

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

Lecture - 56
Canonical Correlation

Welcome back, my dear friends and students. A very good morning, good afternoon, good evening to all of you and this is the DADM which is Data Analysis and Decision Making course and one DADM one course under NPTEL MOOC series. And as you know this course total duration is the use of 12 weeks, total hours being 30 hours that is 60 lectures and each week we have 5 lectures, each being of duration half an hour. And my name is Raghu Nandan Sengupta from IME department IIT, Kanpur.

So, if you see the slide number is the 56 slide number; that means, we are going to start the last week. And in the 11th week we were basically discussing about Canonical Correlation method.

And I was discussing that how you can basically break up this variances depending on the on some set of vectors a , b , c which we choose such that the they are in a way orthogonal to each other and amount of variance which you are able to take out or gleaned from the relationship basically decreases. Decreases means you take out the first set of variance from the first set of relationship and then go accordingly towards the second set linear combinations, second set, third set and. so on and so forth.

Now, generally in canonical correlation what we need to do is basically given some set of variables or which can study we want to find out some other sets of variables has that is they has linear relationship. And you also try to decrease the dimensionality all the problems has that the overall dependence is a mimic by minimum number of variables.

So, if you have p number of random variables x_1 to x_p , we will try to be basically take q of them, not q out of them, but we will basically try to find out q such variables related in such a way that, you are able to mimic the dependence structure to the maximum possible extend. Now when you want to basically solve the canonical correlation coefficient method and I did discuss that you have a set x and then you try to mimic or map it to set y ; set x and set y are the vectors.

And you basically form of combinations taking a vector \mathbf{a} and a vector \mathbf{b} such that, you which is a function of x , v which is function of y , have the correlation coefficient in such a way that it mimics whatever was there between \mathbf{a} and \mathbf{b} will try to basically find it out. Now this canonical correlation method can also be modeled has a simple optimization problem and this is how it is done.

(Refer Slide Time: 03:19)

Canonical Correlation Coefficient
(contd..)

From the optimization point of view
CCA can be stated simply as following

$$\sum \quad \max(\mathbf{a}'\Sigma_{XY}\mathbf{b})$$

$$\text{s.t.: } \mathbf{a}'\Sigma_{XX}\mathbf{a} = 1 \quad \checkmark$$

$$\mathbf{b}'\Sigma_{YY}\mathbf{b} = 1 \quad \checkmark$$

NPTEL DADM I
R.N. Sengupta, IIT Kanpur, INDIA
561

So, from the optimization point of view canonical correlation method can be send this stated very simple like this. So, if you have the variance covariance matrix existing with an x and y as the summation and the dimensionality will be maintained here.

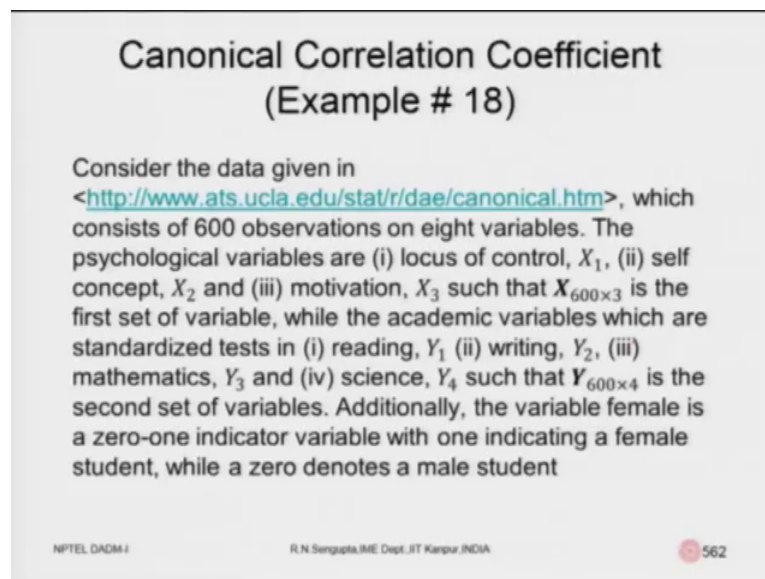
So, what you want to find out is that we want to find out the co variances or a linear combination of this co variances; where the linear combination is found by multiplying on this variance covariance matrix by \mathbf{a} and \mathbf{b} such that, we will try to maintain these 2. And obviously, it will it will continue if you go stage by stage.

