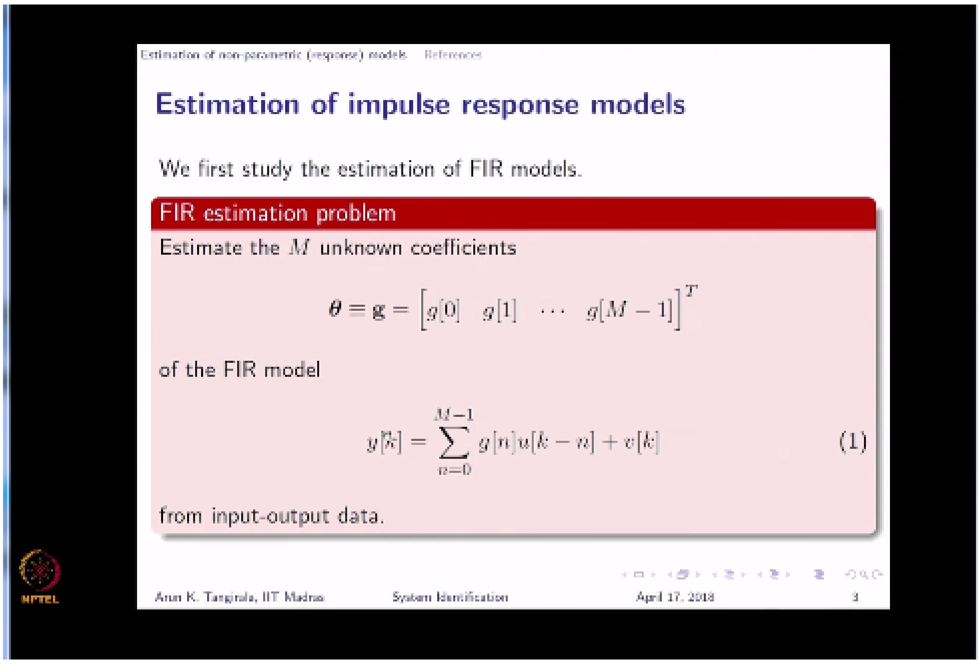
CH5230: System Identification

Estimation of non-parametric model

Part 3

Okay. Welcome to the lecture on estimation of non-parametric models. This lecture of course is a sequel to the previous lecture that we had where we recall the concepts of least squares methods, in particular, in the context of a FIR modelling. Today, we'll get into more technical details on estimating non-parametric models. Where we look at another method based on so-called method of moments of correlation methods. And we will also look at least square methods with regularization, which give us estimates with lowered error but at the expense of perhaps larger bias. So let's get started here and recap that the FIR model has this structure here.

(Refer Slide Time: 1:04)



Estimation of non-parametric (response) models   References

Estimation of impulse response models

We first study the estimation of FIR models.

FIR estimation problem

Estimate the M unknown coefficients

$$\boldsymbol{\theta} \equiv \mathbf{g} = \begin{bmatrix} g[0] & g[1] & \cdots & g[M-1] \end{bmatrix}^T$$

