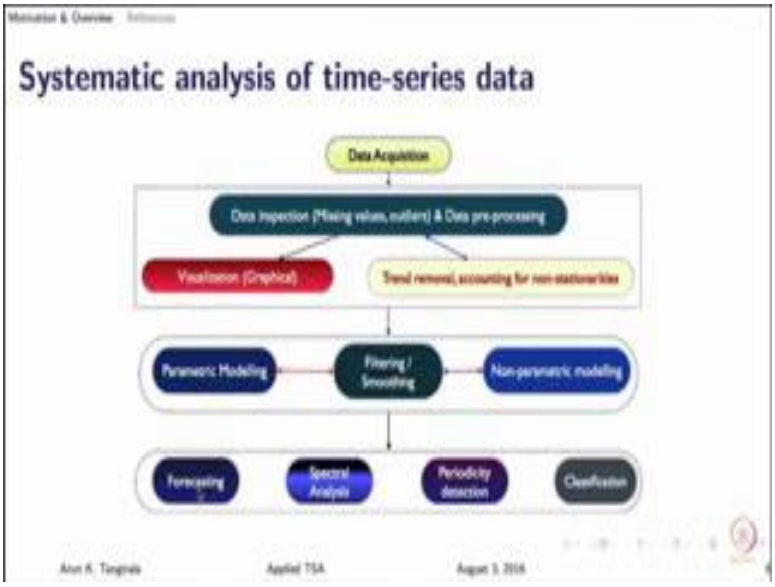
Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 02
Lecture 01B - Motivation and Overview 2

Let me actually quickly talk about what is involved; spend a couple of minutes in what is involved in a time series analysis. Typically it is time series analysis would involve forecasting and that is what comes to anybody's mind, of course, for people working in machine learning or in image processing and so on, it is more about pattern recognition, classification and so on.

(Refer Slide Time: 00:38)



If you look at the bottom, so let us look at the bottom most part here, you have forecasting as one of the branches of time series analysis that is actually one of the most popular exercises in time series analysis and by and large, in this course we will focus on how to build forecast, I use this term forecast, but there is also another technical term which is called prediction. So, you can they are both analogous, forecasting is more of a turn that comes from economics, prediction is a term that comes from engineering.

We will use prediction more often than forecasting as a technical term then you can have spectral analysis as we just discussed in the previous example where we want to detect maybe oscillations, periodicities, frequency content, we want to know what is a bandwidth of a

process and so on that is where we run into a lot of spectral analysis and it has a very set of I mean set of very vast applications not just an engineering ECG, I mean medical field or atmospheric data analysis, meteorological data analysis and so on. And then you have the of course, this periodicity diction which is a separate branch in itself spectral analysis is more of a broad one and then you have classification which we may not discuss.

In this course typically you learn concepts of classifications on in a machine learning course or some other course that is dedicated to a classification by and large, in this course we will focus on the other three that is forecasting, spectral analysis and periodicity detection that is enough to keep us busy for the entire semester. Now there are other things that you see in this schematic here which tell you that there is quite a bit that needs to be done before you start analyzing the data right. Analysis is about trying to discover what is present in the data depending on the objective that you have if your objective is prediction you want to actually see if there is scope for prediction if there are correlations and so on, and if your objective is periodicity detection then you are searching for oscillations and so on.

The objective will determine what kind of approach you will take in time series analysis, but regardless of what your objective is there are certain minimum preprocessing steps that one has to go through before we reach the analysis stage and there is this golden 80 20 rule which says that what is 80 20 rule say? It is applicable to every sphere not only in time series analysis, but every sphere of life what is the 80 20 rule say do you know, what is that? It takes 80 percent of the total time to do 20 percent of the part that is one version of 80 20 rule, 20 percent of the effort goes in actually doing 80 percent of the job.

Now, why do I bring that up here? This data preprocessing all the steps that you see from the acquisition to preparing the data for analysis is you can call this as preprocessing, prefiltering or whatever it takes enormous amount of time what we are interested in is the analysis part, but to get the data up to this analysis part it takes humongous effort and it can consume 80 percent of your time, of your overall time if you say from the time you acquire data to the time you complete your analysis; if you call that as t then 80 per point 80 goes in actually preparing the data and the rest is 20 percent, which is true for anything I mean you take the construction of a of a house or building it is 80 percent of the time goes it is in building the structure and so on and then you have the rest of the things finishing of the interiors and so on. This is true everywhere when you write your thesis also you will feel the heat.

