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

Lecture – 38
PCA

Welcome, back my dear friends. A very good morning, good afternoon, good evening to all of you and as you know this is the Data Analysis and Decision Making course one course and on the NPTEL MOOC series and this course is for 12 weeks, 60 lectures that is 30 hours. Each week we have 5 lectures each being for half an hour and this is the 38th lecture that means, we are in the 8th week.

So, if you remember and by the way I am Raghu Nandan Sengupta from IME Department, IIT, Kanpur. So, if you remember we were discussing about Principal Component Analysis and to give a picture I will repeat it, please bear with me. What we want to do is that we have a set of variables x 1 to x p and we want to find out the so called minimum number of axis amongst them which will give us the maximum set of information that is the main idea.

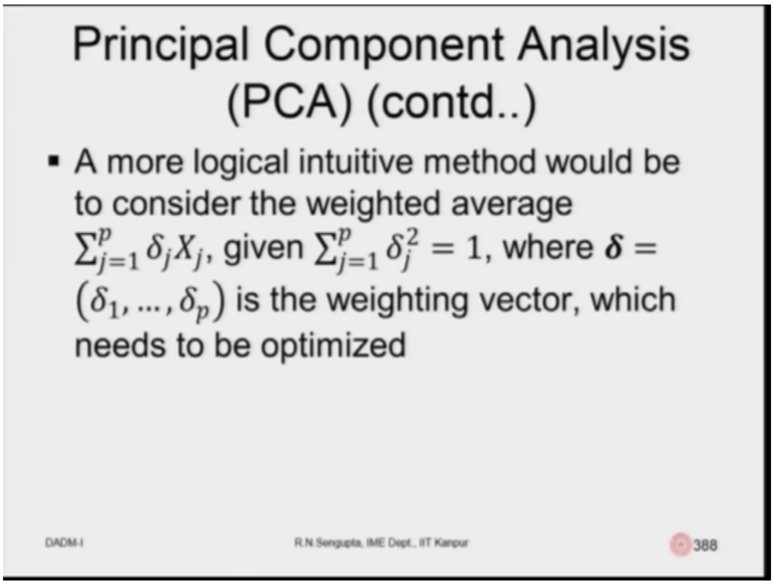
Now, when we are doing that we want to basically pick up one at a time some of the axis, but randomly picking up we do not know which is the best x which will give us the maximum set of information as available, we can visualize, but we cannot give mathematically so, and number 1. Number 2, what we will do is that will also ensure that as we are picking up so called consider we are able to pick up those random variables and go one step at a time that the first set gives us the maximum set information, second set gives the second level maximum set of information, third set gives us the third level of maximum set of information so on and so forth.

Once we are doing is that we will assume that once a variable is out of the scenario of its whole set of information it would not be affecting the later on the calculations which you are doing; which means that as I said I am again repeating they would be orthogonal to each other; that means, the combinations which we are taking for x 1 to x p in the first set would be orthogonal to the second set, again it would be orthogonal to the third set; that means, x the set on the second would be orthogonal to the third and the set for the first would also be orthogonal to the third. When we go to the fourth one fourth set would orthogonal to the third set, fourth one will be orthogonal to the second set, fourth one will be orthogonal to the first set.

So, as we proceed we get the maximum variability, maximum set of information, use them, rank them, take the minimum number of from axis and give us as the maximum information based on which we can give the dependent structure and we will basically project that in the very simple diagram also, ok. Another thing which we did mention is that trying to give equal weightages for x 1 to x p may not work out to be the best for the best possible advantage. So, obviously, we have to find out the best methodology of trying to do that.

Another point which was mentioned was very important that if the scaling factor, scale means to what you needs x 1 to x p I have been measured they would also have an effect when we are trying to take the variance covariance matrix. So, we will basically consider the standardized version of the PCA; Principal Component Analysis and we will basically give example in details about that, so that will make things much easier for you.

