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

Lecture - 55
Loss Function

Welcome back my dear friends and students, a very good morning, good afternoon, and good evening to all of you. And this is the DADM which is Data Analysis and Decision Making I course under NPTEL MOOC. And this total duration of the course is 30 hours, which is a 60 lectures and which is spread over 12 weeks. And we are in the last lecture for the 11th week which is the 55th lecture. And each week we have if you remember 5 lectures each for half an hour and after each week we have one small assignments and my name is Raghunandan Sengupta from IME department, IIT Kanpur.

So, if you consider the discussions about the multiple linear regression, what we did discuss was basically what was the main essence is to find out the betas, which are the regression coefficients to find them use we use the concept of orderly square. This is a conceptual background for that because you want to basically find out the first find out the unbiased estimates for the betas, find out the errors, square them up, sum them up, differentiate with respect to the betas such that we put them to 0, and find out those betas so that will give me the orderly square best estimates.

And if you want to double check, plot the histograms for the betas for different sample sets. See them it will be normal distributed with the mean value equal to betas beta hat this would have a mean value equal to their respective betas. And obviously you can find out the variances which has been given by the variance covariance matrix. Similarly, if somebody is utilised utilising the LINEX loss, we gave the final beta values. So, it would be the beta hats plus minus nuisance parameter depending on the plus minus being depending on the value of a. And you can utilize those values of beta tilde also.

So, using the beta hats you can find out the y hats. Then find out the differences, using beta tilde, you can find out the y tilde, you can find out the differences, compare the errors which are more or less depending on the over estimation, under estimation, penalty cost which have discussed in the three examples. One related to the product, one related to the civil engineering design, and one related to the electrical design.

Now, we will consider something to do with the balance loss function. Now, if you remember the balance loss function, I will repeat it, please bear with me. In the balance loss function as proposed by Zellner. You basically had the precision of estimation, and the goodness of fit. Precision of estimation was basically what we have been discussing so far is something to do with the betas or the regression parameters which you want to estimate. And then we want to basically find out that how good or bad the fits are. So, you leave the precision of estimation and then go into the goodness of fit.

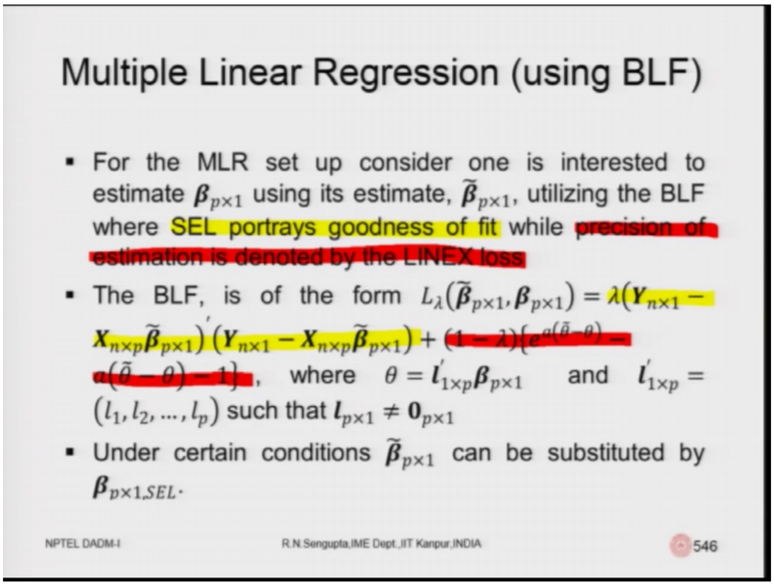
When you go into the goodness of fit, your main concern is basically to find out the differences of y and y hat provided y hat is the value calculated based on the beta hats where beta hats are being calculated using the concept of orderly square. Now, the question comes like this, well, if you have the precision of estimation, and the goodness of fit, why not change the background of the problem keeping the main thrust area that being precision of estimation, and goodness of fit in order to basically calculate different sets of beta such that we can find out under what circumstances on what type of betas we should use.

Now, coming back to that what you can do is there, and what we will now discuss is basically a balance loss function where there are 4 different combinations. In combination 1, the precision of estimation and the concept of this goodness of fit both are squared error. And we use the betas as betas coming out from the orderly square that is case 1. Case number 2 we considered that the concept of a LINEX loss is being utilised. So, the LINEX loss is being utilised in the sense that we still continue using the beta hat which is the beta estimated using the orderly square.

In the third example we use again the different combinations of precision of estimation and the goodness of fit, but in this case we try to utilise the concept of this betas coming out from the LINEX loss function. Beta hats come from the LINEX loss which is basically beta tilde. And final again we use different combination of precision of estimation, and the concept of goodness of fit, but still it can continue using the beta tilde which is under the LINEX loss.