The introduction chapter is the most difficult to write in a thesis that is a 20 percent of actually your thesis, but that is the most difficult chapter to write, the rest of the chapters are easy to fill because you would have already done your analysis and so on, but whereas, in the introduction chapters, you will have to give perspectives and so on, so you will spend 80 percent of the time in finishing up the 20 percent of the term and so on.

The same goes for time series analysis and it is very important to go through these steps here before you head for time series analysis do not ever jump straight into analysis, say I have data, I am going to throw into software called R or MATLAB or whatever and then I am done and here you go this is the report then you are not needed such a such a task can be automated and you can easily lose your job, but what that therefore, tells us is time series analysis is both a science and an art.

What we will learn of course, is a science part of it that is how to analyze, but the practice will help you realize what is the art in time series analysis, why do I say art is in any data analysis not just time series analysis in any data analysis there are certain things that are written in a cast in black and white. So, you know exactly you have to do that, but there are certain other parts of the exercise where there is subjectivity.

The user has to actually take some decisions as to for example, what should be the order of the model that I fit as an example? If you are actually building a model for forecasting one of the things that you will repeatedly encounter is making a decision on the right, so called order of the time series model that you will fit that there is no exact method or method that will give you an exact answer for the order. You will have to use a mix of statistical tools and your own judgment to figure out what is an appropriate order and that requires art which comes from experience. As you analyze more and more data you will you will have a gut feeling of what order should be chosen what are the steps how can you actually determine the order in the quickest way and so on.

(Refer Slide Time: 07:46)

So, we will learn the theory part which will tell us a science part of time series analysis and then the practicing part will teach you the art part of it. So, what I will do is I will not go further and we will continue this presentation tomorrow. In the remaining time that we have I quickly want to go over the course outline so that you have the details of the textbooks and how the course is going to be conducted and so on. So, let me pull up that information for you - I know probably the font size is a bit small for you this course outline is going to be posted on the modal website. So, all of you will have access to this course outline I will not go over the course objectives even if I read out now, it would not make too much sense tomorrow when we go through the remainder of the introductory presentation then the objectives and goals will be a lot more clearer to you of course, you are most welcome to read this from the modal site.

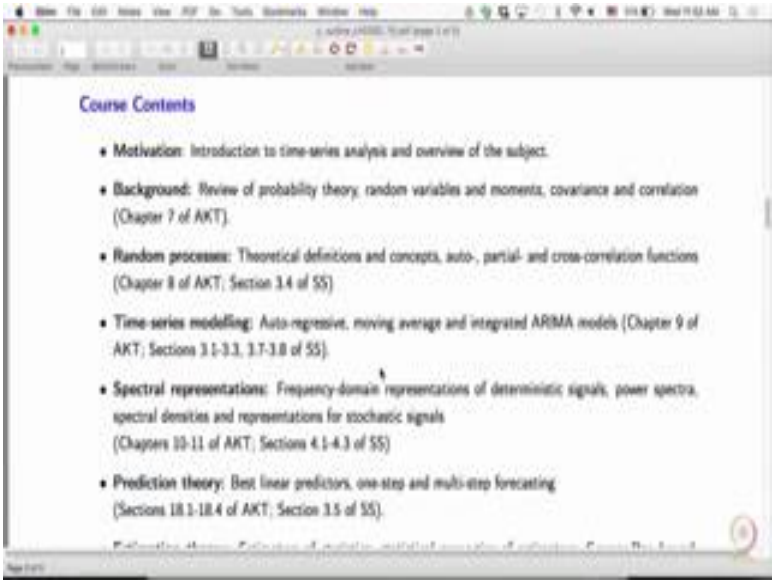
Most of you must be knowing that the prerequisite for this course is for you to sit through the 10 hour or roughly about 12 hours NPTEL MOOC course that I had conducted in the previous session on introduction to statistical hypothesis testing.

(Refer Slide Time: 08:27)

