CH5230: System Identification

Estimation of non-parametric model

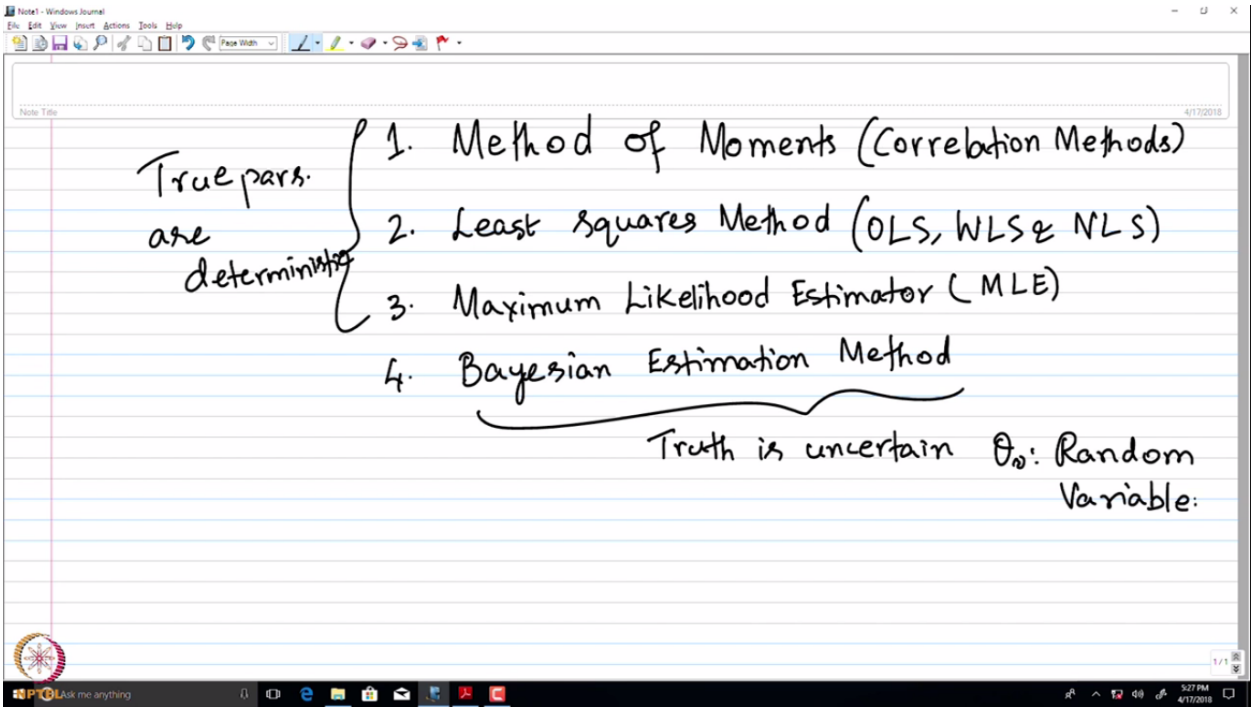
Part 1

Welcome to this lecture on estimation of non-parametric models. We've learned already what no-parametric models are. And what we mean by non-parametric models, essentially if you recall is that we do not assume any structure on the deterministic and the noise models as well. So until in the journey that we have had, we have learned the different aspects of estimation, what is all about? And what are the different methods of estimation is what you have learned recently, namely four different methods. So let me actually summarize that for you. One you've learned what are methods-- method of moments, this is a first method that we have looked at briefly. In the-- in certain parts of the literature as you know, these are also known as correlation methods.

And the second method that we have learned is one of the most celebrated methods, which is the least squares method. And within the-- under the least squares you have learned different variants. The ordinary least squares, the weighted least squares and the non-linear least squares that is abbreviated by OLS, WLS and NLS. Of course, there are other versions as you know of the variance of the least squares method. Then early on we had learned the maximum likelihood method. Maximum likelihood estimation or estimator MLE. We were introduce to this MLE as early as when we were talking Fisher's information. You remember that Fisher's information and Cramer-Rao's inequality is based on the notion of likelihood and the approach of maximizing it.

And the other method, of course, that we have looked at is the Bayesian estimation method. Which is quite different from these three here. So the top three are the classical methods, which rest on the assumption that the unknowns are the parameters or fixed quantities that, in other words the truth is assumed to be deterministic. So in here you can say that truth or the true parameters are deterministic. Right? Whereas the Bayesian method rests on the assumption, that the truth is also uncertain. In other words theparameters are random as well. So theta 0is a random variable or a vector of random variables as the case may be.

(Refer Slide Time: 03:35)



So what we will learn today, of course, I'll give you a very quick summary of this-- at least the least squares method and the Bayesian estimation. MLEis something that we've visited in detail, but mostly I'll summarize the least squares method to begin with and as-- and when the need arises, I will also quickly recap the Bayesian estimation method. And primarily we learn how to apply these methods to estimate non-parametric models. So if you recall the class of non-parametric models that we have learned or the FIR models, the finite impulse response models, the step response models, and the frequency response models. These are the three classes of non-parametric models that we're interested in.So let's actually recall what the least squares methodis all about, just very quickly. The concept in least squares method is very clear. You have your given a vector of observations y k. And genetically speaking you're given a bunch of regressors.

So you're given k running from 0 to N minus 1, in which case of course this vector is not needed. You can stack these N observations into a vector y, if you want. And sometimes we may have also used an upper case Y. You are given this and you are also given the values of the regressors. And you may recall that we assume in general that p regressorscan explain y, so i runs from 1 to p.And the goal is to-- of course build a model between y and psi or the psi i. If you assume a linear model then we run into linear least squares. If you assume a non-linear model between y and psi i then you run into a non-linear least squares formulation. So the linear least squares commonly known as the OLS if you recall, is based on this model y k is-- well you have-- if you consider the vector of observations psi then it is based on this model.