of the FIR model

$$y[k] = \sum_{n=0}^{M-1} g[n]u[k-n] + v[k] \qquad (1)$$

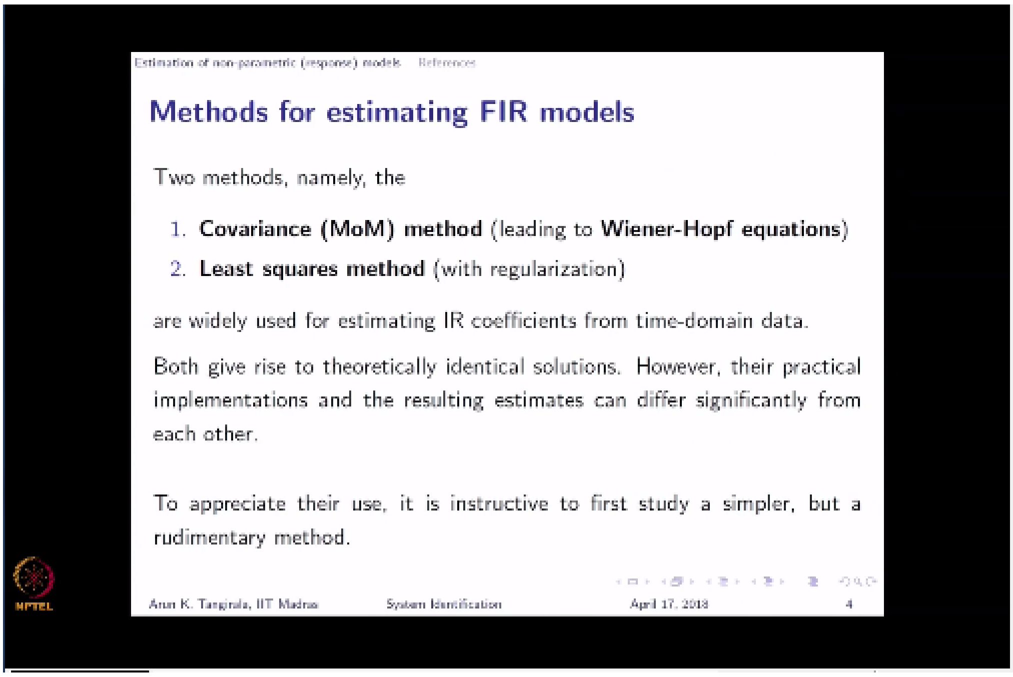
from input-output data.

Arun K. Tangirala, IIT Madras   System Identification   April 17, 2018   3

So if you recall from the discussions that we had on non-parametric models, this is what we have here. Y[k] is sigma g n u k minus n plus v[k]. This is what is assumed to be the data generating process. And we have given the observations of y and we have given the knowledge of the inputs. The goal is to estimate this impulse response coefficients which we have collected together in this vector called theta. So theta is unknown that we are going to estimate. V[k] is the standard noise that we have in y. That's the objective of estimating FIR models.

Now as far as the methods for estimating FIR models are concerned there are many but broadly speaking you can -- what you will find in literature are two methods.

(Refer Slide Time: 1:57)



Estimation of non-parametric (response) models References

Methods for estimating FIR models

Two methods, namely, the

1. **Covariance (MoM) method** (leading to **Wiener-Hopf equations**)
2. **Least squares method** (with regularization)

are widely used for estimating IR coefficients from time-domain data.

Both give rise to theoretically identical solutions. However, their practical implementations and the resulting estimates can differ significantly from each other.

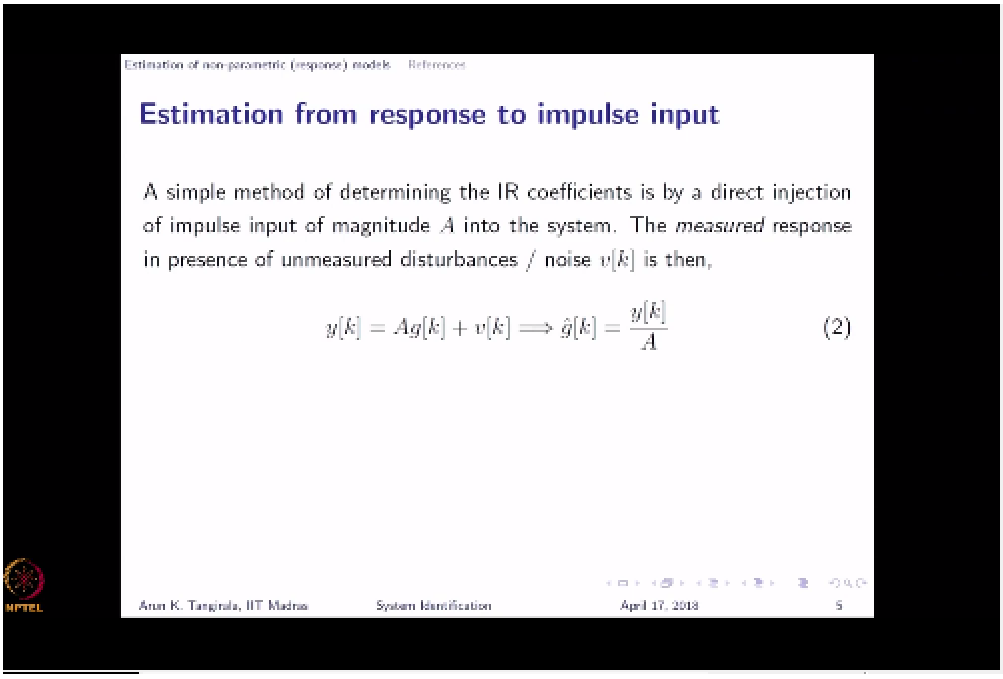
To appreciate their use, it is instructive to first study a simpler, but a rudimentary method.

