

**Introduction to Research**  
**Prof. Arun K. Tangirala**  
**Department of Metallurgical and Materials Engineering**  
**Indian Institute of Technology, Madras**

**Lecture – 3 A**  
**Modelling Skills**

Welcome to the course, once again, on Introduction to Research. In this lecture, we are going to learn certain aspects of modelling; in particular, data driven modelling. So, the lecture is titled as Modelling Skills, but it's not just about skills that we will discuss; we will also discuss what types of models we have, and what are the different ways of developing a model, and so on.

(Refer Slide Time: 00:44)


Modelling Skills

## Objectives

To learn the following:

- ▶ What is a model?
- ▶ First-principles (mechanistic) vs. empirical models.
- ▶ Systematic procedure for building models from data.
- ▶ Few critical aspects of data-driven (empirical) modelling.

With two hands-on examples in R<sup>®</sup> (a popular software for data analysis) ...

 NPTEL

Arun K. Tangirala, IIT Madras

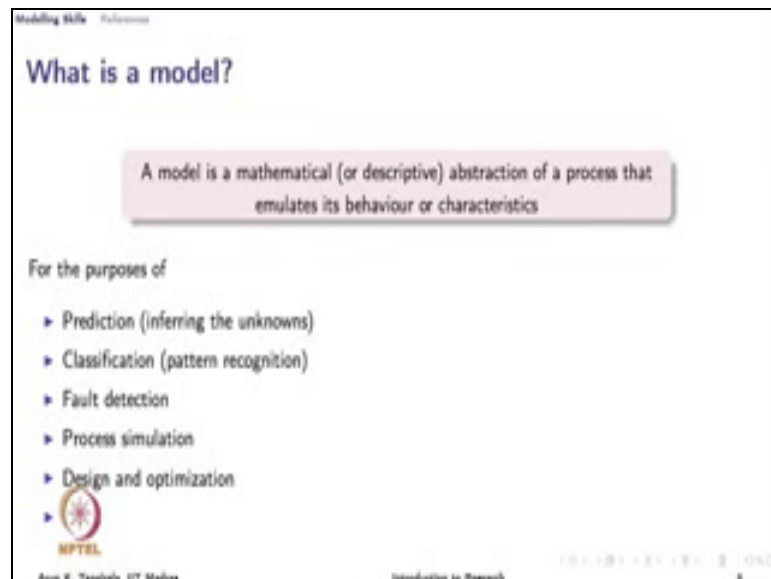
Introduction to Research

So, in particular, we will learn first what is a model? Of course, all of us know what is a model, and this term - model - is used in many fields with almost similar connotations, but it's good to really clarify it upfront, so that we are all on the same platform, and then, move on to learn what are first principles and empirical models. Now, essentially, these are not just models, these are also modelling approaches. The first principles approach is a fundamental approach, whereas empirical approach is a data driven approach. And then, we shall quickly go through a systematic procedure for building data driven models. As I said, a few moments earlier, we will largely focus on data driven modelling

because **that's** where a lot of expertise is not available or lot of beginners find it very difficult. **And** therefore, I would like to concentrate on the same.

And also, talk about a few critical aspects of data driven modelling - what are the things to watch out for, how to go about handling certain aspects such as noise in data, and over fitting or doing the right kind of experiment, and so on. We will also embellish this lecture with a couple of hands on examples like we did in the last lecture on data analysis in R. I hope now, with the previous lecture, after the previous lecture you are familiar with R.

(Refer Slide Time: 02:23)



So, **let's** begin our discussion by asking - what is a model? In general, we know, we can answer qualitatively **what's** a model. **It's** some entity that allows us to emulate a process, the process behavior, **process** characteristics and so on; this term - model - is not only used in data analysis or research and so on, you can see this term being used even in fashion industry and elsewhere, or even models of houses and so on. So, we have a good feel of what **is a** model, but **let's** try and define what a model is. It is a mathematical or a descriptive that is quantitative or quantitative abstraction of a process. It allows us to describe a process in mathematical terms, so that we can emulate or simulate the process behavior.

**And** why would we need this model? There are several reasons. We know, again, but **it's** good to list some of the prime end uses of a model. One of the most popular uses of a

model is in prediction. That is as we call as the forecasting, like we mentioned in the last lecture. Once I have a model, and I know the inputs to the model - remember that a model consists of certain inputs from the user - and then, the model makes a prediction of how the process would respond, and we will talk more about **it** shortly. So, models are heavily used in prediction or inferring certain unknowns and also classification - we discussed this last time pattern recognition.

There we talked about models, not necessarily mathematical form; they are more of models in the form of classes. So, we say when the data falls into certain class, then the process belongs to a certain set of operating conditions and so on. So, models are heavily used in classification as well.

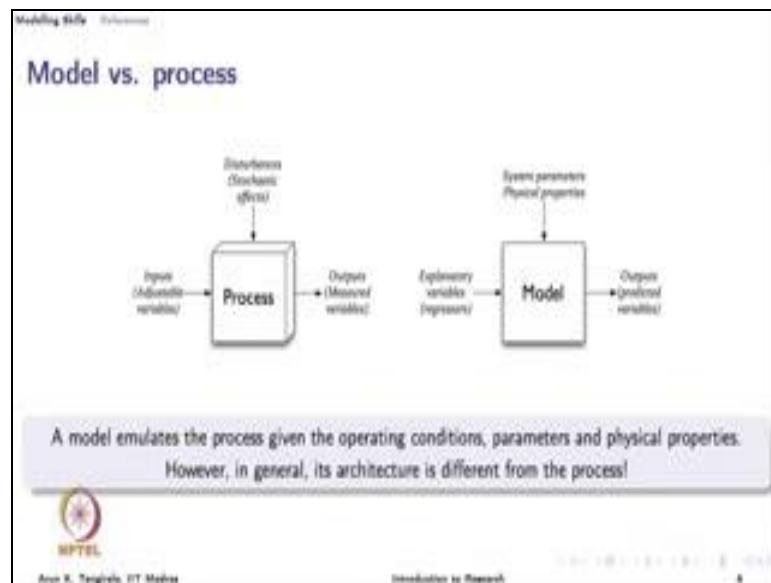
And the third application of modelling is in fault detection, where I build a model of the process under normal operating conditions. For example, I know how a friend of mine would talk, and sit, and so on when he or she is normal, but when something is wrong, may be something is mentally troubling my friend, then I know that something is definitely troubling him by observing the behavior. Now, **what's** happening underneath is I am projecting my friend's behavior against a normal behavior that I have observed over a period of time, and then, comparing both, and seeing well there is a huge difference, and then, finally, coming to a conclusion that something is abnormal. This is pretty much the same idea in fault detection as well. We built models of process under normal operating conditions from historical data or through first principles approaches and then, keep comparing the measurements that come out of the process against what the model is predicting. If there is a significant difference between the prediction and the measurement, then we raise an alarm, and probably conclude that there is a fault, and then, take it up for further diagnosis.