That means, the correlation coefficient existing or the variance covariance existing between the XX and YY given you multiply them from by \mathbf{a} is related to 1, that means you considered the concept of orthogonality to be true. And also you mentioned maintain that orthogonality for the set of vector \mathbf{b} also considering for the set of vector the random variables Y .

So, we will consider, as I was saying this vectors a and b would be considering such a way that the covariances of variances existing between X and X and that should be related in such a way that when by multiplied by vector a and that is X you are multiplying of vector a and Y your when you are multiplying by vector b , which is 1. So, what you are trying to do is that you are trying to basically orthogonally break up the relationship, orthogonally means 90 degrees; such that, the relationship between X and Y are mimic by the variances stage by stage. And obviously, you will again use the concept of eigenvalues and eigenvectors.

So, they broke broken up into sets such that, you will normalize the co variances of X 's multiplied linear combination of X 's and linear combination of Y 's to 1 and try to basically maximize the correlation coefficient which is existing between X and Y , considering the vectors a and b which are used to form the convex combination of u from X and of v from Y . So, let us consider I am not going gives the details of the data. The data can be found out from the net and the URL is given.

(Refer Slide Time: 05:53)



Canonical Correlation Coefficient
(Example # 18)

Consider the data given in <http://www.ats.ucla.edu/stat/r/dae/canonical.htm>, which consists of 600 observations on eight variables. The psychological variables are (i) locus of control, X_1 , (ii) self concept, X_2 and (iii) motivation, X_3 such that $X_{600 \times 3}$ is the first set of variable, while the academic variables which are standardized tests in (i) reading, Y_1 (ii) writing, Y_2 , (iii) mathematics, Y_3 and (iv) science, Y_4 such that $Y_{600 \times 4}$ is the second set of variables. Additionally, the variable female is a zero-one indicator variable with one indicating a female student, while a zero denotes a male student

NPTEL DADM-I R.N. Sengupta, IIT Kanpur, INDIA 562

So, let me read the problem. Consider the gate data which is given at university California Los Angeles, so this is under the statistic department. And it consists of 600 observations in 8 variables. The psychological variables are, locus of control, so this part of the medical studies. Self concept or the general concept level, which a person has this 600 observations of pertaining to individuals.

The motivation level such that, we want to basically have x as considering matrix of size 600 cross 3, where 600 with the observation pertaining to each of this few variables. And is the first set of variables while the economic variables which are of interest based on which we are going to study which is basically y are as follows. Y_1 is the reading capability, Y_2 is the rewriting capability, Y_3 is the mathematical capability or quantity capability, Y_4 Y_1 talk about Y_3 and Y_4 is basically the scientific equipment capability.

So, that these 600 cross 4, which is basically 600 observation for the 4 new variables is the second set variables based on which we are trying to modeled the canonical correlation method. Additionally the variables female or male which is there, so as I mentioned if you remember there were 8 variables.

So, the first 3's are basically X 's, next 4 are basically the Y 's and the 8th one is basically whether a male or a female; one and obviously, will get to the variable values as 01 and 1 or 1 depending on whether the person is a female or a male. So, let me continued reading it. Additionally the variable female is a 01 indicator variable with 1 indicating a female student while as 0 denotes a male student.

(Refer Slide Time: 07:53)

Canonical Correlation Coefficient
(Example # 18) (contd..)

Solving the CCA problems yields

▪ $A = (a_1 a_2 a_3) =$

$$\begin{pmatrix} -1.2501 & 0.7660 & -0.4967 \\ 0.2367 & 0.8421 & 1.2051 \\ -1.2491 & -2.6360 & 1.0935 \end{pmatrix}$$

1st. $a' \sum_{xx} a = 1$

2nd. $b' \sum_{yy} b = 1$

NPTEL DADM-I R.N. Sengupta, IIT Kanpur, INDIA 563

Now, when we solve the canonical correlation method so obviously, you will get these variables A . So, if you remember we had 3 of this X 's, so X was a size 600 cross 3 or 3 cross 600 whichever way you denote. So, X_1, X_2, X_3 were the variables and the values of A which is basically a 1, a 2, a 3, which you will combine with X , so that means, you

will try to find out x transpose this variance covariance matrix of so, I am I am trying to take the first constraint as this.

