Data Analisys and Decision Making - II
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 59
Metaheuristics techniques

Welcome back my dear friends and dear students; a very good morning, good afternoon, good evening to all of you wherever you are. And this is the DADM II which is Data Analysis and Decision Making II course under the NPTEL MOOC series. The total duration for this course as I keep saying it before the starting of the class it is for 12 weeks and we are in the 12th week. If we see the slide later on it is we are in the last, but one lecture which is the 59th lecture and with 60 we will wrap up this course.

And the this total duration as I mentioned was 12 weeks which is basically 60 lectures, each lecture being for half an hour and each week we do 5 lectures of half an hour each. And, you have already done 11 assignments with this twelfth one you will complete the 12th assignment and then take the final examination. And my good name is Raghu Nandan Sengupta from IME Department at IIT Kanpur.

So, if you remember in the, I did mention that again I will repeat in the 57th lecture the time duration was much less because it was 21 minutes because we wrapped up the concept of artificial immune system with an example. How prediction can be done from the financial point of view of bankruptcy and we discuss about clonal selections, negative, positive and combination of them. Then we have been into artificial neural network starting the 50th lecture and there I discussed very briefly though and if I have not been able to explain from the actual biologist or a doctor's point of view because that was not my intention I wanted to only give you a snapshot.

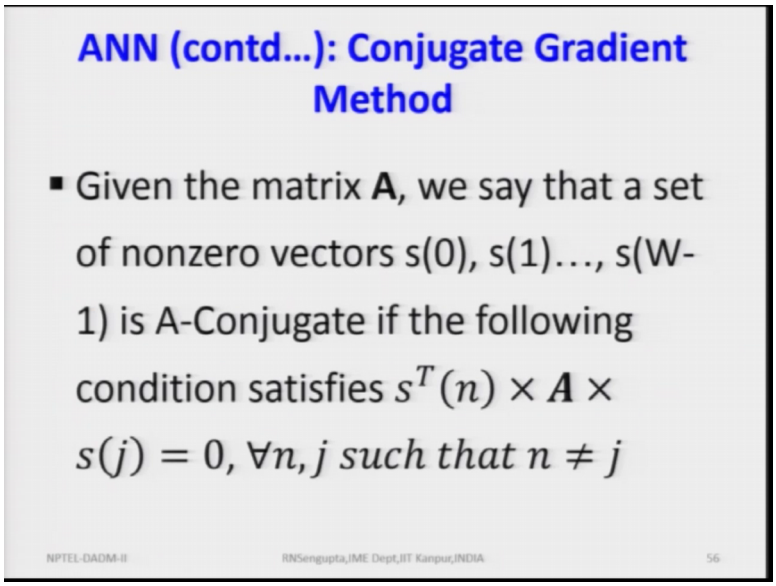
So, there are basically even a neural system, they are dendrites they pass on the information to the neurons, neurons then passes on the signals to the synapse to the connection they are not connected. So, they give this electrical pulses on in signals and brain processes it. So, that can basically be analyzed using the concept of the nodes; nodes has a sigmoidal function they consider the weights; weights are w 1 to w n and depending on the number of nodes you have and there are different layers the first input layer then hidden layer 1 2 3 4 till the output layer.

And the function and for the each nodes they would marked as 1 1 to 1 n for the first layer similarly 21 to 2 n for the second layer and so on and so forth where n is the number of nodes. So, I am not considering than n number of nodes for each the input layer the hidden layers need not be equal. This the function based on which the inputs are analyzed the sigmoid function and the parameters the x values can change. And also remember there is a supervised learning in which the feedback loop pumps in order to basically give a gift the whole neural network system or learning concept.

So, that it learns itself and basically fine tunes is prediction values or the output value I would not use the what prediction because if it is an objective function it is not prediction. So, it gives the output in such a way that we try to find out how good or bad is if it is not good then it learns itself and if it is good based on the parameters which I have said with the threshold we have set for ourselves, it basically passes on the output into the next layer and it goes on so on and so forth till we get the final output.

