sigma to the power of four, which means there has been a proportion increase.

(Refer Slide Time: 26:46)



Fisher's Information and Properties of Estimators References

Example 2 ...contd.

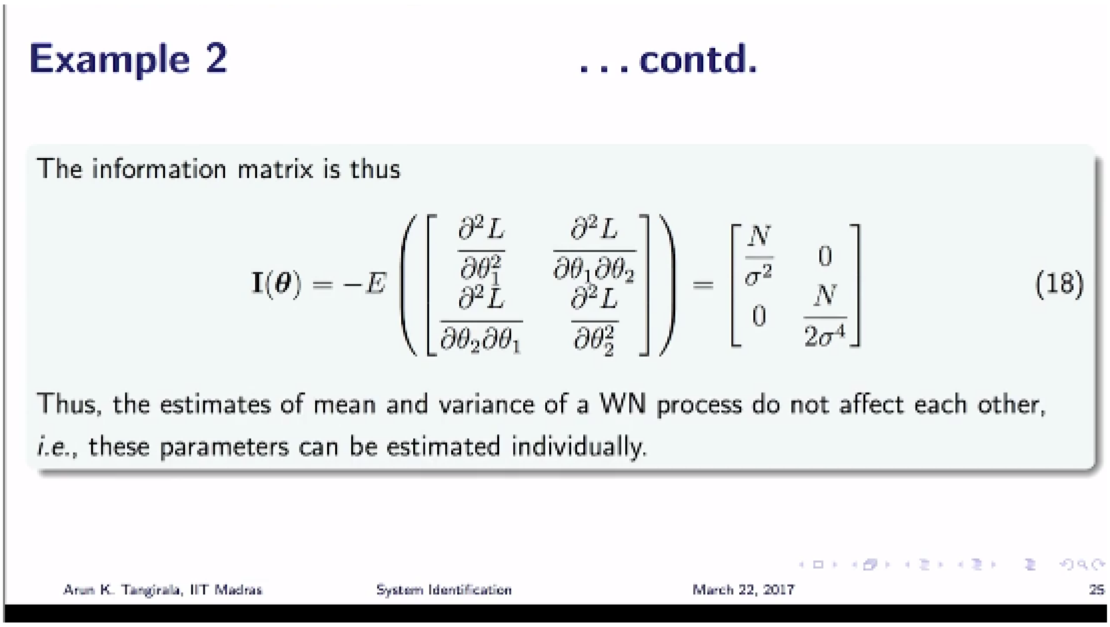
The information matrix is thus

$$\mathbf{I}(\boldsymbol{\theta}) = -E\left(\begin{bmatrix} \dfrac{\partial^2 L}{\partial \theta_1^2} & \dfrac{\partial^2 L}{\partial \theta_1 \partial \theta_2} \\ \dfrac{\partial^2 L}{\partial \theta_2 \partial \theta_1} & \dfrac{\partial^2 L}{\partial \theta_2^2} \end{bmatrix}\right) = \begin{bmatrix} \dfrac{N}{\sigma^2} & 0 \\ 0 & \dfrac{N}{2\sigma^4} \end{bmatrix} \tag{18}$$

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 24

Now the interesting thing is, when I estimate both sigma square and mu jointly, remember I said now you have a matrix Fisher's information is a matrix. In general, this matrix is non- diagonal, but only in this example, it is diagonal. What does it mean? What does the diagram nature mean? What is the nature of the Fisher's information matrix tell you? You have to understand what the matrix is about, right? What are the diagonal elements telling me? So if I were to estimate them individually, what would be the official information, right? But since I'm estimating jointly, there must be some error, right? That may be different from estimating them jointly. Because remember, date, it's the same data, but now you're extracting more juice out of it. So you may end up actually getting more error. In fact, the diagonal elements are also not exactly what you will get if you're a testament and individually, but it says in the individual, what, it gives you an idea of what are the errors in the individual parameter estimates, when you're estimating them jointly.When you're estimating them jointly, there are going to be two kinds of errors. One with the which is error in the individual theta hats, other is, joint one. Because you're estimating them jointly, there is an error associated with jointly estimating them. This matrix tells me, that look, there is no error in estimating there is no additional error in induced because you're jointly estimating. It is as good as estimating them individually. In a different scenario, in a general scenario, if this matrix is not necessarily diagonal at all, which means that there will be an additional error induced because you're estimating jointly, vis-à-vis estimating an individually. So that is how you read the Fisher's information matrix, right? So I just repeat when you're estimating many parameters, you should expect more error in the individual parameter estimates, because now you are diverting some of the information towards estimating additional parameters. It's the same data that you're using. And the off diagonal terms will give you an idea of that.But in this case, in this example, specifically, it says whether you estimate mu and sigma square jointly or individually, the same story. It doesn't matter, okay. And that's why the diagonal elements coincide with what we had obtained individually. In some other problem, suppose I'm estimating, let's say, I have a FIR model. I have a FIR model and I'm estimating the parameters of the FIR model. Suppose I FIR model containing of length two, that means I have two parameters. Then, estimating B one and B two, let us say those are the parameters. Estimating B1 and B2 jointly would be different from estimating B1 and B2 separately. But that depends on the kind of input that you use. Let's see if I can actually set up such a simple problem for you in assignment, so that you understand what difference does it make in estimating the parameters jointly, vis-à-vis individually.