Arun K. Tangirala, IIT Madras System Identification April 17, 2018 4

One is the covariance method also known as the method of moments method or correlation method and the other is the least square method with and without regularization. And the other thing that I should draw your attention to is when we talked of Cramer-Rao inequality. In the textbook as well there is an example on estimating impulse response coefficients efficiently where we don't have the regular objective function such as minimizing prediction errors and so on. But we have the objective of estimating the unknowns efficiently. That is the impulse response coefficients efficiently. And it turns out that under certain conditions that is when the errors are white, that is v[k] is white and Gaussian distributed then the most efficient estimator is none other than the least squares estimator. So that way least squares methods have certain advantages.

So let's actually go to the crudest way of estimating an impulses response coefficient which will then allow us to appreciate the other two methods that is the covariance method and least squares method much better.

(Refer Slide Time: 3:13)



Estimation of non-parametric (response) models References

Estimation from response to impulse input

A simple method of determining the IR coefficients is by a direct injection of impulse input of magnitude A into the system. The *measured* response in presence of unmeasured disturbances / noise $v[k]$ is then,

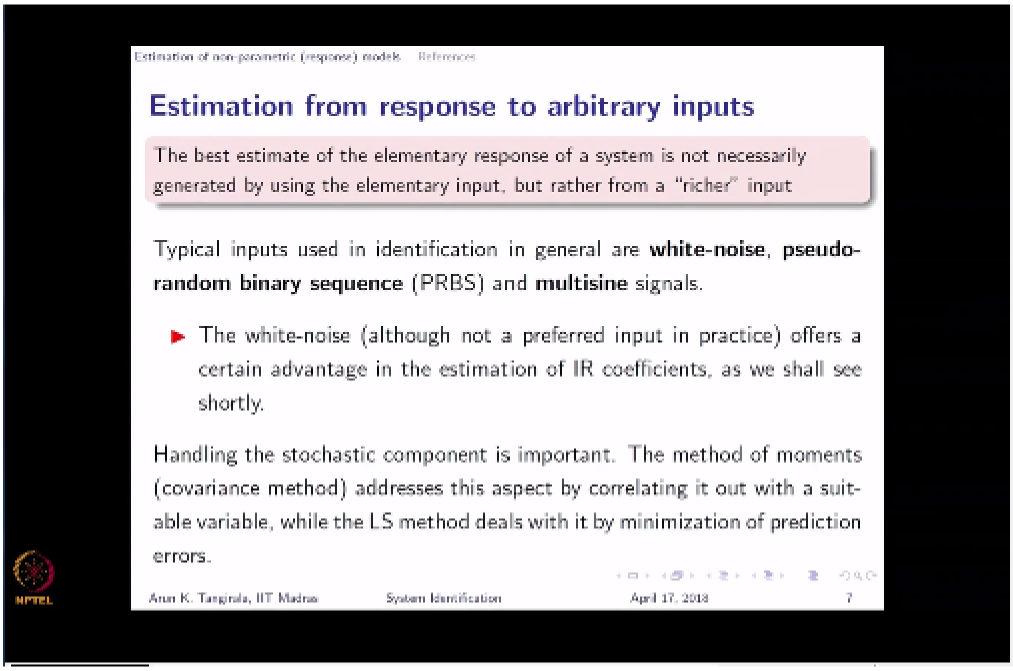
$$y[k] = Ag[k] + v[k] \Longrightarrow \hat{g}[k] = \frac{y[k]}{A} \qquad (2)$$

