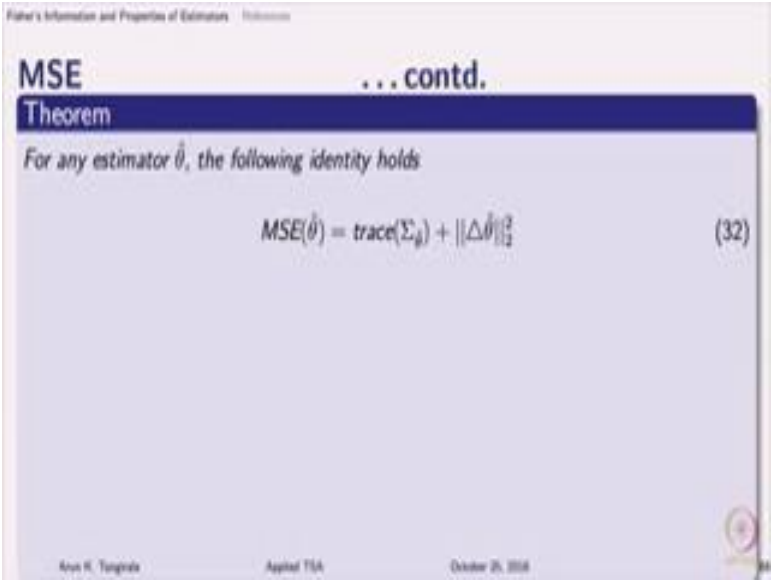
Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 87
Lecture 38B - Goodness of Estimators 2 -2

The Cramer-Rao's inequality tells us about the existence of an efficient estimator and what is a bound on the minimum Varian, what is the bound on the variance that you can get for an estimate. More better much better property of an estimator is the mean square error, because the mean square error in two different ways you can argue that mean square error is a better estimate: the first argument is that it is looking at the variability of theta hat with respect to the truth, it is an different matter of fact whether you can calculate it in practice or not, but theoretically it is saying it is asking for the variability of the estimate around the truth and if that is low then I am happy.

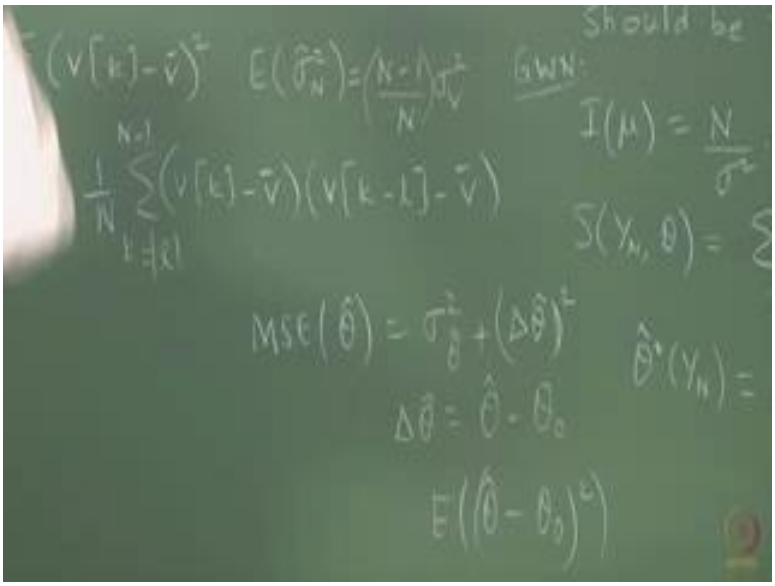
Because ultimately I want that to be low, bias and variance as we will see now are nothing, but two different components of MSE.

(Refer Slide Time: 01:15)



In fact, the relation the equation that I am showing you here is for a vector of parameters, suppose I am considering a scalar parameter.

(Refer Slide Time: 01:47)



So, what this results says is this mean square error is a sum of two terms: one that it is a sum of one which is the variance of theta hat and the other it is a square of the bias, for the scalar case we do not say it is variance we talk about the trace of the covariance sorry for the vector case for the scalar case the trace of sigma theta hat is simply sigma square theta hat itself.