Where this vector here psi transpose as you know in fact psi k at any instant is a vector of regressors that you have, psi 1, sorry it's not i up to psi p. Therefore naturally theta is a p by 1 vector and y is assumed to be a scalar, that's what-- that is a kind of framework that we have dealt with, but of course y can also be a vector will not worry about that right now.Now, when I say OLS here, what I mean here is two things, one is this model, so this is what we call as the model. And the cost function is simply y k minus y hat of k given sums of the-- summation of the squared errors, y k minus y hat of k, the whole square and k running

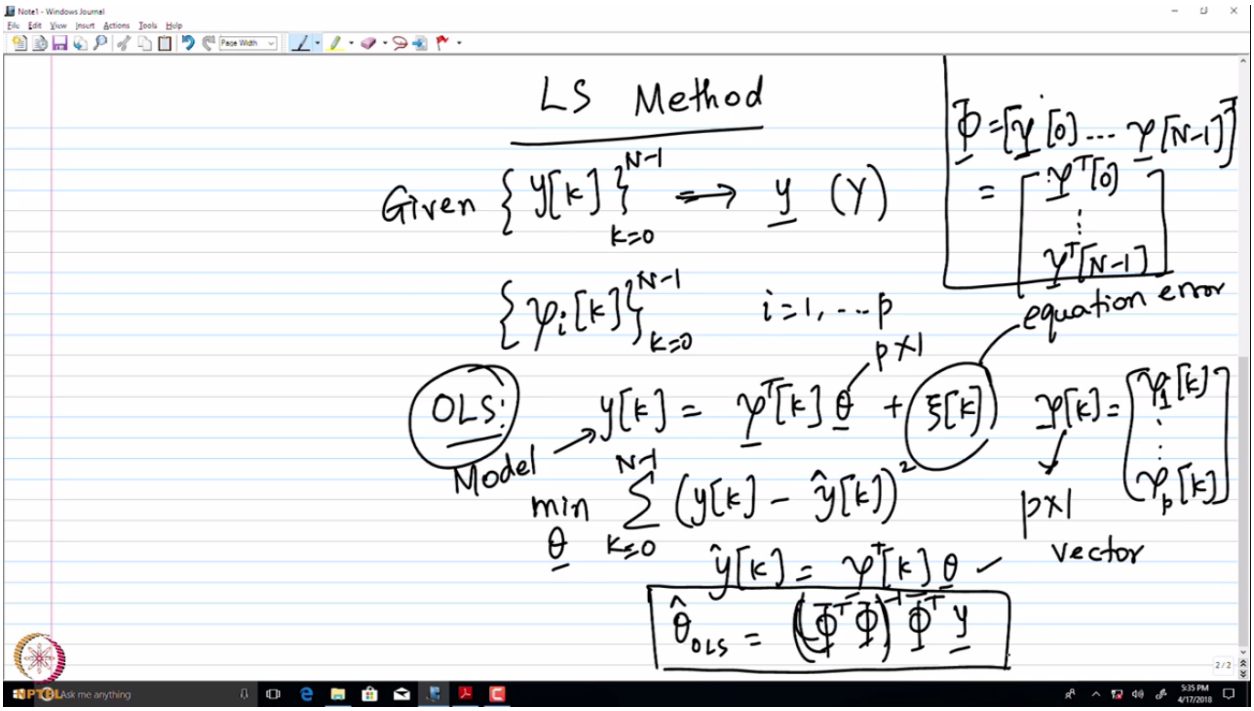
from 0 to n minus 1. We have seen this kind of a formulation even in the simple introductory estimation exercise.

So to recap what we do is in ordinary least squares, this y hat of k is constructed as simply psi transpose k theta. And obviously the implicit assumption is that the z of k that you have here, which is also known as many a times observe-- equation error. So this is also known as the equation error. In using this predictor here, this predictor, an implicit assumption although not stated is that these equation errors are white. Of course, if they're not white then you would factor them into your prediction.So if you recap what we have learned is how thetas are estimated optimally. So this is what we minimize. And we know the solution to this problem. So the optimal solution is theta hat OLS is simply phitranspose phiinverse y. And what about phi? So we know what phi is all about. So we can-- if I can write here, what is phi, you remember.

That phiis essentially an N by P matrix, right? And the way it is constructed is, it is a psi 1, sorry. So it is psi transpose 0, starting from psi transpose 0 to psi transpose N minus 1 and transpose of that. So notice that there is a transpose here and there is a transpose here, right? So notice that. Which means, remember that each psi k is a P by 1 vector. So what would be, sorry, so there is a mistake here? So there is no transpose here, I'm sorry, psi transpose. Correct. So your phi vector is simply psi'sstacked from 0 to N minus 1 and a transpose of that.So each psi 0 is a P by 1. And you have N such columns and you're taking a transpose of them. You could alternatively write this as a stack of psi transpose, phi transpose 0s.

For example, you could write this as psi transpose 0 up to psi transpose N minus 1.Where? I am so sorry, correct, you're right. You're Right. Times phi transpose y. Correct, you're right. Okay, good. That's a-- that shows you guys are alert. Of course, it was unintentional, unintend-- unintended. Now, this is what is the 10:53] solution. Let me just erase this here. Right, this is-- so phi transpose times phi transpose y. This is the OLS solution. And, of course, we talked about in the least squares lectures how good this estimate is and always when we talk of the goodness of an estimator, we have to have a reference. Otherwise there is no point. So as an example, if I devise a new sensor and I claim that the center is very good in measuring some variable. I need to have a reference sensor to show that it is very good, right?