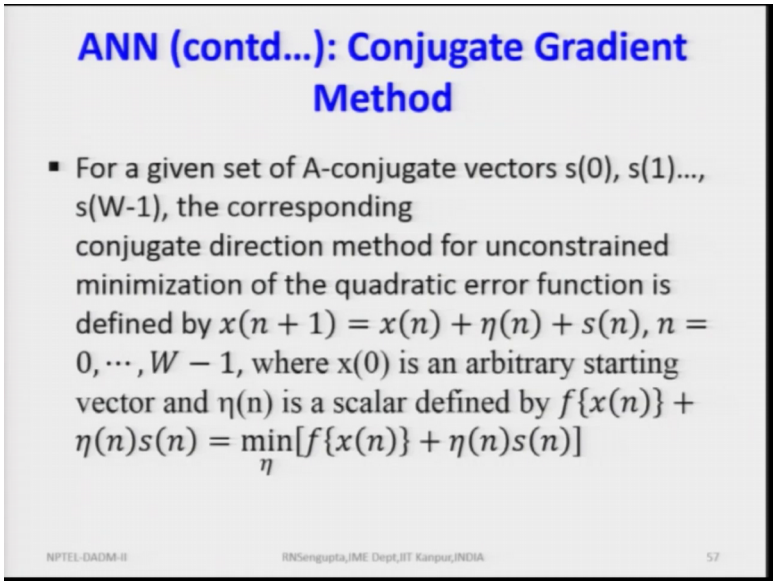
Then I had also discussed the back propagation the node functioning and then I also mentioned very briefly and I did mention it twice during the discussion, during the initial discussion of ANN and also at the last few minutes of the last class about change point detection and why it is important; important from the point of view of how we are trying to utilize it. So, the matrix which we have so, the objective function was half a transpose a x transpose a into x plus b transpose in x plus c b c's and a's values of minus plus are immaterial here, the objective function is more important.

(Refer Slide Time: 00:34)



Here a we say there is a set of non-zero vectors. And is conjugant if the following condition is satisfied. So, these are the mathematical terms which we have such that a s transpose depending on n n is the size of the parameters into a into s j where n and j are the values depending on the number of parameters which you have and n and j are not the same. So, generally the parameters for each node are of each layer may change what will basically have corresponding nomenclature to discuss that.

(Refer Slide Time: 05:09)



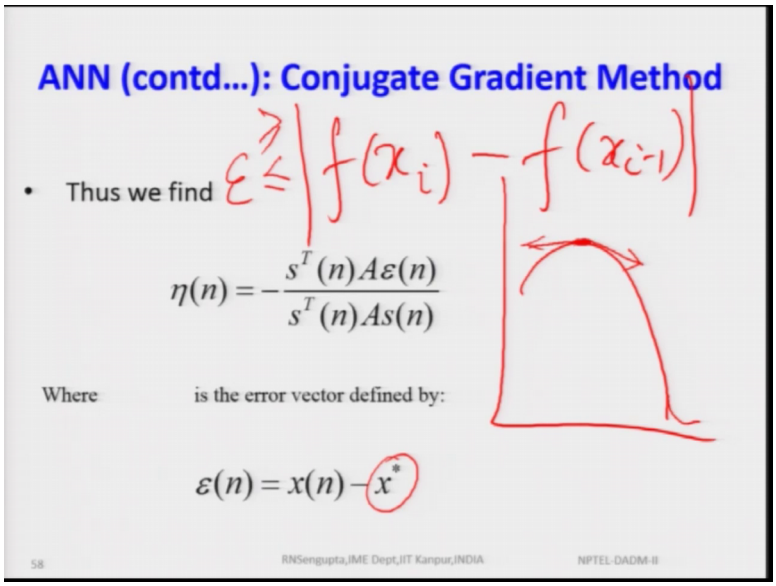
For a given set of conjugant vectors s naught to s W 1 W minus 1 the corresponding conjugate direction would give me for the unconstrained minimization and the maximization in which direction we should go. As I mentioned that partial derivative is generally utilized to find out the rate of change of the function and; obviously, the second derivative the hessian matrix would basically give us some information that whether is the maximum the minimum.

Obviously, we are considering that the second derivative is greater than 0 or less than 0 is not equal to 0 because if it is equal to 0 it will be the point of inflection. Now the with the quadratic error function which will basically expand depending on the functional form. So, it will give me in which direction we are trying to minimize or maximize, I am not going to go into the mathematical detail this is not important. Because if I generally do we try to basically go into the depth then; obviously, a whole set of lectures about 20 to 25 should be divided to ANN and similarly for AIS and so on and so forth.

But this DADM II is basically a set of lectures which only would basically give you a brief idea in a very simple way about the non-parametric method of optimizations which I have considered the P I will come I have you know that where P TOPS is ELECTRA then the concept of this utility function which have been utilized forecasting methods so on and so forth. So, we have also considered AIS and the ANN method is being discussed now. So, in the gradient descent method what we are doing his base basically I will repeat it.

We are trying to take the direction in which the maximum descent is their optimization problem from the point of view optimization problem or will go in the direction where the prediction is the best. So, there would be some criteria.