Similarly, I will change the color, I will have b as force as subject to constraint as the second one and based on that will try to basically maximize the variance covariance matrix linear combination of variance covariance matrix between X and Y considering a and b as the vector. So, once you find out A the values are as given, minus 1.2501 till I am reading the first 1 comma 1 comma 1 value and the 3 comma 3 values 1.0935.

(Refer Slide Time: 09:18)

Canonical Correlation Coefficient
(Example # 18) (contd..)

▪ $B = (b_1 b_2 b_3 b_4) =$

$$\begin{pmatrix} -0.0440 & -0.0016 & 0.0883 \\ -0.0551 & -0.0904 & -0.0961 \\ -0.0194 & -0.0030 & 0.0878 \\ 0.0038 & 0.1242 & -0.0885 \end{pmatrix}$$

NPTEL DADM-I
R.N. Sengupta, I.M.E Dept., IIT Kanpur, INDIA
564

Similarly, when you find out B using the methods as we have already discuss, so if you remember optimization is the other method, but if you remember we have broken down and found out the correlation coefficient step by step and took out the maximum correlation coefficient in the first step and then basically try to modeled the second one considering the orthogonalities maintain.

So, we will go step by step. In the second step we take out the variability to the maximum possible extent considering it is orthogonal of in the first stage then and also it is basically the orthogonal to the later stages which are going to happen. Then we can take out the row 3's; that means, we found out we will find out u 3's and v 3's such that, it will be orthogonal to u 2's or v 2's and also be orthogonal to u 1's and v 1's.

So, once we have that, so correspondingly to we find out the A vector and B vector. So, here is basically v B vector. So, obviously the B vector would be 4 cross 4 because, if you remember I meant we have mentioned when we have basically discussing in the problem the number of variables y would is basically 400 cross 6; 44 cross 600, where 600 is basically the with the number of observation which you have.

(Refer Slide Time: 10:33)

Canonical Correlation Coefficient
(Example # 18) (contd..)

▪ $\rho^{*2} = (0.4464 \ 0.1534 \ 0.0225)$

NPTEL DADM I R.N. Sengupta, IIT Kanpur, INDIA 565

The correlation coefficient maximizations values would be coming as I will only read till the second place of decimal and there is a reason for that because and also note down the values, so how they are decreasing because, that will be applicable as we draw the diagrams.

So, the rho square star square mean the optimum values are 0.45, 0.15 and 0.02. So, they would basically give you the correlation coefficients of the of the variability which is there in the first set u 1 v 1 then the second set u 2 v 2 then third set u 3 v 3 and continue in the ways such that variable to glean the maximum set of dependence of or relationship such that there orthogonal to each other.

(Refer Slide Time: 11:21)

**Canonical Correlation Coefficient
(Example # 18) (contd..)**

This means that the set of linear combinations of the variables are

- $U_1 = \mathbf{a}'_1 \mathbf{X} = (-1.2501 \ 0.2367 \ -1.2491) \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = -1.2501X_1 + 0.2367X_2 - 1.2491X_3$
- $V_1 = \mathbf{b}'_1 \mathbf{Y} = (-0.0440 \ -0.0551 \ -0.0194 \ 0.0038) \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = -0.0440Y_1 - 0.0551Y_2 - 0.0194Y_3 + 0.0038Y_4$

NPTEL DADM I R.N.Singupta,IIT Kanpur,INDIA 566

Now, we need to find out the new sets of linear combinations of X's, linear combinations of Y's which will give me the actual variability which you want to study.

So, if you remember, so this means and I am reading it. This means of the set of linear combination the variables which are applicable for our studies is based on X 1, X 2, X 3, which was the first set of variables which were of interest to us. But we want to find out there new set of random variables which we think would be much better way in order to basically study the relationship; such that, it makes some practical sense. That means, we are trying to basically transform from one coordinate to other coordinate, maintaining the orthogonality and also maintaining the relationship which we want to find out. So, this means the set of linear combination of the variables are given by U 1.

So obviously, U 1 as I mentioned is the convex combination of a and X. So, a is we have already found out. So, a is if you remember they would be a1, a 2, a 3 depending on the rows. So, technically a is basically vector of 3 cross 3. So, the values of a 1 would be if you concentrate on the first row, it will be minus 1.2501, 0.2367 and minus 1.2491, so that would be multiplied by the vector column vector of X 1, X 2 and X 3.

