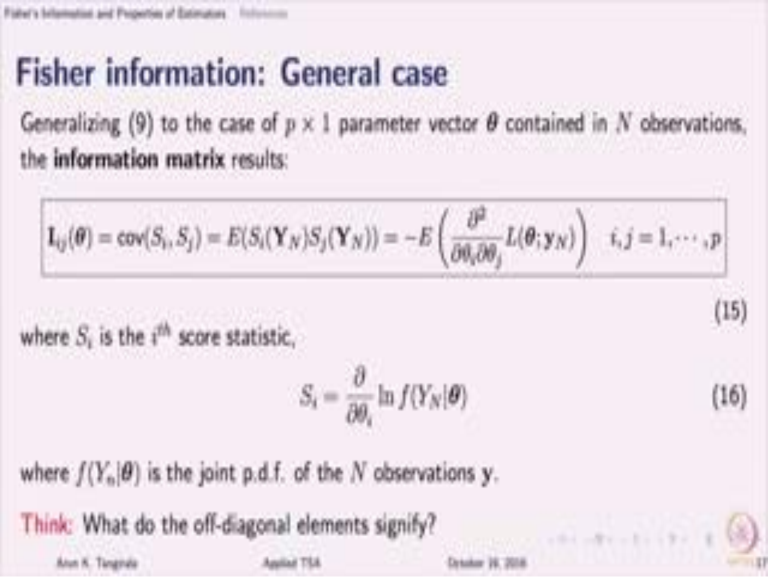
Applied Time-Series Analysis
Prof. Arun. K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 84
Lecture 37B - Goodness of Estimators 1 -3

Now, we move on to the general case, what is a general case? The general case is that I have N observations that is one, and then I may have p parameters to estimate.

(Refer Slide Time: 00:20)



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,

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

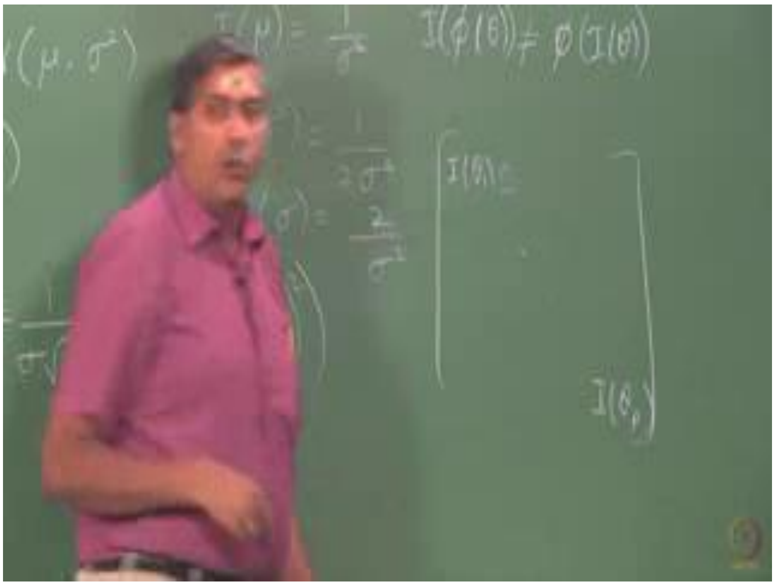
where $f(Y_n|\boldsymbol{\theta})$ is the joint p.d.f. of the N observations $\mathbf{y}$.

Think: What do the off-diagonal elements signify?

Arun K. Tangirala   Applied TSA   October 16, 2016   17

The moment I have p parameters by information now becomes a matrix, until now the information matrix has been a scalar, but now it becomes a matrix and what does this matrix contain.

