CH5230: System Identification

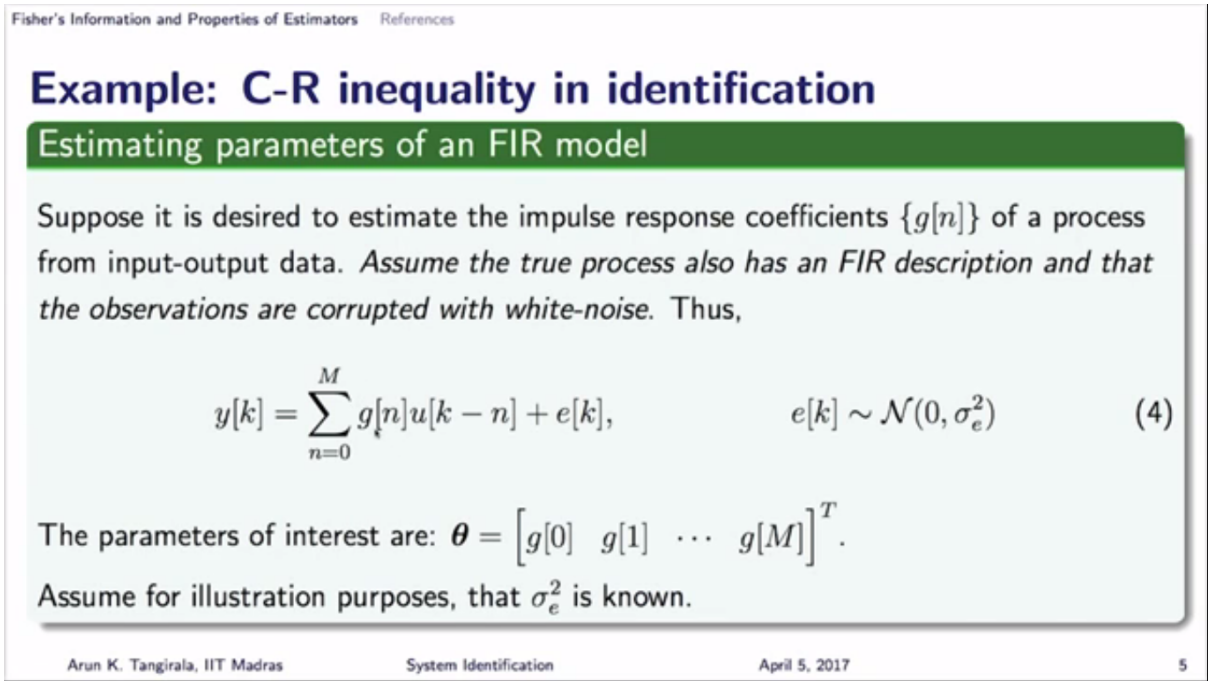
Fisher's information and properties of estimators

Part 06

So, let's look at this problem of estimating parameters of an FIR model, right. A very relevant problem in system identification, I am given an FIR model. I assume that the data is generated in particular way. Remember, what is the Cramér–Rao's inequality say? "The minimum variance that you achieve

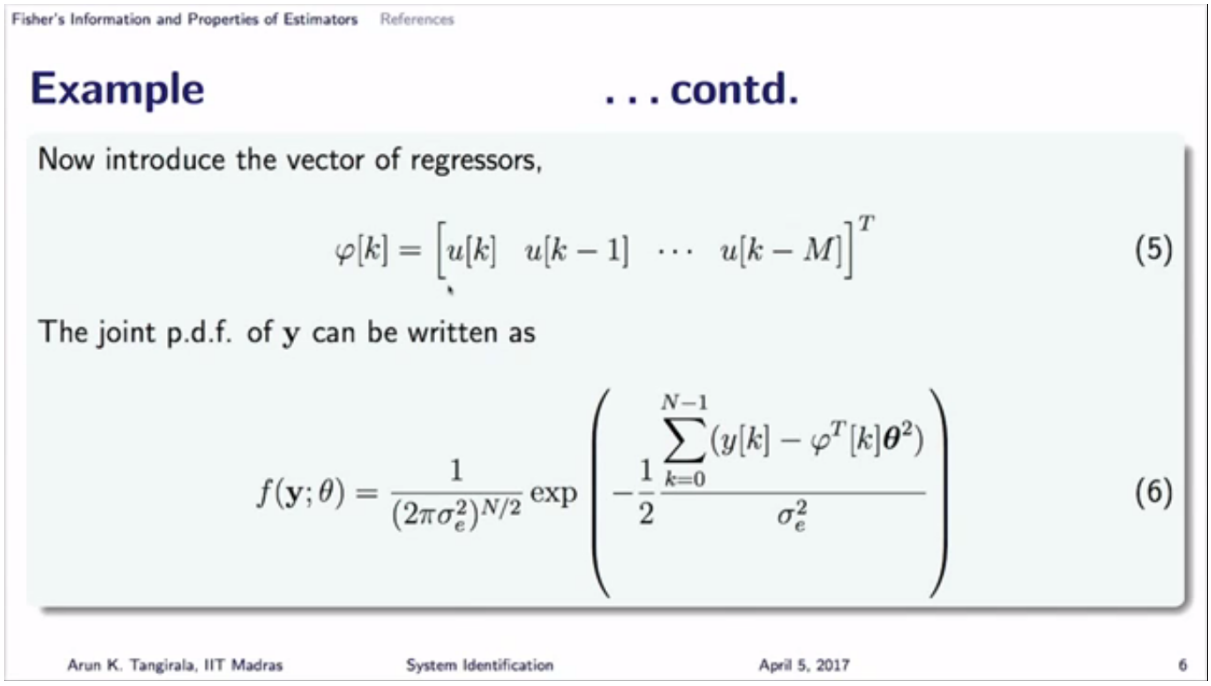
is an inverse of the Fisher's information." Okay. Which means now to calculate what is the minimum variance FIR estimator, I have to first calculate the Fisher's information. So the problem setting is as follows I am given data and given N observations. I assume the data to be generated in this way. Okay. That means for the moment I don't assume a model plant mismatch. I assume that whatever model I have is the same model that the data has, that the process has, e(k) is the white-noise, standard white-noise. Now I want to ask what I want is an estimated that gives me minimum variance estimates off the impulsive response coefficients. Given what? Given input output data, N observations of input output data. Further, to simplify matters we'll of course, assume input to be deterministic and assume for now that sigma square is known. So how many parameters do we have here? We have m plus 1, parameters are unknowns and that's your theta vector.