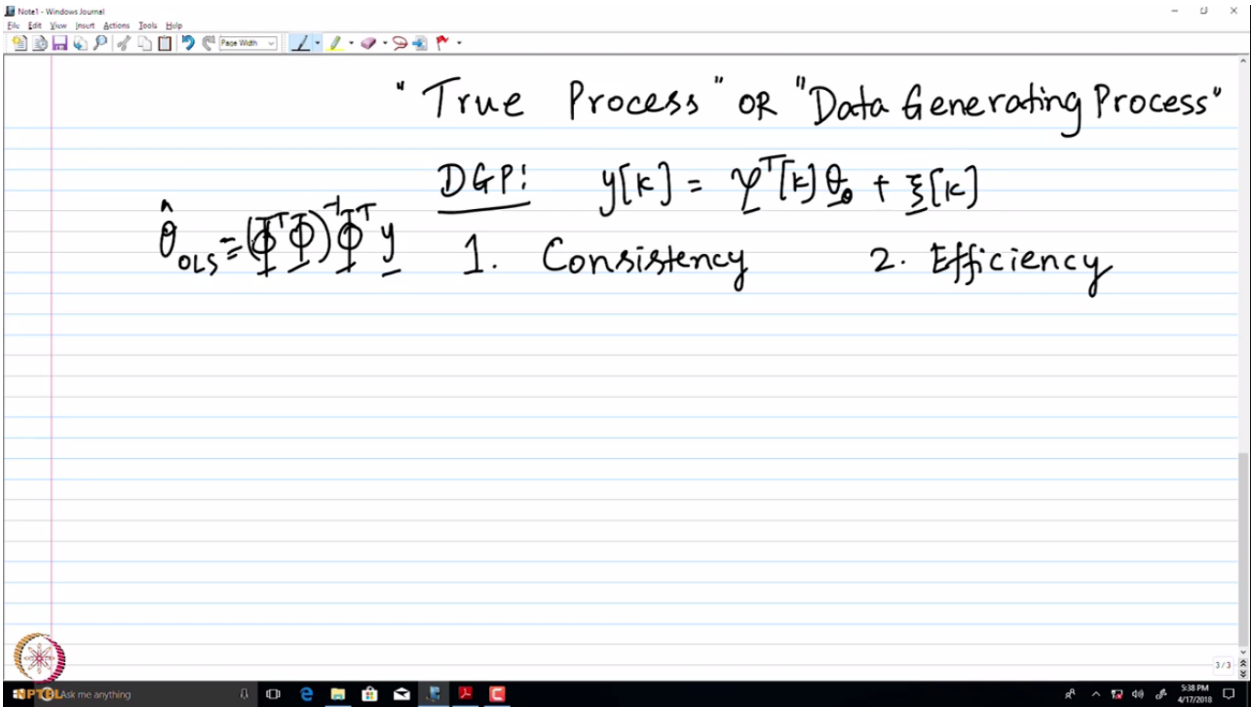
(Refer Slide Time: 11:32)



Likewise when we talk of estimation problems, we need a reference for speaking of the goodness of OLS estimator. So normally we assume model or some kind of a truth. Typically this is either called the true process or a more commonly used term is the data generating process. Data generating process, DGP. Which tell-- which basically states clearly how the data has been obtained? And then you ask how good your estimator is based on the estimates that you have obtained. So in short we call this as DGP. Which serves as a reference and we say if the data were to be generated according to this equation, now since we are referring to this as data generating process, we use a true value here. So instead of theta, we use theta 0. So we assume that the data that is coming to us is being generated by this process. In reality, of course, this is not true. The-- In reality the process-- the data generating process is a lot more complicated than this.

But if you keep talking about the reality then there is no way I can comment on the goodness of any estimator, leave alone least squares. So, we say that assume that even under these conditions, when the data generating process has this structure, where theta 0 is a true value and your z's are the equation errors. Then how efficiently does the least squares method recover this theta 0? Is the question that we ask, we ask and the other question is, whether least squares method consistently estimates it'stheta 0. Remember we are always worried about one kind of efficiency-- consistency and efficiency. Two is efficiency. Consistency examines the asymptotic convergence and efficiency looks at the error incurred in your estimate. So assuming that this is how the data is generated and that we are given N observations of this data and that I am going to use a least squares estimator, such as this. There isordinary least squares.

(Refer Slide Time: 14:27)



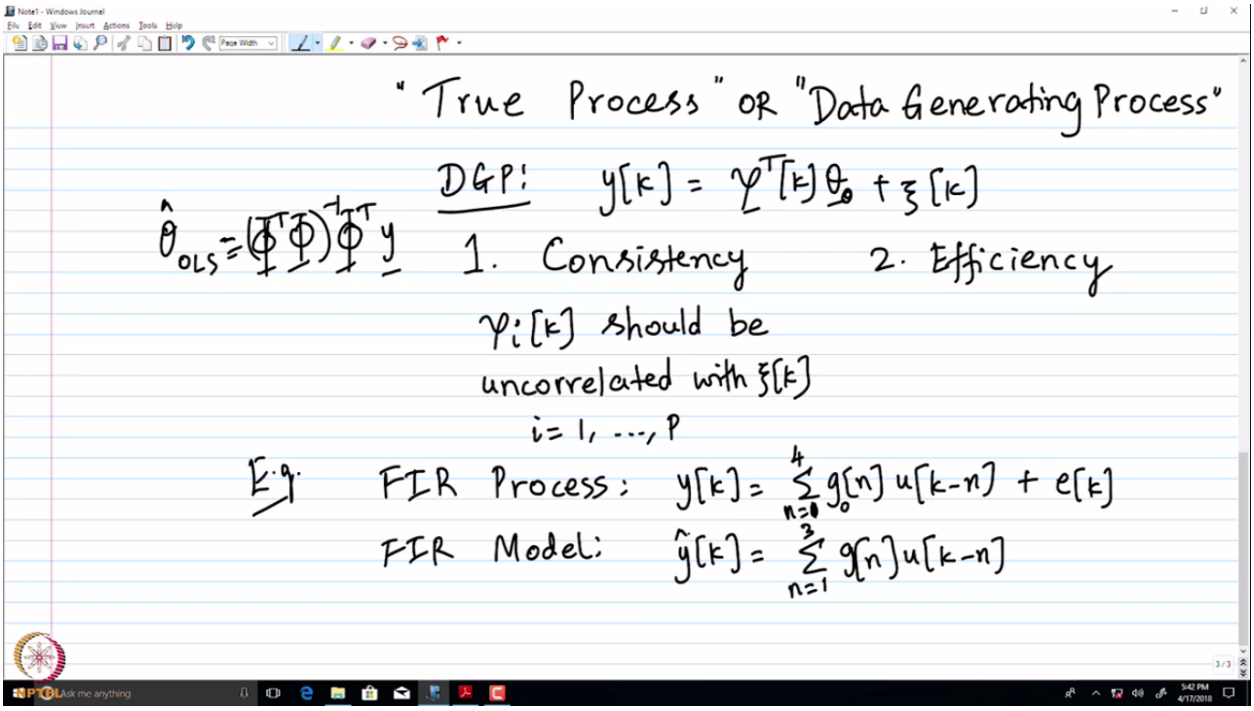
Then what can we say about the consistency? Do you recall, what are the conditions for consistency of the least squares estimator? What is a prime condition? Do you remember? Is that the requirement? The, you mean, the z's should not, so sorry there is novector here. No, so we assume that the data generating process is this, right? And then we state conditions on z. That is if the data were to be generated by this, what kind of assumptions you need to make on z? And perhaps the relationship between z and the regressorsto obtain consistent and efficient the estimates of theta 0.Sorry? Which should be perpendicular? So what is a better way, statistical way of saying that? Uncorrelated. So for consistency, of course, the prime requirement is that the regressors should be uncorrelated with the equation errors. Now that is phi transposekor in fact we don't need transpose here, we'll say that, psi should be uncorrelated with the equation error.