(Refer Slide Time: 03:37)

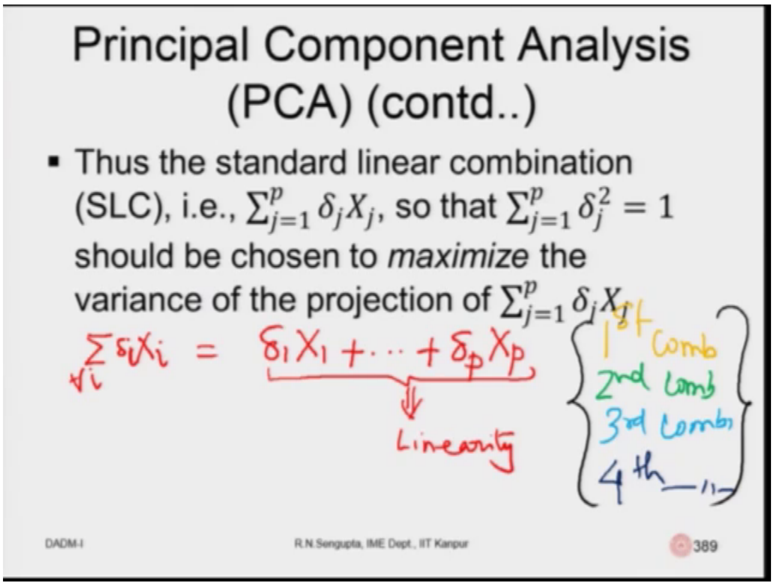


So, what we want to do is that rather than basically giving at equal which to x 1 to x p that means, of weights 1 by p, a more logical and intuitive method would be to consider the weighted average. So, a weighted average what we will do is that we will give weights x 1 a lambda 1, lambda 2, lambda 3, lambda 4 to x 1, x 2, x 3, x 4 and so on and so forth till x p. So, lambda 1 is for x 1 lambda 2 is for x 2, lambda 3 is for x 3, till lambda p is for x p.

Now, we will ensure in that way such that the sum of the squares of them is equal to 1, with means that we are not even if they are positive or negative we will basically ensure that the level of dependent structure which is happening between them is given by the lambdas the squares sum of the squares of the lambdas is equal to 1. So, this is something to do with the concept of variability deduction and so on and so forth which we even though we have not discussed in detail it for some idea we when we did the concept of unbiasness and consistency mainly later by the consistency concept.

So, where lambda and the vector consists of the elements lambda 1 to lambda p and these are the weights vector which needs to be optimized such that we are able to give the maximum set of information coming out from the minimum set of variables; that means, dimension it is being reduced the.

(Refer Slide Time: 05:27)



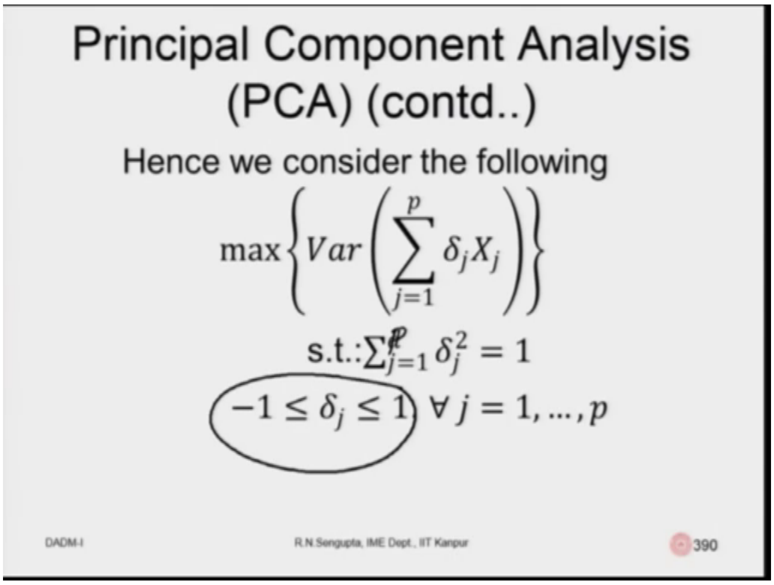
Thus the standard so, this is the standard linear combinations for the case of principal component analysis would be such that the we will basically give weights lambda 1 to lambda p to x 1 to x p such that the square sum of the squares of the lambdas basically equal to 1. And, what we are trying to do is that you want to basically maximize the variance of the projections of the convex combinations of lambda 1 x 1 plus lambda 2 x 2 till lambda p, x p.