(Refer Slide Time: 01:52)



So, now first introduce this vector of regressors, so that it becomes easy to write the likelihood. So we introduce this regressor vector you should slowly get used to because then we go to least squares you will be needing this.

(Refer Slide Time: 02:07)



So that means we want to write this equation in the standard linear regression form. Theta we have already define it to be this. And now this is the vector of regressors, so that I can write this equation here for the data generating process as y equals side transpose k times theta plus e(k), right. Now what do I need to do, I want to find out the minimum variance estimator, which means, I need to find the Fisher's information matrix. What do I need to do to calculate the Fisher's information matrix? I need to first set up the likelihood, right. So I will set up the joint likelihood of N observations I'm not given a single observation here. From a single observation is not possible. So, because we have m plus 1 parameters. We'll assume we have sufficient number of observations. Now, what do we observe here, y(k) is, so we have y as side transpose k theta plus e(k). It's what we have. And what do I have to do. I have to set up the joint likelihood of these N observations, right. Now, assume u to be deterministic, all right. What is the expectation of y? See I need to first set up the joint PDF. Can I straightaway say that the joint PDF is a product of the PDS or I have to check something? How do I proceed from here? From here to the likelihood how do I proceed? This is a likelihood for all practical purposes. It's a joint PDF that I'm interested in. What do I do next? How do I write the joint PDF? You've done this before. Yesterday's example we have written, right. How do you write the joint PDF? Any ideas? You should be comfortable with writing the likelihood. So if you have any difficulties you should raise your hand. What is the difficulty? If you don't have any difficulty then you should come to the board and write it. What is the difficulty? You have to tell me what the difficulties? You don't have any difficulty. Then I will ask you to write on the board. Then what is the difficulty? What is the difficulty spell it out? You don't have any difficulty?

No, I don't know.

You don't know. Then what is a difficulty why you don't know? How do you write the joint PDF? What do you know? No, seriously what do you know in this problem? You don't know anything in this problem. Absolutely, you don't know anything about y (k) anything. Seriously you don't know anything I've written what is y(k), right. I'm telling you how the observations are being generated. Forget about everything. I'm giving you how the observations are being generated. I'm asking to write

the joint PDF. What's a difficulty? What happened? What is a joint PDF? How do you write the joint PDF?

Product of marginal PDFs.

Product of marginal PDFs. But is it always the product of marginal PDFs.

If they are independent.

If they are independent. Good. At least, you know when life is simpler, right. At least you should know that. You simply say I don't know. Life is simpler. Life is simpler. At least you should know when my life becomes simpler. Is that the case here? Is that the case of simple life here? Are your observations independent? Given this situation. Assume for simplicity that the input is white. It's deterministic. Let's assume input is deterministic, it's a known quantity. How do you check if y(k) is are independent?

Correlation.

Correlation. But correlation only gives you linear dependency, right. It can tell you whether linearly independent of. What else do you need to check if the observations are independent?

[07:25 inaudible]

Yeah. So other joint Gaussian, if you have two pairs of I mean if you have a pair of observations any you take any pair of observations, if they are Gaussian, jointly Gaussian and if they are uncorrelated, then you should expect y is to be independent. So now you have to do that. So tell me quickly. First if observations are uncorrelated. What do you think? Are they uncorrelated or not? What do you think?

[08:19 inaudible]

Yeah, even otherwise, even if input is not right you can still check.

How do you check?

For covariance correlation. See for everything you have to go back to the basics until you get used to it. And until you can say by looking in the equation yeah, there are uncorrelated. So go back to the definition of covariance and ask what is a covariance between y(k) and y(k) minus l, for example. What do you think? What do you think? They are correlated or not? So you have covariance y(k), y(k) minus l, which is expectation of y(k) minus Mu k times y(k) minus l. See I am only writing the definitions which you know. I'm not introducing any new definition. This is not a new definition. Now, what do you think? Now you should be able to answer. What is Mu k? What is the expectation of y(k) man?

