CH5230: System Identification

Fisher's information and properties of estimators

Part 02

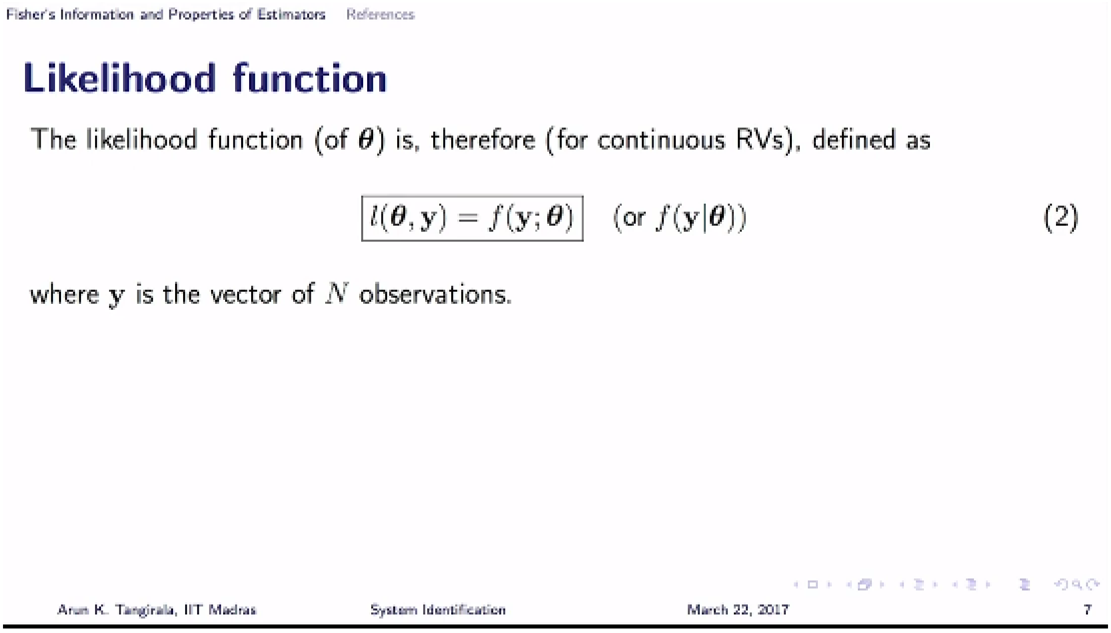
Now you can see yesterday, we looked at optimization based approach. Today also we're looking at optimization based approach to estimation. But what is the difference? Between those two approaches that we looked at yesterday, we could have also done the same problem; we could have used the same idea MLE for yesterday's example.

Right, we have assumed e[k] to be Gaussian white noise. What will be the PDF, jointPDF of y? In yesterday's example, I have y[k] falling out of a Gaussian white noise process. What would be the joint PDF of y? Gaussian because they're uncorrelated Gaussian processis also independent. So the joint PDF of those nobservations will be the product of the individual PDFs. And then I can set up my likelihood. What is theta for us yesterday?

In yesterday's example. C but also Sigma square e, which we did not talk about. So those two other parameters and they would enter the PDF of y, because y would have a mean of C and variants of, y[k], the mean of y[k] is C. Variance of y[k], sigma square e.Correct? So straightaway the parameters are there in the PDF, I can use the MLE. But we used a different approach. We use a least squares approach. Now you may ask, is there a huge difference between least squares approach and maximum likelihood approach. The answer is yes and no. Yes, philosophically. The approach, the assumptions everything, see here, if I were to use the MLE approach for yesterday's problem, I have to begin by assuming what e[k] is? That means whether it's Gaussian white or uniform white or some other white noise? Did I need that for the least squares?

All I needed is that it's a white noise. I didn't need to know the distribution, all I needed to know is e[k] is unpredictable. So that I could construct the prediction, but here I need additional assumptions. Later on, what we will see is a beautiful connection between MLE and Least squares. We will realize that under some conditions, specifically the Gaussianity assumptions that MLE and Least squaresare the same. They are identical. Okay? Fine, so let's proceed now and understand what Fisher's information quickly. Now that we have understood what is likelihood, always keep telling yourself this, likelihood is nothing but the PDF itself. So what one does is one assumes the PDF and then proceeds.

(Refer Slide Time: 3:02)



Fisher's Information and Properties of Estimators References

Likelihood function

The likelihood function (of θ) is, therefore (for continuous RVs), defined as

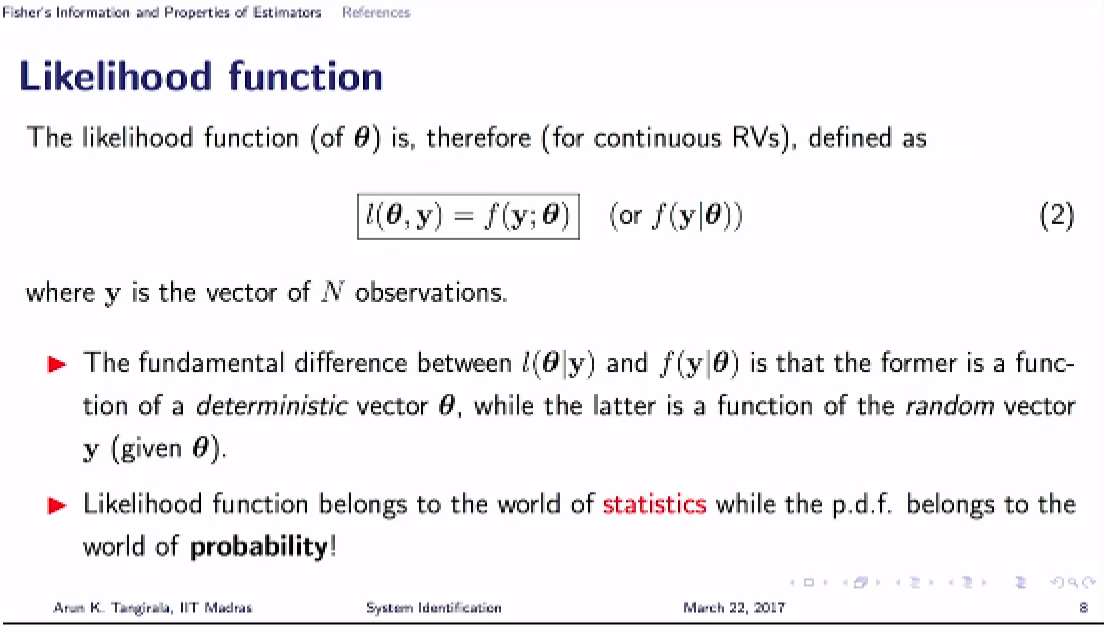
$$\boxed{l(\boldsymbol{\theta}, \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta})} \quad (\text{or } f(\mathbf{y}|\boldsymbol{\theta})) \tag{2}$$

where $\mathbf{y}$ is the vector of N observations.