Now, here I would like to basically draw your attention on two parts; number 1, maximizing the variance is basically we are trying to take out the maximum set of variability in the first set of information, keep it separately such that one the set of information is taken out from that relationship between x 1 to x p and the dependence structure which is there, it would not be affecting the second reading which is about the orthogonality, point 1. Point number 2 is that one we once we take the combinations as a lambda 1 into x 1 plus lambda 2 into x 2 dot till lambda p into x p we are basically going to consider the linear combinations.

So, it is like this combining, so, this would work in this way we. So, this is equal to so, this is linear. So, linearity would be maintained, number one and the first stage the first combination this is the combinations which I am giving. So, this is combination means what combinations of lambda 1 to lambda p I am considering. The second combination again we will consider some set of other lambda 1 to lambda p, the third combination, the fourth one. So, we will go these in a way that the combinations will give us the variability; variability in stepwise fashion reducing one step at a time and also trying to get take out a maximum set of information and each step such that their orthogonal or ninety degrees to each other.

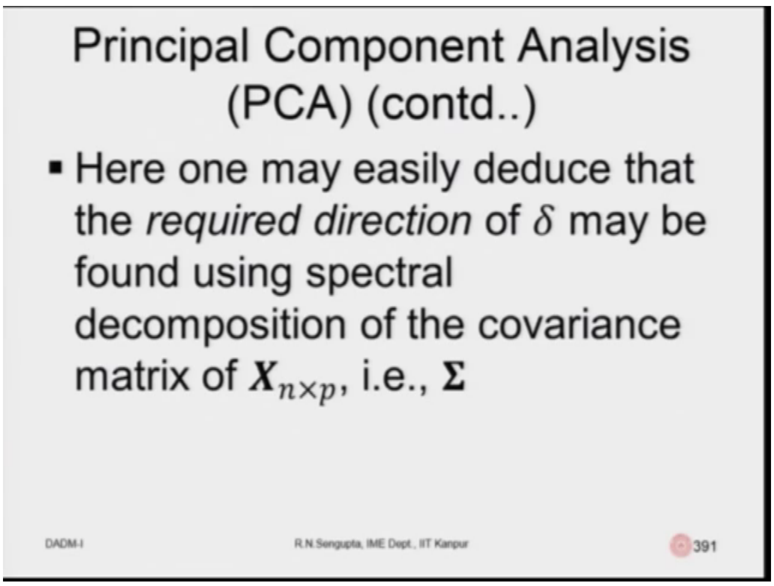
So, this is the optimization problem. We would not do the optimization I am just giving you the picture remember that. So, hence we consider the following optimization, you want to basically go in stepwise.

(Refer Slide Time: 08:46)



Maximize the variability of the combinations of lambda 1 into x 1 plus lambda 2 into x 2 plus lambda 3 into x p dot till lambda 3 into third element was lambda 3 into x 3 plus dot lambda p into x p. Subject to the condition that the variables, so, this should be p, my mistake. So, the variables which we have lambda j square some of them lambda 1 square plus lambda 2 square plus lambda 3 square dot till lambda p square is equal to 1, such that we are able to take out the first set of information for the variability as soon as possible and obviously, it would mean that lambda 1, lambda 2 till lambda p are between minus 1 to plus 1, obviously, the sum should be 1 square and j is equal to 1 to p.