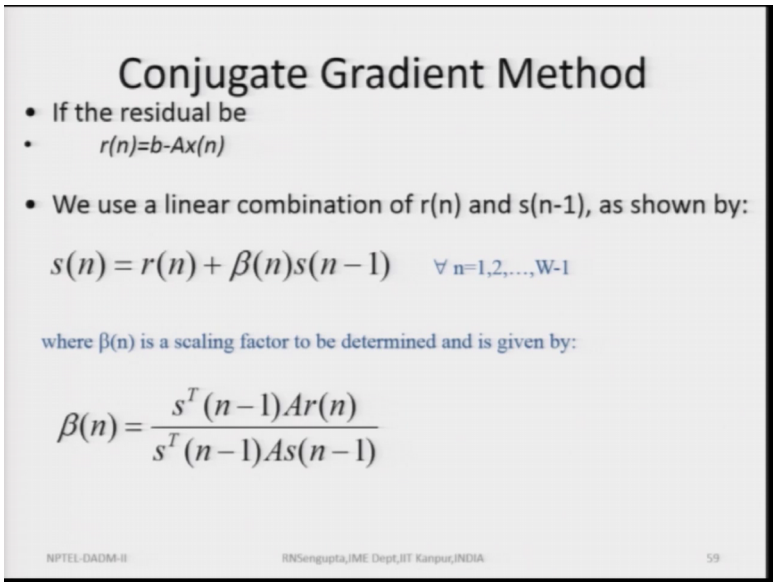
(Refer Slide Time: 07:15)



So, the in this case the actual value is extra which you have already discussed and if you consider from this point. So, here in a 2 dimension one either the tangent in this direction or in this direction I am trying to draw it as simply as possible simple. So, if it is in is a minimizing problem; obviously, you move in that direction where the negative change is happening the maximum and hence the distance would be for the fastest in the positive it will be just the other way round in the maximization problem.

And we will find the errors such that the errors; which is epsilon would be found out depending on the difference. So, what we are trying to do is actually we find out x in the ith stage depending on way how you are moving based on the fact x minus 1 and we try to find out whether this value is less than equal to epsilon or greater than equal to epsilon, but the moment it is epsilon value has been fixed from our side. So, this is the epsilon value based on that we stopped the process.

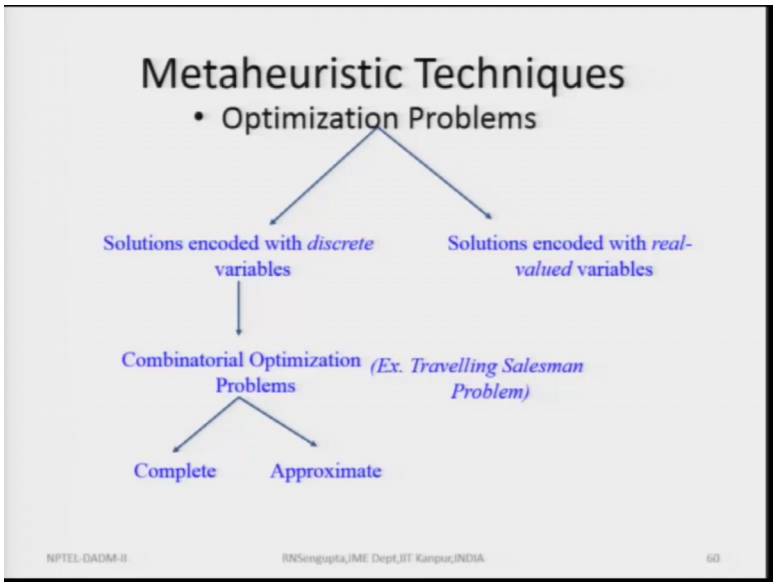
(Refer Slide Time: 08:22)



Now, the residuals would be given. So, then that that its residual concept is something to do with the multiple linear regression concept the errors which are there. So, we try to basically minimize the errors in the multiple linear regression concept using the rate of change of that values which is basically we try to differentiate with respect to all the alpha alpha beta 1 beta 2 beta k then put it to 0 and find out the values.

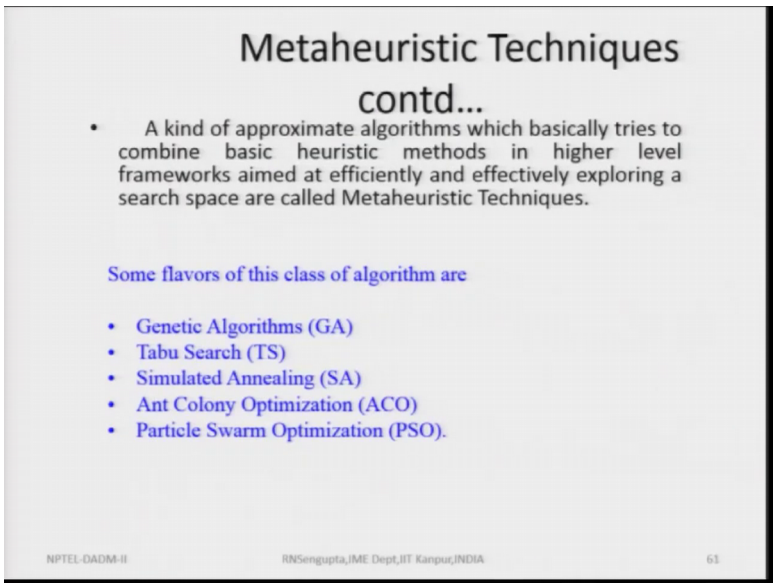
Now here the linear combinations would be given such that based on the linear combination we can the scaling factor can be found out. I am not again I am repeating I am not going to the each and every fundamental details of the equation. I am just giving you the idea based on which I how I can build up the model singularly as I did for AIS.