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 7

Of course, we have quite a few examples to throw light on this.

(Refer Slide Time: 3:09)



Fisher's Information and Properties of Estimators References

Likelihood function

The likelihood function (of $\boldsymbol{\theta}$) is, therefore (for continuous RVs), defined as

$$\boxed{l(\boldsymbol{\theta}, \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta})} \quad (\text{or } f(\mathbf{y}|\boldsymbol{\theta})) \tag{2}$$

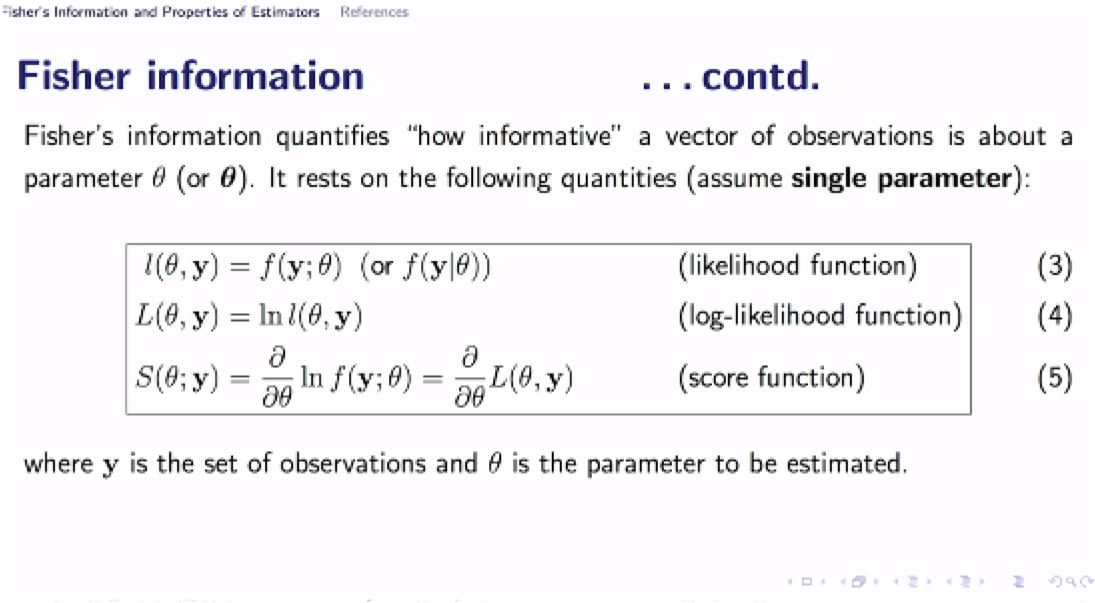
where $\mathbf{y}$ is the vector of N observations.

- The fundamental difference between $l(\boldsymbol{\theta}|\mathbf{y})$ and $f(\mathbf{y}|\boldsymbol{\theta})$ is that the former is a function of a *deterministic* vector $\boldsymbol{\theta}$, while the latter is a function of the *random* vector $\mathbf{y}$ (given $\boldsymbol{\theta}$).
- Likelihood function belongs to the world of statistics while the p.d.f. belongs to the world of **probability**!

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 8

And we'll go past and come straight to Fisher's information. We already talked about maximum likelihood principle. So what is Fisher's information?

(Refer Slide Time: 3:19)



Fisher's Information and Properties of Estimators References

Fisher information . . . contd.

Fisher's information quantifies "how informative" a vector of observations is about a parameter θ (or $\boldsymbol{\theta}$). It rests on the following quantities (assume **single parameter**):

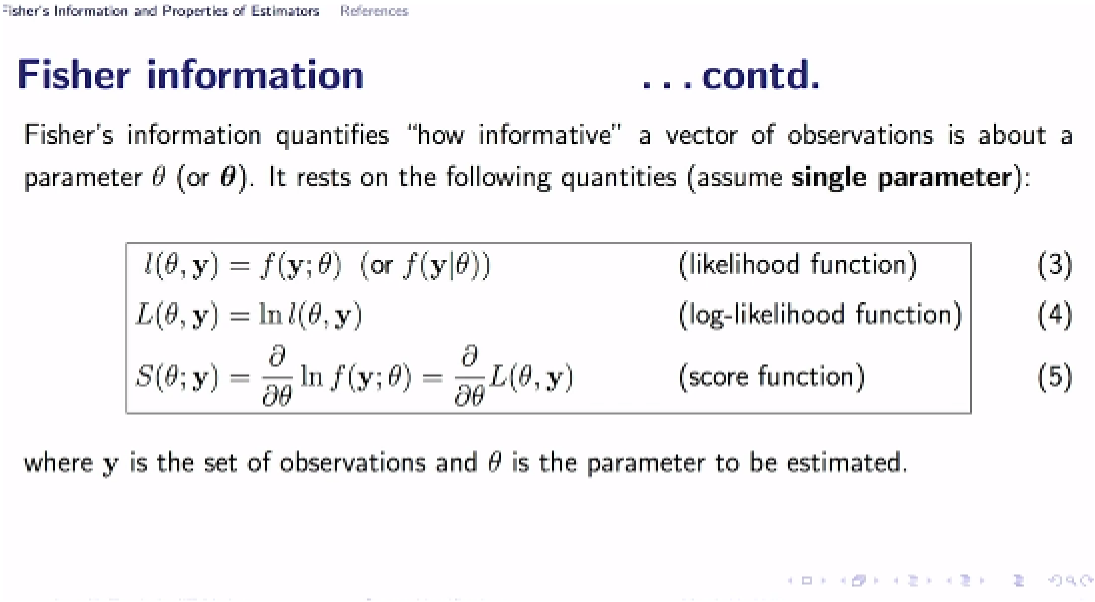
$l(\theta, \mathbf{y}) = f(\mathbf{y}; \theta) \quad (\text{or } f(\mathbf{y}\vert\theta))$	(likelihood function)	(3)
$L(\theta, \mathbf{y}) = \ln l(\theta, \mathbf{y})$	(log-likelihood function)	(4)
$S(\theta; \mathbf{y}) = \dfrac{\partial}{\partial \theta} \ln f(\mathbf{y}; \theta) = \dfrac{\partial}{\partial \theta} L(\theta, \mathbf{y})$	(score function)	(5)

where $\mathbf{y}$ is the set of observations and θ is the parameter to be estimated.