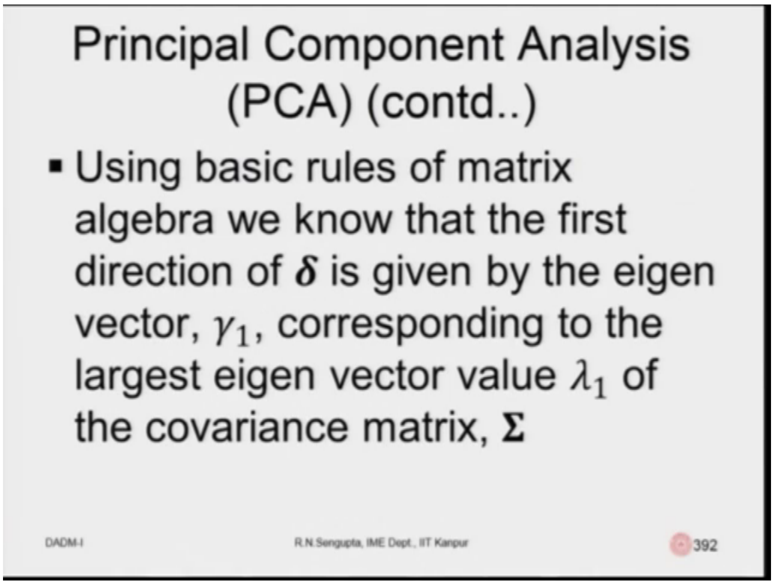
(Refer Slide Time: 09:45)



Here one may easily deduce that the required direction of lambda may be found out using the spectral decomposition. So, you are basically trying to break the direction of the dependence structure into different orthogonal planes that is all. So, orthogonality is very important.

And, use this using this spectral decomposition we are basically trying to break down the structure not the structure as the dependence informations of the covariance variance matrix which is basically for x 1 to x p that is the main reason.

(Refer Slide Time: 10:18)



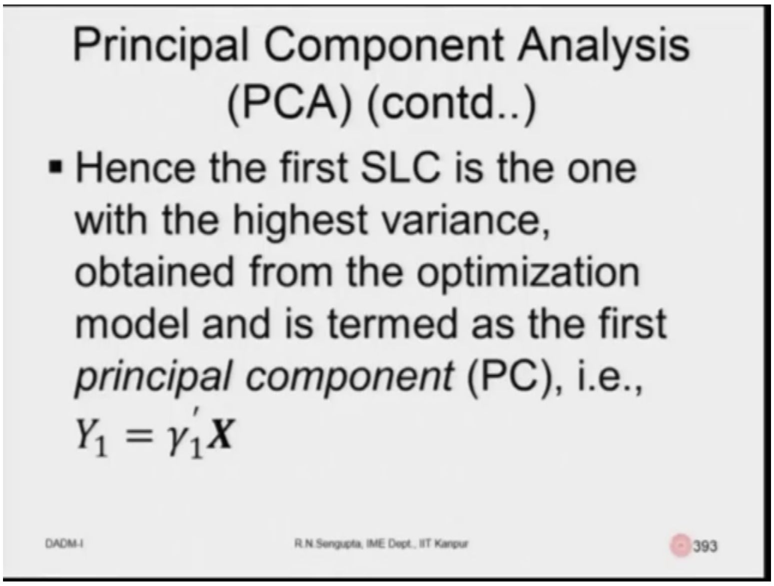
Using basic rules of matrix algebra we know that the first direction of lambda is given by the; obviously, so, if you remember in the Eigenvalues case. So, the Eigenvalues lambda 1 this gamma 1, gamma 2, gamma 3, gamma 4 would be orthogonal to each other.

So, what we are trying to do is that for the linear combinations break the relationship of the covariance structure and put them like till the tilt, the dependence structure and the first set in the first Eigenvalue direction then pull the rest of the information in the second Eigenvalue direction continue doing it such that it is by choice by the design of how you find out the Eigenvalues this gamma 1, gamma 2, gamma 3 would all the Eigenvalues would always be orthogonal to each other such that putting that set of information in the Eigenvalue plane will ensure that they are orthogonal to each other that the relationship.

So, using basic rules of matrix algebra we know that the first direction of lambda of the values of the deltas which we have sorry, sorry I am. So, the deltas which we have is would be given by the Eigen vector gamma one corresponding to the largest Eigenvalues which we have which is lambda 1. So, in the direction of the Eigenvalues and the Eigenvectors will basically break up the first set of the combinations of the delta or the weighted which we have for the first case for the first combinations.