So, with this I will give you the 4 different backgrounds I would go into the rehearsals, but these are basically can we considered in different type of a problem solution in the in the actual practical case, which would definitely gives you much better results in than in the theoretical results. Obviously, theoretical results would be very nice. In the theoretical sense they will good give good results in the theoretical point of view, but from the point of view it is better way use the LINEX loss using the concept of whether under estimation is more finalise or over estimation it is more finalised. So, let us discuss these four different simple ideas in the corresponding 4 slides.

(Refer Slide Time: 05:41)



For the case of multiple linear regression set under the conditions. We are interested to find out the beta vector which is basically vector of regression coefficients using the estimate beta tilde those here beta tilde is being used very generally. So, we use the balance loss function where the squared error loss portrays the goodness of fits. So, this is where the, so the squared error portrays the goodness of fit. While precision of estimation is denoted by the LINEX loss.

So, now, you have two loss functions squared error LINEX squared error being used for goodness of fit that is with respect to finding of the differences between y, and y hat. And finding out the square of that and trying to find out the errors, while precision of estimation would be utilize with respect to LINEX loss when you are trying to find out the betas them self. So, the balance loss function would be of these forms you are going to weights.

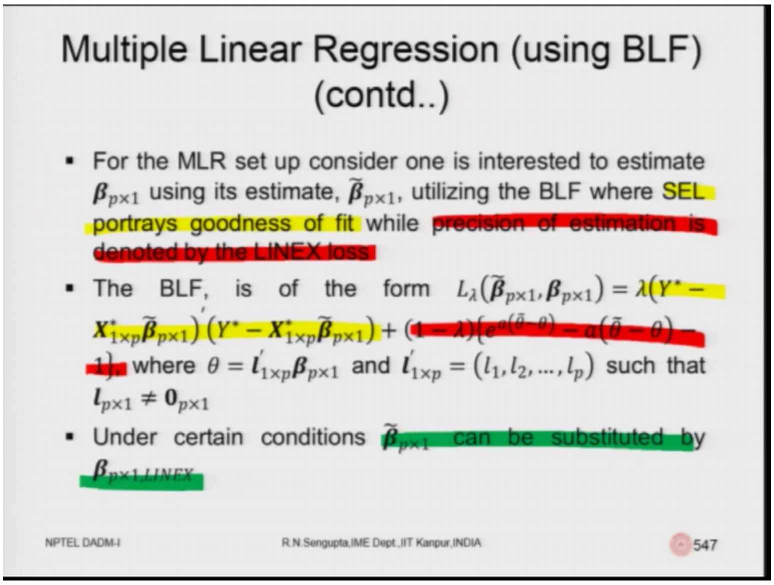
So, weight should be given such a way you will give a weight of lambda to one set of errors and 1 minus lambda to the other set of errors. So, what is the error for lambda. So, let us go to the goodness of fit. So, use the colour yellow. So, goodness of fit lambda where you are try to find out the difference between the actual value and the predicted value. Actual value is y that bold and predicted value is basically x multiplied by the beta tilde which you found out using the concept of squared error loss

While if I go to the precision of estimation is you give a weight of 1 minus lambda, and here you want to basically calculate the value of theta. Theta is basically a vector which basically theta is a value which is basically multiplied by 2 vectors l, l is an arbitrary vector consisting of some numbers so that would depend on the type of problem we are going to do. So, if you basically have p number betas, so l would also be a p vector, but they would be multiplied by column versus row or row versus column, basically to give you the scalar value which is theta.

Now, remember one thing if l 1 value which is call responding to beta 1 value is the 1, and the rest are 0, obviously you find out beta 1. If say for example, l p is 1, and rest are 0, you will find out basically beta p, so it can be changed accordingly. Now in this case continuing this discussion you give basically the LINEX loss in order to basically estimate the theta. Theta is basically the precision of estimation which you want to do.

And here you consider theta is equal to that l transpose into beta. And this l transpose is basically vector or a scalar or a column vector or a row vector in which ever you denote as l 1, l 2, l 3, till l p an l p the that p cross 1 vector which is there not all cannot be 0. So, under certain conditions you can basically replace beta tilde which is the estimate which you are utilizing for beta with the squared error estimate. Squared error estimate is basically the beta hat which you have used. So, you replaced that and you can do the calculations accordingly

(Refer Slide Time: 09:07)



Now, consider the second type of formulation on the same concept. But we will basically being presented in a different way for the MLR setup consider one is interested to estimate beta again beta p cross 1 beta is a vector using beta tilde. So, beta tilde is being used generally, but here the balance loss function has been modelled accordingly. So, what is this squared error loss portrays the goodness of fit which is as it is the precision of estimation is denoted by LINEX loss which is as it is, but the mean difference which I should I highlighted in the last diagram.