(Refer Slide Time: 09:12)



Now, here it was a very brief discussion very very brief discussion of ANN. Now I am going to utilize this concept of ANN in change point direction. So, it can be solutions can be encoded with discrete variables an optimization problem you can basically have the complete an approximate solution in this case of the approximate solution we are going to use the heuristic methods.

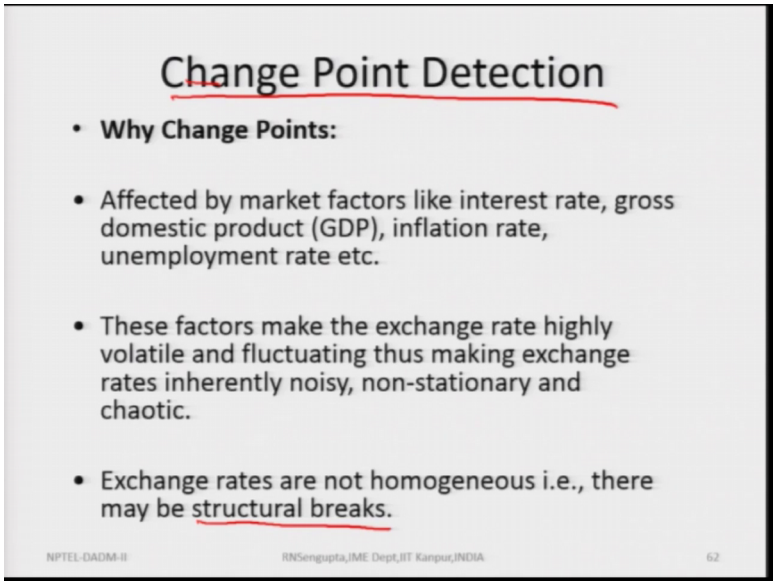
(Refer Slide Time: 09:36)



The metric techniques is basically as I already said is a kind of approximate algorithm it basically tries to combine basic heuristic methods in higher level of framework aimed at efficiently and effectively finding out the solutions.

If you remember I did mention about plan and colony optimization, simulated annealing, artificial neural network, genetic algorithm and all these things were there. So, these are particle swarm optimization. So, these are class of problems which we have generic algorithm tabu search simulated annealing and all these things.

(Refer Slide Time: 10:10)



Now, comes the concept of change point detection now change point detection is basically the concept which will try to utilize from the point of view of statistics. So, consider this you have a data set and the data set underlying distribution if it is disillusion means I am talking about the PDF or the PMF or in general the distribution function which is capital F.

So, if I have a normal distribution the distribution function is known. So, the parameters are actually a distribution function or in statistics would have basically three parameters alpha beta gamma on location scale and shape parameters. Now, the parameters which are there in the normal distribution are mu and sigma similarly for exponential distribution it will be alpha and a or it can be theta a depending on where it starts.

Now, fundamentally a data set which we have basically may have its underlying distribution is evolving with time. So, they can be two types of evolvement; one is the distribution per say remain same that is it is normal it remains normal if it is exponential it remains exponential, if is gamma it remains gamma, if it is beta it remains beta, but the parameters themselves change. That means, if it is a normal distribution we are getting the normal distribution, but say for example, alpha this mu is changing or sigma is changing or both of them not changing.

So, our main aim in change point detection is to basically predict the values of alpha beta or gamma which in the normal distribution cases the mu and sigma and predict also when they are changing two things predict the values predict the time when they are changing. Number 2 problem which is one step higher is basically the distribution function by itself is changing that it was normal changing to exponential I am giving you very very simple examples.

So, it was f 1 x PDF value now it is changing to f 2 x PDF function. So, per say the parameter is also changing the functional form is also changing. So, we need to predict what is the PDF value of what is changing? So, f 1 changing to f 2 or f 2 changing to f 3 and when it is changing. So, change point detection is basically would be utilized in a very simple sense the points at which the change points are happening and to predict the change points we will try to utilize the artificial new neural network system along with one of the heuristic methods.

So, this is the general background. So, consider I will come to this change point detect in the problems using the data point I will come to that now why change points? So, they are affected so, consider if you are considering the finance concept or a marketing concept or the flow concept of fluid whatever it is a heat movement through 2 conductors. So, if suddenly if the heat changes is not happening because there is some congestion flow fluid flow is not happening because this and there is some congestion.