So, what this results says is that MSE let us look at just a scalar is nothing, but sigma square theta hat plus your delta theta square, where delta theta is theta hat minus theta naught; that is the bias essentially right you can say delta theta hat. If I minimize the MSE, I am done with my job because I have made sure that the variability of the estimate around the truth is minimized, but I can achieve that in a number of different ways, this is what that is what this result is telling me if I minimize the MSE, it says I can adjust the bias and variance in many different ways to achieve that minimum MSE.

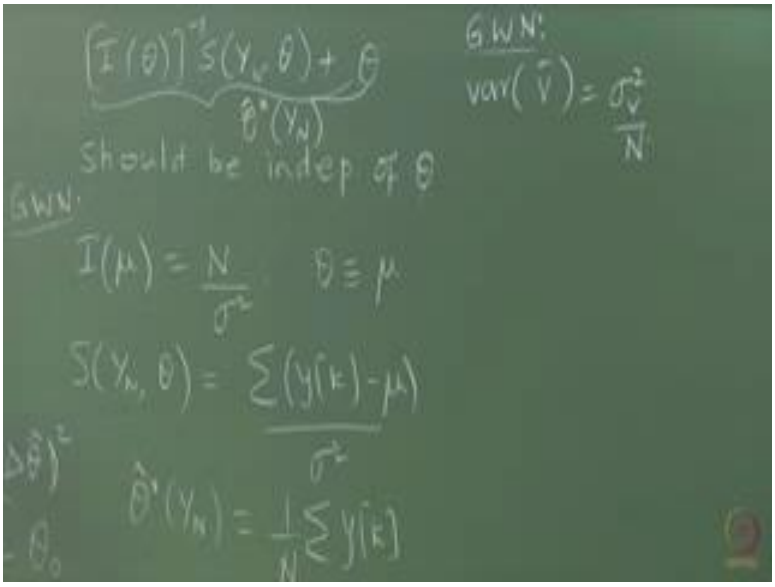
For example I can drive the bias to zero and put everything in variance, or I can have non zero biases and lower the variance right both will get me same MSE, then it is a question of whether you are going to you want a you are willing to work with an unbiased or a biased estimator and so on.

But the ideal thing that we should be talking about is minimum MSE, until now we have talked about variance and bias separately MSE fuses them together, but the difficulty with working with this measure at least in early days was that people said that I do not

know theta naught. So, there is a very difficult thing to compute also in practice, even if I were to compute this from data I do not know the truth, at least in variance what is the advantage I am defining the variance with respect to it is own average and I can compute that and we have vanished to derive expressions, we have shown that the variance of the sample mean for a Gaussian white noise process sigma square over n, it was easy for me to derive that theoretically.

You may ask you may say that sigma square is not known, right.

(Refer Slide Time: 03:57)



When we look at variants of Y bar or V bar it does not matter, we have shown this for a Gaussian white noise process. I can still use this in practice because I can estimate variance and then in place of the theoretical value I can use the estimated variance. So, this expression for variance is usable you can is practically amenable whereas, for MSE I may not be able to derive expressions in general.