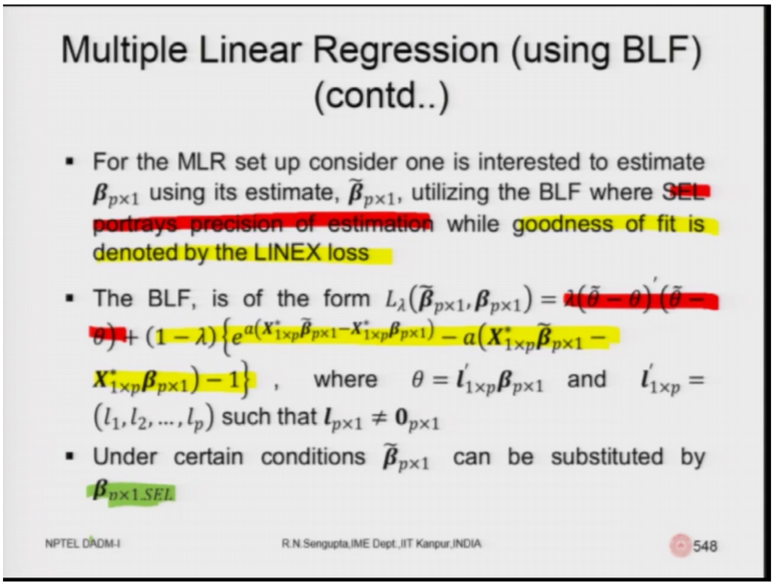
But, I also highlighted with the green colour main difference being beta tilde being substitute by beta from the LINEX loss estimate. So, just watches the colours squared errors for the goodness of fit, precision estimation LINEX loss, and beta tilde, which we utilise to calculate would basically be coming for the next loss some background.

So, the balance loss function is now and highlight it the yellow colour being. The goodness of fit which you have one minus lambda is the weightage which will give, for the precision of estimation. And again here theta will be equal to the vector l column or row which ever you want to denote multiplied by beta. And theta is basically scalar quantity l as I said none not all can be 0 under certain conditions we will use beta tilde being replaced by beta corresponding to the LINEX loss.

So, first case was the goodness of fit was squared error. The precision of estimation was LINEX, and use beta the estimated beta has to be coming from the orderly square setup.

Second which is there in front of you. Again squared error loss for the goodness of fit LINEX loss for the precision estimation, but now you basically use beta the estimated value of beta coming from the LINEX loss setup. Let us consider the third set up.

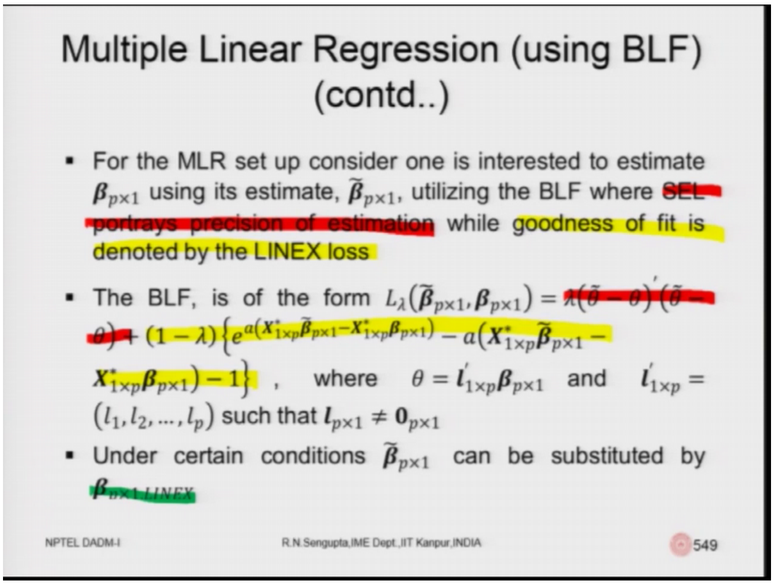
(Refer Slide Time: 11:24).



Again I will highlight using the colours I will remove all them. So, goodness of fit was, now which was initially squared error now it is using the LINEX loss. So, goodness of fit was basically LINEX loss. So, I will highlight that. So, this is the LINEX loss now this is the LINEX loss sorry my mistake my mistake though this is the goodness of fit. And the precision estimation and basically come out which here I am using the red colour squared error loss portrays precision estimation.

So, this is the precision estimation. Now under this conditions again theta I will repeat it theta is basically the scalar which is the multiplication of l vector multiplied by beta vector is are whether scalar or which are column or row you can descend accordingly. And l values are not all can be 0, and under certain circumstances we use beta to be replaced by the orderly square which is beta hat which I will again highlight using the red colour green colour whichever I decided.

(Refer Slide Time: 13:00).