(Refer Slide Time: 00:41)



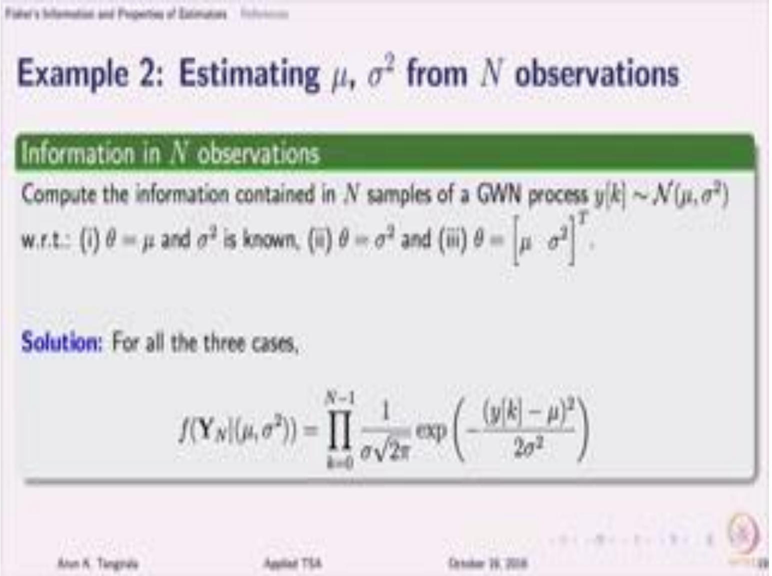
Well, it is going to contain information about the individual parameters along the diagonals and then it is going to contain information about estimating two parameters jointly and so on, it is not going contain information product of theta 1 and theta 2 and so on. But essentially what this information matrix is going to tell you is suppose I pick the half diagonal elements let us say 1 comma 2 in this information matrix it is going to tell me what is a information contained in 10 observations with respect to those estimating the two parameters jointly with respect to those two parameters. So, it looks at the pair of parameters and the diagonal elements will contain the information about the individual ones you can say so, but these individual ones are being estimate in the sense that when they are being estimated jointly, when will let me ask this question what happens if this information matrix is diagonal, what does it tell you. Suppose it turns out for a problem that the information matrix turns out to be diagonal.

So, we call this as orthogonal parameters problem that is you can the estimate of one parameter does not influence the estimate of the parameters only in such situations you can say that the parameters are orthogonal, but it is quite rare to have an information matrix that is diagonal.

How do you calculate this information matrix? Well it is a straight forward extension of what we have seen earlier, earlier we said the information is variance of this score and what is the score? Score is a derivative of likelihood with respect to theta, but now I have many parameters therefore, now I have to turn to partial derivatives and as you have seen in the expression here in 15, equation 15 - the information contain or this information matrix information ijth element of the information matrix is the negative expectation of the still the second derivative, but now it is a partial derivative. And it is a sigma square, dou square by dou theta i dou theta j earlier we had dou square the likelihood by dou theta square, but now we are going to evaluate the partial second derivative of a likelihood function with respect to theta and theta j. Do you expect this information matrix to be symmetric? It is symmetric, Ok.

Because the second derivative if you change the order of the derivation thing should not change that is if I look at the ijth element and jith element the only thing that is going to happen is order of derivate is going to change, but that is not going to change the answer. As usual you have the score function being the partial derivative here or with respect to the ith parameter of the log likelihood function. Let us under understand this through an example.

(Refer Slide Time: 03:53)



Fisher's Information and Properties of Estimators

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 Applied TSA October 16, 2016

Now, we will move from single observation to N observations which is the general problem that we encounter and now we want to ask what is the information contained about this mu, suppose I want to estimate the mean. This is the problem that we are studied earlier also I have given N observations I am suppose to estimate mean, when we use the least squares approach we get the sample mean as c estimator right. But now we are not referring to the estimation procedure itself we are just asking given N observations how much information does it contained about the mean.

First we will begin with single parameter and then we will move on to two parameters, what is the procedure now? First consult the likelihood function, now I am given that this N observations are coming out of a Gaussian white noise process correct. So, the first step as usual again is to construct the likelihood function, but this time I do not have a single observations I have N observations, right.

(Refer Slide Time: 05:04)

So, in other words now I have to construct the joint pdf y of 0 to y of N minus 1 and this can be an intimidating task in general, but this example is nice because it says that the observations are uncorrelated and further more they fall out of a Gaussian process. So, we know from the property of Gaussian processes that a Gaussian white noise process is also an independent IID process right. Therefore, the joint pdf of this N observations is simply the product of the marginal pdfs. So, in this case information I can write as the product of k equal 0 to N minus 1 f of y k of course, there is theta. So, I am going to write here and in general theta could be all parameters or partially known. So, that is the big advantage of working with white noise process.