What is that? Side transpose theta. So what do you think now are they uncorrelated? What is y(k) minus Mu k? Correct. y(k) minus Mu k is e(k). So what is the difficulty? The difficulties in your brain in your mind. You think, Oh my God this is a beast. I cannot handle it. No. If you think it's a beast you actually start okay, these are its hands, these are its feet, this is the mouth. You can actually, slowly dissect and do an anatomy and you will be okay. So don't get intimidated. So the decision is that y(k) and y(k) minus l are uncorrelated. Okay. So we have verified that observations are going to be uncorrelated. The difficulty would have been if this data generating process has was different in the sense that the instead of e(k) if I had a coloured noise then, we would have felt the heat. That's why to

keep things simple, I'm assuming that the data is generated by as an output error kind of model, right. Okay. What about the second part? We are verified that the observations are uncorrelated. What about the second part? Gaussian, jointly Gaussian. We assume that e(k), each e(k) is coming out of a Gaussian process. So what would be the joint PDF of any pair of observations? It also be Gaussian, right. They are uncorrelated individually Gaussian. So the joint also has to be Gaussian. Now I'm assured that yks are all independent. Therefore the joint PDF is equal to the product of marginal PDFs. Is that clear? So now I have verified that my life is actually corresponding to a simple scenario. So what is the marginal PDF of y(k)?How do you write the marginal PDF of y(k)? What is a mean and what is its variance? See what is a form of the PDF of y(k)? Is it Gaussian uniform pause on what is it?

Gaussian.

Gaussian, then no doubt about it. Now, that means I need to find out what the mean of y(k) is variance of y(k). What does it mean of y(k)?

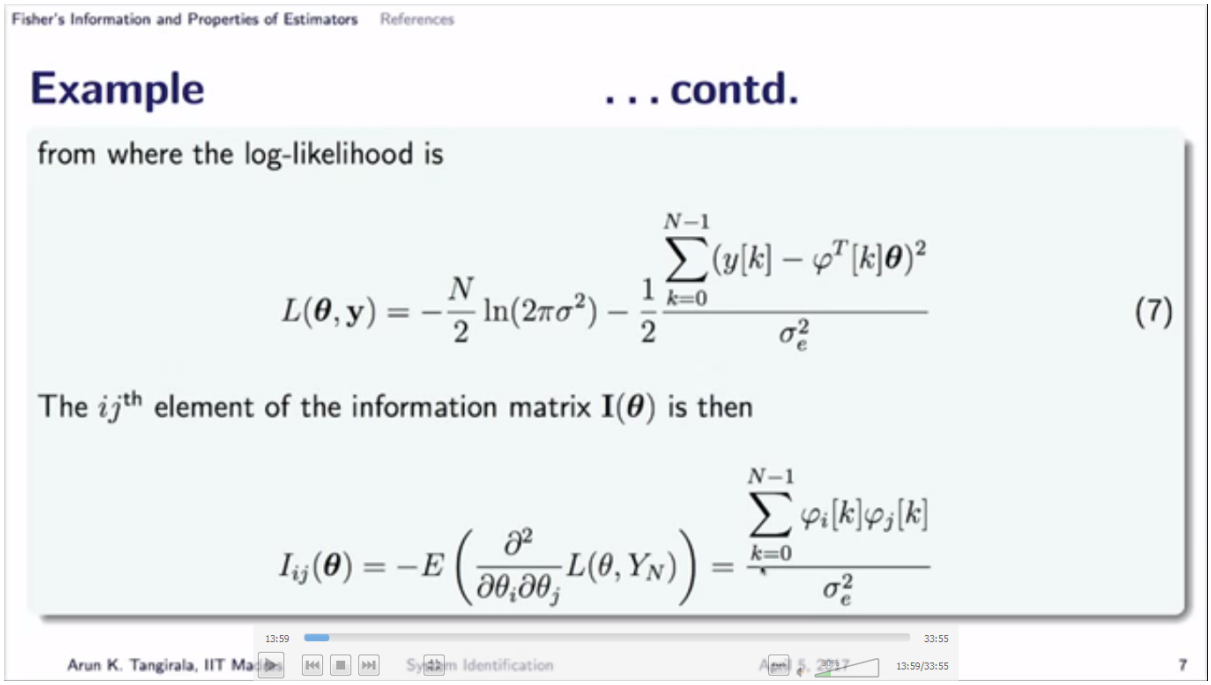
Side transpose theta.

Very good. Side transpose theta. Correct. What about variance?

Sigma square e.