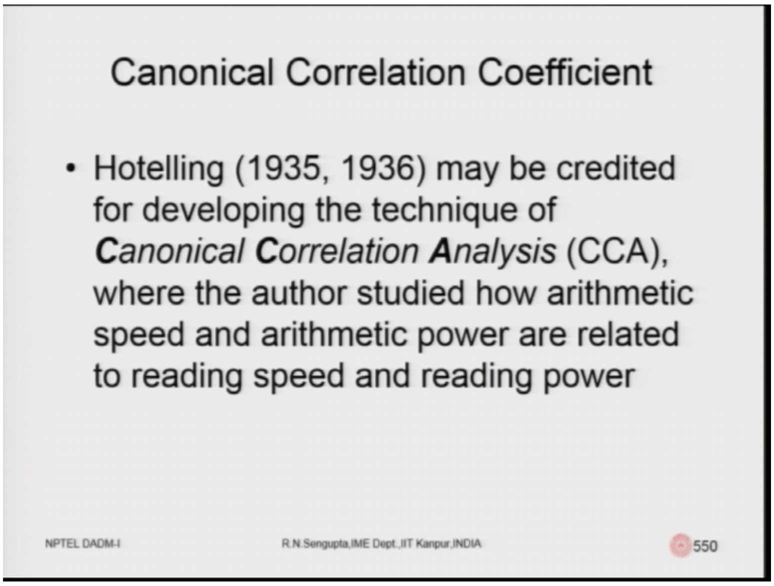
Now, come to the fourth example. Here I will use goodness of fit again to be denoted by LINEX loss. So, the problem is for the multiple linear regression 1 is interested to estimate beta, using beta tilde again beta tilde is being used very generally. And we have basically the goodness of fit, and the squared error loss portraying one error, and goodness of fit being portrayed by other errors. So, goodness of fit let me use the green yellow colour goodness of fit is being denoted by LINEX loss.

So, goodness of fit will basically denoted by the LINEX loss which is here, and the concept of squared error loss being used for precision estimation would be denoted by here where we precision estimation are based estimation is basically related to finding out the best estimate for theta which is the function of beta because beta is been multiplied by l so; obviously, is a linear combinations of beta and a such our general beta.

So, another condition again theta is basically multiplied to multiplication of l into beta in into in the column of row vector format. And all else cannot be 0 l values. And here also we can basically substitute beta with beta tilde where beta tilde is not basically with the LINEX loss. So, in these two cases you have four combinations goodness of fit being squared error and precision of estimation being LINEX. Under these combination you had beta tilde being either beta from the squared error or beta from the LINEX.

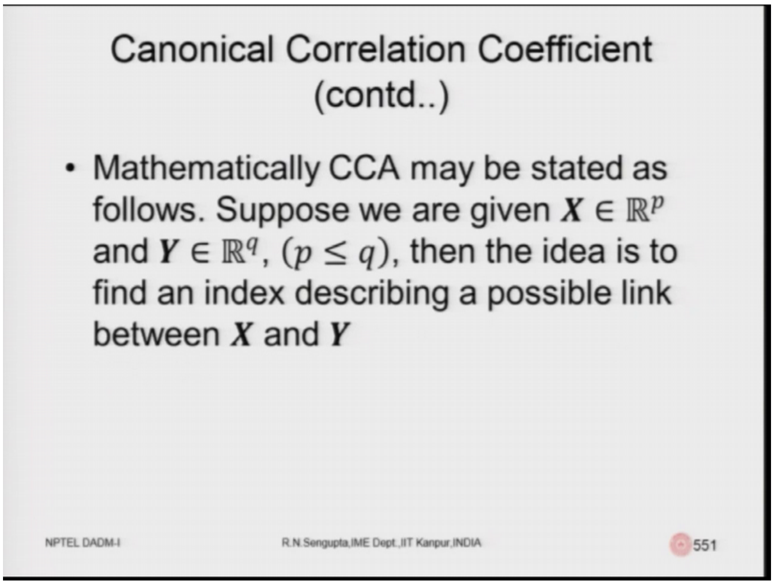
So, there are two cases here. Then in the next 2 cases again you have basically the squared error loss being for the precision of estimation and goodness of fit being the LINEX. So, in for third and fourth case you have the background same, but the estimates being utilised for beta would be in the third case it would be beta under the ordinarily squared, and in the fourth case it is beta under the LINEX loss condition

(Refer Slide Time: 15:15).



Now, we will start you can solve the problem accordingly, but I am not going to go to the details for that because that will consume a huge amount of time. So, we will consider the canonical correlation coefficient method. So, this was basically propose by Hotelling in his papers in 1935 and 36. So, this is the concept of canonical correlation coefficient where the authors studied the arithmetic speed, and the arithmetic power how they are related to the reading power, and reading speed. So, if you are reading power and reading speed is higher, does it basically signify your arithmetic concepts of powers and speed to under do problems in arithmetic front are higher and lower or they have no relationship to basically that was basically the concept based on which the study was done.