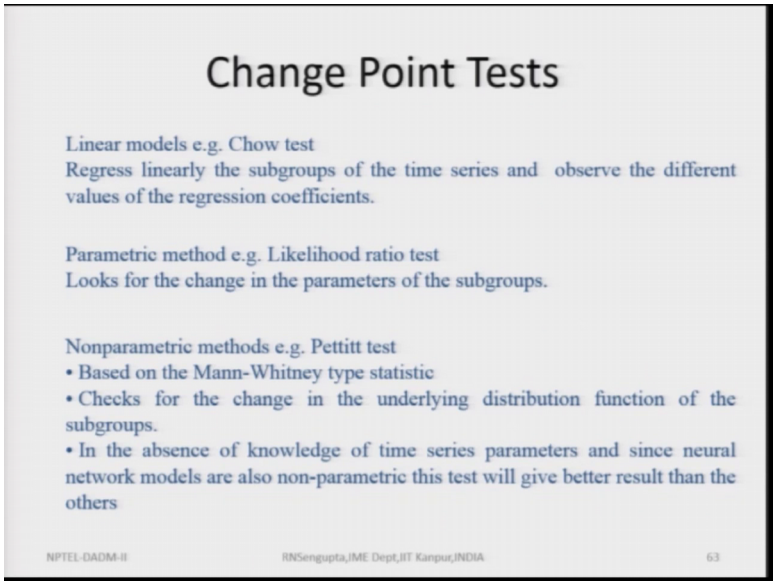
The pressure is increasing decreasing the change the distribution by itself or the rate of change is changing. So, you want to predict it. So, in this case of economics finance this change point detection is heavily used. So, why because the effect on the marking factors could be their or interest rate we want to find out the rate of change in the interest rate the gross domestic product is changing inflation rate is changing unemployment rate is

changing we want to predict when they are happening these factors make the exchange rate.

So, this affects which are happening for the exchange rate because our main aim would be to a final exchange rate for this study because, exchange rates are highly flow volatile and fluctuating through thus making extended inherently noisy non-stationary and chaotic and we want to basically find out when and where the changes are happening.

Exchange rates are not homogeneous that there are may be structural breaks. So, change point detection another word for change point detection is structural breaks. So, in many of the books if you study; obviously, in the area of statistics the concept of structural breaks or change point in detection are chosen interchangeably.

(Refer Slide Time: 14:33)



Now, there are different tests for change point I just mention them without going to the mathematical details because not as a statistical course. So, there you have the chow test which is the linear models then you have the regression methods are basically when they are utilized then you can use the linear models of this chow test.

Because we remember in regression model the model is linear; linear in the sense of the parameters. So, it can be alpha into x square depending on x square has a distribution which is a normal distribution if you remember amongst the 7 different assumption in

multiple linear regression one of the assumptions is basically the dependent and the independent variables all should be linear and they should be independent also.

Because if they are not independent then that concept of cook this covariance and all these things would be coming these were among those 7 set of important assumptions. You have the parametric methods also the likelihood ratio test which are there for the non predicting methods we have the Pettitt test which is based on the Mann Whitney type statistics depending on non parametric values happening at the rate such that the changes of this parameter values are non-linear in nature.

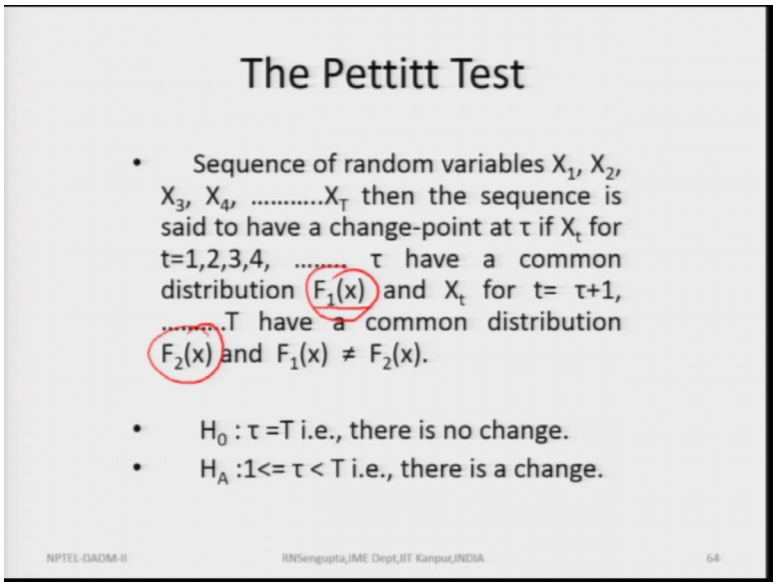
So, they will check for the changes in the underlying distribution function of the subgroup. So, we will what we will try to do is that technically it may be possible that the distribution f of x or f 1 of x is changing and as it is if evolving it may. So, happen that the changes are happening randomly; obviously, they would be happening at the random points which you do not know any we want to predict at the time. So, they would give it they would basically given as two areas which has discussed the change of the parameters are change of the functional form itself. Now, if the underlying distribution function you divided into subgroups.