One of the prime uses of models is also in simulations. We have heard of simulators. Some of you must have worked with different simulators in chemical engineering, mechanical engineering, aerospace, and so on. You have heard of air craft or flight simulators. There, the primary role of the model again is in predictions. So, we give certain inputs to the model, the same inputs that we would see when we operate the process, and ask how the process would respond. So, the model would make a prediction. We need high fidelity models in such applications, whereas when we use models in control, although I **don't** list that here, models are heavily used in control,

where the model makes a prediction of where the process is heading, and then, a controller takes an action to keep the response of the process close to the set point. There, in such applications, **that's** in control, we may not need high fidelity models. We can work with fairly approximate models. And finally, we do find uses of modelling in design and optimization and so on.

**So**, obviously, the list is a bit more than what I have given here, but what should be remembered is the end uses of models vary a lot, and therefore, the type of model, the kind of accuracy, the nature of the model whether you want to build a steady state or dynamic model or you want to build a time domain or a frequency domain model and so on, really it depends on the end use. So, you have to work backwards, and then, make a decision on what kind of model you want to build.

(Refer Slide Time: 07:36)

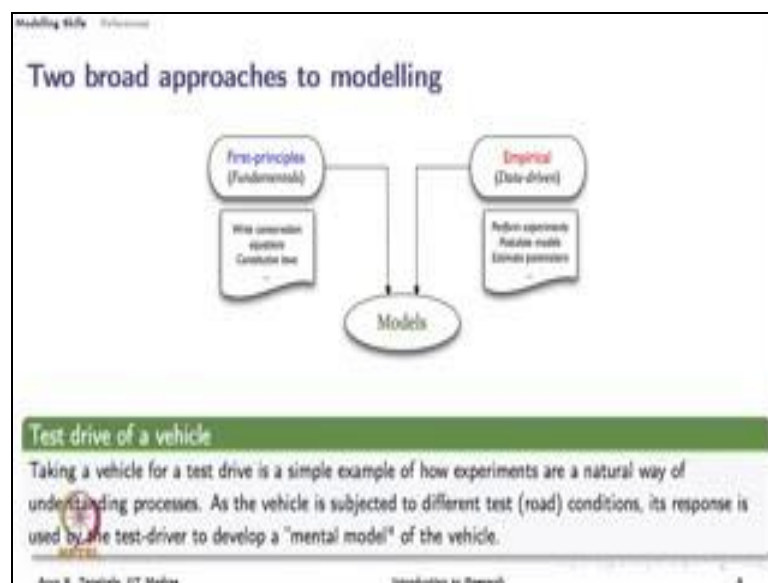


Okay so, just to even make this point very clear - the point on distinction between a model and a process and also a similarity - I am showing you a schematic here of the process and the model. So, if you carefully look at the process architecture, there are inputs which are causal and physical inputs going into the process, and then, there are disturbances acting on the process. You can think of... you can take any process and actually cast it into this architecture. And then, the process responds, which we call as outputs in the engineering terminology.

Now, models also give you the outputs that are of interest to you. In fact, typically, the output of a model is same as the output of the process, but typically the output of a model is nothing but the variable that you want to predict. And the inputs to the model need not be necessarily the physical inputs that go into the process. The inputs to the model are generally more or the same as the inputs that are going to the process, but for example, if you looking at a dynamic model, in a dynamic model, the output is modeled as a function of the present and the past, because transients are important to us. So, the inputs to the model are not only the present input but also the past; whereas the process is operating in real time, and it keeps receiving only the present input at any instant; the past has occurred but at that time. On the contrary, in the model, you do feed the past inputs and so on, and then make a prediction, because model is after all a mathematical abstraction.

And then, there are also certain user defined parameters **and** or user specific inputs that you will have to provide to the model along with the system parameters. So, the architectures are different, but the final use of the model is in prediction – basically, predicting the variable of interest to you. So, **that's** very important. So, do not get confused with the inputs that go into the process and the inputs that go into the model.

(Refer Slide Time: 09:52)



**Alright** so, **let's** look at how models are developed because we want to really gain some insight into how to develop and build models **right**. There are two broad approaches to

modelling. One approach is to start from fundamentals, where you invoke the laws of physics, chemistry, biology, and so on; essentially science based or mechanistic models. And here, we invoke the laws of conservation primarily mass, energy, momentum, and use a few constitutive relationships may be from thermodynamics and fluid mechanics and so on; and finally come up with **the** model; the set of equations essentially. which we call as a model. Now, **that's** one approach.

The other approach, which is quite contrasting, where we **don't** rely on science as much as we rely on data. There is a data science, but we **don't** rely on the science of the physics to begin with; at some point in time maybe we can incorporate, but to begin with, we rely on data. And this approach is called an empirical modelling approach. And **it's** also called a data driven approach, where I will use the data to identify the relationship between the variables of interest; typically, the input and output and so on or sometimes only to build a model for the output which we call as a time series model. **And** here data is the primary **food** for identification. Without data, there is no empirical approach at all. And the kind of models that come out of empirical approaches are called black box models, typically, where you **don't** incorporate necessarily any physics of the process, explicitly. You work with a minimal understanding of the process, but that does not mean that there is no provision for incorporating the physics of the process or whatever you know about the process a priori; you can. And as you keep incorporating the prior knowledge into your empirical model, the black shade turns into gray, and there is some transparency that sets in and such models are known as gray box models.

So, in that respect, the first principles models are actually called white box models because they are very transparent. If I look at a model, first principles model, I will be able to associate every term in that model with some physical characteristics of the process; whereas, **that's** not necessarily the case with an empirical model. An empirical model is some mathematical fit between the input and output. So, to give you a simple example, when we go out on a test drive, let say to purchase a vehicle, the common sense thing that all of us do is take the vehicle, sit in the car and apply certain inputs to rotate the steering wheel and pedal, apply breaks, supply fuel, and so on, and you know, give all different kinds of inputs that we want to really test the vehicle on; and then, collect the response of the vehicle. So, what we are doing there is, we are applying inputs, and observing the response of the vehicle, putting it all together in our brain, and building a

mental model. We may not be able to write an equation there. We are building an empirical model. We are not really building a model of the car from first principles. I am sure that would be a very deadly approach indeed, but we never do it and we have not seen anyone doing it.

On the other hand, when we really sit in courses on **in** automobile engineering, mechanical engineering, and so on or in engineering design, we do learn what are the mechanics of a vehicle through equations, through first principles understanding and so on. So, that has its place, while empirical modelling has its place; where increasingly a lot of people are turning to empirical modelling, primarily because many processes that we are looking at, trying to understand, are quite complicated, quite complex, for us to write a first principles model. So, the experimental or the empirical approach is a natural recourse and that will continue; **it's** here to stay; **it's** been there since times immemorial; it has been there from the time man has started to build models, try to understand processes from observations.