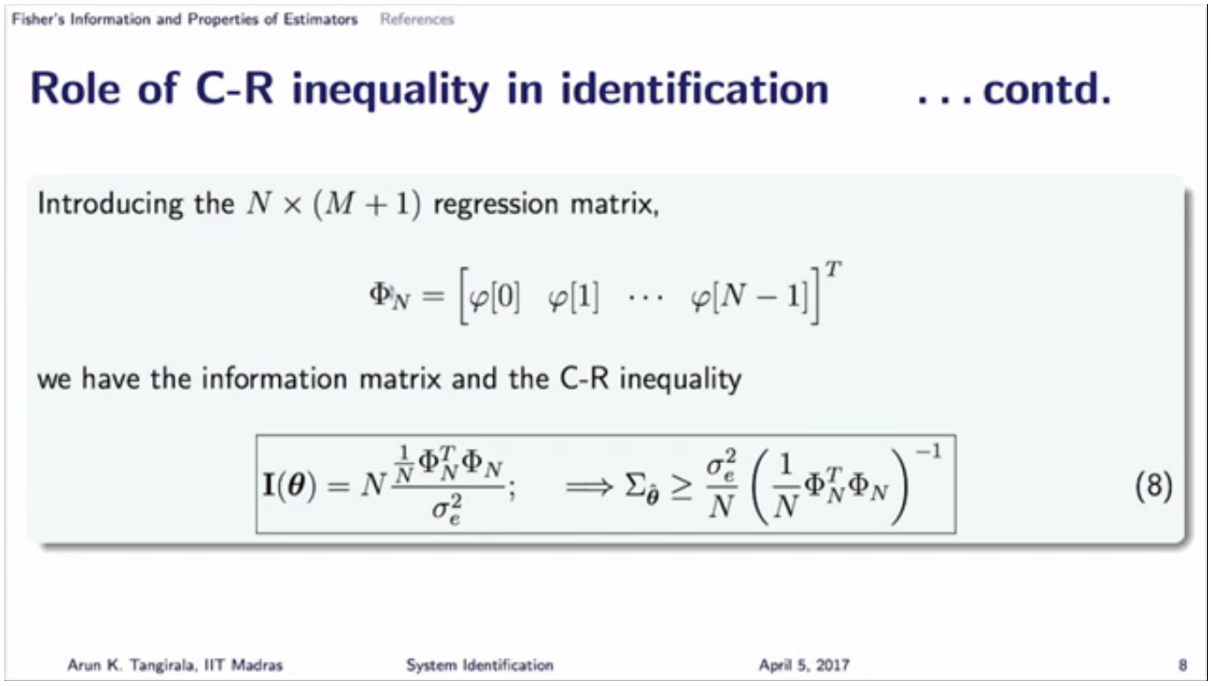
Sigma square e, because variance is about that's it. So therefore I straightaway know write the joint PDF. That's all. If you can, all you have to do is in your mind take things step wise and argue logically at each step and set up the final problem. That's it. So now, I have the PDF or the likelihood and then from here on out how do I find the information matrix? Take the second derivative, right. And then what do I do? Negative expectation. Remember, negative expectation of second derivative will get me that. When I do that, of course, I had to take the logarithm and then work out the thing because I'll be looking at log-likelihood. When I do that I end up with this term here. Okay. When I take the second derivative expectation take the negative of that this is what I end up with. Now the nice thing is here, I can write this. So what do I have here? Sigma k equals 0 to, n minus 1, Psi i(k) times psi j(k). Remember I have Fisher's information matrix. Correct. It's no longer a scalar. I'm looking at a vector of parameters. This is the I, jth element of that matrix.

(Refer Slide Time: 14:01)



When I put the entire matrix together, I get this result. By first defining a phi matrix, so this is where some algebra is involved nothing you just have to go back and verify. If you have understood how to setup the likelihood beyond that it's all algebra. So the point is now like we define the regressor vector earlier. Now I'm defining a regressor matrix. Okay. What is this regressor matrix stacked regressor vectors. From 0to N minus 1, the only thing you have to worry about in practice which I'll tell you in Psi later on is, that strictly speaking you cannot construct Psi of 0, why is that. Can I construct Psi 0 in practice? Look at what is Psik. So I need inputs at negative times. So in practice I will not be able to stack, I will not be able to construct this regressor matrix like this, it will be of lower dimensions. But that we'll worry about later on. For the moment, let us say I have the input at negative types, don't worry about that. So this is called a regressor matrix we should get use to this, because in least squares we'll construct this matrixes quite often.

(Refer Slide Time: 15:23)



Now, I can rewrite this result. This is for the I, jth element of the matrix, information matrix. This is the information matrix itself. I have deliberately multiplied with 1 over N and N, but the actual result is that the Fisher's information matrix is Phi transpose Phi by sigma square e. You should verify, what is the size of Phi? This Phi matrix?

What would be the size of Phi?

[15:58 inaudible]

What is the size of a Psi?

[16:02 inaudible]

Right. So you have N cross N plus 1 for Phi. What will be size of Phi transpose Phi? N plus 1 times N plus 1. Which is indeed the size of the Fisher's information matrix, correct? So what does this result straightaway tell me, the lower bound is sigma square e over N times 1 over N Phi transpose Phi inverse of that. It's a very fundamental result in parameter estimation. We're talking of FIR model parameter estimation. But some of the inferences that we draw from here apply to other parameter estimation problems as well. So what does this result tell me? The lower bound on sigma theta hat is dependent on what values? What are the factors? Sigma square e right. Now you may say I have artificially multiplied N here, but there is a reason why I have done that multiplied and divided by N. The reason is 1over N Phi transpose Phi, what is it? It is an estimate of the variance covariance matrix of your regressors. 1 over N Phi transpose Phi. If you were to expand, if you were to deflate Phi transpose Phi. What do you expect to see along the diagonals?

