

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

**Lecture – 39**  
**PCA**

Welcome back my dear friends. A very good morning, good afternoon, good evening to all of you and this is the Data Analysis and Decision Making course which is one DADM – 1 course, under the NPTEL MOOC series lectures and we are in the 39th lecture; that means, we are going to complete with another one complete the 8th week and you know this total course is for 12 weeks which is about 30 hours, we will have 60 lectures and each week we have 5 lectures is being of half an hour and I am Raghu Nandan Sengupta from the IME department, IIT Kanpur.

So, we were discussing about Principal Component Analysis and the main idea about principal component analysis was that you have a set of random variables  $x_1$  to  $x_p$  and they are not independent that is what I want to highlight from beginning and you want to find out the relationship structure of what is their total out of this  $p$  what is the total number of such variables which are there which will give us the maximum amount of relationship or the effects which we can be found out by a linear combination of them.

Now, this linear part is very important to understand because as we said as we mentioned that we will consider the linear combinations of this  $x_1$  to  $x_p$  at each stage and basically combine them in such a way that the groups of the linear combinations would be orthogonal to each other point 1. Point number 2, also we said that we will consider only the standardized form because the units of this  $x_1$  to  $x_p$ , if they are different then scaling would be would be affected. So, we would not consider the units to be quite different in the sense that scaling factor is assumed not to be there and we will consider the standard form.

Now, our main aim is you if you remember which I had did mention about few seconds before and we I did harp on that fact in the 13th lecture when I was conducting it that the concept of linear combination should besides that they would be orthogonal to each other in those sets. In the sets, the first set of the linear combination would be come orthogonal to the second set. Second set would be orthogonal to the third set and obviously, we will

form the third sets in such a way that it will automatically linear combination of the first set. So, when we go to the fourth set first is linear to second, second is linear to sorry, first is orthogonal to second, second is orthogonal to the third, third is orthogonal to fourth and obviously, you have to mean that all of them fourth would be orthogonal to first and all of them would be orthogonal to each other.

Now, later on we I did show in the pictorial form that how it could be done that considering it is only possible to show pictorially in a 3-dimensional case that if you have a  $x_1$  and  $x_2$  which is  $p$  is equal to 2 and there was a  $x_3$  also. So, we want to basically combine them in such a way the principal component 1, then 2, then 3; obviously, we will consider three principal components which would be combined by the linear combination of  $x_1$ ,  $x_2$ ,  $x_3$  they would be independent.

Now, further on I mentioned and which is very important is that the concept of Eigenvalues and Eigenvectors will be used because the Eigenvalues and Eigenvectors which are basically orthogonal to each other would give us the background of how we are going to be basically conduct the principal component analysis concepts.

Now, given the Eigenvalues and the Eigenvectors we want to basically combine them. So,  $y_1$ , then  $y_2$ , then  $y_3$ , then  $y_4$  till  $y_p$  considering there are  $x_1$  to  $x_p$ . So, these would be the principal components which are basically being formed; these means  $y_1$  to  $y_p$  are being formed in such a way that they are orthogonal to each other and we will basically utilize the concept which I just mentioned the Eigenvalues, Eigenvectors concept would be utilized in order to find out those principal components accordingly. So, in the problem which I did not finish in the last class I will continue discussing that.

Now, if you remember we had mentioned that we can find out the components PC 1, PC 2, PC 3 in the way where we multiply the Eigenvalues and Eigenvectors corresponding to the fact that each rows and columns would be multiplied by the corresponding  $x_1$ ,  $x_2$ ,  $x_3$  and so on and so forth. So, this is what we are doing.

(Refer Slide Time: 05:09)

**Principal Component Analysis  
(PCA) (contd..) (Example # 06)**

Hence the PCA axes are:

- $Y_1 = 0.8719X_1 + 0.3697X_2 - 0.3210X_3 - 2.0 \times 0.8719 - 3.0 \times 0.3697 - 2.5 \times (-0.3210)$
- $Y_2 = 0.4311X_1 - 0.8905X_2 + 0.1452X_3 - 2.0 \times 0.4311 - 3.0 \times (-0.8905) - 2.5 \times 0.1452$
- $Y_3 = 0.2322X_1 + 0.2650X_2 + 0.9359X_3 - 2.0 \times 0.2322 - 3.0 \times 0.2650 - 2.5 \times 0.9359$

DADM-I R.N. Sengupta, IIM Dept., IIT Kanpur 399

So, what we had actually? So, we had I will just write it down. So, this less space, but please pay attention you will understand. So, the Eigenvalues, Eigenvectors are given. Consider there are three and these are the X's; X 1, X 2, X 3. Now, obviously, they are normalized; that means, X 1 would from X 1 we will subtract the mean value of X 1, from X 2 we will subtract the mean value of X 2, from X 3 we will subtract the mean value of X 3. So, once we have this, so, the first value X 1 will be multiplied by the corresponding value here which is here multiplied by X 1. The second value the would be multiplied by X 2, that third value would be multiplied by the value 3.