(Refer Slide Time: 14:33)

Modeling Skills Reference

### First-principles vs. Empirical modelling

| First-principles  | Empirical  |
|---|--|
| Causal, continuous, non-linear differential-algebraic equations   | Models are usually black-box and discrete-time   |
| Model structures are quite rigid - the structure is decided by the equations describing fundamental laws. | Model structures are extremely flexible - implies they have to be assumed/known a priori.          |
| Can be quite challenging and time-consuming to develop  | Relatively much easier and are less time-consuming to develop                                      |
| Require good numerical ODE and algebraic solvers.   | Require good estimators (plenty of them available)   |
| Very effective and reliable models - can be used for wide range of operating conditions                   | Model quality is strongly dependent on data quality - usually have poor extrapolation capabilities |
| Transparent and can be easily related to physical parameters/characteristics of the process               | Difficult to relate to physical parameters of the process (black-box) and the underlying phenomena |

MITEL  
Prof. R. Teagiri, IIT Madras  
Introduction to Research 8

**And** for the rest of the lecture, we will focus on empirical approaches per se, but before we do that, I just want to draw your attention to a few salient differences between first principles and empirical modelling approaches. There are people who really vouch for first principles models, and say, **that's** the best and so on; and then, there are people who argue in favor of empirical models and so on, but there is no hard and definitive rule as

to which modelling approach is the best. It is a situation that really determines the answer to that question, and one has to take a very common sense approach through it, and a neutral approach. But in order to do that we have to understand what benefits or pros and cons that each of these approaches have got to offer.

With the first principles approach, we do get causal, physically meaningful models and so on, and also, its ability to predict the process behavior is good over a wide range of operating conditions. However, they are difficult to solve analytically. Typically, we end up with non-linear differential equations ODEs and PDEs and so on, and those models may take a lot of time and computational effort to solve; whereas, the empirical approaches offer a lot of flexibility in choosing the models.

In a first principles approach, I don't have any say on the structure of the model that I have. Whatever the laws that I am applying, give me, I have to live with those kinds of models. Of course, I can choose to build some approximations, but even those approximations can be quite complicated. Whereas with the empirical model, I am trying to find a mathematical fit that helps me understand or map the relation between the variables of interest, and also a map that helps me make good predictions; that's it. So, there may be many solutions of which I may pick the most simple one, and not really a very simple one, but simple enough model, and that flexibility really gives the empirical model a big, what do you say, you know, vote of favor. A lot of people really prefer that, particularly in control and so on. And therefore, relatively much easier to may be simulate, and less time consuming to develop, and so on. However, they require good estimation algorithms, because you are going to estimate parameters of the model and so on, and there can be considerable computational effort there. So, there is no escape there as well.

The first principles models give us effective and reliable models, whereas the empirical models are as good as your data. Now, that is a very common criticism of an empirical approach, but you have to be careful when you take that criticism into account. Clearly, training a model is pretty much like training a student in a subject. Whatever questions that you are going to ask later on to the student or whatever you are going to really test the student, depends on what you have taught. If you are going to ask a question to a student of a course, on a completely different topic that you have not taught, then obviously, the student may not be able to answer in most probability. And that's the



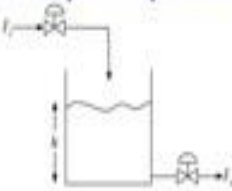
same with an empirical model. You are going to develop the model from data. So, you have chosen a model structure and data, bring them together with the help of an estimation algorithm, and then, force the model to understand the data. If you have chosen the right model structure, and right estimation algorithm, things will work out well, and it would have captured the essence of the process for the operating conditions that are present in a data, that is under which the data has been generated.

Now, if you are going to present a completely new data set from a different operating condition, then it's very likely that the model may not work well, unless the process is linear and so on. So, to speak, the extrapolation capabilities of an empirical model are always questionable, but it depends on how you have trained the model. If you have shown the model a wide range of operating conditions, then, yes. So, the model will do a good job for you, but in any case, if the model is too complex, then, there is no other choice for us. We have to choose, we have to resort an empirical model.

The other aspect that is not reflected here is, when a process is random, that is stochastic - and we had discussed this part in the last lecture - in such situations, first principles models hardly come to your rescue. There is no choice, but to build models from data, which we call as time series models. So, empirical models have a lot of points in favor of them as long as you are aware of the limitations. The prime limitation being model quality is strongly dependent on the data quality. If you remember that, then, you will not really give excessive importance to the empirical modelling approach, really tread with a lot of wisdom, so to say.

(Refer Slide Time: 19:49)

**Example: Liquid level system**



|   |   |  |
|---|---|--|
| <p><b>Case 1: Steady-state model</b> between <math>F_1</math> and <math>h</math>:</p> $F_2 = C_1 \sqrt{h}$ <p>(<math>C_1</math> is estimated from data)</p>   | <p><b>Case 2: Dynamic, non-linear model</b> of <math>h(t)</math> for changes in <math>F_1(t)</math>:</p> $A_1 \frac{dh}{dt} = F_1 - C_1 \sqrt{h}$ <p>(requires numerical solvers)</p>                             |  |
| <p><b>Case 3: Approximate linear, dynamic model</b> about an operating point</p> $A_1 \frac{dh}{dt} = F_1 - \beta \bar{h}, \quad \beta = \frac{C_1}{2\sqrt{\bar{h}_s}}$ $\hat{F}_1 = F_1 - F_{1s}, \quad \hat{h}_1 = h_1 - h_{1s}$ <p>Typically, <math>F_{1s}</math> are steady-state values.</p> | <p><b>Case 4: Approximate empirical, linear, grey-box, dynamic, discrete-time model,</b></p> $h[k] = a_1 h[k-1] + b_1 F_1[k-1] + c[k]$ <p>(Parameters <math>a_1, b_1</math> estimated from experimental data)</p> | <p><b>Case 5: Black-box, dynamic, discrete-time model</b></p> <ol style="list-style-type: none"> <li>1. Model structure based on ease of estimation and end-use</li> <li>2. Model may not be physically interpretable, but designed to give good predictions.</li> </ol> |

Annex B, Tejaswini, IIT Madras Introduction to Research 18

So, let's move on now and just get a feel of what are the different types of model that one can develop for a liquid level system. Now, this is a very simple system that we see in our households everywhere, and also in industries, where there is a flow coming into a vessel or a tank, if you like it, and there is some storage of the liquid in a tank, and then, there is an outlet flow. Both the in and outlet flows are regulated by valves. And let us say, I am interested in knowing how a change in inlet flow affects the liquid level. This is very important, because look at the flush systems in our toilets and so on; they are based on an understanding of how the flow affects the liquid level. There, of course, you have a mechanical device controlling the liquid level, but in industries there are automated controllers controlling the liquid level. In all cases, we need to know how the liquid level changes, when there is a change in the inlet flow.

Now, the first scenario is probably that I want to understand its steady state. How the outlet flow and liquid level are related, because that actually becomes a part of the model eventually between the inlet flow and a liquid level. So, at steady state, we know from Bernoulli's principle, fluid mechanics, and so on, we can derive a relation between outlet flow and the liquid level of the head that people talk about. And we know that the outlet flow is linearly proportional to the square root of the liquid level. This is called a steady state model. This is true at steady state. All you have to do is perform an experiment at different steady states, measure the outlet flow rate and the liquid level, plot the outlet flow rate verses the square root of the liquid level; you should see more or less a straight

line. Well, if you have access to an experiment, you should try it out and you will see that this is not perfectly right, but this is a fairly good model. Now, we just now said eventually I want a dynamic model; that is that tells me how the liquid level changes when not only at steady state but between two steady states, when there is a change in the inlet flow.