Now, look at this objective function here. What are we trying to do? We are trying to find theta such that this is maximized. Fisher said that a metric of information in data can we obtained follows? First step, you ask how sensitive your objective function is with respect to your theta? Suppose your objective function is not sensitive with respect to theta, what happens?Right? The more the objective sensitive that is, it has to be sensitive; theta has to be a part of your objective function. That's a minimum requirement. Look at the sensitivity of the objective function with respect to your theta.

So if you look at it, we have something called a score. Fisher called this as a score. The score is the derivative of the log-likelihood with respect to theta. So, that's a first step. But remember-- by the way the bold faced y is nothing but y[n] vector.

(Refer Slide Time: 4:31)



Fisher's Information and Properties of Estimators References

Fisher information ...contd.

Fisher's information quantifies "how informative" a vector of observations is about a parameter θ (or $\boldsymbol{\theta}$). It rests on the following quantities (assume **single parameter**):

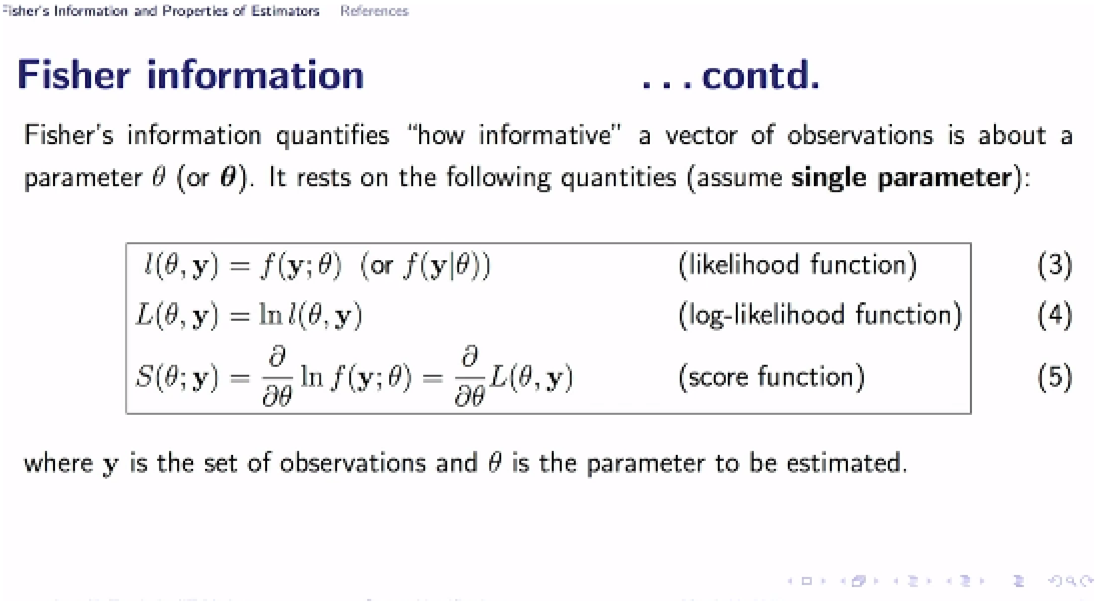
$$\boxed{\begin{array}{ll} l(\theta, \mathbf{y}) = f(\mathbf{y}; \theta) \;\; (\text{or } f(\mathbf{y}|\theta)) & (\text{likelihood function}) \\ L(\theta, \mathbf{y}) = \ln l(\theta, \mathbf{y}) & (\text{log-likelihood function}) \\ S(\theta; \mathbf{y}) = \dfrac{\partial}{\partial \theta} \ln f(\mathbf{y}; \theta) = \dfrac{\partial}{\partial \theta} L(\theta, \mathbf{y}) & (\text{score function}) \end{array}} \quad \begin{array}{l} (3) \\ (4) \\ (5) \end{array}$$

where $\mathbf{y}$ is the set of observations and θ is the parameter to be estimated.

But remember that, whatever score I'm going to compute is for a single realisation. Whatever score, I'm going to compute is not a single realisation. Now I have to look at how this sensitivity varies with realisations?

Correct? And that means, when I'm evaluatingSfor-- yes I'm fixing y. It is like your conditional, I think, right? I'm conditioning it on the given data. But now I want to evaluate this overall possible data records.

(Refer Slide Time: 5:12)



Fisher's Information and Properties of Estimators References

Fisher information ...contd.

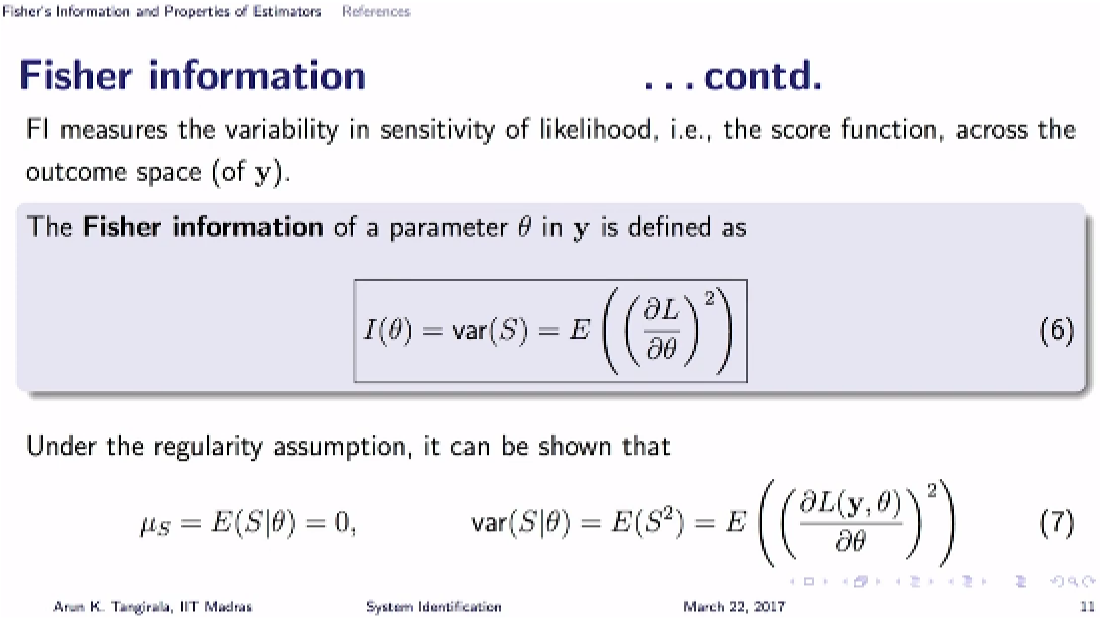
Fisher's information quantifies "how informative" a vector of observations is about a parameter θ (or $\boldsymbol{\theta}$). It rests on the following quantities (assume **single parameter**):