Then in the second combinations we will have another set of delta. So, that delta 1, delta 2, delta p whatever they are they are basically combined in the second set of Eigenvalues and Eigen vectors direction. Similarly, the third one, fourth one, fifth one, we continue such that we are able to find out the maximum set of information in the minimum set of such Eigenvalues and Eigenvectors.

(Refer Slide Time: 12:41)

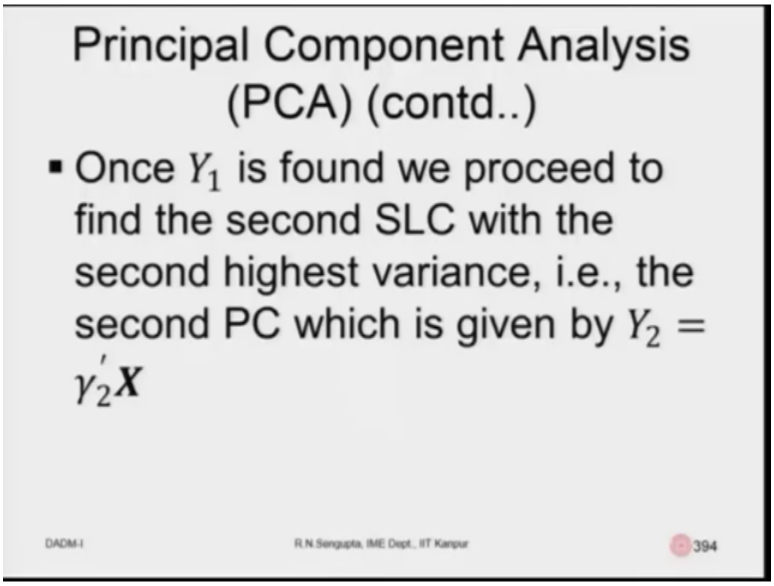


Hence, the standard form is the one when the highest variance obtained from the optimization model would be turn termed as the principal component the first principal component, then will basically have the second principal component and likewise we will basically break them into stage by stage in order to understand them in a much better way.

So, the first orthogonality I am using the word orthogonality for the first set would be given by gamma 1 into x, when x is basically the vector; that means, we are putting weights based on the Eigenvalues such that the weights which will have will give us the maximum set of variability in the first case. Then, obviously, when we multiply this second set of Eigenvalues and Eigenvectors whatever we have they are orthogonal to the first set. So, orthogonality is already presupposed and once we basically multiply the second Eigenvalues Eigenvectors in the vector or the column vector of the row vector with the x values which we have this will basically ensure the second set of combinations for the lambda 1 into x 1 plus lambda 2 into x 2 till lambda p into x p. So, that will give us the second set of information.

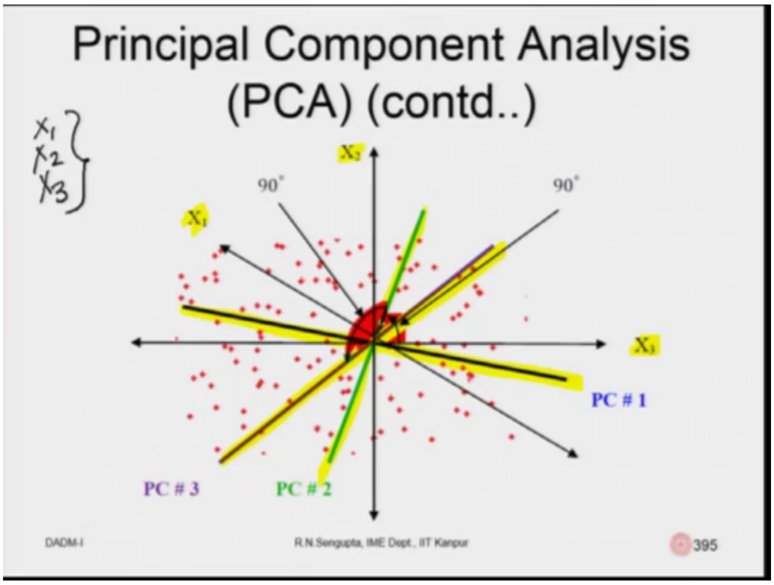
Then, we basically multiply the third set of Eigenvalues and Eigenvectors with the x matrix or the x values of the random variables ensuring the third combination such that we take the third set of variability pursued in this way till will basically get the maximum set of information which is available.