Now, for this we will have to write the material balance, that has apply the mass balance for the law of conservation of mass, and we assume incompressible flow, and we end up with a differential equation - first order differential equation – unfortunately, non-linear in nature. So, the equation that you see here has come about by applying two things. One, the conservation law of mass, and two, the steady state model that we just developed in the case one. Now, I can simulate this. Obviously as I mentioned earlier, this is a first principles model. There is no way to analytically solve this. I will have to use a numerical solver; essentially use numerical integration techniques to determine the liquid level profile - for a profile, input profile - which is the inlet flow.

Now, what I could also do in order to have an analytical solution, I could approximate this non-linear model with a linear one, assuming that the changes in the inlet flow are not going to be too wild, and therefore, I can think of a linear relationship between the liquid level and the inlet flow. A fairly reasonable assumption under many conditions in many situations.

So, what we do is, we approximate the non-linearity, the source of the non-linearity in the ODE in case two is a square root. So, we can approximate that with a linear one using Taylor Series Expansion. So, what we have done is, we have approximated with a first order approximation of the square root, and whenever we approximate, typically approximations are within the vicinity of an operating condition of a nominal point, and therefore, we rewrite this ODE as a linear approximate model. It is still a first principles model, but it is an approximate model, but now the model is in terms of what are known as deviation variables. That is how far you are away from your reference operating point, typically chosen to be as a steady state. On the other hand, I may say, well, you know, I **don't** know any of this laws of conversation, and the valve equation and so on. I will choose to fit an empirical model, and then, there are two possibilities: a gray box and a black box model. Suddenly, I discover that in this empirical approach, no, no, no **it's** conservation mass is not so difficult to write. So, I do write, and I have some idea of how

the model should look like as in case three; but now the problem is we are no longer in continuous time, because we are going to build models from data, and data is available only at sampled, in a sampled instance in time; **it's** not available at every point in time. So, we move from time  $t$ , which is a continuous time variable, to  $k$  - discrete time  $k$ . This  $k$  here, in case four, stands for the sampling instance, the  $k$  sampling instant.

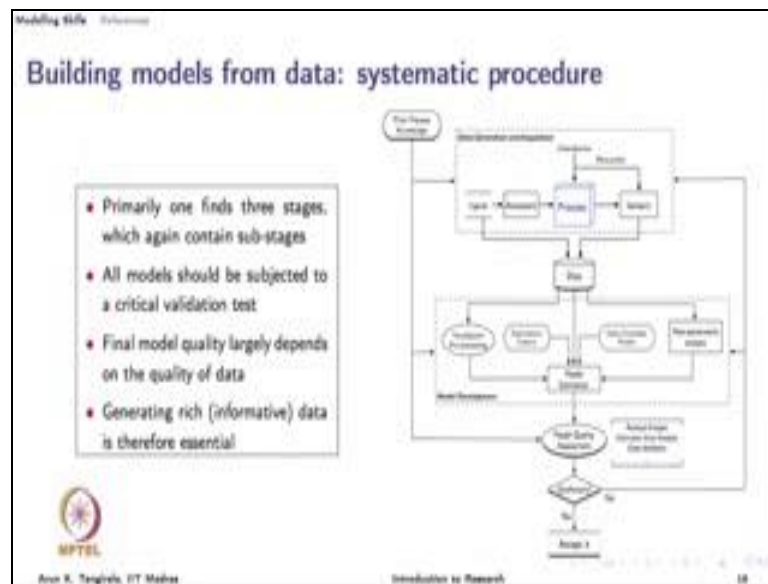
That is I am observing now. I have observed the process in order when I was collecting data, at a regular intervals **s** in time which we call as a sampling instance, and now from some prior knowledge of the process, I know that there is a first order linear relationship between the liquid level and the inlet flow rate. So, I write an approximate linear equation, but in discrete time, and **it's** a first order difference equation essentially. So, because I moved from continuous time to discrete time, I moved from a differential equation to a difference equation. And now, the goal in this kind of a modelling approach is to estimate these parameters  $a$  and  $b$ , and the epsilon that you see is **a** error that you have made in approximation. Of course, we also would like to know what the magnitude of this error is and so on, but the primary goal is to estimate these parameters  $a$  and  $b$ . So, here, we say this **is a** gray box empirical linear discrete time model, a lot of qualifiers, of course, but to be very clear, the gray box nature comes about because I have incorporated prior knowledge of a linear model and first order dynamics.

Suppose, I **don't** know any of that, and I just want to build a dynamic model; I can do that as well. Then, we end up with a black box model and I **don't** have any equation written up here because the choice is yours. Of course, through a systematic study you will be able to discover that eventually the relationship between the liquid level and the inlet flow rate is that of a linear one and a first order one. Typically, when we start off in black box modelling, we start off with simple models. So, we would start off with a linear model, and because this is a dynamic process, we have to choose the order of the dynamics first order, second order, and so on. **And** therefore, through a careful study we can converge to a first order linear model. And this case study is discussed in detail in the book - one of the books that **have** given as references at the end of this lecture **alright**.

So, this hopefully gives you a feel of the different kinds of models that one can develop for a process, and as you have seen here, the nature of the model that you develop depends on what you want to know, and what kind of rigor that you are looking for, and what you have with you - what kind of knowledge is with you. In an empirical approach,

you have data and minimal process knowledge. In a first principles approach, you **don't** have data to begin with, but you have a pen and paper, and you have, you are equipped with good understanding of the physics of the process **right**, but a good solution - working solution - is always a gray box approach, which is a marriage of these two, so that you **don't** have to really take sides whether on the first principles or empirical.

(Refer Slide Time: 28:18)



So, I just mentioned a few minutes ago, that a systematic procedure will help you get a good model, and in particular, I am talking about an empirical approach here. So, I give you a flow chart for a systematic way of identifying the model, and let me quickly go over this. Primarily, you have three stages. One is data acquisition and we **don't** talk about that here; we assume that data is available to us, and so, the first stage we really **don't** discuss much. And then, the second stage is, of course, model development, which is at the heart of this procedure. And the third stage is model assessment; **that's** very important and that applies to all models that we develop, whether **it's** a first principles or an empirical model, **it's** important to assess the goodness of the model; is the model capable of predicting the process over a good range of operating conditions? Is it doing well on a fresh data set? Even a first principles approach has to be validated. So, please **don't** be under that impression that this is not required for a fundamental model. And maybe we want to ask, if in an empirical approach, if the parameter estimates that I have obtained in a model, do they have large errors in them. That is a something of interest.

And the third thing that we have to watch out for an empirical modelling is over fitting. Remember I said building a model from data is pretty much similar to a student learning a subject. The student is presented with the text book, and the course material, and then, the student tries to understand the concepts of the subject, eventually, through a proper interaction with the course material and the instructor. Now, in the end, you have to ask if the student has over learnt; that may seem very strange - what is meant by over learning? Now, over learning is, let us say, I as a student I am trying to solve an assignment problem and the assignment problem is based on a certain concept. If I have understood the concept, that is, if I at the end of the problem solving exercise, if I have gained mastery of the concept on which the question has been based, then the goal of solving the assignment problem is more or less achieved, but if I start paying attention to the numbers, the very fine details that are very specific to that question, but has very little to do with subject itself, and I am trying to really memorize all of that, then I am over learning **right**. So, that is probably a simple analogy of over fitting in modelling as well. And that occurs primarily because of presence of noise in data and I have an example to show you later on. So, remember that there are three stages: data acquisition, model development, and model validation.