$$\boxed{\begin{array}{ll} l(\theta, \mathbf{y}) = f(\mathbf{y}; \theta) \;\; (\text{or } f(\mathbf{y}|\theta)) & (\text{likelihood function}) \\ L(\theta, \mathbf{y}) = \ln l(\theta, \mathbf{y}) & (\text{log-likelihood function}) \\ S(\theta; \mathbf{y}) = \dfrac{\partial}{\partial \theta} \ln f(\mathbf{y}; \theta) = \dfrac{\partial}{\partial \theta} L(\theta, \mathbf{y}) & (\text{score function}) \end{array}} \quad \begin{array}{l} (3) \\ (4) \\ (5) \end{array}$$

where $\mathbf{y}$ is the set of observations and θ is the parameter to be estimated.

And therefore Fisher looked at, the variance of thisSin thespace of y. What is variance? Here with respect to data records.Because theta is fixed. The only-- here when you are evaluating as initially it is only for a single-- conditioned on the data. Suppose I change the date record, yourSwill change and so on. So how doesSvary across the date records? How does your score vary across the records, and that variance Fisher called asinformation.

(Refer Slide Time: 6:00)



Fisher's Information and Properties of Estimators References

Fisher information ...contd.

FI measures the variability in sensitivity of likelihood, i.e., the score function, across the outcome space (of $\mathbf{y}$).

The **Fisher information** of a parameter θ in $\mathbf{y}$ is defined as

$$\boxed{I(\theta) = \text{var}(S) = E\left(\left(\frac{\partial L}{\partial \theta}\right)^2\right)} \qquad (6)$$

Under the regularity assumption, it can be shown that

$$\mu_S = E(S|\theta) = 0, \qquad \text{var}(S|\theta) = E(S^2) = E\left(\left(\frac{\partial L(\mathbf{y},\theta)}{\partial \theta}\right)^2\right) \qquad (7)$$

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 11

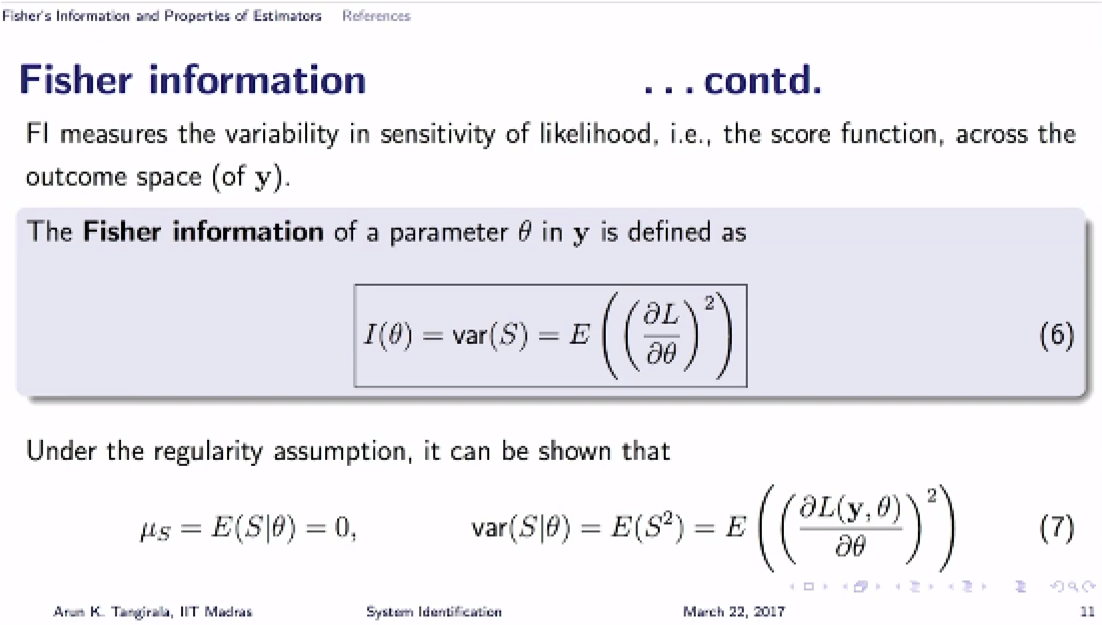
It's a bit counterintuitive. We may expect that the sensitivity should not change with the realization. Right? We may expect that, yeah, you know, whatever sensitivity I have for one realization should be the same, as for other data record and so on. But, it's a counterintuitive. What Fisher's information says is, more the variability of this score across realizations, more information you have. In other words, if you're looking at a deterministic process, for a deterministic process, there is no notion of information. Whatever data record that you have is what you have. Whereas with the stochastic process, you have no different realizations that are possible and you're looking at how this score actually changes.

Now, eventually our intuition works its way backward, eventually. What do youmean byeventually? When I want to answer the question, what is going to be the variance in theta hat sees if this score is actually changing from realisation to realisation, doesn't it tell you that thetahat also is going to change from realization to realization. Correct? So which mean, higher this i, so it's going to actually work its way back backwards.The higher the variability, more the variability in theta hat.Higher this variance of s, more the variability in your theta hat.Clearly right?

If I give you one data record, I have one s, I give you another data recorder I have another s. And if it's too much variability theta hat also is going to change drastically, because after all what are you going to do, in order to find the maximum, what are you going to do, here? You want to take the derivative and set it to zero. And the solution of that will determine your theta hat. So if that derivative which is a score is changing from record to record, theta hat is also going to change from record to record. Which means what, higher the variability of this court, more is going to be the. variability in my theta hat. Later on, we will study something called Cramér–Rao inequality, or Cramér–Raobound, which is one of the milestone results in estimation theory, which will come back and tell us that higher this variability, higher this variability actually what happens is, you will in fact,there also you'll going to have now, counterintuitive, there in fact what you are going to see is, it says that the bound is inverse of I of theta. So I'm sorry. Here also it is counterintuitive.

