CH5230: System Identification

Fisher's information and properties of estimators

Part 08

So to quickly summarize what we have learned and the properties of estimator, first we have learned this concept of Fisher's information which tells me how informative the data is. Then we turn to estimators, where we ask if the estimator is accurate, precise, if it is precise, is it the most precise among-- we know that, we cannot have variance0 for finite samples. But among all the given estimators, I would like to pick the most efficient one. And that there is this notion of mean square error, which strikes-- tries to strike a tradeoff between bias and variance. Those are all finite sample properties. Then you have asymptotic properties among which consistency is a key property. Ultimately we want to work either with a minimum variance unbiased estimator, well that's for sure, but also consistent estimator. So we want consistency and efficiency. And then finally once I have estimated, I conduct hypothesis test by constructing confidence regions and for constructing confidence reasons I need to know the distribution of theta hat. Once I know the distribution of theta hat I can construct a confidence region, from the confident region I can conduct significance test. So this is a story of your properties of estimators. Now, we should move on to the methods of estimation, it's high time that we learned how to estimate. And again the video lectures will give you full details on the methods of estimation. Basically if you look at the different methods of estimation, you have fourdifferent methods that we discuss in this course. Which are also the fourdifferent ones that you'll encounter in the literature, broadly speaking, right? One is the methods of moments. So you have fourdifferent ones. One is a method of moments, we called as MoM, and two is a least squares, and three is your maximum likelihood estimators, and four your Bayesian. So these are the fourdifferent approaches to estimation. And for every method that they estimate obtained from every method, we have to ask the questions that we have just asked. Which is that, whether the estimators are biased, whether they are most efficient, whether they are consistent, and what is the distribution of theta hat? I have to ask four questions.

You should remember that, right? In fact you can think of five. One is whether it is biased. The estimators are going to be biased. Two, what is the variability, what is a variance of theta hat? Where I ask a sub question, whether it is the most efficient and then what else do I ask. Consistency, right?Followed by distribution of theta hat.So these are the standard questions, standard things that you have to ask for any estimator. Tomorrow you may come up with your own method of estimation. You have to ask the same questions. Otherwise nobody will use your estimator. Or somebody has to do that for analysis for you. Now, the video lectures explain to you what is the philosophy behind four-- thisfour different estimations: method of moments, least squares, MLE and Bayesian. Of which MLE I have already explained the principle in the class when we were discussing Fisher's information. There is nothing much to know about MLE as far as the underlying philosophy is concerned. Once you have understood the notion of likelihood, you have understood almost what the MLE procedure? But what needs to be known is how to set up the likelihood? We have dealt with some basic very simple problems, where always the errors are white, but the errors need not be white. Our observation errors can be colored. That is when you have ARX models and so on. How do you set up the MLEproblem for estimating parameters of an ARX model and so on? Is something also discussed in the video lectures, but I'll show you later on as well, very simple example. The method of moments also known as like the correlation method sometimes, when you use second order moments, rests on a very simple philosophy.

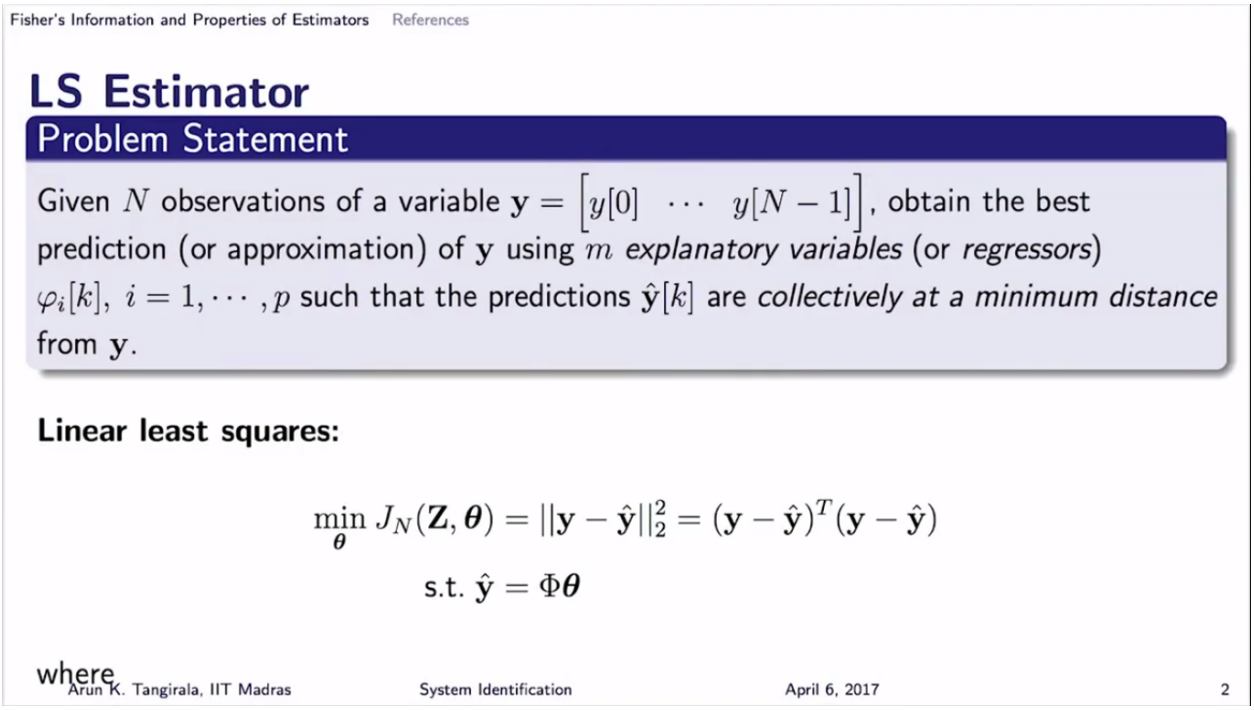
It says that there is a relation between the parameters of interest and the moments of the PDF. Alight, which means straight away it assumes the data to be random. That means the observations to have randomness in them. It sets up as many equations as a number of parameters. As you will learn in the lecture, video lectures. And assumes that the parameter estimators, that is onceI replace the theoretical moments with the sample moments, what do you mean by replacing theoretical moments with sample moments? So the more simple one as I can give you, very quickly is, suppose I want to estimator the mean of some data. We know that, so the parameter of interest is mean. We know that the mean is related to directly the first moment of a random variable Y. This is what we mean by relating the parameters to the moments. This is very simple,by definition mean is a first moment. I want to estimate mu.