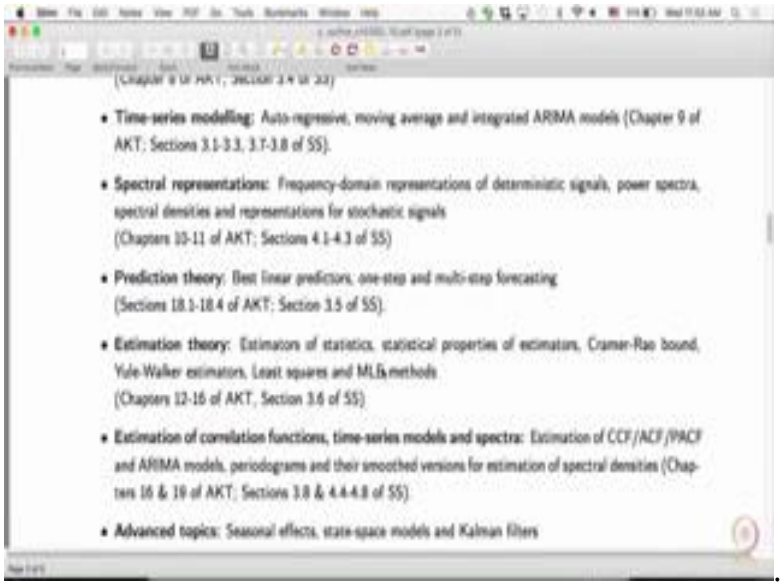
Now, this is the first time I am putting such a prerequisite in place because normally I spend about two to two and a half weeks going over or reviewing the fundamentals of probability and statistics and a bit of hypothesis testing, I would like to avoid that now that we have technology where we have video lectures and so on. So, if you can spare some free time from your Whatsapp and Facebook sessions and so on spend about 10 to 12 hours on this introduction to statistical hypothesis testing, you will benefit a lot because it will actually give you the solid foundation that is required for time series.

The MOOC course that I am talking about gives you an introduction to probability, discusses in ocean of a random variable, what is expectation concepts of correlation and, all that is got to do with purely random variables and probability theory, there is no time series concept that is explained in the MOOC. So, you will still have to attend these classes. So, nevertheless what I will do is I will have a very quick review of what is there in the MOOC, but that will be like a flash maybe about 2 lectures are the most that I will review. I would like to spend that time rather in going over and more important topic which students have been requesting for years to talk about Kalman filters and so on. Every time I plan to do that, but I have been unsuccessful hopefully with this arrangement I will be able to do it that is the idea.

(Refer Slide Time: 10:09)

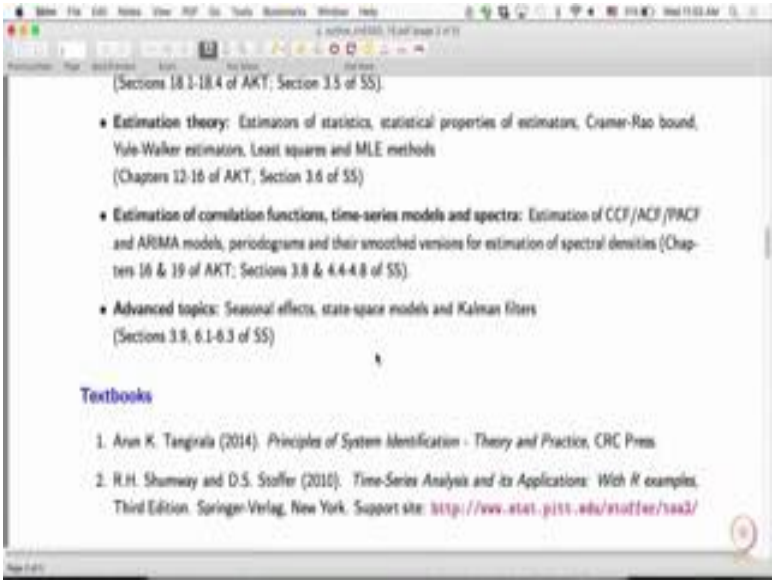


(Refer Slide Time: 10:11)



If you have any questions of course, we will answer that on the prerequisites part and here are the course contents as I said, it does not make too much sense for me to talk about the course contents today I will leave it for tomorrow because I have not discussed the basics of this course that is what are we planning took out, but I have told you, I said we will focus on univariate time series analysis, linear processes, will predominantly talk about how to build make predictions, how to build forecasts and how to carry out spectral analysis periodicity text detection and so on. And I will talk about the reference book very quickly. As you can see there is a set of advanced topics where I talk about state space models and Kalman filter.