What happens is the bound on your theta hat, is dependent on this information. So, earlier I said that we'll come back and it fixes with our intuition, here also you get a counterintuitive result. Through the Cramér–Rao inequality, which states that higher the information,lower will be the variability in theta hat. So it's going to be a bit counterintuitive here, but we'll understand that a bit later. For now, we'll remember Fisher's information as variance of this score that we are looking at. Okay? Now let's look at the mathematical part of it will come back and talk about intuitive nature of this.

(Refer Slide Time: 9:44)



Fisher's Information and Properties of Estimators   References

Fisher information ...contd.

FI measures the variability in sensitivity of likelihood, i.e., the score function, across the outcome space (of $\mathbf{y}$).

The **Fisher information** of a parameter θ in $\mathbf{y}$ is defined as

$$\boxed{I(\theta) = \text{var}(S) = E\left(\left(\frac{\partial L}{\partial \theta}\right)^2\right)} \qquad (6)$$

Under the regularity assumption, it can be shown that

$$\mu_S = E(S|\theta) = 0, \qquad \text{var}(S|\theta) = E(S^2) = E\left(\left(\frac{\partial L(\mathbf{y},\theta)}{\partial \theta}\right)^2\right) \qquad (7)$$

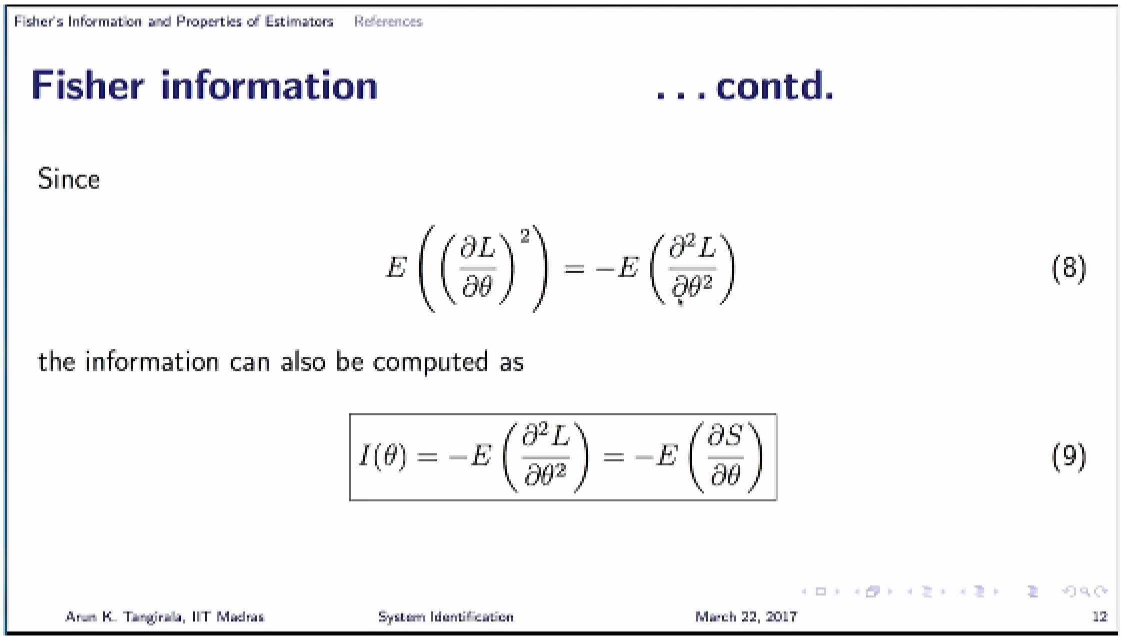
Arun K. Tangirala, IIT Madras   System Identification   March 22, 2017   11

The score is nothing but the first derivative of the likelihood. Therefore it is [9:54 inaudible] L by [9:54 inaudible] theta. Now, yourSis no I now a random variable. What is the randomness-- S due to the randomness in y.

So what is the variance of any random variable? Expectation of x square minus mu square. You can show that expectation of score is zero. Okay? So you can actually show that, expectation of score is zero. As a result, variance ofSwhich is expectation of S square minus Mu square mu S and as a result it is expectation of [10:37 inaudible] L by [10:37 inaudible] theta, to the whole square. Am avoiding the proof that the expectation of score to zero. But you can look up any textbook. As a result you get this expressioninequation six. Now, there is a nicer way of evaluating this variance of S, which is-- so instead of taking the expectation of [11:01 inaudible] L by [11:01 inaudible] theta to the whole square, we can actually take negative expectation of the second derivative of L. Okay?

There is a difference between squaring the first derivative and taking the second derivative. Again I'm avoiding the proof here. This proof, this identity that you see in equation eight comes by virtue of the PDF assumption. That is a PDF is regular. I'm avoiding the proof here. So here you have expectation of [11:33 inaudible] L by [11:33 inaudible] theta to whole square as a negative expectation of [11:38 inaudible] square L by [11:38 inaudible] theta square. This is expression that we use in practice.

(Refer Slide Time: 11:38)



Fisher's Information and Properties of Estimators References

Fisher information ...contd.

Since

$$E\left(\left(\frac{\partial L}{\partial \theta}\right)^2\right) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) \qquad (8)$$