What does method of moments assume? Suppose I replace mu with mu hat that means the estimate. That estimate satisfies this equation. So I say mu hat, which is on the right hand side what do I do, I replace, this is the theoretical moment. This called a population moment. I replace a population moment with sample moment. What do we mean by sample moment? This is aaverage across outcomes. Sample moment would mean averaging across time. So straight away method of moment gives me this expression. There is no optimization or anything. Of course, the very simple thing, but you should understand. What I have done is,I have replaced the theoretical moments with simple moments. So tomorrow if I want to estimate variance, whatdo I do? Variance expression would be if I-- for any random variable sigma square is expectation of y square minus mu square. Suppose I want to estimate variance given mean, I'm given mu then what do I do? Mu is known. According to method of moments the estimate of sigma square satisfies this equation. I'm going to replace the theoretical moment with the sample moment. But we say theoretical moments that means, the statistical averages are going to be replaced by time averages.

So I always write these equations, which relate what, parameters to moments. And I can write any equation as long as the parameters of a Gaussian PDF, what are the parameters of a Gaussian PDF? Mu and sigma square.I don't have to necessarily write the first and second moment relation. Here I have written that, but I can write the fourth moment, see third moment is 0, for a Gaussian PDF.I can write the relation for the fourth moment. The fourth moment also relates the moments to sigma and mu. I can use that relation also. Which means method of moment is not unique. I can-- as long as I have a bunch of equations that relate the parameters to moments, I'm fine, I'm through. And then what do I do? I replace theoretical moments with simple moments and then replace theoretical parameters with their estimates. Solve those equations and get the estimates. Now all-- when I say moments here, of course, I've written here for a single random variable, when you have more than one random variable, when you have let us say two random variables, then you can think of correlations, you can think of second moments being correlations. So we will talk about these correlation estimators a bit later, which are also method of momentsestimators. Where we demand that the-- again we write equations relating parameters to correlations straightaway. And require and make and require something's then you will get what are known as method of momentsestimators.

Now the beauty with this method of momentsestimators, I don't need to know the PDF necessarily, the form of the PDF, right? Those equations that I've written on the board are true for any PDF. But they are not necessarily most efficient all the time. And they are not necessarily unique. There are of course improvements or method of moments and so on will not talk about that. There is something called generalized method of moments and so on. Instead, we turn to the two big datas in estimation theory, parameter estimation methods, which are the least squares and MLE. Both give you efficient estimates and MLEwill give you in general efficient estimates in the large sample cases. Least squares will give you the efficient estimates only under some conditions. Many people use least squares methods without even knowing whether the conditions for obtaining efficient estimators are satisfied for the data. In fact least squares methods can give you inconsistent estimates. What does it mean? As n goes to infinity, the estimates will not converge to the truth. Under some conditions, so you have to be sure that your data, the way you're fitting everything meets the consistency requirements. You cannot simply use blindfolded latest squares methods. Okay. Which is what of course we'll review tomorrow quickly with examples, but I just want to show you the basic least squares formulation than we'll adjourn.