So, today I find one say for example, the in this foreign exchange is same and it has and it does not fluctuate. Say for example, I want to predict it 1 year down the line, but in between that 1 year it may have change and again come back to its origin ourselves. So, subgroups are important to find out predict that if I basically have the invest consider from the point of view investment if you invest and you want to go out come go enter the market or come out of the market in between that 1 year.

So, what are the sub groups in the sub groups what are the change changes and what points they are changing, what is the rate of change of the interest rate? Such that I do not enter the market where I have to pay a higher fee I do not come out on the market where I get the lower fee that means, I want to basically minimize my loss. In the absence of knowledge of time series parameters in this case for the mainly for the in this exchange rate because exchange rate are dependent on many things; it can depend on the GDP, it can depend on the gold price petroleum price inflation interest rate of other country's population whatever it we do not have the parametric form.

Hence, as parametric forms are not possible we will try to utilize neural network concept along with that will also try to utilize some of the heuristic methods in order to do the optimization; optimization means the prediction of the change point as good as possible.

(Refer Slide Time: 17:54)



Now, here is basically what I told you very simplistically verbally here is basically the concept of the change point in the Pettitt test consider a sequence of random variables are X 1 X 2 X 3 till X T mark the word capital T in the suffix. Now, what we have is basically the time, so, I measure X I find out X with the suffix one at time period 1 X 2 at the time period 2 so on and so forth. So, then the sequence is said to have a change point tau if X T; that means, T starting from 1 to T small tau will have a common distribution of F 1 X this n 1 does not mean have any significance with 1 2 3 4 in the time series basically functional form.

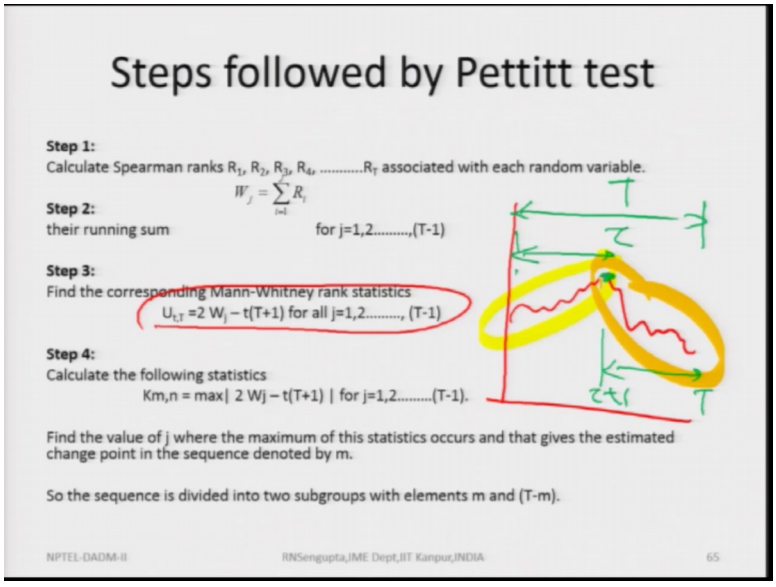
And starting from tau later on that is tau plus 1 till capital T they would be another distribution F 2 such that F 1 and F 2 are different; that means, I am considering per say the distributions also is also changing I am not considering per say the parameters values are changing. So, it is a simplistic form I am considering that normally changing to exponential or explanation is changing to gamma so on and so forth.

Now, if I have; so we will try to basically have the Pettitt test as an hypothesis testing point where H naught would be the case that if tau is equal to T there is no change. For the time period which I am considering if there is no change that H naught would be

ratified or we agree with H naught depending on the test rule. And if H A is ratified; that means, H naught is rejected you will find out that at point tau where this point is changing we will mark that as the change point.

Obviously, it may change you may say that they can be many tau's in between capital T. So, it can be tau 1 tau 2 tau 3 till tau n or tau k, yes it is possible and we can find out using the Pettitt test accordingly. So, we will basically check very simply mention the algorithm based on which you do.

(Refer Slide Time: 20:03)