the information can also be computed as

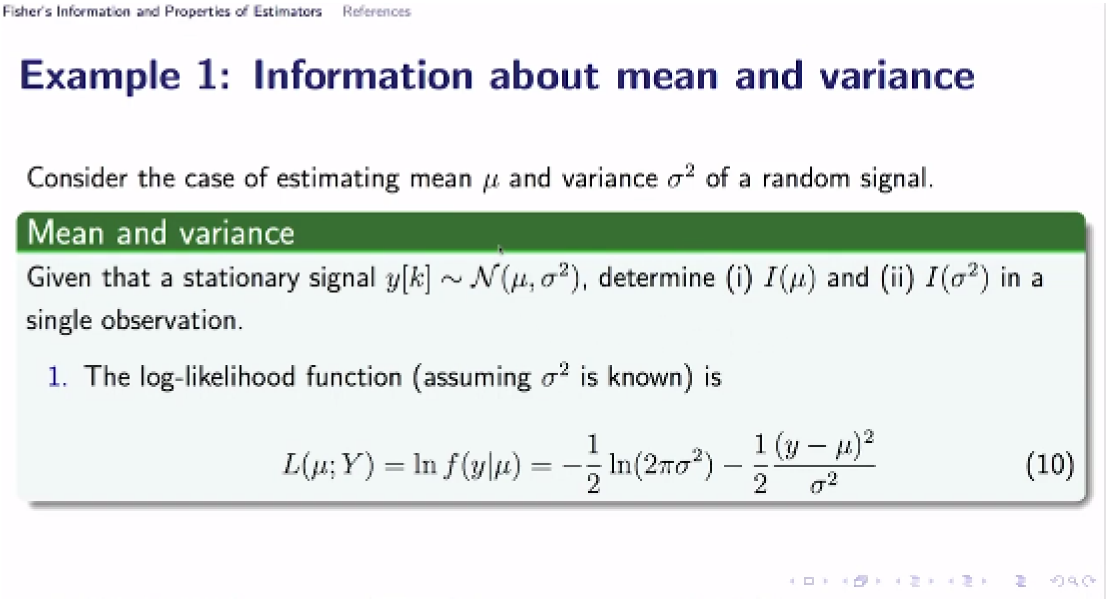
$$\boxed{I(\theta) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E\left(\frac{\partial S}{\partial \theta}\right)} \qquad (9)$$

Arun K. Tangirala, IIT Madras System Identification March 22, 2017 12

Let's quickly understand. Look at an example to understand Fisher's information is. So here we have a very simple example, where I have a single observation. Only one observation, from a Gaussian white noise process and I want to know, how much information is contained in the single observation with respect to mean? Right? You wouldn't expect much information, right? Single observations, what can it give you? Remember in all of this, although we explicitly say that I'm not bothered about how you estimate, but implicitly Fisher's assumes that you're going to use MLE. That is how the estimator, you can say is not really the focus, but Fisher has fixed the estimator to be MLE. And it's okay, because MLE has good properties at least for large samples. So let's look at this information contained in a single observation.

I'm already given that the single observation is falling out of a Gaussian process. So which means the PDF is fixed, correct? And therefore, what is the first step in Fisher's information setting up the likelihood? The PDF is fixed, therefore the likelihood is fixed. Right? So what is it?Gaussian PDF, what is the expression for Gaussian PDF?

(Refer Slide Time: 13:16)



Example 1: Information about mean and variance

Consider the case of estimating mean μ and variance σ^2 of a random signal.

Mean and variance

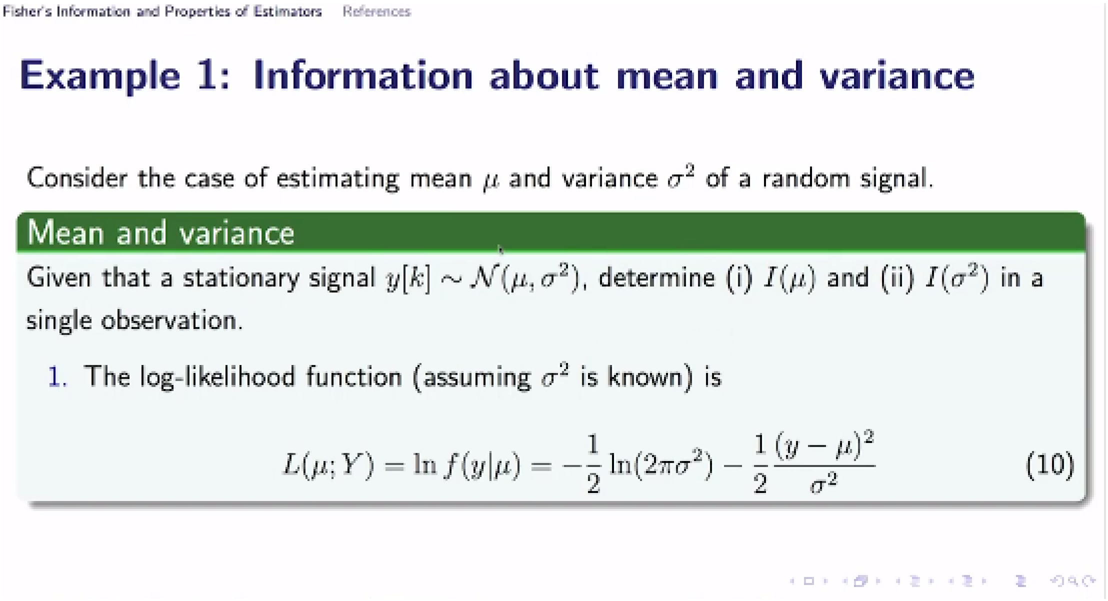
Given that a stationary signal $y[k] \sim \mathcal{N}(\mu, \sigma^2)$, determine (i) $I(\mu)$ and (ii) $I(\sigma^2)$ in a single observation.

1. The log-likelihood function (assuming σ^2 is known) is

$$L(\mu; Y) = \ln f(y|\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} \qquad (10)$$

I have f of y[k] given my and sigma square, what is it? One over sigma root 2 pi to the whole square. Okay? This we know very well. This is nothing but our likelihood [13:47 inaudible] likelihood is a function of theta and given y, that's all. But the mathematical form doesn't change. Remember we work with log-likelihood, correct? So the log-likelihood is given here. I've written here as minus half log two pi sigma square and then minus of course, there should be a y[k] here but I'm given a single observation so I'veomitted k. So what is the question that we haven't had? Given sigma square, how much information do I have in a single observation? So what is the next step?From here what is the next step, your just one step away or rather two steps away, depending on how you look at it.

(Refer Slide Time: 14:32)



Fisher's Information and Properties of Estimators References

Example 1: Information about mean and variance

Consider the case of estimating mean μ and variance σ^2 of a random signal.