And i runsfrom 1 to P, that means for all the regressors you should hold good. Now the other thing that I would like to reiterate, which I have already done in the least squares lectures is that, we are assuming that there is no structural mismatch between the model, I mean, in the sense that I have use the same the regressors. I'll-- In both my model and the data generating process. In reality I will never know what regressors have participated. But what we are doing here is, we are saying that if psi's are the regressors that we have used then and then we say y k is psi transpose ktheta 0 plus z k. If there is a structural mismatch, then that structural mismatch will go and sit in the equation error, right?I think there is an example that I talk about, ARX's model and ARMAX model or even if you have, for example, taken FIR model,what happens when I take an FIR model?

And it's a very relevant discussionbecause we are gonna talk about estimation of non-parametric models. Suppose I take an FIR model. Suppose this is the FIR process as an example. And I assume-- and I fit an FIR model. Let's see the data generating processes is FIR, right? And the model that I fit is also an FIR. Now there are two possible. Let's first fix the process. Let's assume that we have here, fourth order FIR model. We'll call this as g0, let's say n runs from 0 to 3 or we can say from 1 to 4 assuming unit delay.

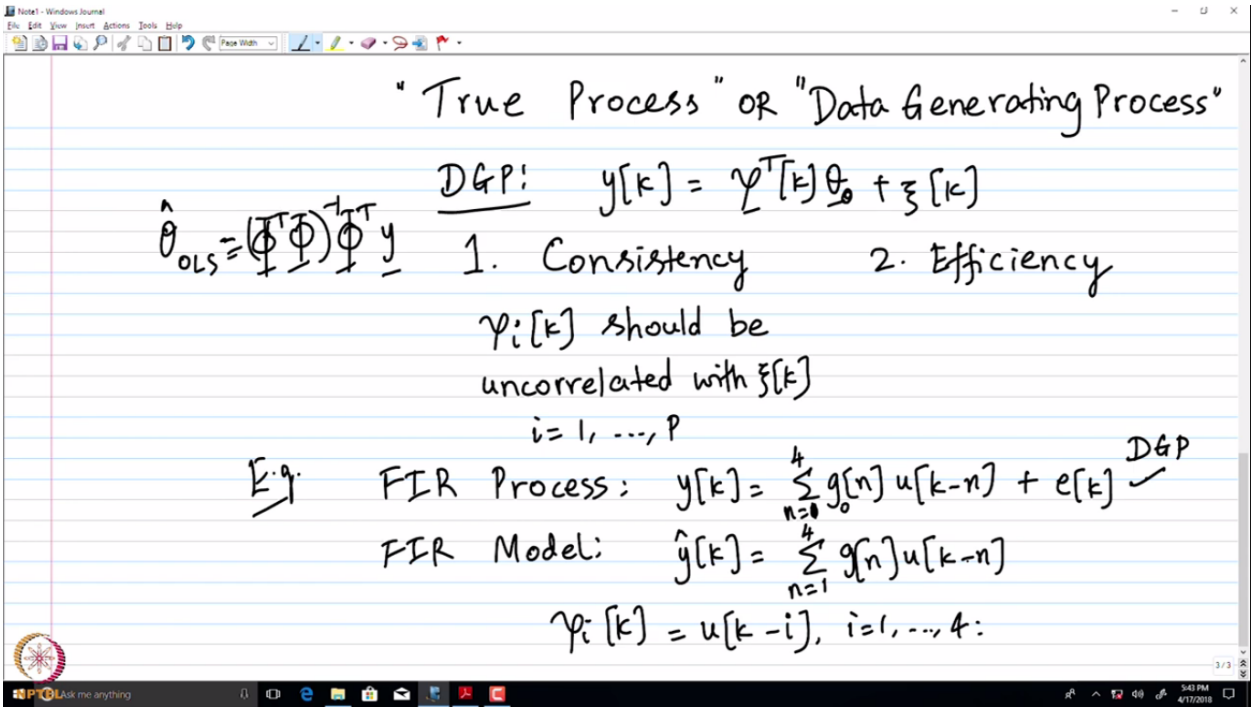
Plus let's say this is how it is the data is being generated, okay? And let's also assume that in my model I have fit for-- somehow let's say I know the delay and I go ahead, but I fit a third order for some reason, I've fit a third order model. So this is my model. Although use the same notation, please don't get confused here, alright. So let us say this is what I have fit.

(Refer Slide Time: 18:37)



The-- Remember in the FIR model, we assume a convolution model like this. Now, if you compare notes, what are the regressors here? If you just compare quantities, the regressors are simply past inputs, right? Now, first thing that we'll ask is let us consideredthe case of-- even a simpler case where I know exactly that the order is 4. Okay? I know exactly that the order is 4. Now what are the regressors that we have, uk minus 1, uk minus 2, up to uk minus 4, those are the regressors, right. So the psi i k is simply u k minus i and i runs from 1 to 4. Now what do you think? Do you think that, if the data were to be generated using this model here, this is your DGP. And if I fit a model like this will I obtain consistent estimates of g or not.