Later on when we talk of MLE formally then we will learn, we shell learn there how to construct the log likelihood for a correlated process it is not that easy, but there is way out. What is a difficulty if I have correlated white, Gaussian correlated process.

Student: (Refer Time: 06:31).

So, there is no auto covariance or first of all I cannot use this relation, I cannot write the joint pdf as the product of the marginal pdfs, here it is very straight forward, that is ok. So, now we have the independent criterion kicking in and therefore, I can write the joint pdf very easily. When I do that the log likelihood becomes easy. I will give you a minute and see if you can come up with the answer for the information contained about a single parameter mu. Though rest of the procedure is straight forward this is your likelihood function right, then you are take a logarithm of that do not forget to take the logarithm otherwise the derivatives can massive. Ones you take the logarithm then the rest of the story is same you have to differentiate with respect to the single parameter mu and come up with your and then take the expectation; second derivative of a likelihood with respect to mu, take the negative expectation and then that is your answer. So, what would be intuitively what would be the answer?

Student: (Refer Time: 07:40).

What is into n? So let us actually erase this. So, this is the information contained in mu using a single observation right that is 1 over sigma square, now we are asking what is a information contained in mu when I have a vector of observations collection of N observations, intuitively what would be the answer?

Student: (Refer Time: 08:19).

Why, any arguments?

Student: (Refer Time: 08:34).

I do not want mathematics here.

Student: (Refer Time: 08:38).

No information you have here, but why should be proportional to n.

Students: Because all observations are (Refer Time: 08:47).

Each observation is giving you is same information, but they are uncorrelated or independent and that is why you think the informations will add up. So, shall we lock that answer can you check now if you get that answer, already got it; what would be your answer intuitively if this series was correlated, this process was correlated?

Students: (Refer Time: 09:27).

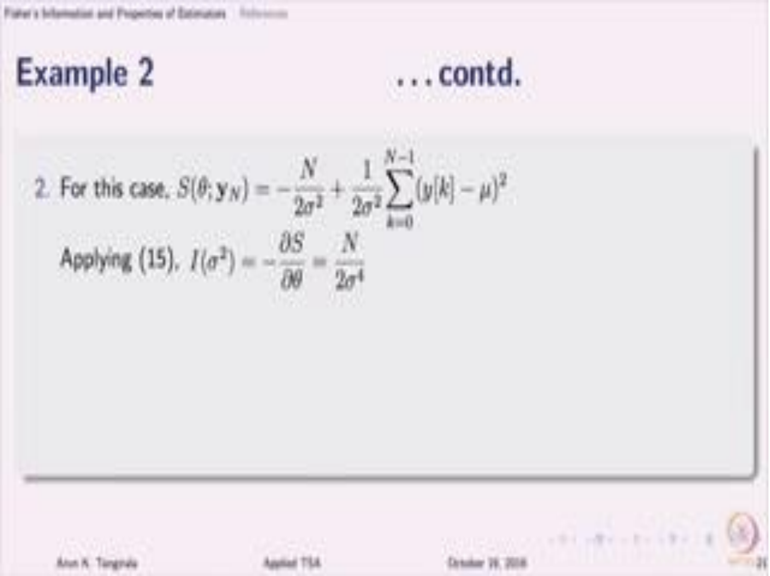
Why?

Student: (Refer Time: 09:30).

There some; that is correct. So, the information contained in N observations of a Gaussian white noise process scales with the number of observations straight away and as you have rightly said the because the process is uncorrelated the information simply add up and they are uncorrelated. So, each observation is bringing in some new piece of information it turns out that it is a same level of information bringing in, but it is new information because what it contained in one observation is not contained in the other. When the series is correlated one should expect this, propose this factor here to drop down. You can now arrive at a fresh perspective on the degrees of freedom here; when you look at the degrees of freedom here you said there are n degrees of freedom you can say in estimating mean. What we mean by degrees of freedom here is there are N independent sources of information with respect to mu I mean mu and that how it turns out to be.