Mean and variance

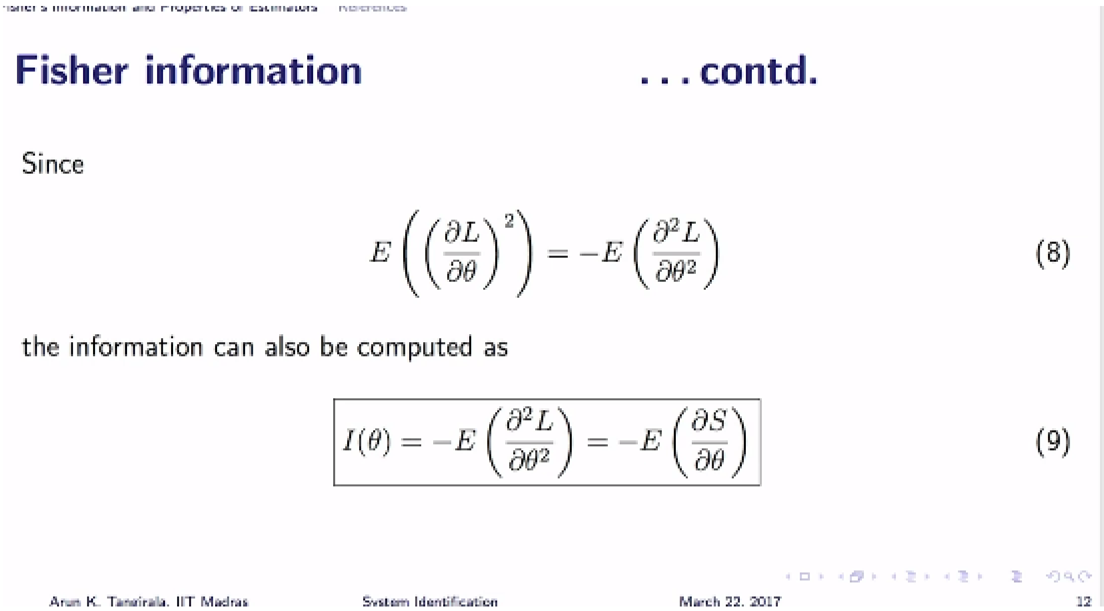
Given that a stationary signal $y[k] \sim \mathcal{N}(\mu, \sigma^2)$, determine (i) $I(\mu)$ and (ii) $I(\sigma^2)$ in a single observation.

1. The log-likelihood function (assuming σ^2 is known) is

$$L(\mu; Y) = \ln f(y|\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} \qquad (10)$$

We look at equation eight, right? Because variance of-- this is information for us. Variance of S. And Variance of S, is this. Right?

(Refer Slide Time: 14:51)



Fisher information ...contd.

Since

$$E\left(\left(\frac{\partial L}{\partial \theta}\right)^2\right) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) \tag{8}$$

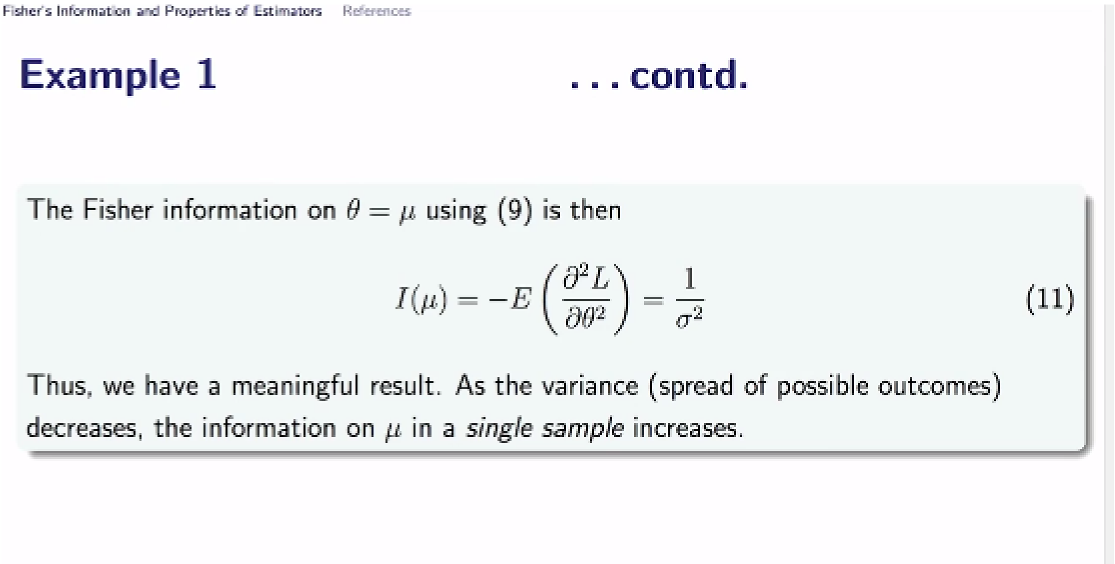
the information can also be computed as

$$\boxed{I(\theta) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E\left(\frac{\partial S}{\partial \theta}\right)} \tag{9}$$

So what do I have to do now? Second derivative of L with respect to-- what is the theta now for us? Mu alone. So what is the second derivativeof this log-likelihood with respect to mu? You can look at the PDF and straight away give me the answer. Or you can look at log-likelihood andgive me the answer, second derivate with respect to mu alone. What is it? So long. You have the answer. [16:12 inaudible] very good. Does anyone else get this answer? One by sigma square. Got it. So we have the answer there for the fish--

Now of course you have take the expectation of that but, now one over sigma square is a deterministic thing. In fact they should essentially get minus 1 over sigma square as the derivative.

(Refer Slide Time: 16:37)



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.