(Refer Slide Time: 19:49)

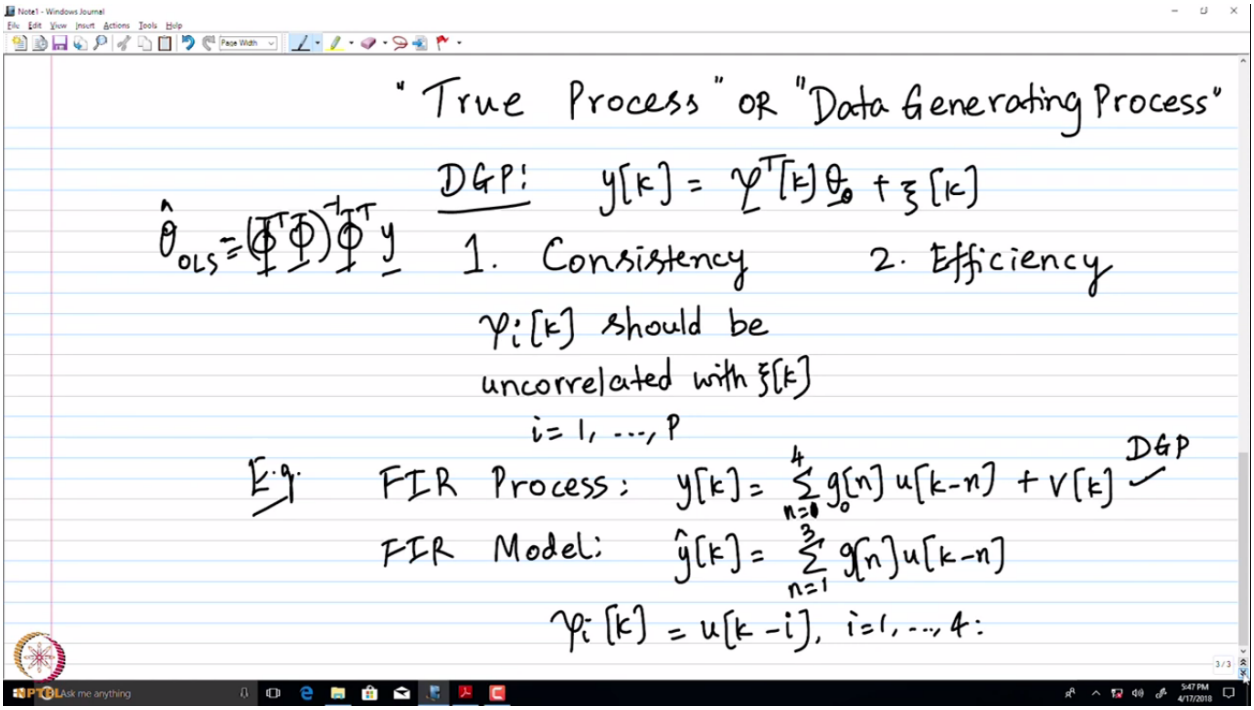


What is your thought? Yes or no?What do you think? No. Why is that? Where is error there? What is a z for us? So how do you compare what is z? How do you come to a conclusion as to what is z? You will-- in order to figure out what is z, you will have to look at your model also, right? Whathave you assumed in your model sigmai running from 1 to 4 g n uk minus n and DGP is also the same, but plus an e k. So the error zk is nothing but e k in this case. Right? Is that clear? What we have is with respect to the model, what you have left out is something that you should call as zk. With respect to the model, what is being left out? You have e k only.

Now what are the regressors that we have? u k minus 1 u k minus 2, so in fact, I can even remove this e k and assume a general colored v k also. I don't have to assume it to be white. So the regressors are the past inputs and z is v k. Are they correlated? Yes or no? So if you assume open-loop conditions, if you assume open-loop conditions then the inputs and vk are-- remember what is v k for us, right now, v k is everything that you would actually leave out in your model. Always remember that, it is not just-- although in your DGP vk appears to be a disturbance, but you have to look at it from the model's perspective as well. Anyway even either way we know that vk is uncorrelated with the past inputs therefore in this case you should expect consistent estimates of the FIR model.

Now suppose I made a mistake. Suppose I-- so when the model structure has been captured exactly, I know that I will get consistent estimates. Now, suppose I fit an FIR model of order three. Then what happens? Sorry? So now you can say that from-- what is about information that I have left out? The term corresponding to u k minus 4 plus v k, right?So when I am writing this model, what I have left out is g 4 uk minus 4 plus v k. So you have to look at it that way. Now what can you say about the correlation between the regressors and the-- whatever you are left out? Correct, it depends on the autocorrelation. If u k is white, then you'll get consistent estimates, right? So when you have this kind of a situation then the consistency depends on the nature of the input. And white input will guarantee consistent estimates. Even though you have left-- you have there is a model mismatch. Right?

(Refer Slide Time: 23:31)



So let us actually summarize that. Again I'll just write the DGP. So DGP is sigma g n u g0 of n true value u k minus n, n running from 1 to 4 plus v k. And this is y-- this is data generating process. And typically I can write here that the model that I'm model that I'm assuming is y kequals sigma n equal to 1 to 3, if I have made a mistake. Let's say this is a kind of mistake I made. And let's assume plus e k, because that is what amounts to saying y hat of k is that much, right? So in this case consistency depends. If u k is white, then consistent estimates are obtained of g 1, g 2 and g 3 are guarantee.