Let me quickly talk about the model development part. You see that there is a pre-processing - data pre-processing - we talked about at in the last lecture. We have to watch out for missing values, outliers or any other anomalies, get rid of them and so on. And a big part of that is visualization; involves visualization of data. **And** we had an example in the last lecture highlighting the importance of visualization. Once we have understood the data well and **it's** ready for modelling, we should not straight jump away necessarily to build a model unless I know the model structure very well; that is, I know **it's** a first order or I know that **it's** a linear model, non-linear model, and so on of this type, and so on. So, an intermediate step involves what is **known as a** non-parametric analysis, where I make minimal assumptions on the process, and try to gather as much information as possible from the data, so as to make a good guess of the model structure; **that's** called a non-parametric approach. And this step, can be skipped if I already know the structure of the model that I am going to fit **okay**. So, in many situations the non-parametric analysis may not be even present.

Finally, I have some good decent guess of the model structure, and then, I have to estimate the parametric of that model, where I apply estimation algorithms which are essentially optimization algorithms to estimate **the** parameters, and then, I have a model with me. So, **that's** pretty much the model development in a very, very simple way that can be possible, but that can be possibly explained, but remember there is so much there, **that's** a huge ocean in itself, and in any of these, at any of these stages, there are no definitive answers or formulae for you to really go through, but there are very good guidelines based on theory of data sciences. **It's** very important to know at least the basics of those principles, and have certain guidelines in place, so as to minimize the effort.

Now, one very important thing that you would see is that we are able to incorporate prior knowledge; that is, there is a provision for incorporating domain or prior knowledge at each of this stage, and I have also emphasized this aspect in the previous lecture as well. In any data analysis exercise, we should incorporate the domain and prior knowledge as much as possible.

And the last, but not the least important aspect of data driven modelling is that it is an iterative exercise. **It's** very unlikely that you will be able to get the best model in just one round of this iteration, one pass of this flow chart here. It is very likely that the model that we develop is not satisfactory in many respects, maybe **it's** not predicting well on a fresh data, maybe the parameter estimates have large errors in them and so on.

Now, why would this occur? Perhaps because the data quality itself is bad, **that's** very likely, which means we may have to go and repeat the experiment or the data quality is good and I have chosen a wrong model structure, which means I have to re-examine the models that have assumed, or that I have chosen a wrong estimation algorithm, in which case, I have to go back and choose a better estimation algorithm and so on. So a systematic study will really help us minimize this back and forth steps, and also, be able to pin point, with a fair degree of accuracy, as to what the source of problem is when the model is not satisfactory. So, please keep that in mind.

(Refer Slide Time: 35:19)

Modeling Skills

## Two broad classes of models

### Time-series models

- ▶ Suited for modelling stochastic or random processes (e.g., stock market index, rainfall, sensor noise)
- ▶ Causes are either unknown or also random
- ▶ Usually dynamic models (in a few applications, steady-state as well)
- ▶ Challenges: choosing model structure, making the right assumptions on process characteristics, non-linearities, etc.

NPTL  
Ann K. Tangirala, IIT Madras  
Introduction to Research 17

Now, within the empirical models, there are two broad classes of models that one encounters and I just want to briefly discuss those. I am not referring to the linear, and non-linear, and so on. This classification is based on more or less the same lines as we discussed in the data analysis lecture. The deterministic versus stochastic models and so on. So, the first class of models that one would see prevalently in the literature, predominantly in the literature is a time series models, which cater or which are suited for stochastic processes, where the causes are either unknown or known with error; that is, they are actually random themselves.

And a simple example for that would be that of an atmospheric process. Suppose I want to build a model that predicts the atmospheric temperature. Now, what do I do? Yes, I probably know that there are several factors affecting the temperature, but I have not measured them or probably I don't know the complete list of causes that influence the atmospheric temperature. So, a natural recourse is to take the historical data and hope that there is something in the history that will repeat itself; not exactly, but there are patterns, and correlations, and so on, and then, build what is known as a time series or a dynamic regressive kind of model. These are very common in many, many fields. On the other hand, let us say, I want to model the relative humidity of air. Now, I know that the relative humidity of air is significantly dependent on the temperature - atmospheric temperature. So, I have the relative humidity measurements, I have the temperature measurements and I can build a model between these two; that still counts as a time



series model in the general literature, primarily because both the temperature and the relative humidity are random in nature. So, we can build what are known as multivariate time series models, but then, you know, **it's** a matter of perspective. You can also call it as an input-output model. So, the time series model, when you look at the terminology, typically, it means you are building model based only on the response of the data, but many people use it with a larger **connotation**.

Now, some of the challenges in building time series models are choosing the right model structure; that is how much in the past, for example, affects the present, and that, of course, again, guidelines and some mathematical as well as statistical methods are available to help us, but again there is no definitive answer. And making the right assumptions. For example, in stochastic processes, do I assume the process is stationary or non-stationary or it is stationary with a deterministic trend like the one that we saw last time, the carbon dioxide emissions, we saw a trend. **There's** a linear trend, and then, on top of it there were oscillations and so on. So, we **don't** know. Or if the underlying random process is linear, non-linear and so on. And, of course, the unwritten challenge is estimating the parameters; **that's** anyway challenging in empirical modelling.

(Refer Slide Time: 38:37)

Modeling Skills

## Two broad classes of models ... contd.

### Input-output or causal models

- ▶ Suited for modelling relationship between a variable (or more) and other **explanatory variables** (a.k.a. **regressors** or **factors**) (e.g., power and current in a wire, temperature and coolant flow)
- ▶ Regressors may be known accurately or with some error
- ▶ Models could be steady-state or dynamic.
- ▶ Challenges: sufficient variability in factors, selection of regressors, measurements and regressors available at different sampling rates, choosing the order of dynamics, etc.

**Remember**

In practice, all modelling exercises demand a careful handling of uncertainties both at the experimental and modelling stages!

NPTTEL

Prof. R. Teagiri, IIT Madras

Introduction to Research

18

Now, on the contrary, we have what are known as input-output models, which are causal models, where I try to explain one variable using other variables that I think are causing or influencing the variable of interest. You may like the temperature relative humidity

example or may be if I am measuring the power versus current in a wire or the temperature of a reactor versus a coolant flow and so on. There I can think of an input-output relationship between these variables and will regressive models. Of course, multivariate models are also possible.

Now, one of the challenges here, in this input-output models, is **this** so-called Regressors or explanatory variables as we call them, may be known accurately or may not be known accurately. In the temperature, relative humidity example which we called as a multivariate time series models can also be considered as an input-output model, but the difference is that the temperature which is the regressor for relative humidity, is a measured variable, is not something that I am able to adjust. Unfortunately, I am not able to, I **don't** have the ability to adjust the atmospheric temperature. If people had that ability, then it could be a chaos **right**. So, the temperature is a measured variable and every measurement has error in it. So, the regressor, in that kind of a situation is known only with in error; whereas in a coolant flow versus temperature of a reactor, I probably have the provision to induce changes in a coolant flow and measure the temperature.

So, there I know exactly what kind of changes I am inducing in the coolant flow and only temperature is a measured variable. So, the coolant flow is considered a deterministic signal or a variable in which case the regressor is known accurately.