And the first U1 is basically given by this linear combinations of X's utilizing the s. The second set are given and user lighter shade. The second set means of corresponding to Y's, so the V1's are given where you multiplied the first rhos of V's with the corresponding Y value, so Y if you remember there are 4 variables.

The mathematical speed, the science concept and all these things, so once we have them we have the V1 as given. So, just note down the U 1 and V 1 are given, they are the linear combinations such that, we basically get the maximum variability. So, the maximum variability if you remember we have mentioned please note the rho star square.

So, 0.44 and all the values they would become important later on when we study. Then after finding out the orthogonality in the first step, we would basically go into the next step where orthogonality is maintained, variability is maximum gleaned out depending on the fact that we find out the relation to the maximum degrees possible. Then we need to find out the new set of linear combination of excess which is basically U 2 and linear combination of Y's Y vectors which is basically V 2 provided that we have basically the informations of vector A 2 and vector B 2.

(Refer Slide Time: 14:29)

**Canonical Correlation Coefficient
(Example # 18) (contd..)**

- $U_2 = a_2'X = (0.7660 \ 0.8421 \ -2.6360) \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = 0.7660X_1 + 0.8421X_2 - 2.6360X_3$
- $V_2 = b_2'Y = (-0.0016 \ -0.0904 \ -0.0030 \ 0.1242) \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = -0.0016Y_1 - 0.0904Y_2 - 0.0030Y_3 + 0.1242Y_4$

NPTEL DADM-I R.N.Sengupta, IIT Kanpur, INDIA 567

So, these vectors a 2 and b 2 as you know are given, I am not going to highlight, I am just going to market is 0.766, 0.842, minus 2.636. I am only reading the first 3 phases of decimal when multiplied by X 1, X 2, X 3 which are the random variable based on which you are doing the studies.

The second set would be given as minus 7.66 X 1 plus s 7.66 X1 plus 8.4 X2 minus 2.63 X 3. Similarly, when you find all the linear combinations of b's and b 2's and Y, the second set of V 2 which is formed by the combination of b and Y comes out like this;

minus 0.0016, Y 1 minus 0.0904, Y 2 minus 0.0030, Y 3 plus 0.1242 Y 4, so there are 4 variables here.

(Refer Slide Time: 15:36)

**Canonical Correlation Coefficient
(Example # 18) (contd..)**

- $$U_3 = a'_3 X = (-0.4967 \ 1.2051 \ 1.0935) \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = -0.4967X_1 + 1.2051X_2 + 1.0935X_3$$
- $$V_3 = b'_3 Y = (0.0883 \ -0.0961 \ -0.0878 \ 0.0885) \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = 0.0883Y_1 - 0.0961Y_2 - 0.0878Y_3 + 0.0885Y_4$$

NPTEL DADM I R.N. Sengupta, IIT Kanpur, INDIA 568

Finally, we go into U 3 and V 3. U 3 again is found out by multiplying the vector a 3 and X. And values comes out to be as this minus as which have highlighted minus 0.4967 X 1 plus 1.2051 X 1.2051 X 2 plus 1.0935 X 3. And the values of V 3, which are found out utilizing the convex combinations of b and, we are going to multiplied b with Y's. So, that is V 3 into the Y vector. So, that comes out to be reduce the color as we have utilizing it comes out to be 0.0883 Y 1 minus 0.0961 Y 2, minus 0.0878 Y 3 minus 0.885 Y 4.

(Refer Slide Time: 16:36)

Canonical Correlation Coefficient
(Example # 18) (contd..)

- The corresponding linear combination graphs for (U_1, V_1) (blue dots), (U_2, V_2) (green dots) and (U_3, V_3) (red dots) are can be illustrated
- Though not apparent but one can easily discern that the value of correlation coefficient or the slope of the set of blue plots are the maximum, followed by green and then red

NPTEL DADM I R.N. Sengupta, IIT Kanpur, INDIA 569

Now comes the main part. The corresponding linear combinations which you have found out; that means, the set U_1, V_1 set U_2, V_2 and set U_3, V_3 because, we are trying to find out the orthogonality, at the same time try to find out the maximum variability which is taken out at each stage.