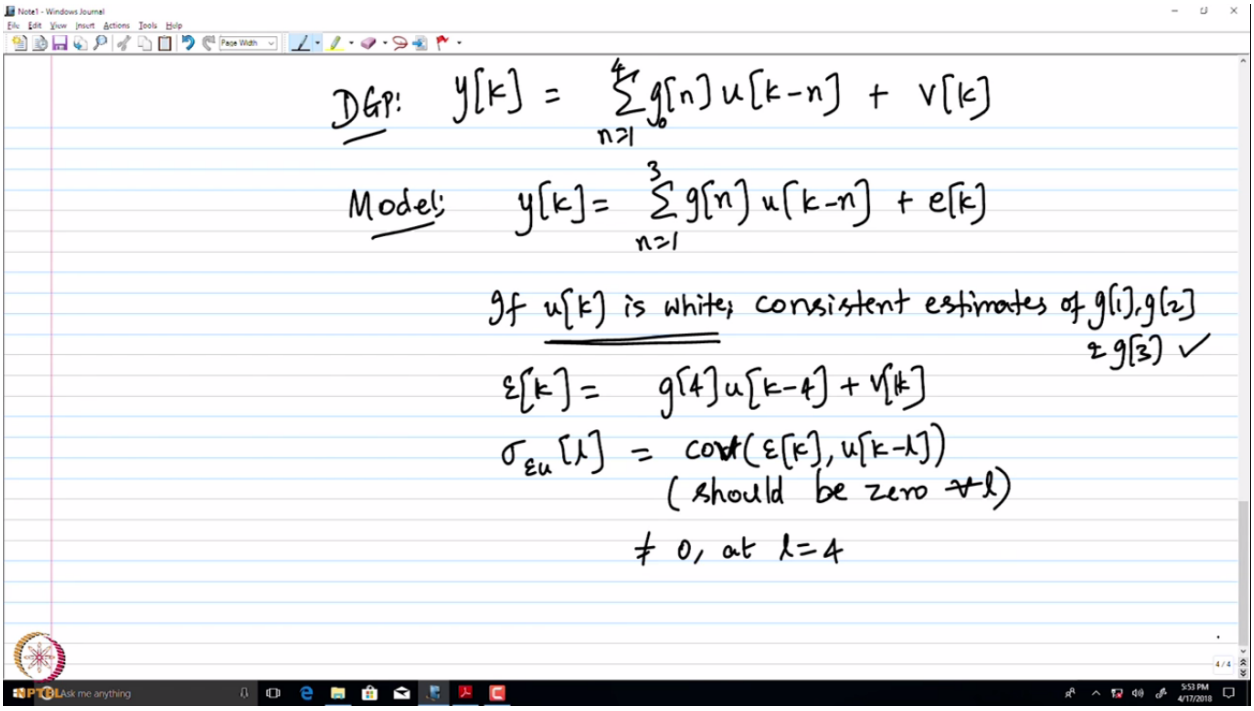
But-- so the next question is naturally, how will you detect that you made a mistake?The one thing that we noticed is, this white-- whiteness of the input is advantageous in some respects. If u k is colored, then what happens is whatever you have left out is correlated, with what you have included. That is how you should remember for least squares. Don't think in terms of equation errors and so on. The best way to remember is whatever you have left out in your model, if it is correlated with what you have included then consistency is not guarantee. Right?So white inputs have that kind of an advantage, but there are some other disadvantages of a white noise input, which we shall talk about later on, from an experimental viewpoint and so on. Fine

So now going back to the question that I asked a couple of minutes ago, how do you figure out that you made a mistake? How can you figure out? You have estimated, let us say, the coefficients. How do you go back and figured out first that you have made a mistake? Then you can refine your model. So there is no way residual analysis will reveal that you made a mistake, that your model is inadequate. But variants of residuals now will be variants of whatever v k plus the variance of what you've left out, right? So the cross correlation between the residuals and the input won't give you anything. Is that what you're saying? Are you sure, what you are saying? So what is left out here? If I were to write the prediction error, what would be the prediction error?

What is a theoretical prediction error?g4 uk minus 4 plus v k. Because that is, what is a truth, right? This is what is the theoretical predict-- one-step ahead prediction error. And in cross correlation, what do we do? When I look at the cross correlation between prediction error and the input, I'm evaluating the correlation between epsilonk and u k minus l, ideally if the model is good, what should be revealed? What should be-- sorry? Correct. So ideal-- if the model is not an under fit, then what should I see for the cross correlation block or the cross, there should be cross covariance. Or you can replace this with a rho if you like. But let us actually retain this as sigma, okay. So what should be the ideal covariance it should be 0.Right? Will you get a 0 at all lags?What'll happen?

Correct, so even though the input is white, you'll still be able to detect, that you've made you are under-fit. So the earlier assumption that,if the input is white I will not be able to figure out, it's not correct. So in this example sigma epsilon, so this should be ideally 0, should be 0, for all l. But it won't be, in fact this will be not 0at l equal to 4, again assuming open-loop conditions and so on. So which means a cross correlation will pick the inadequacy of the model and maybe even suggest that you have missed out the fourth order term. Right? If the input is white, exactly it'll peak at 4. Which is nice and then you go back and do it, okay, good. So hopefully now you've understood the consistency part of it.

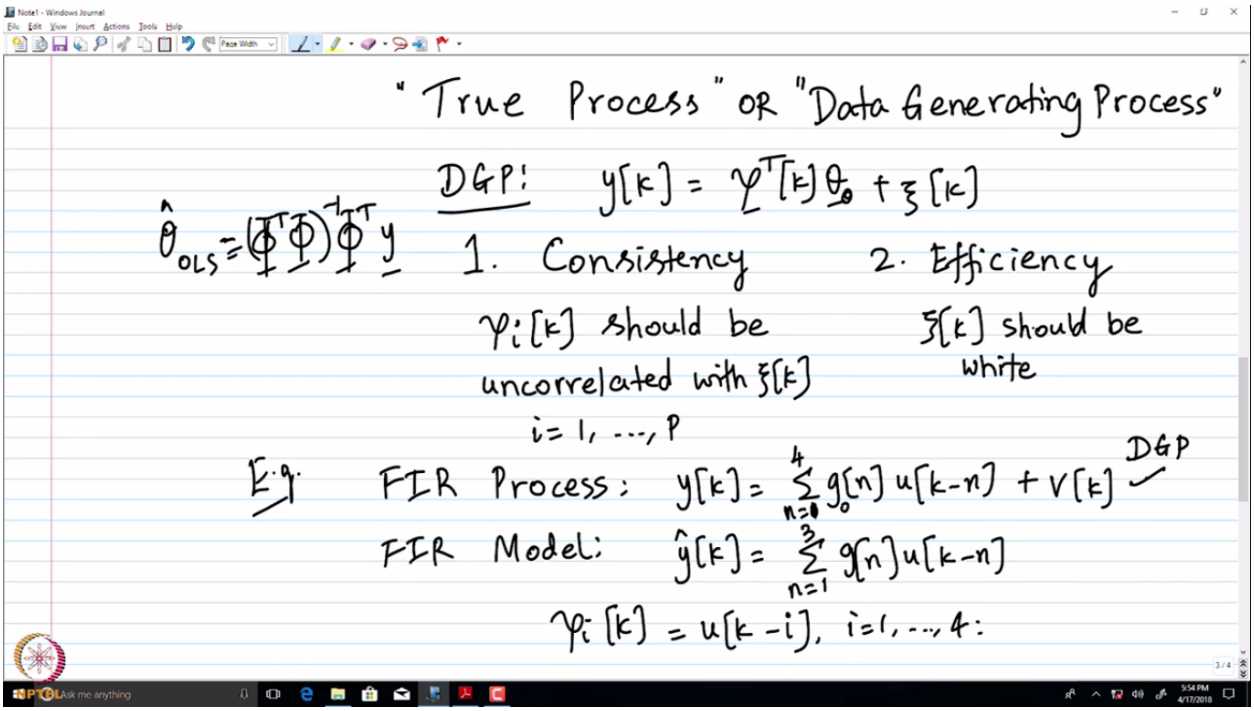
(Refer Slide Time: 29:01)