So, you have to look at a situation and determine how to treat each variable. So, it takes us back to the same thing that we learnt in last lecture, deterministic or stochastic. And of course, you can have different models and there are several challenges again here. For example, in cases where I have the privilege of performing an experiment where I change the factors or the explanatory variables - how should I change them? what should be the level of excitation and so on? Or which regressors? In multivariate regression there may be different factors, several factors affecting the variable of interest, then which variable should I factor in or which set of variables should I factor in? And there are also situations where the measurements and regressors are available at different sampling rates and so on.

Now, in practice, all modelling exercises involve a mix of both input-output and time series modelling, because even in an input-output modelling exercise, we may not be able to explain all the variations in the variable of interest. So, for instance, in the

temperature and coolant flow reactor example, I may be able to explain most of the changes in the reactor temperature using the coolant flow, but there is something in the measurement that I cannot explain using the changes in coolant flow and that something is probably sensor noise. So, how do I model that sensor noise? There, I have to take a time series approach. So, sometimes there may be no predictability in that sensor noise, then I only have to estimate the statistics, but in any case, I have to address both the deterministic and the stochastic portions of any **any** model. There is no escape to it. So, we will just quickly go through some of the critical aspects of empirical modelling.

(Refer Slide Time: 42:35)

Modeling 44/6

### Excite the process sufficiently!

**Example**

Consider a steady-state model:

$$y[k] = b_0 + b_1 u[k] + b_2 u^2[k]$$

- Three unknowns
- Therefore, data corresponding to three steady-states is required
- A more general statement: The regressor matrix

$$U = \begin{bmatrix} 1 & u[k_1] & u^2[k_1] \\ 1 & u[k_2] & u^2[k_2] \\ 1 & u[k_3] & u^2[k_3] \end{bmatrix} \quad (1)$$

should be non-singular, i.e., of full rank.

MPTEL

Prof. K. Srinivasan, IIT Madras

Introduction to Research

**And the** first aspect in empirical modelling where we perform experiments or even if you don't is that of the excitation in the regressor or in the factor or input, whatever you want to call it. So, **let's** look at a simple example where I have a steady state model. **The** y is the output and u is the input and y is a quadratic function of the input. So, I have three unknowns, and therefore, I need data corresponding to three steady states, clearly. In other words, if I write the equation for estimating the unknowns - three unknowns - in a matrix form, at these three instants - different instances - in time k 1, k 2, k 3, remember we only have sample data. Therefore, we use this notation. When I write the equation of the model in a matrix form, then the big U matrix that we see here comes into picture, which maps basically y to the parameters b naught, b 1, and b 2. This big U matrix has to be non-singular or a full rank. Of course, in discussing this example, we have kept noise out of the picture. So, **it's** a noise free condition. We are not making any noise here.

When it comes to noise, there are other things to worry about, but even in the noise free condition, it's important to remember that we have sufficiently excited data.

(Refer Slide Time: 44:04)

Modeling slide

**For dynamic systems**

**Example**

Consider a (deterministic) process:

$$y[k] = b_1 u[k-1] + b_2 u[k-2] + b_3 u[k-3]$$

Suppose that the process is excited with  $u[k] = \sin(\omega_0 k)$  (sine of single frequency). With this sine wave input,

$$y[k] = b_1 \sin(\omega_0 k - \phi) + b_2 \sin(\omega_0 k - 2\phi) + b_3 \sin(\omega_0 k - 3\phi)$$

$$= \left( b_1 + \frac{b_2}{2 \cos \omega_0} \right) \sin(\omega_0 k - \phi) + \left( b_3 + \frac{b_2}{2 \cos \omega_0} \right) \sin(\omega_0 k - 3\phi)$$

$$= \left( b_1 + \frac{b_2}{2 \cos \omega_0} \right) u[k-1] + \left( b_3 + \frac{b_2}{2 \cos \omega_0} \right) u[k-3]$$

Only two of the three parameters can be identified! Why?

SIFTSL

Amr K. Tayeb, IT Madras

Introduction to Research

In fact, for a dynamic system, very interestingly, let us look at this example, where the output of a system is dependent on the past input, the one input beyond a past and the second input beyond the past. So,  $u[k-1]$ ,  $u[k-2]$  and  $u[k-3]$  are essentially lagged inputs, but that 1 there, means essentially one sampling instant. This is the hallmark of a dynamic system. It has a memory essentially. So, I want to once again identify these three parameters.

Suppose I perform an experiment where the input is sinusoidal. This is very common in mechanical systems or all systems which have a kind an oscillatory characteristic and so on or which are characterized only at specific modes or frequencies. There, if suppose I perform an experiment with a single sinusoid, then, what i mean by single sinusoid is a single frequency sine wave, and then, it turns out that the process manifests as a two-parameter model. You just have to work through the trigonometry here. So, what I have done is, I have taken this input and plugged it into the equation here and asked how the response would look like. It turns out that the response appears as a two-parameter model to the user, to the experimentalist vis-a-vis the reality which is three parameter models. So, what has happened here? What is the consequence now? The consequence is only two of the three parameters can be identified.

In other words, going back to the same story that we had in the steady state case, now the big  $U$  which would now consist of  $u(k-1)$ ,  $u(k-2)$  and  $u(k-3)$  in the first row and so on for the remaining two rows that becomes singular when input is a sine wave of single frequency. Therefore, I do not have enough information in the data to estimate all the parameters and that is the key. Always data should contain sufficient information to estimate the parameters. On the other hand, if I have an input which is made up of two frequencies - sinusoids of two frequencies - then I have sufficient information. You can show that, it's very easy, check for yourself. Construct this matrix - big matrix  $U$  - that I showed in the last slide, but now with the first row being  $u(k-1)$ ,  $u(k-2)$  and  $u(k-3)$  likewise at  $k-1$ ,  $k-2$ ,  $k-3$ , for two different situations, and you will find that the matrix is singular in the first case, that is when you use a single frequency sine wave, and it is non-singular or a full rank when you use a multiple or two frequencies at least in the sine wave **alright**.

So, for dynamic systems the story looks different, but the bottom line is the same. Have sufficient information in the data. We will conclude with two things. One with understanding, and understanding of how noise effects our modelling, and then, finally is a bunch of questions that we want to answer in any empirical modelling exercise.

(Refer Slide Time: 47:29)

Modeling Data

### Effects of randomness (noise) in data

The second challenging aspect of empirical modeling is the **presence of random or stochastic effects**.

Randomness (uncertainties) in the process influences identification in several ways:

- ▶ Accuracy of predictions
- ▶ Errors in parameter estimates
- ▶ Goodness of the deterministic model

NPTEL  
Anur K. Tongole, IIT Madras  
Introduction to Research

So, how does noise affect our data? Well, of course, in many different ways. It affects the accuracy of predictions; we will not be able to predict the output accurately. It brings

about errors in parameter estimates. And three, it affects the goodness of the deterministic part of the process that we want to estimate.