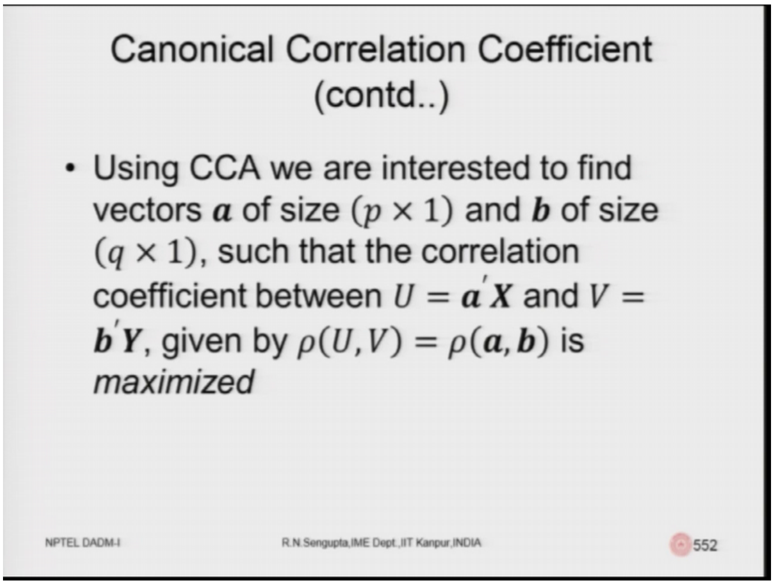
(Refer Slide Time: 16:07).



Mathematically canonical correlation coefficient would be studied by like this. So, there are X number of a random variables X 1 to X p an what you want to do is that if you want to find out the relationship between X 1 to X p and Y 1 to Y q, where technically p would basically be less than equal to q. So, what you are trying to do is that trying to basically synthesize, and find out the information of the relationship between the access using the concept of co-variances and using the concept of correlations are such that you can find out some relationship between what is intended, what is given, and based on that what you intend to study.

So, you want to basically find out the relationship, and find out what degree or what power, you are able to predict or basically try to give some useful information. So, the idea is to basically find out index such that that describes a possible linkage between the X vectors and the Y vectors, where X and basically maybe they are the observable one, and Y is basically what you want to study or vice versa, whichever you want to basically put it in the studies.

(Refer Slide Time: 17:19).

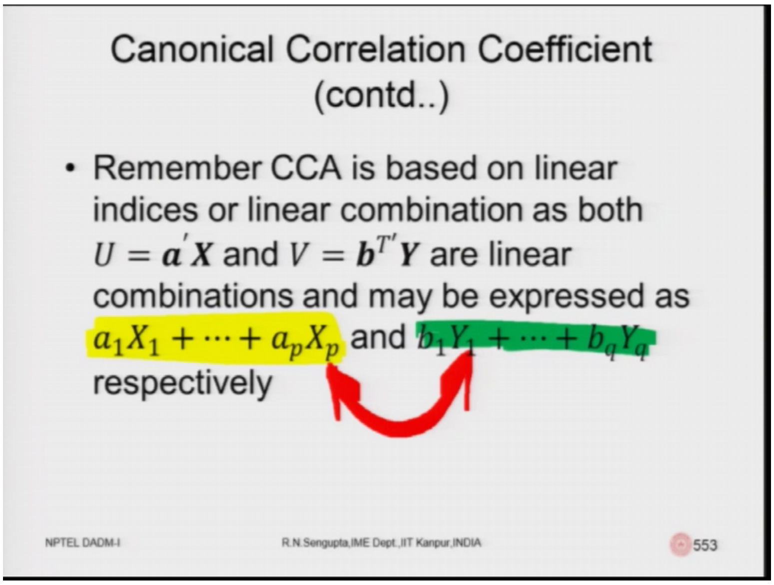


So, now in using the canonical correlation coefficient, we are interested to find a vector. So, a so vectors a would basically would be numbers, which would technically be a size of p cross 1, because if you remember X is a size of actual random variables of size p.

And also simultaneously, we want to find out set of vector b, which is b cross 1, which would basically correspond to the Y random variables, which are of size q such that this vectors a and b would give me the relationship, and different stage what is that between X and Y such that I can find out that given some observable information's, and given some theoretical information, which you want I want to basically not predict, I would not use the what predict you want to basically find out that what is the relationship between the X and Y, and to what degree.

Now, we want to find out the correlations such that the correlations with which is existing between a and b or U and V. U and V would basically be the new sets of a random variables be formulated by the combinations of the so called convex combination of a and X. And V would be the random variable formed by the convex combinations of b and Y. So, what you want to do is that given the correlation between a and b, we would we would basically try to predict, what is the correlation coefficient between U and V, which are the new vectors, which has been formed in order for better prediction under CCA, which is canonical correlation coefficient analysis method.