(Refer Slide Time: 10:48)



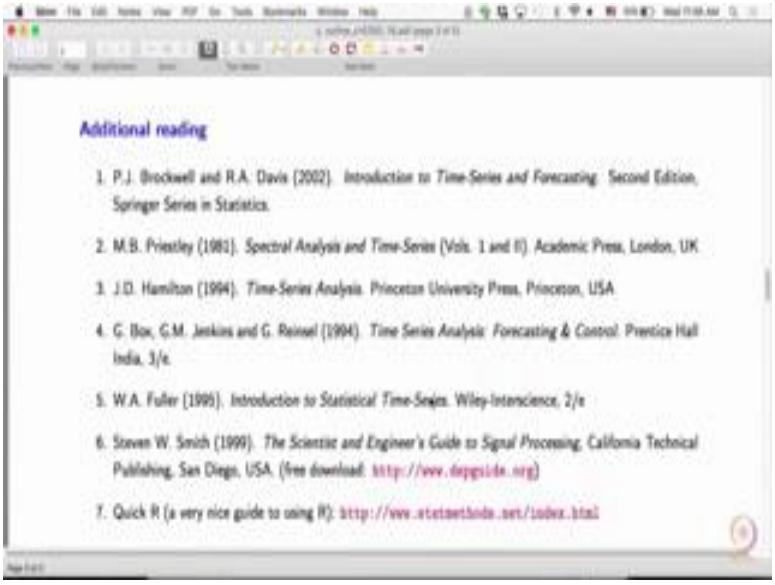
Typically I managed to talk about seasonalities. So, with the MOOC arrangement that we have I hope that we will be able to discuss at least the basics of Kalman filtering and state space models. Now as far as the textbooks are concerned right, as far as the textbooks are concerned there are 2 textbooks, one is written by a person called Arun Tangirala and it the title is principles of system identification, now that is a course that I offer in the even semester, usually time series analysis is considered a prerequisite for that an informal prerequisite which means that this textbook will also actually talk about that and you do not need to read the entire book because if you look at the size of the book you may even think of dropping this course, but do not go by the size of the book, there is a part of this book. In fact, what I have done for your convenience is as you go over the course contents have actually given specific sections of the book that will be useful in this course. That book is a large book and the part 2 of the book is exclusively devoted to so called random processes which is what we will discuss in time series analysis and the only glitch is that the book on principles of system identification uses MATLAB as it is base software whereas, we will use R.

That is not to be worried about because we will have tutorial sessions and the assignments will give you enough opportunities to own your skills in our. Other book that we will also follow, a bit closely is a book by Shumway and Stoffer it is a very classic book on time series analysis and applications. The nice thing about that book is of course, being written by pioneers in time series analysis. It also has throughout the book examples in R. So, the data

sets are available on the author's websites and the R codes are given for you both on the website and within the book and so on, so you will feel quite comfortable with that.

However the book assumes that you are quite familiar with the theory of random variables expectations and so on. So, it begins at a slightly higher level, you will be able to catch up with that again I have given the particular parts of the book that you have to refer for the course contents.

(Refer Slide Time: 13:18)



And there are certain other books for additional reading, now the choice is yours whether you want to do an additional leading whether you will have the time to do that. Mostly the lectures if you are attending the lectures and if you are solving the assignments on your own which is a big challenge, solving the assignments on your own and following the notes then you should be fine, but let me also tell you that this is one subject which is quite different from all the other subjects that you have must have seen in your curriculum because here we are trying to teach you how to model build models from data whereas, the rest of the courses must have taught you how to build models from first principles that is point number 1.

Point number 2, we are actually going to deal with randomness uncertainties and so on, the knowledge of which itself is quite uncertain at the moment for us, how to actually go about doing that so therefore, in every class you are guaranteed that you will be presented with at least 2 or 3 different new concepts that you have to catch up with. I am not scaring you, but I am just cautioning you that that is the state of affairs when it comes to this course and you

should be prepared as long as you like the mathematics and statistics part of it and the application part of it you will enjoy the course and I generally try not to overload anyone with the equations unless they are necessary. So, that is something to keep in mind and a lot of these books are quite useful if you do not like those 2 textbooks you can refer to these other ones.

So much for the course outline, the specific contents of this course and the software of course, I have already said are within the MOOC that I have referred to the introduction to statistical hypothesis testing, there are tutorials in our because fortunately that MOOC also uses R and it begins with a spoken tutorial.

If you have sat through the MOOC then you must have actually gone through a tutorial you can start off based on that. R is open source software, a very nice software for statistical data analysis and computational purposes as well as for graphical analysis that is plotting. It has been developed by people across the globe very powerful set of people are sitting in the R community and R is available in on all 3 platforms windows, Linux and Mac that is a nice thing about it what you also want to download is what is known as R studio which gives you a very nice graphical user interface to R, R comes with it is own GUI, you may not really like that so much as compared to R studio. I strongly recommend that you use a R studio, but you are most welcome to use other GUIs for using R, all right and the again the course outline tells you where to download R from apart from giving you some links on R tutorials videos and so on, quick R is one website where you will find lot of help on how to use R. So, what we will do is we will stop the lecture today.