(Refer Slide Time: 47:48)

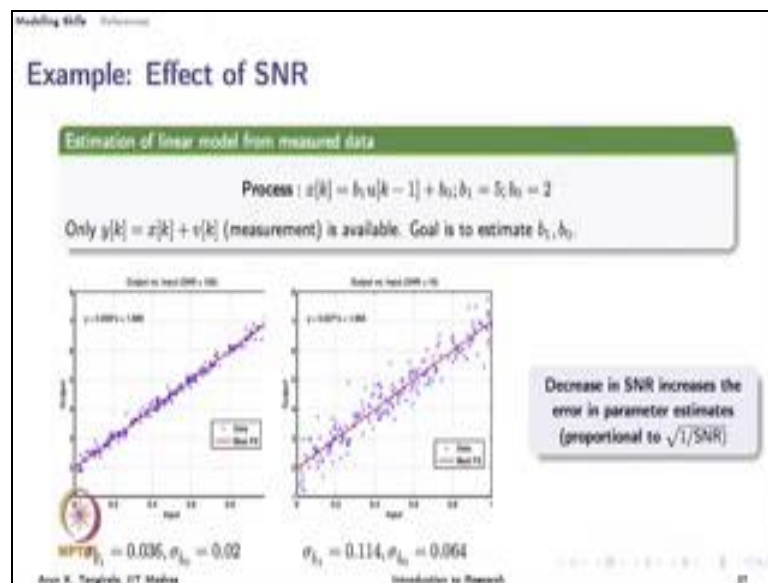
The slide is titled "Signal-to-Noise Ratio (SNR)". It features a red header with the text "SNR". Below the header, it states: "A key measure that quantifies the effects of noise is the Signal-to-Noise Ratio (SNR), which is defined as the ratio of variance of signal to the variance of noise in a measurement." The mathematical formula is given as  $SNR = \frac{\sigma_{\text{signal}}^2}{\sigma_{\text{noise}}^2}$ . Below the formula, it says: "SNR can be interpreted as a measure of the degree of certainty (deterministic portion) to uncertainty". At the bottom left is the IIT Madras logo and the text "Anur K. Tewari, IIT Madras". At the bottom right is "Introduction to Research" and a page number "38".

So, normally, in quantifying the effects of randomness or the noise in data, there is a quantity called a signal to noise ratio that is used widely. It is not just in electrical engineering, but in every data analysis exercise, this signal to noise ratio is a very nice quantity that helps us explain or understand the impact of randomness on the quality of the model or the parameter estimates. And a signal to noise ratio is defined as the variability in a signal. Think of it as the level of amplitude or the power in the signal. The signal, what we mean by signal, here is a deterministic part.

So, imagine that I am measuring, going back to the flow reactor and, sorry, the reactor coolant flow temperature, for example. I am performing an experiment, where I am introducing changes in coolant flow and I am measuring the temperature. There, I induce changes in coolant flow and I measure temperature. And when I measure temperature, the measurement contains two effects. One effect due to the changes in coolant flow and the other comes from the sensor noise or any other disturbance. The signal in that situation would be the true response, that is a response contained in the measurement due to changes in coolant flow only and the rest is all noise. So, obviously, if I want to get good estimates, that is of the model, of the coolant, of the reactor, then the level of response due to the coolant flow should be way higher than the noise right. And this is

true for anything. If you are listening to a speaker in a classroom, the speaker has to speak loud enough - which is hopefully the signal of interest to you - compared to the sources of noise in the classroom, which could be due to a fan or an air conditioner and so on. So, there the signal to noise ratio is the amount of power or the amplitude - you can say squared amplitude - in the speakers' speech signal **vicav ( there is some term , not understandable )** or divided by the amplitude square of the noise contributions coming from the ceiling fans and air conditioners and so on.

(Refer Slide Time: 50:01)



So, higher the SNR, better the parameter estimate **is**; that is lower the error. So, to give you a simple example, I have a process here, which I am going to simulate. So, I am simulating a simple process here.  $u$  is the input and  $b_1$  and  $b_0$  are parameters that I choose, and  $x$  is a true response of the process, but I **don't** have access to the true response. I have access only to the measurement which is  $y$ . So, the way I am simulating is I simulate I generate  $x$  first, the true response; and then, I add some random signal to, some kind of noise to  $x$  to generate my measurement. Now, I pretend that I do not really have access to  $x$ . I only have access to  $y$  and  $u$ , and that I know that the **underlying** relationship between  $x$  and  $u$  is this. So, when I fit a model between  $y$  and  $u$  using the data, what do I get?

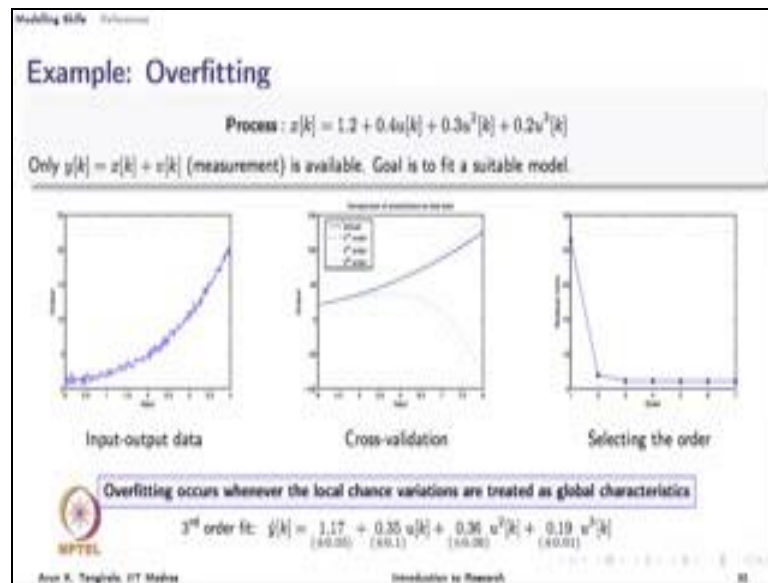
**Let's** look at it quickly here and the two different situations; on the left, you have the situation where the signal to noise ratio is 100. So, I have adjusted the noise level in  $v$ ,

such that the signal to noise ratio is 100 and the codes for this are available on my website. What you see on the right is the case or the situation for SNR 10. Clearly, on the right-hand side, there is more noise as you can see in the scatter plot. In both cases, I am plotting  $y$  versus  $u$ .  $u$  is on the... input is on the  $x$ -axis, and what you see on the left top of the plot is the equation of fit of straight line. So, I have done this in matlab, and it gives me the equation of line, and it does a fairly good job of estimating  $b_1$  and  $b_0$ . The true values of  $b_1$  are 5 and  $b_0$  are 2;  $b_1$  and  $b_0$  are 5 and 2 respectively, and the estimates are pretty close in both cases, but the difference comes about when we look at the so-called **the** standard error in these estimates.

Now, what I have obtained here are  $\hat{b}_1$  and  $\hat{b}_0$ . The hat denotes the estimates. So,  $\sigma_{\hat{b}_1}$  and  $\sigma_{\hat{b}_0}$  are so-called standard errors, which I compute through some theory. **Let's** not talk about that right now, but what is more important for us to notice is the errors that I am reporting for SNR100 are much lower than the errors that I report for SNR10 **right**. In fact, they share a relationship which is essentially that the errors in parameter estimates are proportional to square root of  $1$  over SNR. What this means is, I have a **fall** in the SNR by factor of 10 which means the errors should increase by factor of square root of 10, which is roughly about 3.6. So, you can see here in  $\sigma_{\hat{b}_1}$  when SNR is 100 is 0.036, whereas when SNR is 10  $\sigma_{\hat{b}_1}$  is 0.114, which is roughly about three and half times the error in the case of SNR 100. So, there is theory to tell us how the errors are dependent on SNR, but **what's** more important to understand is wherever possible we should perform an experiment to make sure that the SNR is high enough; of course, respecting other constraints in the experiment.