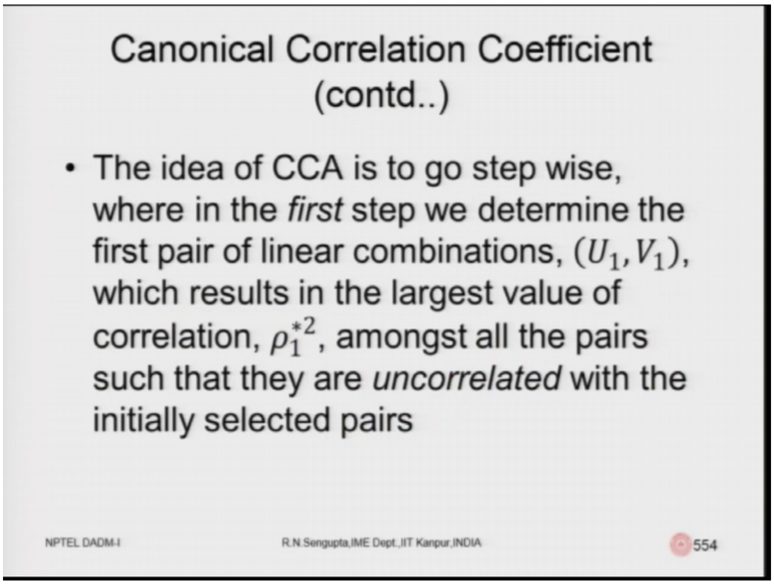
(Refer Slide Time: 18:54)



Or so in this case, so in under canonical correlation coefficient method or analysis method is based on the linear indices or linear combination. So, as if remember, I mentioned that you would be a basically be formed by linear combination of a and X, and similarly we would be basically formed by the convex combination of b and Y.

So, in this case what we are doing is going to do is that given this convex combination now of a and X, and b and Y. You want basically find out this relationship between them, such that the predictability of what is observable, and what is not observable based on which you want to predict would come out in the maximum possible way or maximum degree.

(Refer Slide Time: 19:51)



Now, what we do in the canonical correlation method. So, we go step by step what we do is that we in the first step, we determine the first pair of combinations, which is basically so U would be a vector, similarly V would be a vector. So, you want to find out the relationship between the first element of U, which is U 1 depending on how the combination have been set up and V 1 such that it takes out the maximum amount of correlation, which is existing between U and V.

So, let me read it, as it says. The idea of canonical correlation is to go step by wise, where in the first step we determine the first pair of combinations of U suffix 1, and V suffix 1, which results in the largest amount of correlation, which is can be gathered or glean from this relationship, so which says which results in the largest value of the correlation rho 1 square. Rho 1 is basically corresponding to U 1, V 1 amongst all the pairs such that they are uncorrelated and gives a gives us the maximum amount of it information.

So, once we have found out rho 1 square star means, the best value, you keep it aside. And then, you will basically assume the relationship is going for the next stage, where you find out another two values of U and V, which are given by U 2 suffix 2, and V suffix 2 such that given the first set of information the relationship between U 1 and V 1.

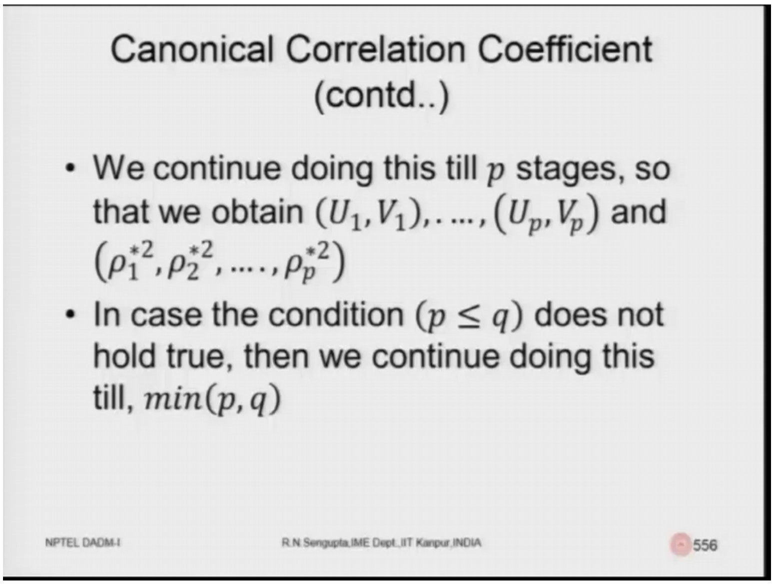
You basic and taking out the correlation coefficient, we will try to find out the next set of maximum correlation coefficient, which is existing such that they would be they would be uncorrelated in the sense in the pair wise such that you are able to find out rho 2 square. And we continue doing it step by step, such that we are able to gain gather the maximum amount of information, which exists between the U’s and the V’s, corresponding to U 1, V 1 then U 2, V 2 then U 3, V 3 and so on and so forth.

In the next step, what I just mentioned I will read it. In the next step, on onwards then one determines the second pair of linear combinations, which is U 2, V 2 that has the largest correlation given by rho 2 square, amongst all the pairs such that they are now correlated with the initial selected pair. So, initial selected pair which was there U 1, V 1, you will find try to find out the second set, such that you will get the maximum correlation provided U 2, V 2 are initially not correlated to U 1, V 1. So, what you are trying to do, you are trying to basically group them. And find out the orthogonal relationship between U 1, V 1 as a group with respect to U 2, V 2, because they are uncorrelated, which means they would not to each other, they would be 90 degrees.

Then you go to the 3rd step, where he considered a 3rd combination of U 3, V 3, such that it will be uncorrelated both with respect to U 2, V 2 and U 1, V 1 such that you are able to find out the remaining maximum amount of correlation which is there.