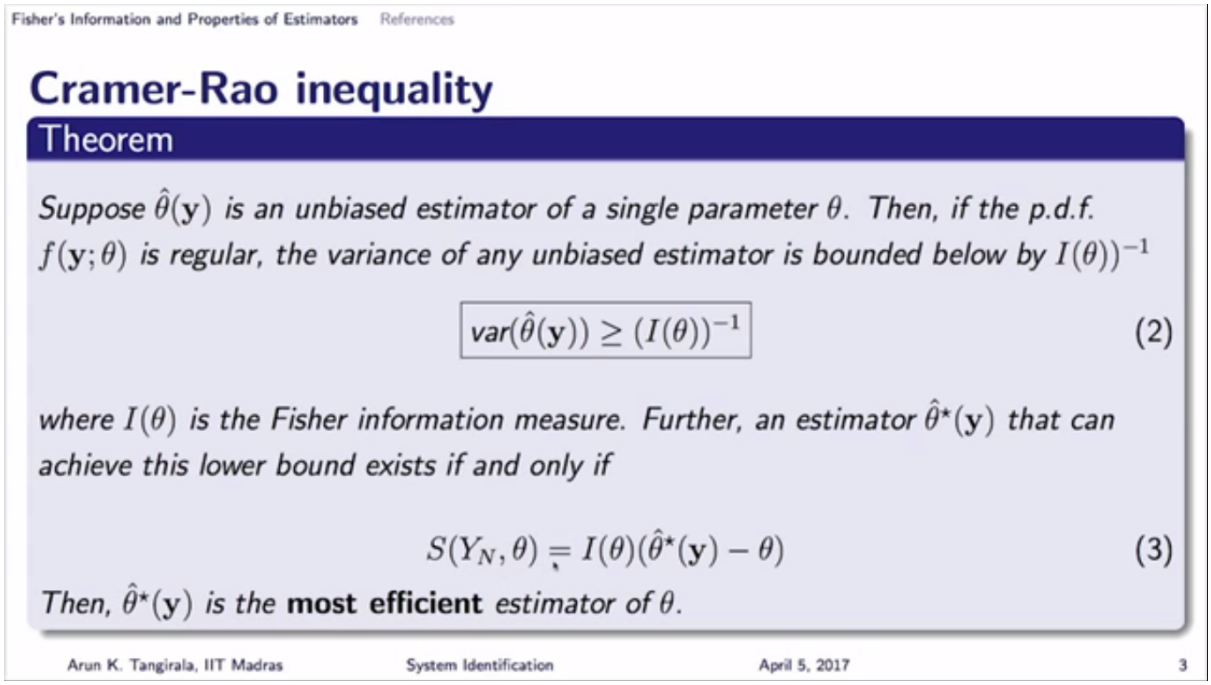
There's some square, some square inputs divided by N, we'll give you an estimate of the variance of the input. Okay. So 1 over N Phi transpose Phi is an estimate of the covariance matrix of your regressors. It gives you and if you think of a single regressor it tells you how much power is present in the input. Okay. Because if there is a single regressor, what will be Phi? Phi will be simply a vector. Phi transpose Phi would be a scalar. And 1 over N Phi transpose Phi will be nothing but the power in that regressor. So, what this tells me is that the lower bound on the variance depends on three things,

sigma square e. The power are in the variance in the regressor and the sample size. That means if I want to get low errors. That means now if I want to control this lower bound. What are the things that I can do? I don't have a say on sigma square e. That's beyond my control. There are two things that I can do. Increase the sample size. So that the lower bond goes down, lower the lower bound better the situation is, correct. So I can increase the sample size or increase the power in the regressor.

In fact you should see the signal to noise ratio coming here. sigma square e times 1 over N Phi transpose Phi inverse. What is it? It is the inverse of signal to noise ratio. What does this tell me if I maintain a high signal to noise ratio, then I can get better and better estimates, because the bound keeps going down. Or if I say no there is a limit on which I can maintain the signal I can achieve a high signal to noise ratio, because if I want to achieve high signal to noise ratio is what I, what do I have to do at least in this case. I have to maintain, I have to give high amplitude inputs. The inputs have to be very high amplitude, so that 1 over N Phi transpose Phi very large, but there is a limit to which I can give and amplitude of the input. What is the danger of giving high amplitude inputs? What can I do, end up doing?

I am billing linear models. So I can run into the risk of going into high severe non-linarites and for which this linear model may fail miserably. So I want to still live in the linear regime. I want higher SNR then there is a compromise. So the next controlling factor is a number of observations. So it says okay. For a fixed SNR is there any other thing that you can do? Yes, the number of observations. Which means by increasing the number of observations, I am able to achieve more and more efficient estimate. Now what is that efficient estimator? So there are two parts to Cramer-Rao inequality. One is the bound. So for the FIR model parameter estimation problem we have shown that this is the bound, right. Now I won't to ask the question. I'm a bit greedy now. I want an estimator. I want a formula. I want a method which will get me this lower bound, right. And what do I do for that. I have to go back to the Cramer-Rao inequality. What does inequality say? It says an estimator exists that will achieve this lower bound. Then that estimator should satisfy this condition. What condition? I of theta times theta hat star of y minus theta should be independent of theta. Okay. So, if and only if actually the statement is incomplete. If and only if the score function is independent of theta.