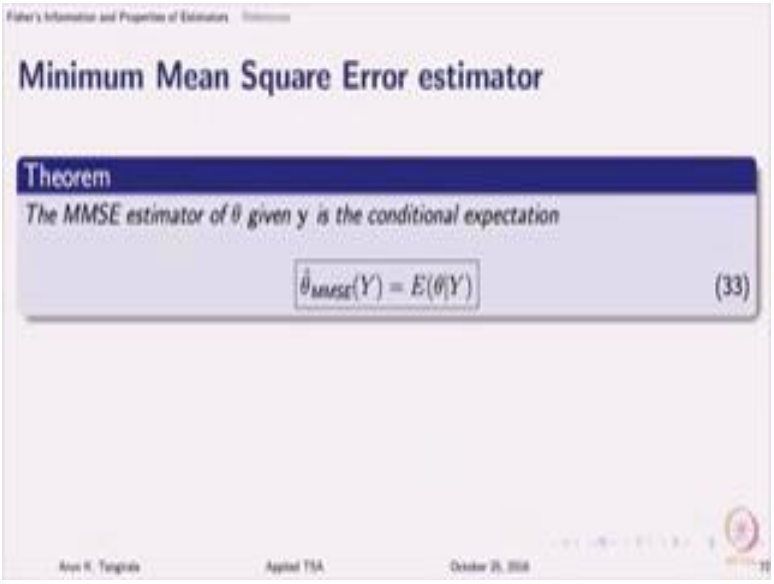
But then gradually as things developed, Bayesian estimation came along and you could show that Bayesian estimators do give you minimum mean square error estimates. Although you cannot probably setup this function that is expectation of theta hat minus theta naught square. Although initially you cannot set which is what we had said at the time of introducing estimation, we said that this is a very difficult thing to minimize because I do not know theta naught and that is why we took the alternative route of minimizing the prediction errors or approximation errors.

But Bayesian estimators came along of course, the wieners filter, the Kalman filter they are all in fact, minimum mean square error estimators; they came along and said although you cannot minimize this I will show you how to do it to set the problem in a different way and arrive at minimum square error estimator. If you have an MMSE estimator you should pick that, if you can if you can find one and implement it you should use that more than minimum variance and so on because MSE is a measure of the spread of theta hat around truth and if an estimator is minimizing that then very well and good.

Now, having said this sometimes people would say well minimum mean square error estimates may be very good, but they may achieve that at the cost of maybe non zero bias because there is nothing in the statement of minimizing this on whether the bias should be zero or not. So, after once you have found a minimum mean square error estimator, it may be a good idea to look at the bias and if there is a bias and if you are working with small sample sizes, maybe you should reconsider using that estimator. In fact, they Bayesian estimator is a minimum mean square error estimator it is very good asymptotic property that is when the sample size grows very large, but when the sample sizes are small the bias in Bayesian estimates is quite high.

So, one has to worry about always small and large sample sizes, which is good because there are some 1000 to 10 maybe 10000 PhD's and other kind of papers coming out of small sample size literature and then there is a whole lot of research for large sample size. God has for every one something is this world and I am just avoiding the proof of the relation for MSE sometimes this relation here in equation 32 is also called the kind of the Pythagoras theorem, equivalent in estimation theory right you can think of MSE as being the hypotenuse square, and the bias and the variance be being the sides of the triangle.

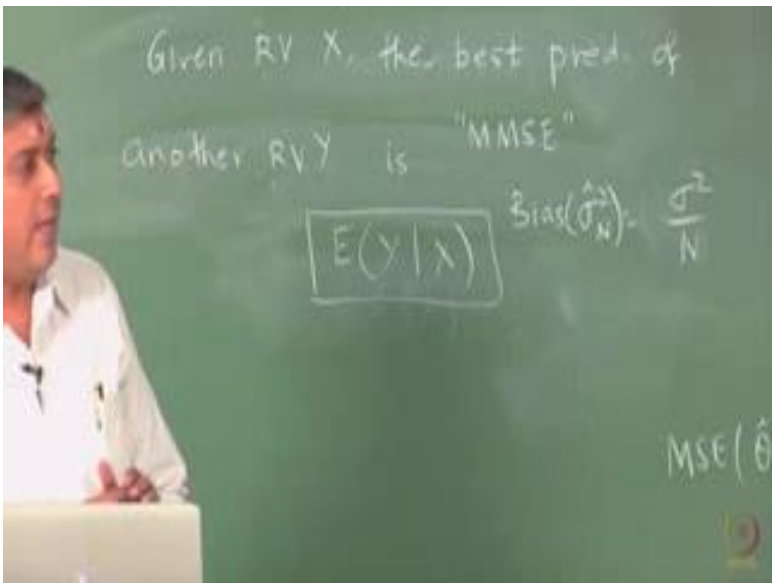
(Refer Slide Time: 07:23)



So as I said now the minimum mean square error estimator which is very hard to compute, you can show that the minimum mean square error estimate is none other than the conditional expectation, long ago we said this, that the conditional expectation is the best prediction of a variable given another random variable.