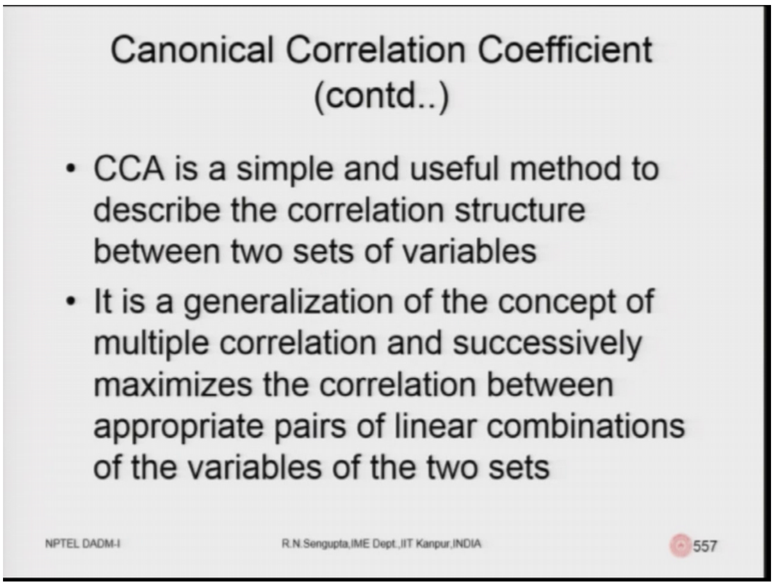
You proceed likewise, you find out U 4, V4; U 5, V5, but in each step you ensure that you are able to get the maximum correlation were provided that the orthagonality between the preceding the sets, which you have obtained is maintained such that it is in a ways. So, I am trying to find out basically the Eigen values and the Eigen vectors, which are there for the studies, where you want to break up the relationship using a new set of V’s corresponding to the fact that you have all the set of information, which is available is in the U’s. U’s and V’s, I am talking about the vector as such.

(Refer Slide Time: 23:27).



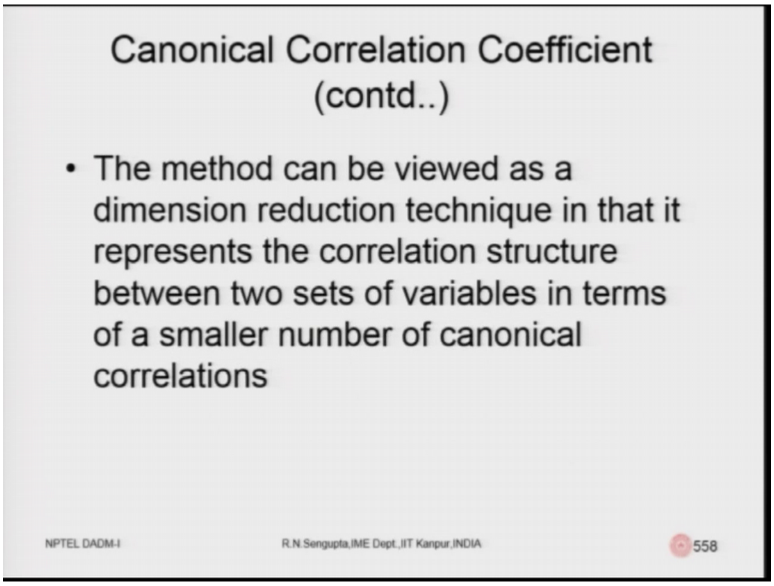
We continue doing this till the p stage, because you remember the p stage is basically depending on the number of such a random variables, which is there for X, so that we obtain U 1, V 1, as a pair U 2, V 2, as a pair and continuity U p, V p. And the correlation coefficients are given the maximum for each state would be given by rho 1 square with a star, rho 2 square with a start, till rho p square that means, rho 1, rho 2, rho 3, rho p the maximum value of the correlations corresponding to the uncorrelated values of the preceding set off of relationships, which you have. In case the condition of p less than equal to q is not there, you take basically take the minimum value of p and q.

(Refer Slide Time: 24:09)