(Refer Slide Time: 14:25)



So, once Y 1 is found out that in which direction we have broken down combining the Eigenvalues, Eigenvectors into multiplied by the axis. So, we proceed to find a second set of standard direction of the principal components such that the second highest variance would be ensured for the principal component which will be given by Y 2 and that will be the multiplication of the second highest level of Eigenvalues, Eigenvectors multiplied by x. Similarly, we will go to the third one, fourth one, fifth one and correspondingly we proceed.

(Refer Slide Time: 15:13)



Now, here comes the diagram and I will try to spend some time because in the 3-dimension case it will be a little bit difficult to understand so I will be. So, consider we have basically the there are three random variables which are x 1, x 2 and x 3. Now, if you see x 1, x 2, x 3 without concentrating on the red dots without concentrating on the green line, violet line and blue line. I have purposefully drawn it in such a way that they are they are not orthogonal to each other. So, x 1, x 2, x 3 may not be orthogonal to each other which is the random variables which I have. So, these are the three random variables.

Now, consider these points are there which are basically the combinations of x 1, x 2, x 3, right. It can be height pressure humidity and temperature and we measure it at different places and we basically plot them in the 3-dimension case. So, they would be a scatter plot and these red dots are the scatter plots. Now, what we do is that we combine x 1, x 2, x 3 corresponding to the fact that the first Eigenvalues, Eigenvectors are there for these random variables x 1, x 2, x 3 such that we get the first principal in this direction, where the maximum variability is assumed. Maximum variance is taken out maximum set of information is taken out.

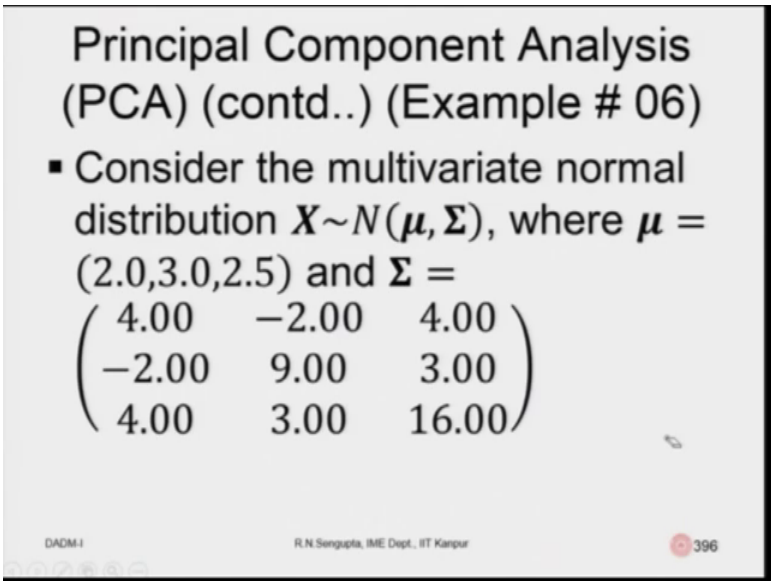
Now, once that is done we keep it fixed and then basically put the rest of the variability whatever the remaining part is of the variability and basically turn them in the direction on the second Eigenvalues which is here. So, these directions have already been presupposed by the Eigenvalues and Eigenvectors concept. Now, see one thing important which I did mention and rhyme again, when I do the principal component for the first stage and then I find out the principal component for the second stage they would be orthogonal to each other.