Arun K. Tangirala, IIT Madras System Identification April 17, 2018 5

So the simplest way of estimating an impulse response coefficient is to go by the definition. Remember an impulse response coefficient is nothing but the response to an impulse input. It's very simple. So all I can do is go by the definition. If I'm given access to the experiment then I go ahead and inject a discrete time impulse which is physically realizable. And I do that. And the response that I get is y[k], I mean, if I assume that there is an amplification by or attenuation by a factor of a. So y[k] is Ag[k] plus sum v[k] where A is essentially the magnitude of the impulse. Okay? When I say, amplification, attenuation here the impulse response coefficient is g[k], could be amplified or attenuated by the magnitude of impulse that you inject. For the unit impulses A is one. Now a crude estimate of g because you're looking at a single point is simply y[k] by A, right? I mean, I inject input of magnitude A, I obtain a response y[k], y[k] by A will give me a crude estimate of g[k]. Obviously, if you look at the error in this estimate, it is v[k] by A, right? That is what it is.

Therefore, small error requires large A. There is no doubt about it. If I want to reduce this error since I have no control on v[k] the only control I have is on A, I need to give in large amplitude inputs but if I do that we do run into the risk of pushing the process into nonlinear resumes. Remember, we are building linearized models. We want to keep the process as much as possible in the linear regime. Therefore, in practice, this method has very low utility. That's simply because you're relying on a single observation. And also it has got to do with the nature of the input. In general, a better way of estimating the impulse response coefficients is to use a non-impulse input and a more richer input.

(Refer Slide Time: 5:19)



Estimation of non-parametric (response) models   References

Estimation from response to arbitrary inputs

The best estimate of the elementary response of a system is not necessarily generated by using the elementary input, but rather from a "richer" input

Typical inputs used in identification in general are **white-noise**, **pseudo-random binary sequence** (PRBS) and **multisine** signals.

- The white-noise (although not a preferred input in practice) offers a certain advantage in the estimation of IR coefficients, as we shall see shortly.

Handling the stochastic component is important. The method of moments (covariance method) addresses this aspect by correlating it out with a suitable variable, while the LS method deals with it by minimization of prediction errors.

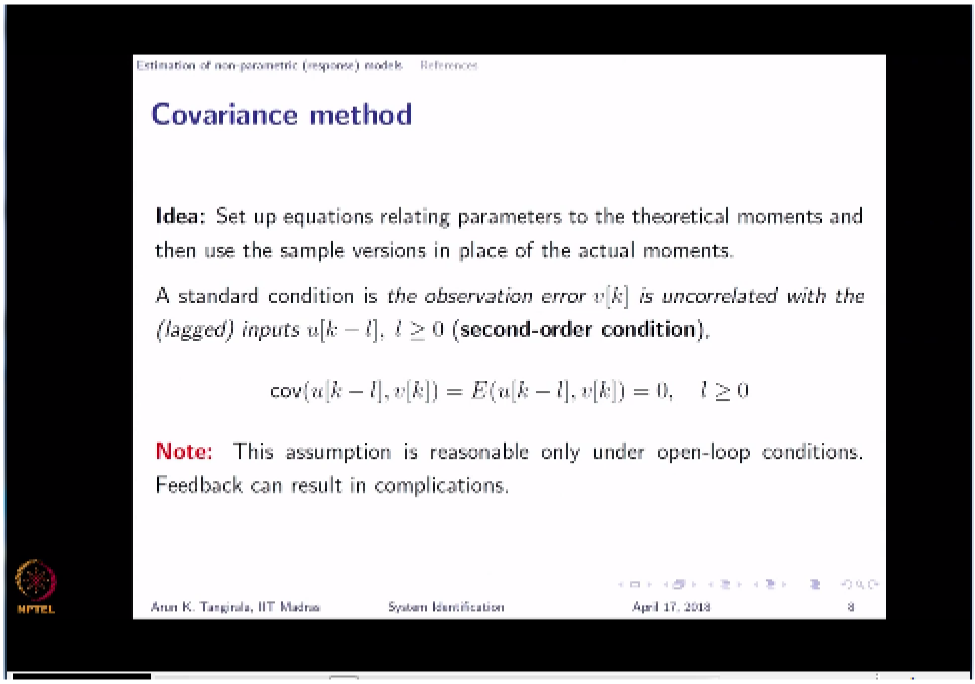
Arun K. Tangirala, IIT Madras   System Identification   April 17, 2018   7