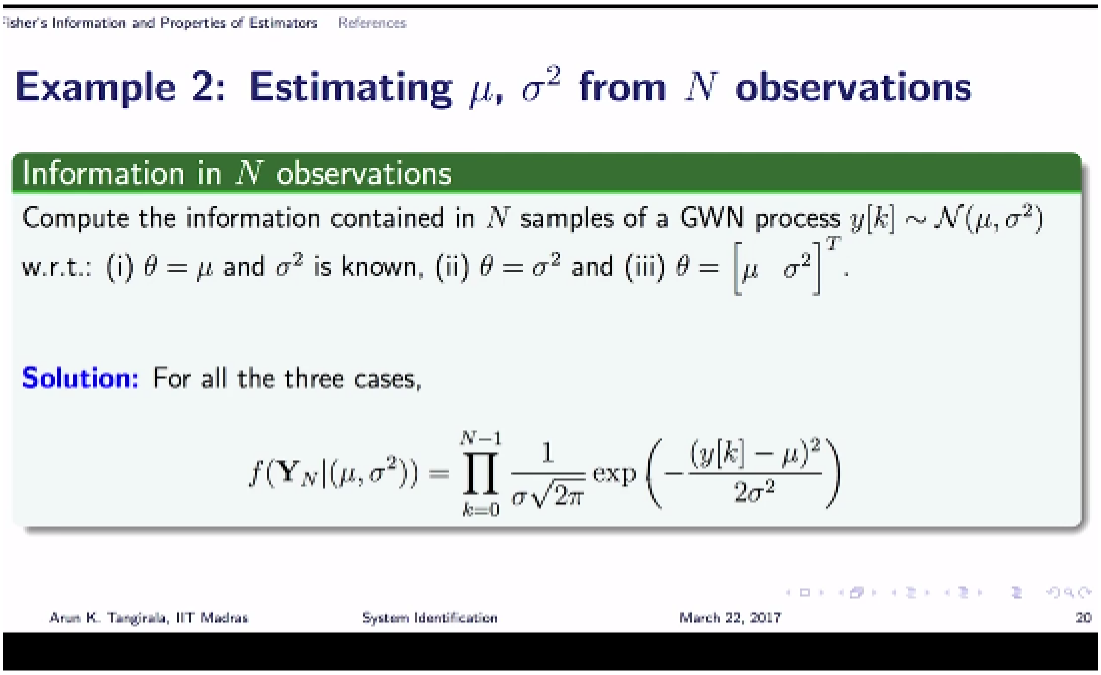
And then you take that negative expectation of that, you'll get 1 over sigma square as theFisher's information. What does it tell us? The information contained in a single observation is inversely proportional to the variability. Now that's interesting. Which means as a sigma square goes down, what is sigma square a measure of? It's a measure of the variability in the process. As sigma square goes down, let's say goes to zero, what kind of a process do we have? Deterministic process. So for a deterministic process, a single observation has a lot of information in them. But for a stochastic process more the variability in your data, correct? That is why now you understand why Fisher's information has been defined this way. More the variability in your process, lower is the information contained in a single observation. Why is this and why does this make sense?

Because if have more variability, then the spread is very large. So I could have attained any value. That means the possible values are very much higher in terms of range. Therefore a single observation may not offer much information with respect to mean. Actually you must ask how do you estimate mean with a single observation? Well, it is that observation itself intuitively. Right? Now, what do you expect if I have n observations? Does the information improve? Yes or no? Instead of giving you a single observation, I'll give you n observations. [18:25 inaudible] collect from 50, 100 whatever. Now I'll ask this question. Is there more information about mu? Potentially do you have or no? What do you think? Intuitively yes, right. But does it increase proportionally, that means if I have one observation it's one over sigma square. If I have n observations will it be n over sigma square?. Do you expect that to happen?Yes or no? Or you can say, it depends. If you say it depends, I'll ask depends on what? What do you think?

Anyone. Do you expect it to improve? Proportionally? We all agree that more the data points, more the information I have about mu. I can estimate better, that's what it means. But in terms of information-- Fisher's information does it improve proportionally increase proportionally by n?.What do you think? Intuitively, what do you feel? Or maybe n by 2.See I have one observation and I have n observations. One observation has 1 over sigma square. What is the nature of these n observations? They're uncorrelated. That means each observation is bringing in some new information that the other observation doesn't have. Correct? If there is correlation, there is some overlap of information. This is a Gaussian white noise process or white noise process. Therefore every observation is bringing me some new information and therefore I should expect the information to grow. So when I have-- we'll come back to sigma square shortly but when I have n observations then let us see what more do I get here.

Well, when I have n observations now, first thing is setting up the PDF becomes important. Have skipped the other part of the example, we'll come back to that later but this is interesting, so we'll address this first. Then I have n observations settingup the PDF is not so easy, because I have to set up a joint PDF. However, the good news is that it's a Gaussian white process. Therefore the process is also independent. It's not necessarily true for all white noise processes,we have already learned. If two random variables are jointly Gaussian and they are uncorrelated they're also independent.So straightaway I can write a join PDF of n observations for this example, as the product of marginal PDFs. And at the joint PDFis Gaussian, marginal is also going to beGaussian. And let marginal is given by here. Okay. Sorry, it should be y. That marginal here, should be is Gaussianitself. So I'm going to simply multiply all of them,that's my likelihood. And then I take the logarithm, right? So I have a log-likelihood. Now, if you look at it here, I have the f and then I take the log-likelihood, of course, I'm not showing the log-likelihood here and then I'm taking the derivative. So when you take the log-likelihood here the exponential vanishes, right?

(Refer Slide Time: 21:55)



isher'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 you will get instead logarithm of, sum of, in fact, sorry, the exponential vanishes and you'll get to sum of this inner bracket here, apart from the logarithm of this multiplication factor. And once you take the derivative, you're left with this. But then you have to take the second derivative right? So when you take the second derivative and then take the expectation, you are left with now the information that you have, negative expectation of the second derivative of log-likelihood, gives from the information and that answer works out to me n by sigma square, clearly telling me that the information content has actually improved by n times. On the other hand if it was a correlated process, it wouldn't improve my n times. It would be, it will actually fall down, maybe n minus 1, n minus 2 or whatever depends on the nature of the correlation. If it is highly correlated, then this proportionality constant n, I mean, will come down. We'll not worry about that, but at least this is a very interesting fact to know and Cramér–Rao's bound tells me which we'll learn later on that the--for the scalar quantity sigma square theta hat is less than or-- greater than or equal to this. Which means, what this theorem says is, for any unbiased estimate of course there is a limit to which you can obtain on the precision of the estimate. Sigma square theta hat is the measure of the precision. And that limit or bound is given by inverse of information. More the information, lower is the bound which means you can obtain more and more precise estimates. Which estimator will give you a sigma square theta, this lower bound we do not know. But that then there is another part of theorem we'll worry about that later on. But what you can see straight away is therefore, the lowest bound that you can get for mu hat according to Cramér–Rao's inequality and the result that we have for a Gaussian white noise process and from n observations is sigma square by n. Which is what you will see in many basic statistics textbook but this is the lower bound, I'm writing here equal to, but which estimator achieves this lower bound we not know. Right. So we'll just conclude now by understanding that the Fisher's information gives me a quantitative idea of how much information is present with respect to a parameter or a vector of parameters. We have just gone through example of a single parameter, but when we come back we will actually study how to extend this idea to a vector of parameters. And then once you have understood all of that we'll more on to bias and variance and so on. So the Fisher's information is an extremely important metric in system identification, particularly in input design. You want to design an input such that your data is informative, remember we have said that earlier. Very often this metric is used to quantify information in the data. So you'll find an input such that the in Fisher's information is maximized. And that's how the input design problems-- a lot of input design problems are formulated in identification literature. We may not go in that direction but I just want to tell you what the connection between Fisher's information and SysID is. Okay, so we'll conclude the class here.