So, when you get such an answer do not just stop it intuitively ask if this is answer makes sense. So, that is good, everything is now consistent. What happens when we ask a same question with respect to sigma square? Earlier with a single observation the information contained sigma square is, what is it? 1 over 2 sigma to the power of 4; what do you expect to see now with N observations, same story. Sure, that is correct.

(Refer Slide Time: 11:19)



Fisher's Information and Properties of Estimators References

Example 2 ...contd.

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

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

Arun K. Tangirala Applied TSA October 19, 2016 21

So, the argument is a same once again, but now also it says that you have n degrees of freedom in general to estimate sigma must because you are given mu. So, there is one parameter that you are estimating which is sigma square and it turned out to be N over to sigma power of 4. Now suppose I want to estimate mu and sigma square jointly, we have an information matrix now correct. Can you now work out the; likelihood function does not change, the likelihood function is a same - the only difference is now the parameters the unknown parameters are now different it is a vector of parameters mu and sigma square. So, what is the information matrix?

Student: (Refer Time: 12:09).

How do you compute it? Same, you just now take partial derivatives with respect to second partial second order derivatives with respect to individual parameters fill the diagonals and then take the joint derivative partial partially and then fill the half diagonals. Already the diagonals are known or not known? They are known, I do not have to revaluate correct. So, only the half diagonal I have to evaluate what is the answer for the half diagonal.

Students: (Refer Time: 12:53).

Zero, does it make sense, what is that zero mean? What does a half diagonal value evaluating to zero mean?

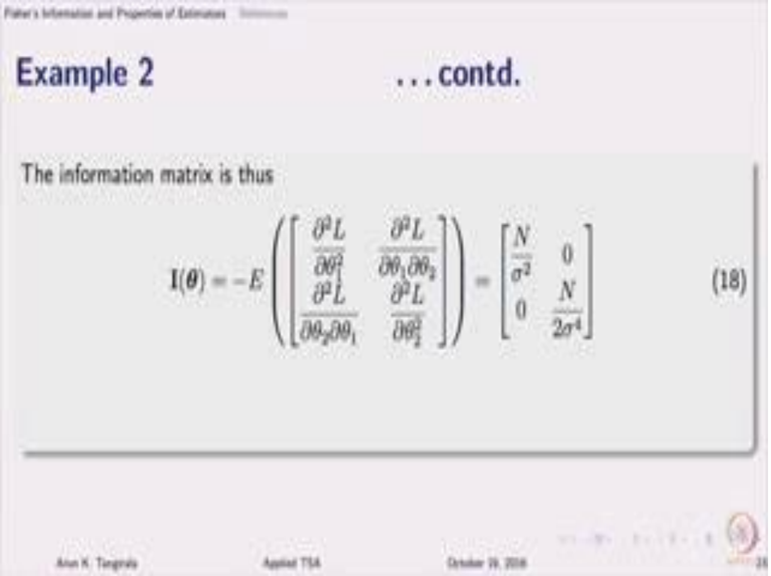
Student: (Refer Time: 13:07).

Can be; they can be estimated independently, does it make sense? But when we estimate sigma square we say I want to I have the estimate of mean; am I right, do not we say that we called the expressions that we use for the estimating variance.

Students: But it is a Gaussian so.

So? What does Gaussianity got to do with this? May be may not be, but I just want to know what is Gaussianity; do not you recall the expression for estimating variance the basic expression that we have been using for many years 1 over N minus 1, sigma y k minus y bar square and y bar is a estimate of mean right. So, unless estimate mean I cannot estimate sigma square types, but here it says that; there is a mistake here it show been not x k by y k in equation 17 I will correct that.

(Refer Slide Time: 14:16)



Fisher's Information and Properties of Estimators

Example 2 ...contd.