Always remember consistency is all about what you have included, whether it is correlated with what you have left out. Whether it is FIR model, ARX model, ARMAX model the story is the same. Now let's talk about efficiency. Recap efficiency. When is the efficiency guarantee? Do you recall? We talk of efficiency only after speaking of consistency, always remember. It should be in that order. If you are unable to recover the truth then don't even talk about efficiency. It doesn't make a sense anyways, correct? Okay, so what is the requirement on efficiency? What should be, z k should not be colored, exactly. In other words

the z k should be white. We will not add the other part that z i should be uncorrelated and so on, because it's understood, you'll-- unless you check for consistency, you will not actually check for efficiency.

(Refer Slide Time: 29:57)



Now, one point I want to make again, this should cross your mind. We--It must have crossed your mind when we were going through the least squares lecturesand I must have reiterated it. Spend some time on it, but I just want to reiterate this point that, in reality the data generating process is going to be much more complicated than what we have written on the screen, right. But then how do we figure out, if I have consistent estimates or not. Right, in reality the data generating process is not going to look this. It is not going to look this simple. The data it could be generated by even some locally linear part of a non-linear process. How do you guarantee that you have consistent estimates? That means how do you guarantee that, the conditions or consistency have been met? I keep saying this and I've said this even the liquid level system. If you recall, when I was reporting the errors of parameter estimates, I said don't look at the errors in the parameter estimates, which is what will help you talk about efficiency, unless you have ensured that the model is not an under fit.

The equivalent of that is what I have said just now, you should talk of consistent-- the efficiency only after ensuring consistency. In reality how do you check if the conditions for consistency have been met? That's all. So first thing you should check for cross correlation. The requirement is that, whatever you have left out should not be correlated with your regressors. Of course, we have discussed only the FIR model example, if I had applied the least squaresmethod to estimate parameters of an ARX model. In the ARX model what are the regressors? Past outputs and past inputs. Nevertheless we always look at correlation between residuals and inputs only. We don't look at correlation between residuals and outputs also. The reason is output themselves are functions of inputs, so I don't have to separately look at

correlation between epsilonthe residual and the outputs. It suffices to just look at the cross correlation between residuals and inputs, unless outputs are generated by some other way, which is not the case. So the important check for consistency is that the cross correlation between the residuals and the inputs should be clean. There should be no correlation.

Then you are-- you can assume that although the process is being generated by a complicated process, but you can still assume that locally it is being-- it can be represented this way. Otherwise, what is the meaning of theta 0in reality? In reality the process is non-linear. What is the meaning of theta 0? There is no meaning to it at all. Even though the process is non-linear, you assume that locally the process is being-- is operating in this way.Okay, so that is something to keep in mind. Because these are the things that come to our mind, when we start practicing system identification, you say, this is a theory that I learned. I know my process is coming from some complicated process. Then how do I know that these conditions are met? Okay. So let's-- now talk about efficiency. And let me ask you, now, will I get efficient estimates in this case? So in this example that we just discussed, will I obtain efficient estimates we've just checked that we can obtain consistent estimates, if u k is white. If u k colored then consistency is not guarantee, right?

You will have to go back and first improve your model. There, yeah, if you do it that way. Correct, that is why-- exactlythat is why consistency is guaranteed. Exactly, so if you were-- you're right. So if you were to-- the question is, if I were to correlate whatever is leftover which is a prediction error, with what I have included, then it will never show up, correctand that's why you'll conclude that estimates are consistent. And that is also pun intended consistent with what we have said, okay? That's an English consistent, but the second question that we ask is, how do I know the model is adequate? And that's when we've said, we look at in sys ID, this is now in sys ID. In a general sense you'll always look at the prediction errors and the regressors, the leftovers and the regressors and you ask if they were-- if they're correlated, if they're uncorrelated then you will obtain consistent estimates of whatever parameters you have including. But now in sys ID particularly, we know that the regressors are not independent they're actually-- they are being derived, they're children of the same signal.

They're just simply lagged versions. Then we ask the question, how do I know if my model is adequate? And that's when we said look at the covariance. But in a general regression framework, you will never know. In a general regression framework that each regressor is perhaps a different physical variable. In that context, I have to just-- I will just check, if whatever I have left out is correlated with what I have included and if the answer turns out to be nil, then I will conclude that I will obtain efficient estimates of, sorry, consistent estimates of what I have. Then we turn to efficiency straight away there. But even there the standard questions, how do I know my model is-- then they do a stepwise linear regression, if there are any other regressors that I know of then I'll included. Okay? This is one such example there are many examples, where you can run into problems with consistency. The classic example is ARX and ARAMX. And we'll talk about that again when we talk of parametric models. Since we have the non-parametric, well, I'm only centering the discussion around the non-parametric model, okay.