But in this statement there is an implicit assumption that theta is a random variable, when we said long ago that conditional expectations give you best predictions, we made a statement but assuming that both variables are random.

(Refer Slide Time: 08:09)



So, given X, one random variable X, the best prediction of Y is; let say the best prediction of another random variable let us let us be very clear here Y is the conditional expectation and best in what sense? It is best in the minimum mean square error sense; this is the fundamental result in prediction which we have which people keep using, but then we said this conditional expectation can be in general a non-linear function, maybe hard to compute and so on. Apart from that if you look at it is the best predictor I mean you can use 100 1000 neural networks and so on everything nothing can beat the conditional expectation.

Coming to estimation of parameters, the minimum mean square error estimator is given Y let us say some random variable Y, the best estimate of theta in the minimum mean square error sense is the conditional expectation of theta given Y, but then you are assuming that theta is a random variable, which is the philosophy on which Bayesian estimation rests; when we talk about Bayesian estimation we will revisit this, but you have to remember that this expression implicitly assumes theta to be a random variable, which is much against the philosophy that we have been following until now. Until now we have been assuming parameters are deterministic they are fixed quantities, but this statement assumes that theta is a random variable, how on earth theta can be random is what? Many people argue and fight and you know spend several evenings and so, on do not have to just relax.

But you just have to understand that when theta is random, a way of interpreting that is that your knowledge of theta is uncertain, theta itself may not be the truth, may not be uncertain; but what Bayesian estimation which is it rests on this result says that, you before data you have some uncertainty about theta, after data you have still estimate, but with uncertainty. So, what is the big deal I mean you begin with uncertainty and you end up with uncertainty admitted, you will never be able to estimate the parameter with certainty with from finite data. So, why do not you live and admit it right before right from the time of experiment to the time you estimate the parameter?

It basically says do not fool yourself by thinking that theta is fixed and then you construct an interval around it, and then you say this is the interval in which truth is present, normally you will see the statement this is the interval in which truth lies, but how can truth lie truth cannot lie right. So, it is exactly. So, Bayesian estimation says this is an oxymoron statement, you can I mean maybe sometimes you can throw away the o x

y, but basically it says that be practical theta will remain uncertain in your mind, the truth will remain always uncertain, what the data is doing for you is reducing that uncertainty for you.

If it is informative; if it is not informative the same uncertainty will prevail before and after. So, you have a B.C and A.D for years, you can say here you have BD and AD before data and after data in Bayesian estimation before data we call it as a prior, after data we call it as a posterior; we will come back to that when we talk of Bayesian estimation.

Let us move on now we are done with the statistical properties, the next in line what are the other class of properties asymptotic properties right what do the asymptotic properties qualify? They actually qualify the behavior of the estimator as the sample size increases, as I collect more and more data will my estimate improve? Which is a very very important question and one of the properties in that line as we have discussed is asymptotic bias; asymptotic bias looks at the bias as n goes to infinity.

So, we wrote for example, two different expressions for variance estimators: one was unbiased, other was biased, but the biased if the variance estimator that we wrote earlier was 1 minus 1 over N times sigma square v this was the bias that we had in the estimator that used 1 over N as a factor which is in the MLE estimate; obviously, as N goes to sorry should be I am. So, sorry it should be sigma square over N as the as N goes to infinity the bias goes to 0. So, then we say that it is an asymptotically unbiased estimate. Why do we look at asymptotic bias? Because we say so for it is to have bias for finite observations, at least as the observation grows if the bias vanishes, I am willing to live with it.

So, you are making certain compromises and you what it says is, if you have large sample, data with large sample sizes it is to work with asymptotically unbiased estimators, although for finite samples it may have a bias; that is all. That is all to the asymptotic bias part, it is a very important requirement.