The information matrix is thus

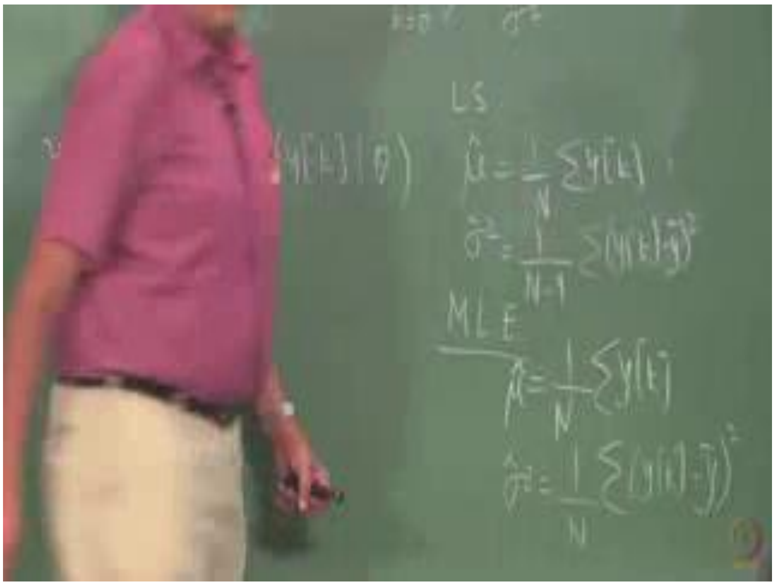
$$\mathbf{I}(\boldsymbol{\theta}) = -E\left(\begin{bmatrix} \frac{\partial^2 L}{\partial \theta_1^2} & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L}{\partial \theta_2^2} \end{bmatrix}\right) = \begin{bmatrix}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

Arun K. Tangirala Applied TSA October 19, 2016

So, does this make sense? Well it depends on how you estimate if you are estimating mu and sigma square jointly, see the classical way of estimating sigma square is first estimating mean followed by estimation of variance there you are not estimating jointly. That is in fact, you can say that that is a least squares approach, if you recall the example the motivational example that we discussed first we focused on estimating c the constant c which is nothing but the mean and then I skipped one part where I said will estimate sigma square e later on - that estimate of sigma square e which is nothing but the variance of y is actually constructed after estimating the c optimally.

So, it is a two step procedure that we follow in general, but that need not be the only way of estimating mu and sigma square for any series for any process, I can estimate them jointly and what Fisher’s information tells us is when you estimate them jointly there is a possibility that you can estimate them independently. In fact, it shows now clearly that even though I am estimating jointly one estimate does not have any effect on the other both have n degrees of freedom that is what maximum likelihood approach will actually give you.

(Refer Slide Time: 15:47)


LS
$\hat{\mu} = \frac{1}{N}\sum y[k]$
$\hat{\sigma}^2 = \frac{1}{N-1}\sum (y[k]-\bar{y})^2$
MLE
$\hat{\mu} = \frac{1}{N}\sum y[k]$
$\hat{\sigma}^2 = \frac{1}{N}\sum (y[k]-\bar{y})^2$

In other words if I were to work with the estimates let us say I use least squares and here information use MLE to estimate mu and sigma square. The least squares estimate of mean as we know is a sample mean and the least squares estimate of sigma square is in fact, you can show that to be 1 over N minus 1 sigma y k minus y bar square. Whereas, will be MLE what do you think is optimal estimate of mean, in other words you have the log likelihood function with you right; all you have to do is set the derivative with respect to mean to be 0 and find the optimal estimate what is the answer? What is the MLE estimate of mu?

Student: (Refer Time: 16:34)

That is it, same sample mean, it is correct. So, MLE also gives you the same estimate as the least square interesting. So, the sample mean is optimal in many different ways, it is optimal in the least square sense, it is optimal in the maximum likelihood sense sample mean is so therefore a very attractive estimator. But if you were to estimate sigma square also, so you have this log likelihood function what would be the optimal estimate or sigma square? You have the log likelihood function with you, all you have to do is differentiate now partially with respect to sigma square and set it to 0 along with already you have done this mu estimate what would you get? We said that derivative log likelihood with respect to sigma square to 0 and solve it, but you have to assume that the mu is also unknown, we are estimating jointly.