(Refer Slide Time: 53:57)



And the other aspect, which is a final aspect, of the noise that I want to talk about, which I have mentioned earlier, is over fitting that occurs primarily when you have noise. Once again what we do to simulate the data, we first generate the response of the true process, the equation for which is given at the top. It's a third order polynomial, and I add some noise to the true response maintaining a certain SNR, decent SNR like 10 and so on, and I generate my  $y$ ; that is the measurement. So, once again we pretend that we do not have access to the true response which is a reality. I have access only to the measurement  $y$  and the input that I have given to the process. Now, the input and output plot is shown here. I can see a polynomial kind of relationship, but I do not know whether it is a quadratic or cubic and so on. So, what I do is, I try out third order, fourth order, fifth order polynomials, and obviously, as I increase the order of the polynomial, here I have generated about 200 observations. Sorry 100 observations. I can fit a 99<sup>th</sup> degree polynomial to it. It exactly explains the relationship that I see in the left plot, but if I do that the danger is that on a fresh data, the 99<sup>th</sup> degree polynomial will fail miserably.

As you can see here in the center plot, the third order polynomial... what we see here in the center plot is the prediction of these polynomials of different orders that I have fit on a fresh data. So, I have reserved certain data for training and another data set for testing. The third order polynomial performs the best, the fourth order performs reasonably better and the fifth order polynomial goes for a toss. It predicts completely different. In fact,

it's unstable and I could have avoided this over fitting. So, this is what we call as over fitting.

If I had looked at the improvement that I have obtained by fitting models of successive orders. So, the plot on the extreme right shows us that. What have done is with each model I have... Remember no model perfectly explains. So, there is going to be some residual. I have taken the variance in the residual or you can say the squared, sum squares of the residuals and plotted it verses the order that I have fit. So, when I start with the lowest order, obviously, the sum square is very high, and as I get closer to the true order, the sum square comes to a minimum, and thereafter, the improvement in the sum squares is very, very marginal. So, by increasing the order of the polynomial I have not benefited much. In fact, what I have lost out on is the ability to predict very well on a fresh data cell. So, a plot like this of how much improvement I am obtaining verses the order or whatever model complexity that I am fitting is always helpful in avoiding over fitting okay.

So, why does over fitting occur? Essentially when I start confusing the local chance variation. So, if you see in the plot here, there is a global trend which is a polynomial trend on the left-hand side plot, but also there are some local fluctuations which are due to noise. So, if I start confusing those local fluctuations with a global trend, then I am over fitting, and that is what should be avoided in all situations. And that can be done with a careful study. So, just to give you a feel of what are the kind of errors that we obtained in the estimates of the third order polynomial, I report the estimates here, along with standard errors reported in the brackets underneath. They are called one sigma standard errors, and you can see that the standard errors in this parameter estimates are quite low compared to the estimates themselves, making us accept the third order model to be a satisfactory one. It has done a good job of predicting well on the fresh data and also the parameter estimates are low in errors. So, this should be the typical approach to empirical modelling.

(Refer Slide Time: 58:03)

Modeling Skills

### Questions for reflection

- ▶ What type of models are possible? Which one(s) to choose?
- ▶ How do we "fit" a model that "explains" the variations observed in experimental data?
- ▶ How to "correctly" account for the deterministic and stochastic effects?
- ▶ Will the experiment influence the model that we fit? If yes, in what way?
- ▶ How do we set up and solve the problem of estimating the unknown model parameters?
- ▶ What kind of experiments should we design to obtain a good quality model?
- ▶ How much data do we collect? (what should be the sample size?)

SPTTEL  
Annex K, Tenggol, IT Madras  
Introduction to Research 33

Let's conclude this lecture with a few questions for reflection, and obviously, we are not going to discuss them in detail, but in any empirical modelling approach, in fact, also to a large extent first principles models, these questions apply. One is always faced with the question what type of models to choose? Again, there is no formula there. We have to go based on the end use, how easy it is to estimate, prior knowledge of the process and so on. And the general guideline is keep the model as simple as possible. Not simpler, then simple, but as simple as possible. Good enough to explain, easy enough to estimate and so on, and of course, also convenient enough to implement, if you are going to implement the model online. And also, we will have to worry about two different models. One model that explains the deterministic portion and the other that explains the stochastic part. How do we correctly account for the deterministic and stochastic? That's a big challenge and one has to go through a careful procedure. There is no time to discuss those, but there are certain systematic procedures in place, and the key is the model assessment stage where you check for over fitting of the deterministic and stochastic model; whether you have nicely segregated the deterministic and the random effects.

Will the experiment influence the model that we fit? Of course, yes. There is no doubt about it. The data quality is highly influential on the model quality. Your model is as good as the data. So, perform experiments with care, think of the class of, range of models that you want to build. We have already seen excitation matters. If I poorly excite

the process in the experiment, then I will have limited information, and therefore, I have not interrogated the process or interviewed the process enough to make a decision on a good model.

**And** how do I, what kind of experiment should be obtained or designed? As I said, we should perform experiments taking into account all the factors, suppressing the sources of disturbances as much as possible, choosing a nice instrumentation which limits the noise levels and so on. And then, there is a design of experiments subject which tells you how you should go about designing the experiments, not only to enhance a signal to noise ratio, but also making sure you have all the factors that affect the variable are excited sufficiently, how to minimize the time, perform in an optimal way, and so on. So, you should refer to the design of experiments subject.

How do we set up the problem of estimating? There are different estimation algorithms, least squares, maximum likelihood, Bayesian methods, so many different methods of estimation - which one should I pick? Again, you will have to understand how these estimation algorithms perform, but the general guideline is - choose the one that is efficient. That means, that gives you estimates with low errors and also computationally less burdensome; and usually these are conflicting factors. An algorithm that gives you efficient estimates need not be the computationally most friendly one.

And of course, how much data to be collected – **that's** a big question. It has a huge impact on the errors in the parameter estimates. So, general guideline is the errors fall down as a function of 1 over root n, where n is number of data points that we collect, if you have chosen the right estimation algorithms. So, obviously, more the data, better the estimate that one should expect.

Another question that **doesn't** crop of here is domain of modelling. I may collect data in time, but I may build a model in frequency domain. **It's** very, very likely, particularly, in periodic... in detection of periodicities and so on. So, **that's** another decision one has to make. Again, that completely depends on the application that you are looking at.

(Refer Slide Time: 01:02:22)



So, with those words I would like to close this lecture and here are a few references. Again, not exhaustive. The case study that I was talking about earlier, the book that I referred to earlier in the lecture is the one **that's** given at the bottom - Principles of System Identification. There is a book, there is **a** website for this book on my web page and you can download some of the mat lab course. For example, for the over fitting and the signal to noise ratio examples that I illustrated. Please feel free to write to me if you have any questions. So, hopefully you enjoyed the lecture and that you have a good modelling session whenever it is.

Thanks.