So, this part which I am trying to draw, that means, PC 1 and PC 2; PC 1 and PC 2 would be orthogonal to each other. So, PC 1 is fixed then what I do is that I the remaining amount of the variability force them in the direction of PC 2 which is the second Eigen vector and find out the orthogonality there which is ensured and I take out the maximum set of information in the second case.

So, first set takes out some information, second set takes out some information, the third set whatever the remaining variability is there I see basically put it in the third principal component direction. And, again if you notice this is orthogonal; that means, PC 3 is orthogonal to PC 2, PC 1 has already been orthogonal to PC 2. Now, they are the Eigenvalues and Eigenvectors; obviously, everyone would be orthogonal to each other. So, we have basically taken the orthogonality based on the fact that Eigenvalues, Eigenvectors are and basically break that combination on the variability in those directions.

Now, let us consider a very simple example and I will give you the main calculations.

(Refer Slide Time: 19:20)



So, consider the multivariate normal distribution. So, where the mean values of the three random variables x 1, x 2, x 3 is given as 2, 3 and 2.5 here. So, mu 1 is equal to 2, mu 2 is equal to 3, mu 3 is equal to 2.5. So, this is done. So, you have understood it.

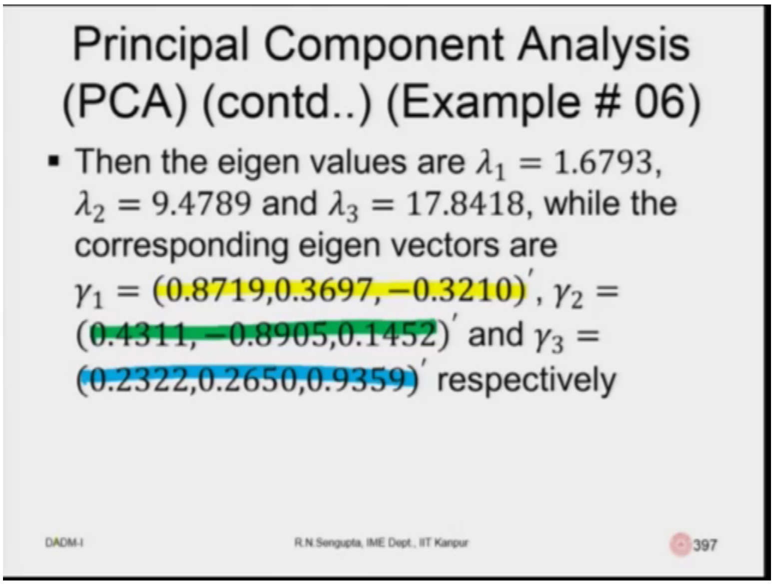
Now, I come to the variance covariance matrix. So, the variance covariance matrix is given by this. The so, obviously, it will be like this sigma 1 1, sigma 2 2, sigma 3 3, sigma 1 2, sigma 2 1, sigma 1 3, sigma 3 1, sigma 2 3, sigma 3 2 so, 3 cross 3 so, now, on the values. So, this one is equal to 4, then this one is equal to 9, this one which is the variance of the third is equal to 16. So, standard deviations are 2, 3, 4.

Now, come to the covariances. So, now, covariances value is not being 0; obviously, they it means there are not orthogonal there is dependence structure. These are the values. So, the covariance of the first to the second or second to the first is given as minus 2. So, is negatively related. Covariance of the first to third, third to first is basically 4, they are positively related because the negative sign is not there. Second to third, third to second is 3 again, they are positive related.

So, it means so this is what we have on these cells. They remove this; once I am drawing it I will remove it for better ease of understanding. I am sorry, I am going and repeating it, just please try to understand. These are the values for the covariance of first to third, third to first. Again, I am I am basically remove it, then I come to first to second, second to first. No, sorry this sigma just removed done and finally, when I come to the principal diagonal which is the variances.

So, this is the information. I went a little bit slow just please bear, this is done. So, now, with this I will basically go to the calculations.