The estimate should come out to 1 over N sigma y k minus y bar square, that is only difference between the MLE estimate and the least squares estimate. The least squares estimate of sigma square as a 1 over N minus 1 whereas, the MLE maximum likelihood estimate of sigma square has a 1 over N as a factor right of course, when N is very large the difference is not much. But degrees of freedom that MLE assumes is N because it is estimating jointly whereas, the least squares method says I have already estimate a two stage approach I estimate mu hat that sorry; estimate mu and then estimate sigma square the I explicitly adopt a two stage approach whereas in MLE I am estimating them jointly and it works out to be the that the degrees of freedom is N, although truly the degrees of freedom is N minus 1, it uses a 1 over N.

Now, we already know from prior discussion that this is an unbiased estimator of the variance that is expectation of sigma square hat is sigma square. If I use a 1 over N minus 1, that is a correct one to use as for as bias is concern or lack of it is concern, if I want to an unbiased estimate of the variance I should use 1 over N minus 1. What does it tell me about MLE? Estimate maximum likelihood estimate of sigma square, is it biased or unbiased?

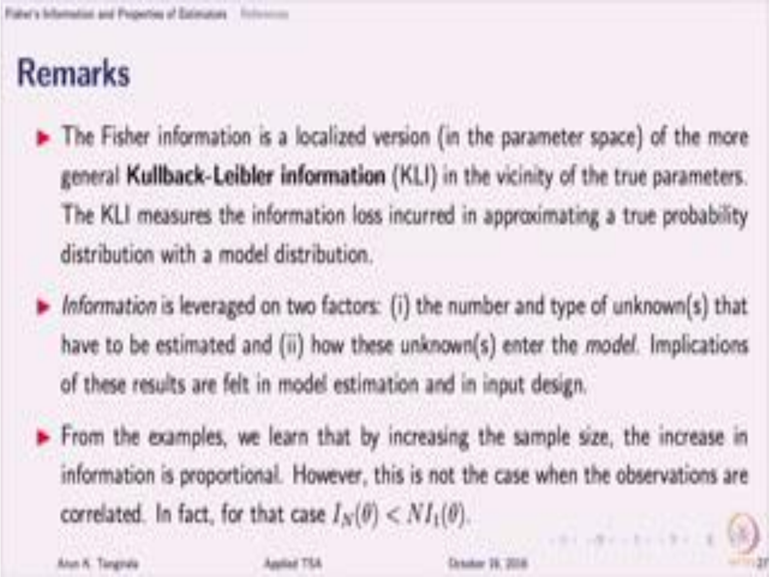
Student: (Refer Time: 19:17).

So, maximum likelihood estimates unfortunately give you in general biased estimates of parameters although and I am talking only with respect to sigma square here in general also it is true that maximum likelihood estimates are biased; however, what happens to the bias as N goes to infinity? That bias vanishes because; obviously, as N becomes very large whether I use 1 over N minus 1 or 1 over N, it does not matter. So, we say MLE gives asymptotically unbiased estimates, that is a difference between your bias or asymptotic bias. Bias looks at the error between your estimate at the average of the estimate and the truth for finite samples, whereas asymptotic bias examines the same thing as N goes to infinity. S

o, in general maximum likelihood estimates give you asymptotically unbiased estimates that is why if you turn to the MLE literature you will find one very common criticism of MLE is that it is only good for large samples, it is a beautiful estimator only when you are dealing with large samples. The maximum likelihood estimates are not so great when it comes to small sample size. So, there is a whole body of literature on small sample estimation where MLE is really put down is a well look MLE does not give me great estimates because it gives me biased estimates. So, these are some of the thinks that we have to know.

Coming to the point here what this result tells me is that the estimate of mean and sigma square are disjoint that I mean the information contained have has no bearing one on the other one and will see also the consequences later on.

(Refer Slide Time: 21:15)



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 $I_N(\theta) < NI_1(\theta)$.