(Refer Slide Time: 30:55)



Example 2 . . . contd.

The information matrix is thus

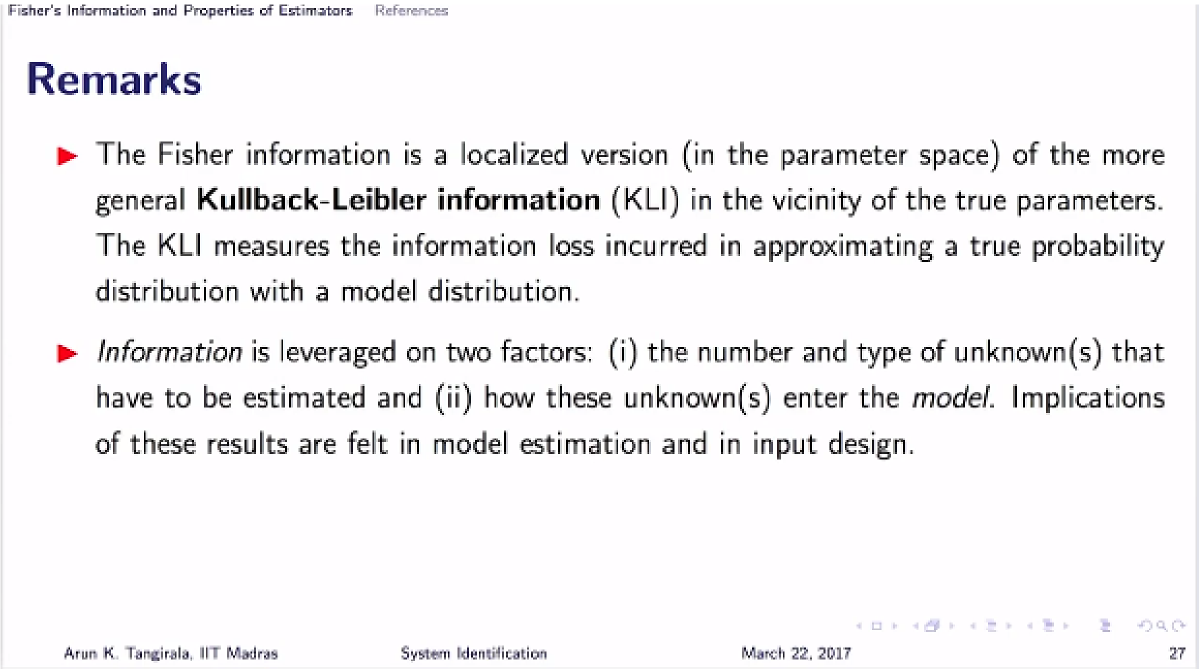
$$\mathbf{I}(\boldsymbol{\theta}) = -E\left(\begin{bmatrix} \dfrac{\partial^2 L}{\partial \theta_1^2} & \dfrac{\partial^2 L}{\partial \theta_1 \partial \theta_2} \\ \dfrac{\partial^2 L}{\partial \theta_2 \partial \theta_1} & \dfrac{\partial^2 L}{\partial \theta_2^2} \end{bmatrix}\right) = \begin{bmatrix} \dfrac{N}{\sigma^2} & 0 \\ 0 & \dfrac{N}{2\sigma^4} \end{bmatrix} \qquad (18)$$

Thus, the estimates of mean and variance of a WN process do not affect each other, *i.e.*, these parameters can be estimated individually.

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 25

Okay. So with that, we come to a close on the Fisher's information as I said, as far as system solid system identification is concerned, Fisher's information plays a central role in input design, because when we design input, one of the key criteria is to generate information data, right. And that's a qualitative term, Fisher's information allows you to quantify that information. And as we have seen in been discussing the Cramér Rao's inequality does tell us that Fisher's information has a direct bearing on the variants of the parameter estimates. Therefore, if you maximize the official's information, you are minimizing the error as per Cramér Rao inequality and that is why you should generate informative data. And technically, Fisher's information is supposed to be a localized version of the Kullback-Leibler information measure that came much later that fishes information was conceived earlier. The Kullback-Leibler is based on p.d.fs straightaway. There is no second moments and, so you can show that the Fisher's information is some kind of a first order approximation. And what we also learn is that the information is leveraged on two factors.