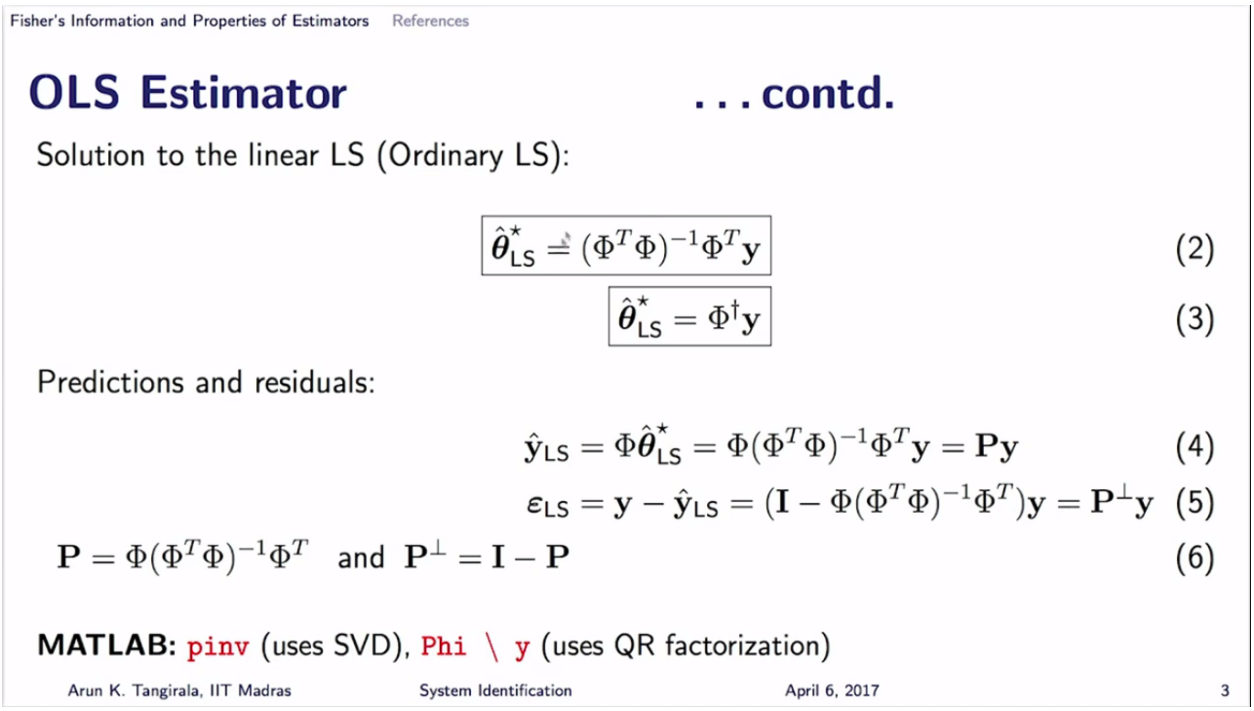
(Refer Slide Time: 11:52)



So this is your least squares estimation problem. I'm given n observations and of some variable y, let say the output. And I have a m explanatory variables. What do we mean explanatory variables? The variables that I'm going to use for predicting y, for explaining y. This-- there is a nice term called explanatory variables, in regression these are regressors. But there is a subtle difference between these two we'll not worry about it right now. So I have in fact not m, I'm sorry here, we'll assume that there are p, I'll correct that, p explanatory variables or p regressors. And least squares problem states that, I would like to explain this variable y or n observations, not a single observation. But all the n observations using these p explanatory variables, such that collectively the vector y had which we call as a prediction is at a minimum distance Euclidean, squared Euclidean distance from this vector y. So we're talking of collectively being close. We are not focusing on specific observations. When I-- so which means here I am not giving undue importance to some observations and more import-- unduly more importance to one set of observations and less to others. I am giving uniform weightage and this is called ordinary least squares. When we give more importance to one set of observations than the other ones, then we have weighted least squares. So within this problem you have linear least squares and non-linear least squares. Again within weighted least squares you will have weighted linear least squares, weighted non-linear least squares. What make something a linear least squares problem? When I choose to work with a linear predictor.So when I say y hat is a linear function of the p explanatory variables then I have a linear least squares problem.

So what is a linear least squares problem, it is a standard optimization problem here. I've discussed in the video lectures how to solve this problem. I'm just giving you the problem statement. Remember now we are minimizing the squared to norm of y minus y hat. In this statement, have we assume y to be random anywhere? Have you assume y to be random anywhere? We have not. We are just saying there is a vector y and I want to approximate that vector y in a-- with a p dimensional space, in a p dimensional space, right? So y hat can be simply thought of as an approximation problem, here. Constructing this best way hat is nothing but a functional approximation problem. It's not necessarily random. So you should remember that least squares problems need not be introduced always in the statistical framework. I can think of this as a functional approximation, vector approximation, whatever you want to call it. All I'm doing is,I'm minimizing the squared Euclidean distance. Many statistics textbooks will present it in a different way. Where y is introduce as a random vector. Okay, here, this is called a sample least squares problem. So y hat is written as a Phi theta, where you Phi is your matrix of regressors, I'll talk about it tomorrow.

(Refer Slide Time: 15:22)



But you should be familiar with the solution, the least squares solution. Phi transpose Phi inverse, Phi transpose y, y is your vector, Phi is a matrix of regressors. This matrix of regressors will change from problem to problem, depending on how you choose your explanatory variables. Tomorrow when we come I'll show you, what Phi is for FIR model estimation? What Phi for ARX model estimation and so on? We've already talked about it, but we'll work out examples in MATLAB, okay?