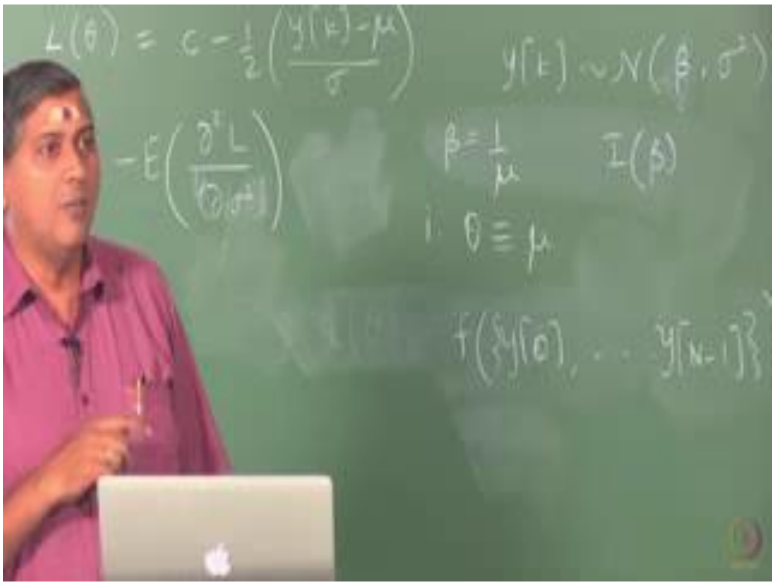
Arun K. Tangirala Applied TSA October 16, 2018 27

So, just a few passing remarks this closes the discussion on Fisher’s information, essentially the Fisher’s information gives you a metric or a measure of how much information is contained in some data set with respect to some parameters that you are interested in and that can changed with the number of parameters that you are estimating also. In the examples you may not have seen that coming out very pronouncedly, but in general as you put in more and more information, suppose I have p parameters and I calculate the information you are going to end up with the metrics it need not be diagonal, but as you pump in more and more information, you say no now I know out of the p I know a few parameters then the information content can changed. So, regardless of all of that the Fisher's information is a good metric of the information contained in a given set of data with respect to parameters.

Later on it was shown that is the Fisher's information was proposed somewhere in 1920s and somewhere in 1950s or 60s it was shown that the Fisher's information is actually a kind of a localized version, what we mean by localized is in the parameters space of a more general information measure known as a Kullback-Leibler divergence measure.

If you have come across even a bit of information theory you would have come across this Kullback-Leibler - KL divergence measure. This KL divergence measure looks at how much loss of information occurs when I assume probability distribution function that is not the same as a truth that is when there is a deviation when there is a mismatch between the assumed pdf and the true pdf, but we will not going to that this is just for your information. And remember that the information that we are talking about is leveraged on tooth on two factors, one is the number and type of unknowns as I just said and how these unknowns enter the model that is very very important.

(Refer Slide Time: 23:37)


$L(\theta) = c - \frac{1}{2}\left(\frac{y[k]-\mu}{\sigma}\right)$
$y[k] \sim N(\beta, \sigma^2)$
$-E\left(\frac{\partial^2 L}{\partial \sigma^2}\right)$
$\beta = \frac{1}{\mu}$
$I(\beta)$
$\theta \equiv \mu$
$f(\{y[0], \ldots y[N-1]\})$

If you re-parameterize what you should ask is suppose I define a new parameter let us say some beta which is 1 over mu and I ask what is information contained in beta. So, instead of parameter is in Gaussian in with mu and sigma square I would say now this is parameterized with respect to beta and sigma square, what is this beta? 1 over mu, but this beta now enters a Gaussian in distribution in a different way than mu.

In other words your ability to figure out very simple analogy is suppose you have a guest coming to your house the parameters are like guest when they come to your house your ability to find out what they are doing depends upon where you place them. If you place them, if you may, if you seat them in the main hall where your also doing your work then you know very well what the person is doing, but suppose is guest is hiding behind some corner there you will have no idea the information contain would be much lesser. That is exactly the scenario here - the parameter, the information contained about a parameter in the model depends a lot on how the parameter is seated in your model where, so this is small branch of study in estimation theory known as re-parameterization. You would want to re-parameterize a model so that thinks become a lot more convenient when it comes to estimation. We will not perceive that, some time I may mention it again.

And now this is a factor this is the last comment is something is that I have said already which is that, when we had N uncorrelated observations information increased proportion to N or in fact exactly by factor of N, but if you had correlated observations the scaling of information is going to be less than N that is something to keep in mind. So, that kind of concludes a Fisher’s information discussion that we wanted to have.