(Refer Slide Time: 15:23)



So let's look at that if it is possible. So let's take this I of theta here. Look at this here. Now statement is complete. The way you read this statement is. If I were to write that statement in a slightly different way, I of theta inverse times the score, score function plus theta. He's what is theta hat star of y. So if I take the left hand side and evaluate it, eventually it should work out to be independent of theta. It should be only a function of observations. Let's see if that is possible. If it is possible only then I'll be able to find an efficient estimator. It say that if there exist and efficient estimator it would satisfy this relation. The way to read this condition is evaluate this expression and see if it is independent of theta. Because the right hand side is only a function of observations. So let's quickly do that and plug in. We know already I of theta we have derived this. We know the score also, right. Because we know the likelihood. So let's put that together. The score function is this. When you just work out the algebra and write in terms of Phi and y. The score function turns out to be this. So write I inverse times S plus theta, you end up with this. Now the question is, if in this-- right hand side of the equation ton. Then is it independent of theta, what do you think? Is it independent? Or is it dependent on theta? Are the parameters appearing on the right hand side or only observations? Only observations. So which means, I have struck gold here. This is, what is this? In fact, this is the efficient estimator of theta that is your theta hat star of I.

(Refer Slide Time: 24:13)

MVU estimator of IR coefficients

First construct the score vector

$$\begin{aligned}\mathbf{S}(\boldsymbol{\theta},\mathbf{y}) &= \left[\frac{\partial L}{\partial \theta_1} \quad \cdots \quad \frac{\partial L}{\partial \theta_p}\right]^T \\ &= \left[\sum_{k=0}^{N-1}\varphi_1[k](y[k]-\varphi^T[k]\boldsymbol{\theta}) \quad \cdots \quad \sum_{k=0}^{N-1}\varphi_p[k](y[k]-\varphi^T[k]\boldsymbol{\theta})\right]^T \\ &= \Phi_N^T\mathbf{y} - (\Phi_N^T\Phi_N)\boldsymbol{\theta} \qquad (9)\end{aligned}$$

Using (8) and (9),

$$\mathbf{I}^{-1}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta},\mathbf{y}) + \boldsymbol{\theta} = (\Phi_N^T\Phi_N^T)^{-1}\Phi_N\mathbf{y} \qquad (10)$$

which satisfies the requirements of an efficient estimator of $\boldsymbol{\theta}$.

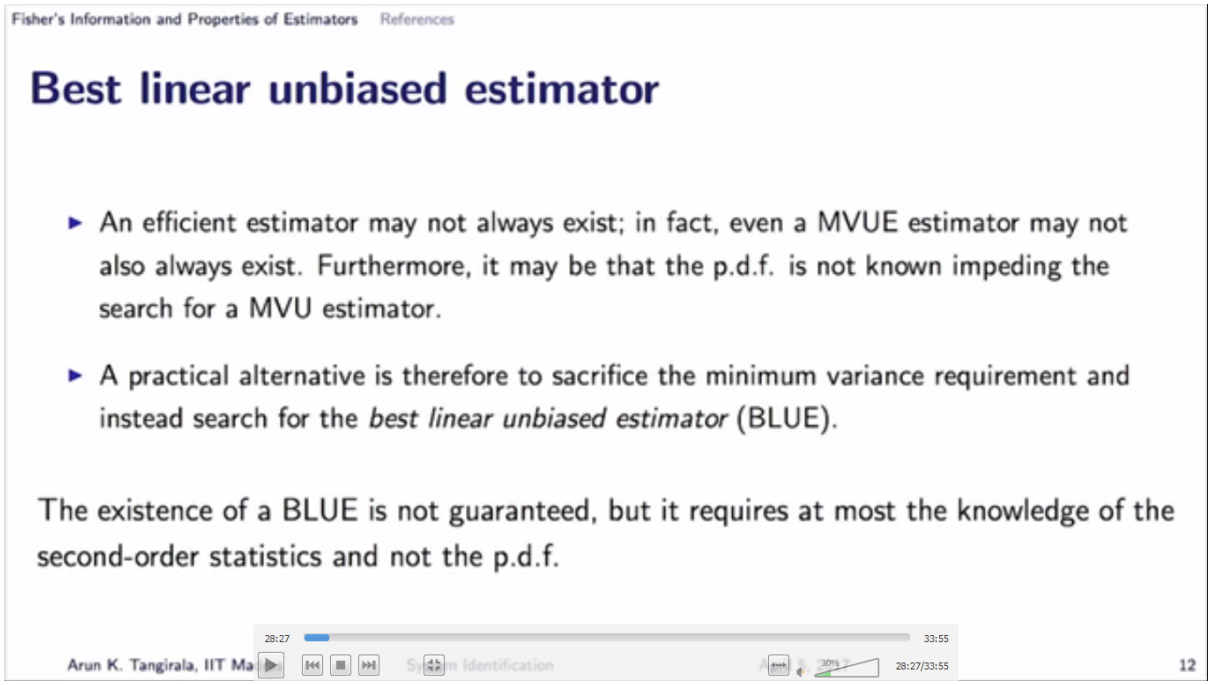
Therefore the most efficient estimator of the FIR model parameters is Phi transpose Phi inverse Phi transpose y. In fact this is exactly the least square solution that we learned about. How do we arrive at the solution? I didn't impo-- I didn't invoke any least squares are any such formula here, method. I said, I want the most efficient way of estimating the parameters. So for that first I determined the lower bound, step one and step two, workout this expression on what I've written on the board and see if it is independent of theta. If it is, then I found the most efficient estimator. So in this case we have managed to but what are the assumptions that we have made? What is a key assumption on which this result rests? Can you point out that key assumption? Under which we have derived this result.