The impulse input is somewhat weak in terms of frequent excite -- the signal to noise ratio, in terms of excitation. It doesn't have that features that other inputs such as white noise or pseudo random binary sequence that you've been using in your assignment and your home works or multi sine signals. So the difference is in this approach is we don't go by the definition per se, we use a different input. And then from the data we infer what is the impulse response coefficients. This is an indirect way, you can say of estimating the impulse response coefficient but it is worth it. Although you are putting in an input that is not really straightaway commensurate with the definition, it's still worth using this input because with this data I can not only estimate impulse response coefficients. But I can estimate step response, frequency response, parametric models and so on. So there is a wealth of information or a host of models that I can build with this data. And typically that is the practice. Among these inputs, a white noise is preferred theoretically in many ways but as I mentioned in the previous lecture it may have certain disadvantages from an input design viewpoint which we will talk about a bit later.

The most important thing that you should observe is handling the stochastic competence that is what is calling for a different input. If there was no noise, I could have just simply injected the impulse and obtained a response and that it, I am done, right? The other thing also to remember is that I may not have access to an experiment all the time, in which case I may have to work with some arbitrary input output data. So from both these viewpoints it's good to know, how to estimate impulse response coefficients from data that is generated by non-impulse like input, so that is the motivation with which we proceed forward.

As I said earlier broadly speaking you have two methods. The Covariance method and the Least squares method.

(Refer Slide Time: 7:20)



Estimation of non-parametric (response) models   References

Covariance method

Idea: Set up equations relating parameters to the theoretical moments and then use the sample versions in place of the actual moments.

A standard condition is *the observation error $v[k]$ is uncorrelated with the (lagged) inputs $u[k-l]$*, $l \geq 0$ (**second-order condition**).

$$\text{cov}(u[k-l], v[k]) = E(u[k-l], v[k]) = 0, \quad l \geq 0$$

Note: This assumption is reasonable only under open-loop conditions. Feedback can result in complications.

Arun K. Tangirala, IIT Madras   System Identification   April 17, 2018   8

In the covariance method the idea is very simple, you go back to the equation that you have the convolution equation. Remember, the main impediment for us is noise. So I want to get rid of noise. As usual, in estimation, we want to get rid of noise or reduce the effects of noise. So what do I do? I correlate y with some variable which actually kills the effect of noise, right. That is the basic idea. In this case, if I assume open loop conditions, I can simply correlate y with the input. That is on both sides of the equation I multiply with u[k minus l] and take expectations theoretically. So that the second term which will have the correlation between v[k] and [k minus l]will vanish, under open loop conditions that disturbance and input are uncorrelated. This idea will not work under closed loop conditions obviously because the disturbance and input are going to be correlated. But we'll restrict ourselves to open loop conditions for now. Under open conditions therefore the idea is to compute the covariance between y[k] and u[k minus l], so that the second term which is the covariance between u[k minus l] and v[k] goes to zero. That is the idea. Okay. As a result, we are presented with this set of equations that you see in three.

When I say, although there is a single equation, the bunch of equations come about by choosing different values of l. And since we have M unknowns, you setup M equations, right? So the equation 3 is now more nicely and elegantly written as an equation 4 which is also known as a Wiener-Hopf equation, crediting to both Wiener and Hopf who came up with this idea independently. So the basic idea is that the cross covariance is also the convolution of the impulse response coefficients with the auto covariance or you can say, the auto covariance of the input with the impulsive response. We have seen this relation earlier also. We have studied this relation where we have said that the cross covariance for a stochastic or even for a deterministic process, the cross covariance and auto covariance have they are related through the same system as the output and input. That is the idea here. The nice thing is now the noise is missing. The noise is absent in this.