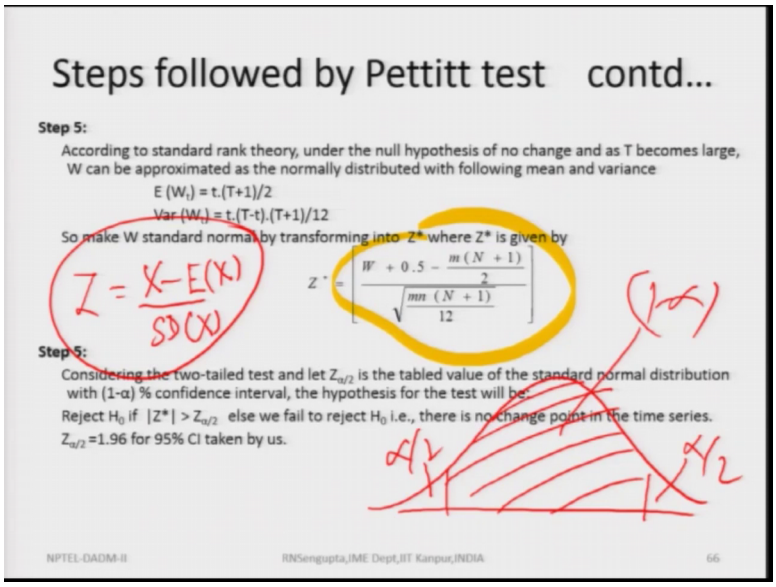
You can calculate the rank spearman rank is basically the one of the correlation coefficient we use the spearman ranks would be found out associated with each random variables and we will find out the sum of the spearman rank coefficient. Then we will basically find out the running some of them spearman ranks for the corresponding Mann Whitney rank would be found out depending on the formula of U t for j is equal to 1 2 T minus 1 for all the points where we are finding on the Spearman rank values. You will calculate the values of Mann Whitney ranks by finding on the maximum of them.

And find this I am just mentioning the bullet points of the steps you will find the value of j where the maximum on the statistics occurs and that gives us the estimated change point of the sequence. So, the sequence would be divided into two subgroups; one from t is equal to small t is equal to 1 to m and another would be from m plus 1 to team t. So, it will be divided 2 groups; the first group and the second group where the changes are

happening. So, if I have; so this is the change point. So, let me mark it with a different color.

So, this is the change point we find out so, the time period. So, this is virtually capital T this is from 1 to tau this is tau plus 1 I have trying to draw it as simple as possible is and this is from 1 to tau. So, they can be more so, this is one subgroup this is one subgroup and similarly this is one subgroup based on that we can do the calculations.

(Refer Slide Time: 22:03)



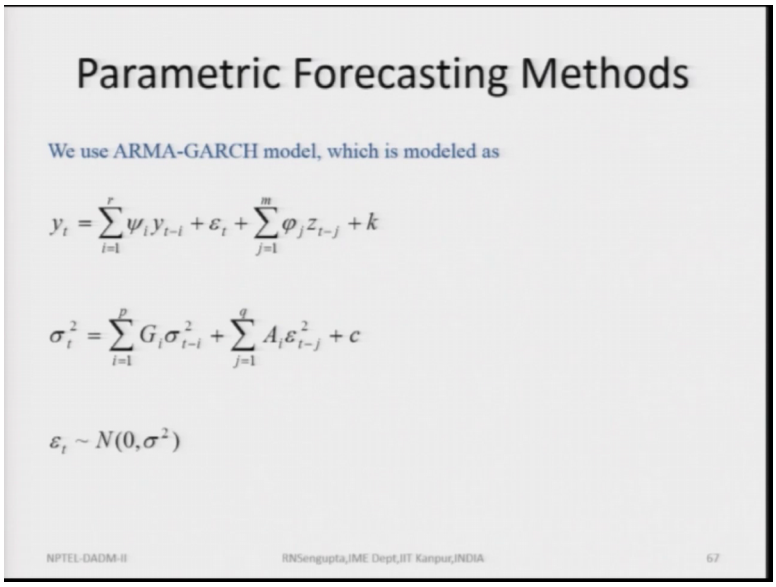
In step 5 according to the standard rank theory we will try to find will under the null hypothesis we will consider there is no change and as T becomes large we can approximate using the normal transformation. So, here is the normal transformation. So, this is the expected value and the variance. So, it is basically exactly the same way. So, you have basically in the Z test what you do X minus mu by standard deviation of X.

So, anything I should mean write it like this so, it is easier. So, this is the expected value of X we are using this concept that is the central limit not in the exact central limit theorem, but the concept of central limit theorem would be utilized. Now considering the 2 tailed test the 2 tailed means on the left hand side and the right hand side and the alpha values which are there.

So, alpha various means alpha by 2 alpha by 2. So, this is 1 minus alpha 1 tail 1 tail test will be on to the left of the right depending on that you can find it out. So, if you

remember in statistics if you have; obviously, done on is somebody who has done DADM I we considered the concept of hypothesis testing greater than type less than type and not equal to and based on that we have three different rules, which were busy coming from the concept of interval estimation. Now, the parametric forecasting methods which we generally have which will try to basically incorporate mean the ANN these are not to go I am just mentioning them.