Okay. Use deterministic that's okay. The observation error is white. If the observation error was not white what would have happened. The likelihood function would have been different, right. First of all independence is spoiled. So setting on the likelihood is going to be a challenge. We don't know what likelihood it will look like. It depends on the correlation in the observation error. v(k) we have assumed v(k) to be white. If it isn't then the course of the derivation everything will be changed. So when the observations are white, the errors are white. Then the efficient estimator of FIR model parameters is Phi transpose Phi inverse, Phi transpose y. But as I said earlier, when you use this result in practice you cannot set up Phi the way we have set it up. You have to start from after a few observations, right. So look at the beauty of this result. We have obtained the most efficient way of estimating FIR model parameters, under the assumption that the observation errors are white. If they are not, this is not an efficient estimator, necessarily, right. But it'll turn out that even in the least squares case, when we work out the least squares estimator, we will use. We will knock a different door. We will not knock the doors of efficiency. We use a different door there. We enter the world of estimation through a different door and after having derived the result, then we ask under what conditions it is efficient. You should see the difference between these two approaches. Here we have knocked the doors of efficiency right away. Right from step one, but when it comes to least squares MLE or BLC or any other method, you will not knock the doors of the efficiency or any such property. You would rather knock the doors of method. What principle you are applying and then arrive at the formula or the estimator and don't ask under what conditions I will get an efficient estimator. So there is a difference, here straightaway we are asking the question and we get the result.

And we also tells us under what conditions this estimator is efficient. In the least squares, we go in the circuitous route. Then also we will derive this result and then we realize that oh, it is efficient only when the errors are white.

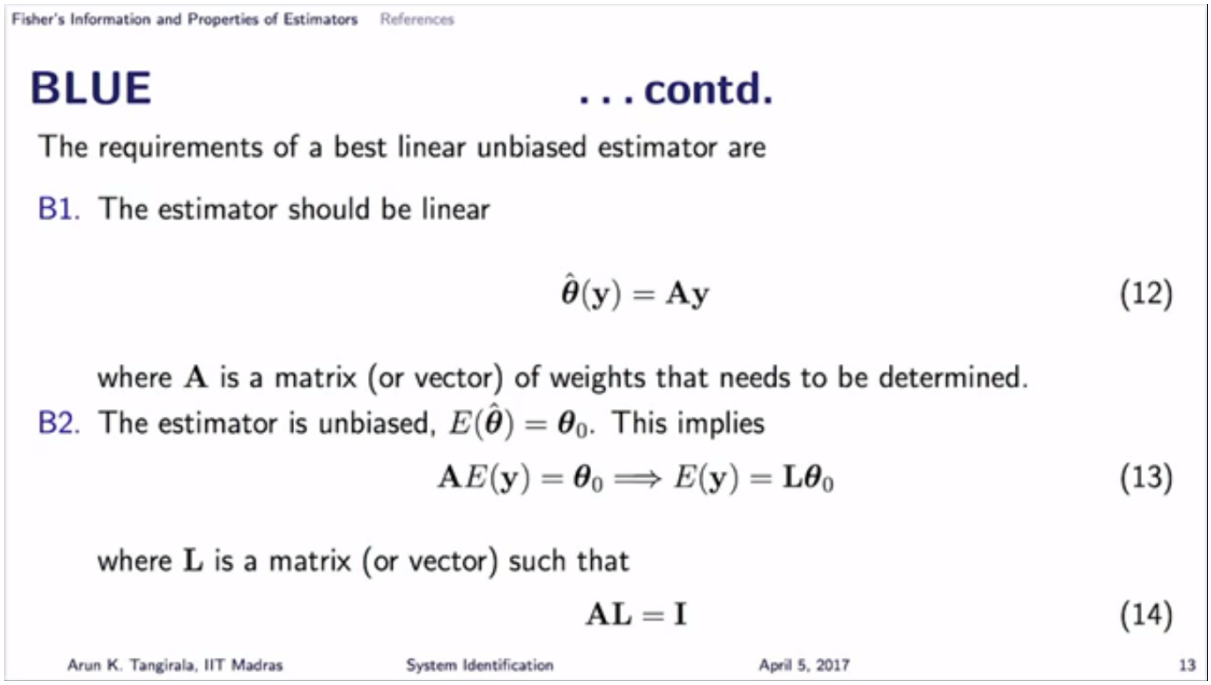
So that's it. There is a difference between these two, anyway. So very often I will just close the discussion with this notion of BLUE. Very often it may happen that you will not be able to find the minimum variance unbiased estimator. In this case we were able to find. So what do I do when I cannot find the minimum variance unbiased estimator. In this case we were able to find. So what do I do when I cannot find a minimum variance unbiased estimator. Means that it can turn out that this thing is not just purely a function of observations but also the parameter, which means an efficient estimate won't exist.