Of course, in reality what do we do? How do we use this equation? We replace -- so this is a theoretical equation. Now we -- when we replace the theoretical covariances, the auto and the cross covariances with the respective estimates then we are applying the method of moments philosophy idea here where in the method of moments if you recall we assume that the sample moments satisfy the same relation as the theoretical ones, right? So that is the basic idea in the covariance method. I'm just continuing with the theoretical equation where I've replaced the auto covariance now, the auto covariances at different lags and so on and written now in a matrix form, working out for all lags you get now sigma phi psi times theta equals sigma Y where this Y is the vector of stacked outputs. So sigma Y psi and therefore you get an estimate which looks like this, theta hat equals inverse of sigma phi psi times sigma Y psi.

And this of course, looks pretty similar to what you see in least squares. In fact, the least quest method also tries to work on the same principle orthogonality of the residuals with the regressors. Here we don't use the term orthogonality, we use the term uncorrelated but the idea is more or less the same. Here we explicitly enforce the orthogonality whereas at least squares the property itself comes out by virtue of the least squares approach, right. Now the other thing to remember at this point in time it's good to know that when the -- So let's actually go back to this equation y[k] equals sigma g[n] u[k minu n] plus v[k]. In the covariance method we correlate y with u[k minus l] so as to kill the effect of noise v[k]. But suppose you're looking at some situation where the noise is correlated with the input. It need not be on the closed loop conditions, under some other conditions also it's possible that the noise can depend on the input. For example, if the sensor noise depends on operating conditions. We know that certain sensors have this behaviour that the level of noise depends on operating conditions. If the input is not at some extreme values then the sensor level noise level is different from those when the input is actually at the extreme extremities of the sensor and so on.

So assume for now that you are looking at a situation where the input is correlated with noise or noise is correlated with the input. In which case, the method that we just used of correlating out the noise by with the past inputs is not going to work because the second time is not going to cancel out. In such cases you may have to actually work with some other variable, we don't know what those variables are. Let's call them as, let's say some eta. Eta [k minus l] and let's actually write down the equation, resulting equation we have u[k minus n] covariance between this plus expectation of v[k] times eta [k minus l]. Going by what we did earlier, Etas were simply inputs. But we have chosen now to correlate with a different set of inputs because we are looking at a scenario where v[k] is correlated with the inputs. Now Etas are some auxiliary variables. Okay, so what are Eta's? Eta's are some auxiliary variables.

What condition should Eta satisfy? The idea is to kill the effect of noise, right? First and foremost what are the conditions that you can think of that Eta should satisfy? Remember I want to kill the effect of noise and I want to estimate g. There are two conditions that Eta should have at least. So what conditions should we place on choosing Eta. Eta is something that you're free to choose. But how would you choose them. What conditions would you place them on? If you choose Etas as input themselves then it's not necessarily going to satisfy the first requirement that it has to kill the effect of noise. Correct. Under open loop conditions and where situations where input and observation error are uncorrelated then it's okay. But in a general scenario, what conditions would be impose on Eta? Any idea? Uncorrelated with v, correct. So the condition is that 1, uncorrelated. Eta should be uncoordinated uncorrelated with v[k]. Good. And then is there any other condition? Is that enough? So Eta should also be strongly correlated with u. Why is that required?

You can't estimate g otherwise. What would happen is, the matrix that if you write this in a matrix form, the matrix of the covariance between u and Eta will be rank deficient. Can be rank deficient. So Eta should also be correlated, so let me write here. should be correlated with u. It's a tough requirement. It's more like a situation of a newly married husband, should be actually sandwiched between two people, right? Or it could be a newly married woman also say, sandwiched between two people. So while it has to be uncorrelated with noise, it has to be strongly correlated with input. How do you actually solve this anguish like situation?