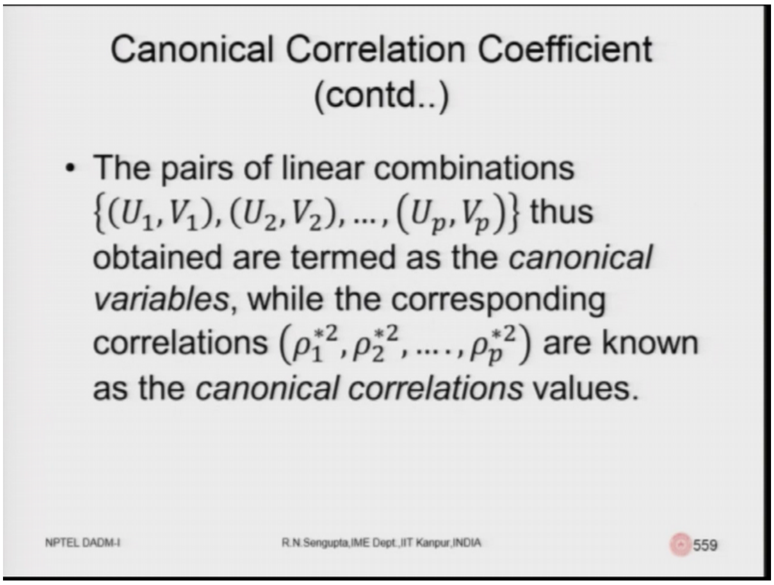
Canonical correlation method is simple useful method to describe two correlation structure between two sets of variables. In generalizations of the concept of multiple correlations and successively, it maximize the correlation existing between the pairs. Linear pairs of two sets at the times such that you are able to divide the correlation coefficient into groups, such that you get the maximum information of the relationship between actual X and Y, which comes which is coming out from U and V.

(Refer Slide Time: 24:39)



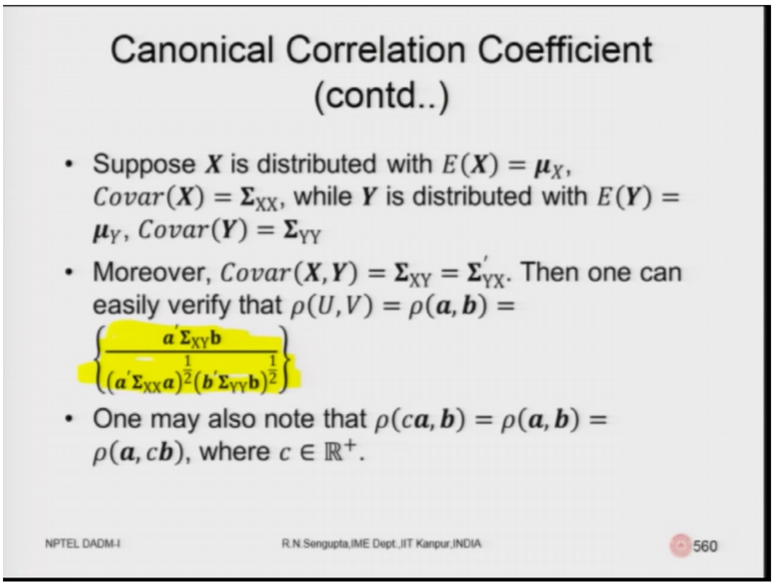
This method can be viewed as a dimension reduction method. So, you are trying to basic reduce the dimension still maintain the level of correlation, such that you are able to lean the maximum set of information. Is a reduction technique in that it represents the correlation structure between two sets of variables in terms of smaller number of correlation coefficient values, which would be coming out from the Eigen values and eigenvectors.

(Refer Slide Time: 25:07)



The pairs of linear combinations that is U 1, V 1; U 2 U 2, V 2 till U p, V p. Thus obtained are termed as the canonical variables, while the corresponding rho 1 square star, rho 2 square star till rho p square star are known as the canonical correlation values, which give me the degree of relationship or to what level the relationships have been summarized utilizing the correlation coefficient values. So, we will basically state it accordingly, I will solve a simple problem later on.

(Refer Slide Time: 25:47).



Suppose, X is distributed with unexpected value of mu suffix X, and Y is distributed with the value, where the expected value of Y is mu suffix Y. Corresponding covariance of X is sigma suffix X X, and corresponding value of covariance of Y is sigma suffix Y Y.

Moreover, we will considered the covariance of X and Y is there, it will be given by covariant of sigma suffix X Y or sigma transpose of Y and X. So, they would obviously co-variances are once you find out the covariance of X to Y and Y to X, they are the mirror image, because the principal diagonal is same, half the diagonal element are mirror image.

Thus one can verify very easily that the correlation coefficient existing between U and p would be exactly be equal to existing between a and b, which is the variables which you want to find out. So, you want to basically find out the covariance of a and b by the concept of this, this is the formula which is given and we will try to replace. So, this is the correlation coefficient.

So, the first the above term, which is in numerator the co-variances, and the below terms are basically standard deviation, this is the same formula which we use. So, based on that we will proceed one by one, and find out the correlation coefficients such that the pair wise U 1, V 1 with rho 1 square maximum, U 2, V 2, rho 2 square maximum star are the maximum values you continue doing this such that you are able to break up the relationship. Break means, partition into and the maximum relationship between say for example, set of observed values with the set of such other set of variables, which may not have been studied at which are important for the study in order to predict.

So, with these I will end of 55th lecture, which is end of the 11th week. And continue discussion with the correlation coefficient methods of canonical correlation, and discuss few other the important topics not from the theoretical point of view, but from simple problem solving, such that you gather a lot of information as required in the DADM 1 course. Now, with this I will end this course and this class.

And have a nice day, and thank you very much for your attention.