(Refer Slide Time: 28:31)



Then the search is always on for the best linear unbiased estimator called the BLUE. Best in what sense, what does linear tell you? That the estimator is a linear function of the observations, unbiased tells me that the estimators are unbiased. Best in what? Minimum variance again. So there are three conditions that we impose. One that the estimator should be linear, upfront itself, I say theta hat should be a linear function of the observations. Theta hat of y is Ay. All right. Where A is some matrix. Now I have to figure out what is that matrix? I am upfront imposing a form. And the second we want unbiased. That means expectation of theta hat should equal theta 0. And that translates to a condition like this. That is if you express expectation of y as L time's theta 0. Then you can straightaway translate this constraint of unbiasedness to a constraint on A.

(Refer Slide Time: 29:38)



And the third requirement is that it should have minimum variance. That is what we mean by best. Okay. So, what is the difference between minimum variance unbiased estimator and best linear unbiased estimator? What is the difference? In both cases I am searching for unbiased estimators. In both cases I'm searching for minimum variance. But the one that I'm requiring here is estimator should be linear. Whereas the minimum variance unbiased estimator need not be linear.

Earlier by the way do you think this estimator is linear or non-linear? How do you say that it's linear? Linear in what? In your observations. When they say observations it's y, Phi transpose Phi inverse Phi transpose. Is it actually independent of y? It's only consisting of inputs, right. So you can think of this as A Phi transpose Phi inverse times Phi transpose as A. So in this case you do get a linear estimator, but not always a minimum variance unbiased estimators are linear. So, essentially what we do is, we solve this optimization problem for BLUE. Minimize the trace of A transpose sigma y A, subject to AL equals Identity.

(Refer Slide Time: 31:09)

BLUE . . . contd.

With the third requirement, the optimization (estimation) problem is

$$\min_{\mathbf{A}} \text{trace}(E((\mathbf{Ay} - \boldsymbol{\theta}_0)(\mathbf{Ay} - \boldsymbol{\theta}_0)^T))$$

Using (13) and the properties of trace, we can re-write the optimization problem as

$$\min_{\mathbf{A}} \text{trace}((\mathbf{A}^T \Sigma_{\mathbf{y}} \mathbf{A}))$$
$$\text{s.t.} \quad \mathbf{AL} = \mathbf{I} \tag{16}$$

The solution to this optimization problem is

$$\boxed{\hat{\mathbf{A}}^\star = (\mathbf{L}^T \Sigma_y^{-1} \mathbf{L})^{-1} \mathbf{L}^T \Sigma_y^{-1} \Longrightarrow \hat{\boldsymbol{\theta}}_{\text{BLUE}} = (\mathbf{L}^T \Sigma_y^{-1} \mathbf{L})^{-1} \mathbf{L}^T \Sigma_y^{-1} \mathbf{y}} \tag{17}$$

And when you work out this problem for the FIR case and I will close the discussion today with this. For the FIR case expectation of y is Phi theta that means L is Phi. Compare with L. What is L? I write the expectation of y and it should be L times theta 0. So comparing notes here L is Phi, right? And then sigma y is sigma square E, we know that. Sigma y is your variance coherence matrix of N observations.

(Refer Slide Time: 31:43)

BLUE of FIR model

Consider the FIR model estimation that we studied earlier. Assume that only first- and second-order statistics of the noise are known.

$$E(\mathbf{e}) = 0 \implies E(\mathbf{y}) = \Phi\theta \quad (\mathbf{L} = \Phi)$$
$$\Sigma_e = \sigma_e^2 \mathbf{I}_{N\times N} \implies \Sigma_{\mathbf{y}} = \sigma_e^2 \mathbf{I}_{N\times N} \quad \text{(uncorrelated errors)}$$

Then, the best linear unbiased estimator of the FIR model parameters are

$$\hat{\boldsymbol{\theta}}_{\text{BLUE}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} \tag{18}$$

which is the MVUE estimator as well. Here, however a distribution of the noise was not required to be known.

So which means here in the solution for BLUE, what do I substitute L as Phi and sigma y as sigma square E times identity. When I do that what is the result that I get? L is Phi. So I get Phi transpose time sigma square E inverse times Phi inverse of that. Times Phi transpose again sigma square E

inverse times y. That is nothing but my estimator that I obtained earlier. In this case the BLUE and MVUE coincide. But in many cases they don't have to coincide. So the best linear unbiased estimator offers a compromise that means it will give you some estimator, which is best in some sense. Does it coincide with MVUE all the time? No. It coincides only when the observation errors are white and Gaussian. That is the famous Gauss–Markov theorem. That the BLUE and minimum variance unbiased estimator are identical, if the errors are Gaussian and White. In all other cases the BLUE is different from minimum variance unbiased estimator. All of this is to give you of different flavours of setting up your estimation problem. When we come back tomorrow I'm just going to finish up this discussion and then we'll move on to methods of estimation. As I said, please do watch the videos at least on the least squares. My TAs are supposed to send the links I will ask them to do this today. I don't know why they have not send the link to the videos, but please do watch the videos on least squares, because tomorrow will focus on least squares estimation, right. I'll show you a couple of examples in MATLAB.