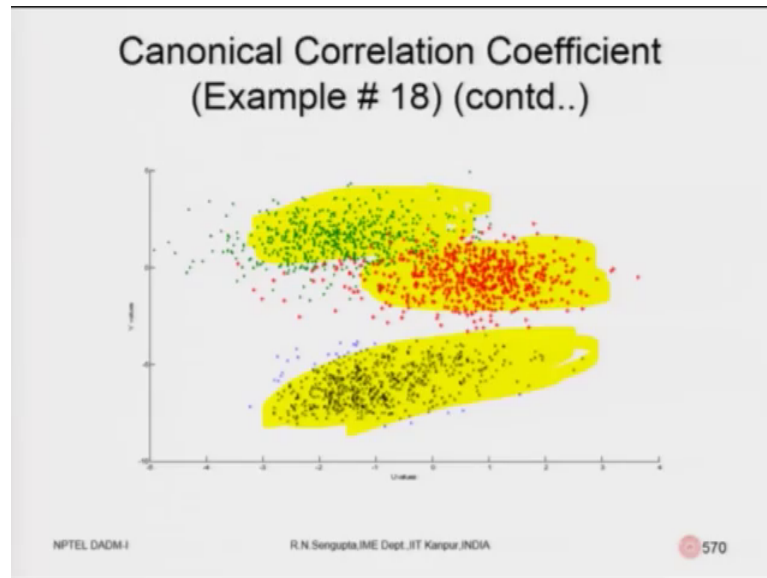
So, concentrate on the on what is written. Will denote the sets of U_1, V_1 , the points as blue dots so, obviously, you will have different sets of points U_1, V_1 U_2, V_2 U_3, V_3 depending on X_1, X_2, X_3 and Y_1, Y_2, Y_3 and $X_1, Y_1, X_2, Y_2, X_3, Y_3$ is basically you are trying to find out the each step, it not that you are only going to utilize X_1 for U_1 or X_2 for U_2 , only you are going to basically concentrate on the vectors a_1, b_1 for U_1, V_1 a_2, b_2 for U_2, V_2 a_3, b_3 for U_3, V_3 .

So, based on that I find out the combination and denote it for our simplicity of understanding as U_1, V_1 with blue dots, U_2, V_2 as green dots and U_3, V_3 red dots and then try to basically an analyze how we draw the diagram. If you remember I did not mention that when the road should start square what being written down. I mentioned please note the values they are the decreasing order 0.44 and till the last value of 0.02.

So, though not very apparent and let me continued reading it in this slide that is given though not very apparent, one can easily this discern or see that the value of the correlation coefficients or the slope of the set of the blue plots are maximum followed by the green plots, which is second followed by the red plots which is the third, which

means that, you are trying to slowly divide the I am repeating this words time and again please bear with me. We are trying to basically find out the relationship step by step such that, the correlation is basically mimic stage by stage maintaining orthogonality and trying to take out the maximum correlation coefficient at first stage then second stage and third stage as depicted by the correlation coefficient values which is rho star square.

(Refer Slide Time: 19:03)



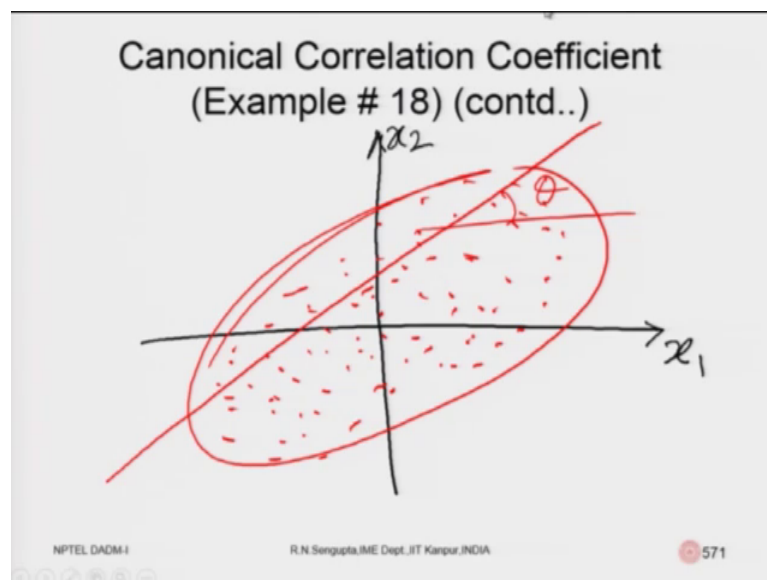
So, this is the correlation coefficient values. So, correlation coefficient values would basically be if you remember, I did discuss if I am not wrong in the first I think in the within the third or fourth week of this course that how would you basically understand the correlation coefficients being between minus 1 and plus 1 and they can be range is between minus 1 to 0 the value of 0 and the value of 0 to plus 1. So, I will again draw it and then basically try to compare the analysis which I am going to do with the diagrams which is there. So, please give me 1 minute. I will basically add 1 slide for a better understanding.