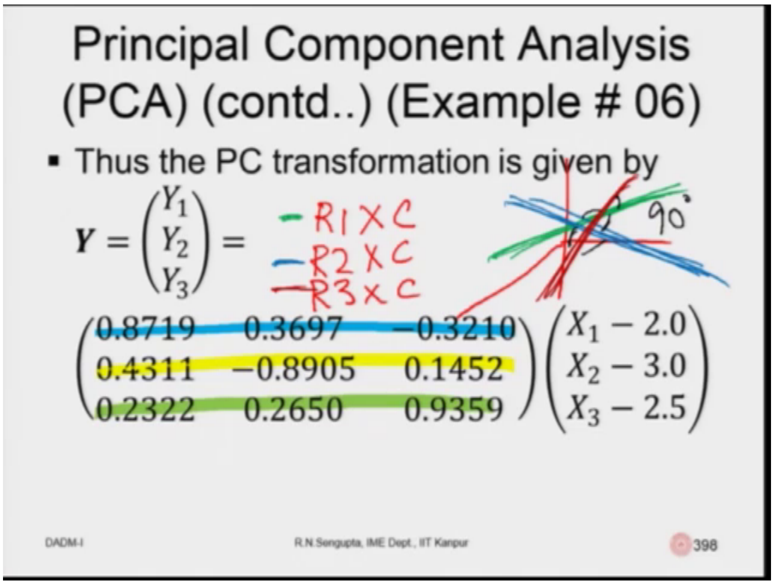
(Refer Slide Time: 24:05)



Then the Eigenvalues we can calculate are given as 1.6793, Eigenvalues 9.4789. I am not going to go to this this detailed calculation can be picked up in any class 11 or 12 textbook and lambda 3 is 17.84. While the corresponding Eigenvectors which is basically gamma 1, gamma 2, gamma 3, are given the first set is given by this where the first set of of Eigenvalues, Eigenvectors would be mapped. Then the second set is this with the second set of Eigenvalues, Eigenvectors in the set of information you mapped, the third set is this.

So, using basic simple calculations given the variance covariance matrix we can find out the Eigenvalues and Eigenvectors for the first, second, third depending on three variables which are there. So, that part is first done that is basic class 10 mathematics which we will try to utilize it.

(Refer Slide Time: 25:19)



Now, we want to do the principal component transformation. Now, this remember is the standard form. So, if it is standard form remember here standard form means that these x 1's, x 2's, x 3's would basically be normalized with respect to the mean values. So, x 1 minus 2 because it is mean value that is why I have highlight at the mean values of x 1, x 2, x 3. So, x 1 minus 2 is mean values, x 2 minus 3 which is mean value, x 3 minus 2.5 which is mean values.

And, the prints the Eigenvectors which you have are basically sorry no I should. So, what we are going to do? We are going to multiply the first one the first row into column, then second row into column, third row into column such that in the 3-dimension case the first one will the principal component corresponding to this. Second one will be orthogonal remember I am not able to draw it here it will orthogonality corresponding to this and the third part would be again orthogonal corresponding to this.

So, we will be mapping the Eigenvectors multiplied with the corresponding standardized x 1, x 2, x 3 because there are 3 x and map them in such a way that they are orthogonal which means angles between them will all be 90 degrees because Eigenvalues and Eigenvectors are done accordingly.

Now, remember that when we are doing is this dimensionality and the concept of of Eigenvalues and Eigenvectors and the linear combination is very important, that we will assume to be true like in multiple linear regression all this thing we assume some of the assumptions which are very simplistic in nature may not be true, but they had give us good results. We will assume the concept of linearity and the concept of dimensionality in the sense of the units which I mentioned are such that the affects of non-linearity would not be there, affects of units would not be there. And, based on that once we use the concept of principal component analysis using the Eigenvalues and Eigenvectors you can break them the overall variability stage wise such that with minimum number or dimension we get the maximum set of information.

So, with this I will end the class and continue this problem in more details such that we are able to get the concept of principal component analysis as far as possible. Drawing it in that in higher dimension is difficult. So, that is why I have taken a 3-dimension case in order to basically portray this information as far as possible.

Thank you very much and have a nice day.