(Refer Slide Time: 32:20)



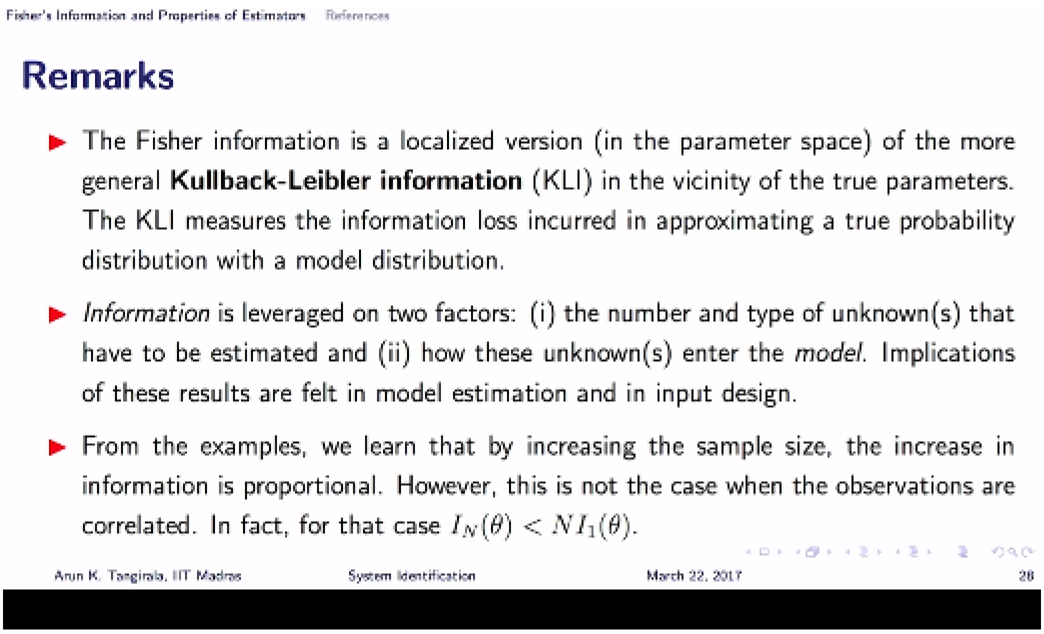
Fisher's Information and Properties of Estimators References

Remarks

- The Fisher information is a localized version (in the parameter space) of the more general **Kullback-Leibler information** (KLI) in the vicinity of the true parameters. The KLI measures the information loss incurred in approximating a true probability distribution with a model distribution.
- *Information* is leveraged on two factors: (i) the number and type of unknown(s) that have to be estimated and (ii) how these unknown(s) enter the *model*. Implications of these results are felt in model estimation and in input design.

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 27

How many unknowns you're estimating and how these unknowns enter the model? The second part is extremely important. What we mean by, how this unknowns enter of is, how you're parameterizing your model. I'll give you a simple analogy. We have always said parameters are like our guests or manpower, I mean, we have been giving this kind of analogies. Suppose, I have a person who is coming and doing some work at home or I have an unknown person in some kind of a guest, who is at home, what this person is doing, what is information that I have about this person depends on where this person is seated in the house? Am I right? If the person is working in a central area, then it becomes easy to watch what happens. But if this person is working in some remote area of the house then you have to give, you have to put in more effort. You have to give special attention, it's not easy to know what is happening. In a same way where your parameters are seated in the model, how they are seated, makes a difference. I don't know if in the previous assignment or the next assignment, there's a question which tells you for the exponential distribution. If you workout the Fisher's information for Lambda which is the parameter for an exponential p.d.f, it is not possible to find efficient estimator of lambda. But if you re-parameter is a model in terms of one over lambda, you know, exponential distribution are characterized by this lambda, which is average rate. Then it turns out that you can find an efficient estimator of them. So this read parameterization that is writing the model in terms of a different set of parameters so that they appear differently in the model. Sometimes there's a lot of magic. And that comes with experience also. And sometimes, if you cannot find a way of efficiently estimating one parameter, then you can, and you should think about rewriting the model in terms of different parameters. All right. So the general thing that we learn is increasing the sample size improves information in general, but you have to be cautious for estimators that are not consistent, this is not true.



Fisher's Information and Properties of Estimators References

Remarks

- The Fisher information is a localized version (in the parameter space) of the more general **Kullback-Leibler information** (KLI) in the vicinity of the true parameters. The KLI measures the information loss incurred in approximating a true probability distribution with a model distribution.
- *Information* is leveraged on two factors: (i) the number and type of unknown(s) that have to be estimated and (ii) how these unknown(s) enter the *model*. Implications of these results are felt in model estimation and in input design.
- From the examples, we learn that by increasing the sample size, the increase in information is proportional. However, this is not the case when the observations are correlated. In fact, for that case $I_N(\theta) < NI_1(\theta)$.

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 28

(Refer Slide Time: 35: 05)

Okay, so until now, we have looked at how to assess the quality of data that goes and sits into the estimator. We've talked about Fisher's information, which tells us how good the data is, whether you're sending in data, or you're sending in garbage. Now comes the second part, which is that of estimator, assume that the data is informative, what can you say about the estimator. How well can it actually extract the juice out of the data, right. So I give a healthy sugarcane, which has a lot of juice, but if I give it to a machine that cannot extract enough juice, or there is a person who was not a pro at extracting news out of this, then you may actually leave, you may throw away sugarcane after one processing and in that process, throwaway a lot of juice, you're not being efficient enough. Because you're putting in a very healthy, nice juicy sugarcane into the machine and you're not getting the juice that you should have obtained from this cane. So there are certain things that we look out for when it comes to estimator.