So, this is the slide which have added the blank 1 with respect to the correlation coefficient concept which I am going to explain for the $U_1 V_1$ combinations linear combinations of X_n and Y is which a coming out by multiplying a with XX vector b with Y vector a n be being a 1 and b 1, then we find out $u_2 v_2$ by multiplying a 2 with X b too with. Why then we find out U_3, V_3 which is basically multiplying a 3 with X and b 3 with Y .

Now I said that the correlation coefficient of this green, red and blue and who were basically giving us the information as the numerical value of rho square star vector was that how the correlation coefficient changes.

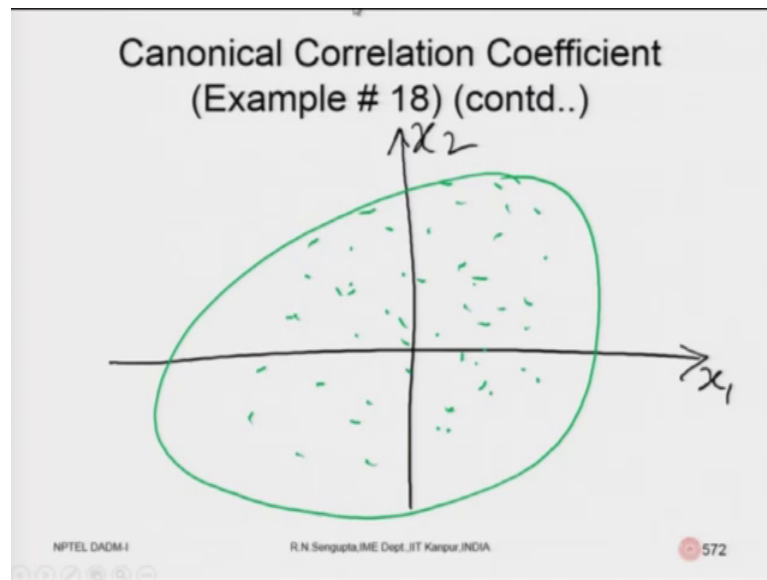
So, let and then I said that, we had discussed this problems when we are doing correlation coefficient. I will try to draw it within this diagram. So, let me draw it.

(Refer Slide Time: 20:59)



So, consider, this is the relationship between x_1 and y_1 or x_1 and x_2 . These are random variables. So, let be denote by $x_1 \times x_2$. So, consider they are positive correlated and correlation is basically between greater than 0 and less than equal to 1. Let we use the red points. So, they would be scattered in this way. So, the whole set would be like this and the tan of this angle theta so with best fit would be positive.

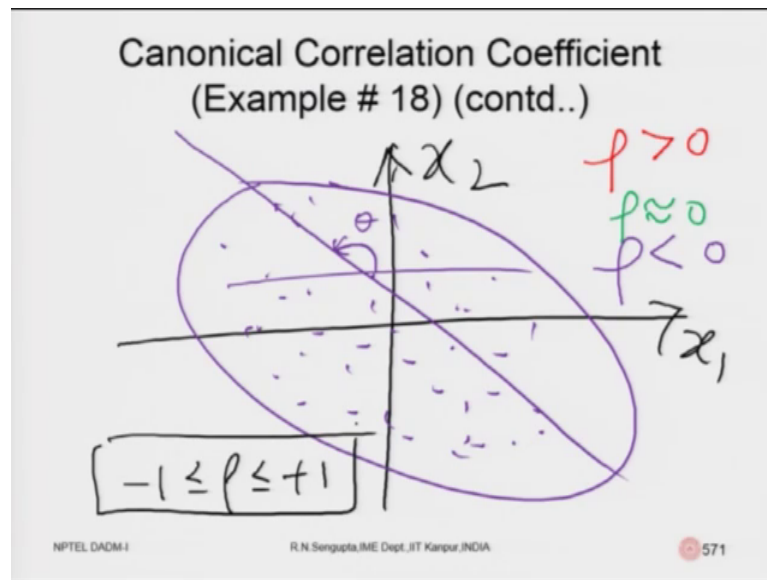
(Refer Slide Time: 22:11)



Now if I go for the so, make an out of this if I go to the value where correlation coefficient is in be in a non 0, so, in the first graph the red one we saw the they were positive, if they were in the first and the third quadrant. If they were 0 again I plot x_1 and plot x_2 and how the correlation coefficient would be scattered for the correlation values being between almost in a non 0.