If I choose Eta to be input then it satisfies condition number 2 but it may not satisfy condition 1. If I choose Eta to be some white noise, some realization of a white noise sequence then it satisfies 1. But it may completely dissatisfy 2. So essentially this Eta should be chosen in such a way that it is some noise free version of u or noise free version of y. Because what these two conditions tell us is Eta should essentially be devoid of noise but also be very strongly positioned with respect to input or output. And there are several methods of choosing this. We will not go further into detail at this point. But what you should know is this entire idea of choosing to correlate with some bunch of new auxiliary variables such that this auxiliary variables satisfy the two conditions is known as the instrumental variable method. Instrumental Variable Method. And it has a nice abbreviation, IV.

And the IV methods resonated initially from econometrics and then gradually percolated to engineering. But these are very popular to obtain unbiased estimates. Remember, we talked about consistency yesterday. If I use the covariance method that we just talked earlier where Etas are simply inputs and it so happens that for that particular application input and noise are correlated then you will obtain biased estimates of G. So to obtain unbiased estimates of G, the instrumental variable method is very handy. It is superior to least squares in that respect because it explicitly now chooses the correlating variables and Etas are called the instruments. So these Etas are now called the instruments.

It's like an instrument that you would use to remove noise. And maybe that's the reason why this name has been coined. Although I have presented IV method in the context of FIR models, IV methods can be used in general for estimating parameters of other model structures as well. We will again invoke this idea for example in the estimation of ARX models. Where the data generating process could be ARMAX, but I still want to fit an ARX model and I want to obtain unbiased estimates of the plant polynomial coefficients, the numerator and denominator in which case I can use the IV method to obtain unbiased estimates. If I use the least quest method then I'll obtain biased estimates and we'll talk about that when it comes to ARX modelling.

It's good to be familiar with the IV methods for many reasons and one of the practical reasons is the SysID tool box. That has a routine are estimating parametric model such as output error, Box-Jenkins and so on, particularly for estimating models that give rise to non-linear predictors. They need the initial guesses and these initial guesses are generated through IV methods. So if you read the documentation on the OE routine or the BJ routine and so on. There is a section which talks about the initialization and there it talks about a method called IV4. IV4 is a variant of IV method. Right. I don't think that four times the IV method is implemented. But there are four stages to it and therefore it's called IV4. The basic idea is this and then some further embellishments on this. Okay. So let's get back to our discussion.

Fine, so we will again confine ourselves to the situation where v is uncorrelated with u and therefore you end up with this kind of an estimate. And this is just a further expansion of what I've just said. You're right g hat or theta hat in this way. And I've given how you set up the regressors. Essentially the psi and phi matrix and the vector y and that's it. And notice one thing that although you have n observations with you to begin with, you end up using only N minus M observations, right. And the reason is that you can set up your equations only starting from M onwards up N minus 1. Our index runs from K equals 0 to N minus 1. And that is why effectively you have only N minus M observations and that can be a problem if you're estimating very high order FIR models. See, suppose I have 1,000 observations and my process is a very slowly decaying one. So that it may require 200 FIR coefficients. Remember, FIR coefficients rests on the assumption that the impulse response decays beyond a certain point, right? If the impulse response coefficients or decay only after M equals 200 then effectively you have only 800 observations which is significant loss in the degrees of freedom that you have. The degrees of freedom is essentially the effective number of observations that you have for estimating the unknowns. So for higher order FIR models you can end up with -- from the estimation theory what do we learn? That the variance of any parameter estimate is inversely proportional to the number of effective observations, right, that are available.

So when you are estimating slowly decaying processes or high order FIR models then this approach can lead to not biased but inefficient estimates. That means the errors can be larger. How do you resolve that is what we do in regularization. And I will talk about that a bit later. So you have to be careful of these special situations. One special situation is, what we talked about when v[k] could be correlated with the input. The other special situation is that your FIR model is of very high length. And that can happen either if your sampling interval is very small, that means you're sampling very fast and although it is decaying off in three seconds your sample's so fast that you require a long length. But in that case you may have a large number of data. Putting it the other way round, if the FIR model length is compatible or is a significant fraction of the number of observations then you have to be watchful. You have to be watchful of the errors that you incur in the estimates. Because effectively you have only N minus M observations. Any questions? Okay. So we'll proceed. We'll come to that reduction in variance or error a bit later when we talk of regularization.