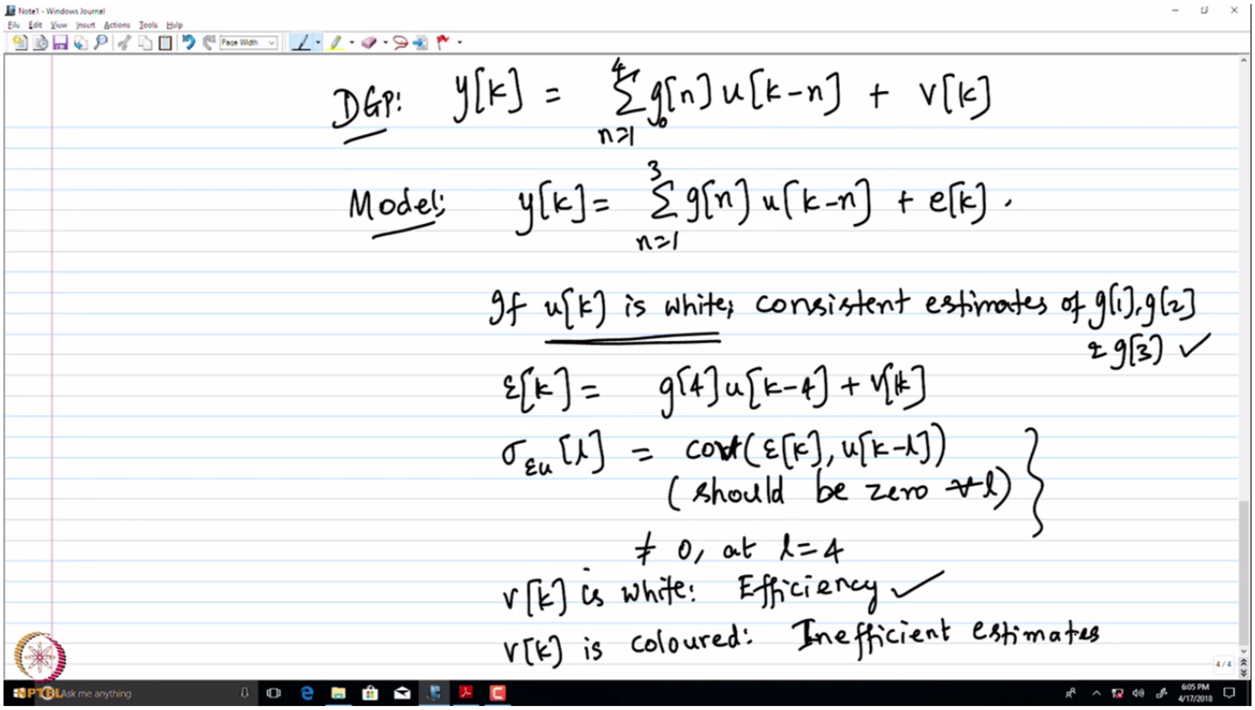
So do you obtain with this model that I have-- would you obtain consistence-- efficient estimates of g or not? So what will happen here, in the case of autocorrelation of epsilon? So you have to look at the autocorrelation of leftovers. Because the requirement for efficiency is that, whatever you have left out should be white. Right?Again it doesn't really talk about over parameterization and so on. That is not the criterion here. Remember in all of this, we are not talking of over parameterization here. So whatever I

left out should be white. Then I'll obtain efficient estimates of those. So what do you think, in this example will I obtain efficient estimates of g, g1, g2, g3? Yes or no? Depends on u k, if u k is white then only with this discussion is valid, right? Because then we have assumed that if u k is white, it's consistent, if it is not, I know that I've made a mistake. So I'll go back and correctand that is why the earlier discussion was necessary. This discussion here was necessary to correct your model. Now we'll have to, so there are two conditions, now.

One is on the input other is v k. If v k is white, then you obtain efficient estimates. If v k is colored then you won't obtain efficient estimates. Right, because whatever you have left out, even if you were to correct your deterministic model whatever you are left out will be v k. In-- And if v k is colored then you'll get-- so if v k white you will get efficient estimates. When I say efficient, fully efficient, efficiency is guarantee. If v k is colored then efficiency is not guarantee. We say inefficient estimates. That means there exists another estimator that can give you more precise estimates than what the least squares can give. Many at times a lot of students and beginners get confused between consistency conditions or consistency and efficiency. But hopefully this example has kind of clear-- clarified all such misconceptions that may exist in your minds. It's not easy to understand this, but once you have understood you know that consistency is all about correlation between what you have included and what you have left out. And efficiency is all about the nature of the left out totally, it doesn't look at what you have included.

But it is assumed that you have already done your consistency check and then you have come to efficiency, otherwise it doesn't make real sense. Any questions?Okay. So that is in a nutshell, how when you apply OLS to estimate FIR models, what happens and so on. And that is why you have to be careful, when you are estimating impulse response models, you should do a check, if you are only going to work with the impulse response model and you're not building a parametric model then you should pay full attention to the efficiency part. Whether you are going to use the FIR model as a stepping stone or as the only model, consistency check should always be done. Which means you'll have to check if you have obtained biased estimates. Essentially what does lack of consistency mean, you'll obtain biased estimates [39:57 inaudible]. You don't want to work with biased estimates of FIR model. Of course, the unwritten requirement is that you have also done a model adequacy check.

(Refer Slide Time: 40:07)



So in the order of checks, doesn't matter whether it's FIR model or any other model. One is you'll have to check for consistency. Right?Which includes sometimes a model adequacy check. So, two and then model adequacy, and three isefficiency. So these are the three checks that you need to perform. Before you really admit a model or accept a model. Now what happens is the test for model adequacy which is a correlation between the residuals and the inputs already subsumes check for consistency in some sense, right? Because what you are trying to do is you're trying to make sure that the model is adequate and so on, until you're happy with the mode, you will not even talk of consistency. Although I've written model adequacy as a second thing, it kind of overrides the consistency check. You want to make sure your model is adequate and so on. And then once you have the model adequacy, you can check further for efficiency.

As an example, we will take up the later in a-- at a later situation of fitting an ARX model to an ARMAX generated data. And you know what happens because there is a difference between the noise model-- modeling assumptions in ARX and ARMAX,you can run into consistency issues. But you may fit a very high order ARX model to match-- to compensate for the discrepancy between the noise model in which case you may be able to go past the model adequacy, but you may get inefficient estimates or you may be doing over parameterizationand so on. So the fourth thing of course you should check for is over parameterization or over fitting. Right? Consistency doesn't mean that you are not over fit. Since only means that as n goes to infinity, you will get 0MLE, the estimates will converge to the truth. Okay. So much about the OLS part, what we have learnt is that the OLS estimates are consistent. If the leftovers are uncorrelated with the included ones regressorsand they're efficient, if whatever you have left out are white.