So, they would be basically on all the 4 quadrant. So, they would not be any best fit line. Finally, I will not discard it, I will erase it and draw it for the sorry my mistake and I will draw do it for the negative one also.

(Refer Slide Time: 23:15)



You have seen it then I will come to the so once I have 1 negative values this is x_1 . This is x_2 and I will not use this. So, it is negative. So, it would be \tan would be in the negative direction.

So, it is in this second and the 4th quadrant. So, in the first case when it was the red one, correlation coefficient was positive that is why I am using the red color, when it was in the so, correlation is almost equal to 0 for the second case and for the third case correlation is less than 0.

So obviously, as we know the correlation coefficient is always between minus 1 into plus 1 so, now, with this if you come to this diagram it will make sense. So, if you see with the correlation for the red one and the green one are almost tending towards 0, while the blue one is positive. And if you note down I am first time, I am just going again in the front to the slides. So, you see the blue green and the red are in the order where blue has the highest green second and red third. So, it means $U_1, V_1, U_2, V_2, U_3, V_3$.

So, blue set if you see and just use the highlighter. So, it is the maximum set is here and if you see the green one, they are less steeped slope and the red one is always almost equal to 0. So, they have broken down the correlation into 3 sets depending on the maximize an concept you are trying to maximize the covariance linear combination of the covariance depending on some values of a and b and you force the concepts of the linear combinations of a into 1 with x and b is to 1 with y and solve the problem.

(Refer Slide Time: 25:54)

Canonical Correlation Coefficient
(Example # 18) (contd..)

- This fact is also corroborated by the values of $\rho_1^{*2} = 0.4464$, $\rho_2^{*2} = 0.1534$ and $\rho_3^{*2} = 0.0225$

NPTEL DADM I R.N.Sengupta, IIT Kanpur, INDIA 573

Now, if you remember I did show the correlation coefficient vector, almost at the beginning and I did mention the values were is note down I am not there in the decreasing order.

So, these values are 0.44 I am I only did until the second place. 0.44, 0.15 0.02. So, these correlation are you are trying to find out by the combinations of U 1 V 1 then U 2 V 2 and U 3 V 3 depending on the fact you can gleaned accordingly.

(Refer Slide Time: 26:30)

Canonical Correlation Coefficient
(Example # 18) (contd..)

- Another way of verifying the values of $\rho_1^{*2}, \rho_2^{*2}, \rho_3^{*2}$ is to have a look at $Covar(U, V) =$

1	0.4464	0	0	0	0
0.4464	1	0	0	0	0
0	0	1	0.1543	0	0
0	0	0.1534	1	0	0
0	0	0	0	1	0.0225
0	0	0	0	0.0225	1

NPTEL DADM I R.N.Sengupta, IIT Kanpur, INDIA 574

Now, you want to verify this task which you have done; that means, you have broken down the correlation coefficient and trying to basically get the maximum in set of information. So, another way to verifying the values of ρ_1^2 , ρ_2^2 and ρ_3^2 is to have a look at the covariances U V . So, if you find out the covariances of U V they are if those 0.44 values. So, here is basically in the first box so called box which have formulated you have the ρ_1 values.

If you go to the second one, so this is the correlation coefficient for the second stage ρ_2 , ρ_2^2 and the third one I will try highlight is the blue one. This is the third value which I have. So, I have found out or delineated ρ_1^2 , ρ_2^2 and ρ_3^2 depending on the covariance of U and V . So, U and V are if you remember, U is basically are matrix depending on x and a combinations and x being of 3 600 cross, 36 100 is basically the random samples numbers of observation and y being of 600 cross 4 hence we see if you count the values 1, 2, 3, 4, 5, 6; 1, 2, 3, 4, 5, 6 depending on the combinations we have been able to form for U and V .

So, with this I will close this 56th lecture and continue the last remaining 4 57,58, 59, 60 depending on correlation coefficient concept and some on the concept of multiplication statistical analysis have a nice day.

Thank you very much.