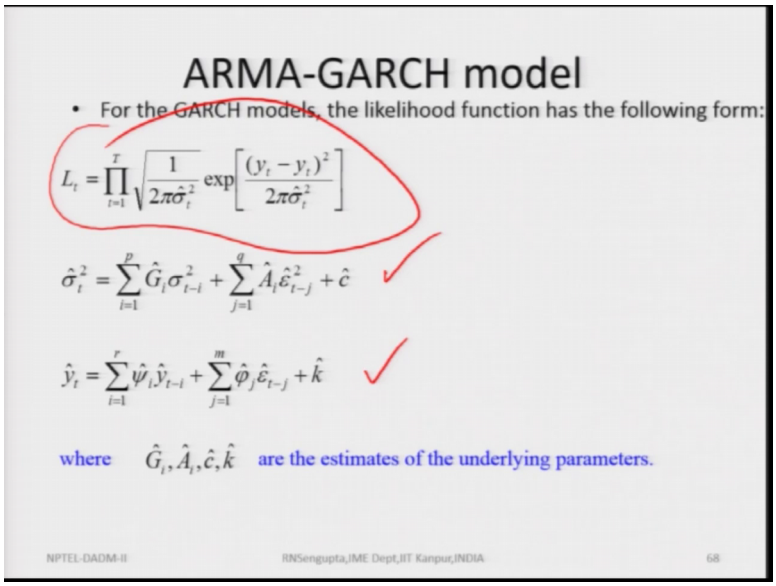
(Refer Slide Time: 24:00)



So, you will basically in the ARMA GARCH method which are in the forecasting method we did consider very simply. The concept of Holt linear and Holt Winter method Holt method then moving average, explanation moving average, adaptive moving average; so, based on that we will try to predict where the prediction would happen depending on some error part and its and its regressed values to what level you will go.

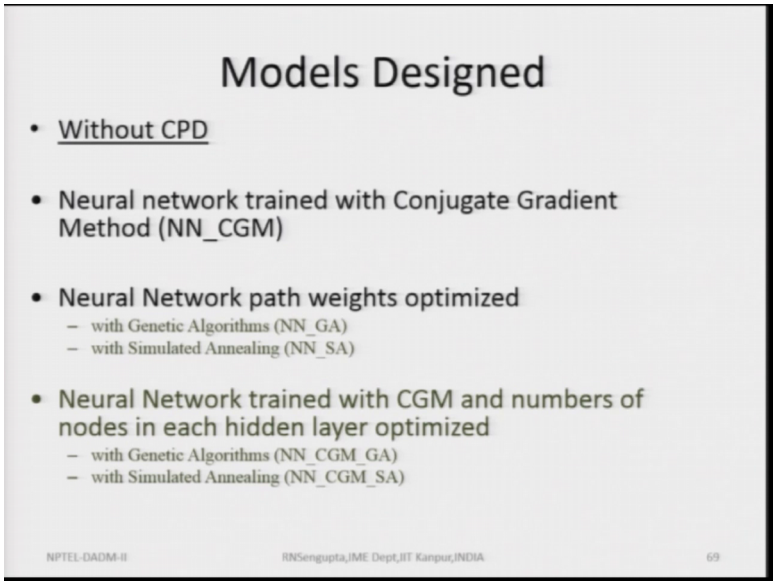
So, if you are predicting basically for y t you can go for regress to y t minus 1 y t minus 2 and so on and so forth. Again these explanations I am omitting they are not required because the main focus for the ANN how you can utilize that for the ARMA GARCH method; obviously, the log likelihood function for the prediction models can for that.

(Refer Slide Time: 24:52)



This is basically I am considering the log likelihood function; likelihood function it will be converted to log likelihood function the best estimates for sigma sigma hat and y which is y hat would be found out using these models; where the hats are the estimates of the underlying parameter values which we are trying to utilize.

(Refer Slide Time: 25:14)



Now, we basically utilize the model for the NN one using these; what are these I will come to that later. We will consider the neural network which will be trained using the conjugate gradient method simply as it is developed in ANN literature will use the

change point reduction for the neural network combining two different methods; one is the genetic algorithm one and one is the similar in handling.

So, it will be mentioned as NN GA NN SA depending on NN being neural network the first towards and G and SA3 being basically for genetic algorithm and simulator handling. Now, you will basically have the neural network with the conjugate gradient method utilizing both GA and SA; that means, in the first stage we are only utilizing the conjugate gradient method. In the second method we are only using the ANN we with the GN SA and the third stage will combine the existing conduit gradient method being we refine more by the use of genetic algorithm and similar training and check how the results are better or worse. Now with the change that was basically without the change point detection.

Now, with the change point detection concept again we utilize the change point detection along with the gradient one. So, I will come I will repeat this overall picture again. So, in the change point detection with those Pettitt tests and all these methods which we discussed the rank coefficient and all these things will consider using a conjugant method we will use the change point detection methodology using the GN SA.

And, then basically combine the change point detection methodology with the conjugant method along with GA one time and SA one time. So, overall picture is like this no change point algorithm we utilized will use the gradient descent conjugate gradient method then we will use the concept of GN SA separately and then combine them.

In the change pointed direction the algorithm being utilized using the statistical point of view will use the conjugate gradient method once separately will GA and separated will SA and then combine them accordingly. So, with this I will end the 59th lecture and in the 60th lecture I will try to give some results for this mythology how we utilize them have a nice day.

Thank you very much.