So, these are the value. So, if we consider this is the first value which is here, then we have second value and we have the third value. Now, remember here they are normalized so, obviously, the second part being this one, second part of the first, the second part of the second, second part of the third. So, this value of two which I am not going to highlight anymore value of 3 and value of 2.5 are the respective mean values of X 1, X 2 and X 3.

Similarly, when I go to find out Y 2 again, so, let me delete it would be. So, I will delete this. So, first row is gone; gone means it has been already taken care of through the calculation. The Y 2 would be this multiplied by X 1 minus mu 1, second value of the vector multiplied by X 2 minus mu 2 and third value of the vector or second vector

multiplied by  $X_3 - \mu_3$ . So, these values would be here accordingly I will just highlight it in order to make things.

So, the first value would be this, second value is this, the third value is this. Now, come to the last row and corresponding to  $Y_3$ . So, again first element of third Eigenvector multiplied by  $X_1 - \mu_1$ , second value of the third Eigenvector multiplied by  $X_2 - \mu_2$  and third value of the third Eigen vector multiplied by  $X_3 - \mu_3$ . So, what are those values to in order to find out  $Y_3$ ? So, these are the first set of values because this 2 is coming and I am again repeating it because as you are normalize it and these 2 value is  $\mu_1$ .

This 3 is coming because 3 is the mean value of  $X_2$  which is  $\mu_2$  and this 2.5 is coming because the mean value of  $X_3$  which is  $\mu_3$ . So, once we do the calculation. So, I will remove this. So, these would give me  $Y_1$ ,  $Y_2$  and  $Y_3$  which are orthogonal to each other. We will prove that, do not worry. We will prove that  $Y_1$ ,  $Y_2$ ,  $Y_3$  which is principal component 1, principal component 2, principal component 3, they are orthogonal.

We will prove that take that into consideration we will also take that other thing into consideration which is important is that the variability which we are going to prove for and find out for  $Y_1$ ,  $Y_2$  and  $Y_3$  would be in the descending order with the variability coming out by the combination to find out  $Y_1$  being the highest, then  $Y_2$  would be the second  $Y_3$  would be the third and so on and so forth. For in this case we have three random variables, obviously. Hence there would be only  $Y_1$ ,  $Y_2$ ,  $Y_3$ .

Now, I want to double check. So, what I mentioned just at the fag end of the last slides I am going to mention verify all the important facts, so that will become clearer to you.

(Refer Slide Time: 11:44)

### Principal Component Analysis (PCA) (contd..) (Example # 06)

- A way of double checking whether the PC transformations are correct is to calculate  $\sum_{j=1}^3 \delta_j^2$  for each of the eigen values. It is very intuitive to note that each of these values, i.e.,  $\{0.8719^2 + 0.3697^2 + (-0.3210)^2\}$ ,  $\{0.4811^2 + (-0.8905)^2 + 0.1452^2\}$  and  $\{0.2322^2 + 0.2650^2 + 0.9359^2\}$  are equal to 1 as the case should be

$\delta_{11}$	$\delta_{12}$	$\delta_{13}$	$Y_1 \equiv PC_1$
$\delta_{21}$	$\delta_{22}$	$\delta_{23}$	$Y_2 \equiv PC_2$
$\delta_{31}$	$\delta_{32}$	$\delta_{33}$	$Y_3 \equiv PC_3$

DADM-I R.N.Sengupta, IIM Dept., IT Kanpur 400

So, way of double checking whether the principal component method transformations are correct is to calculate the sum of the squares of delta 1, delta 2, delta 3; considering there are three random variables for each of the Eigenvalues which you have already found out. So, those Eigenvalues were lambda 1, lambda 2, lambda 3.

It is very intuitive to note that each of these variables which we have should basically add up to equal to one because that was basically the fact what we are doing. So, if we remember what we are trying to do it for the first case when you want to find out Y one we need to basically maximize the variability of the first one with respect to the fact that the sum of the constraints would be the sum of the squares of the delta 1 square plus delta 2 square plus delta 3 square corresponding to Y 1 should add up to 1.

Once the variability is taken care for the first case, then we go to the second optimization second stage optimization second. They are independent of each other in the sense the optimizations would definitely depend the second one depend definitely depend on the answer the first, but in case we want to basically consider them as optimization problems therein they are standalone problems. So, we will try to basically now maximize the variability corresponding to Y 2; that means, summation of delta 1 into X 1 plus delta 2 into X 2 plus delta 3 into X 3 would be maximize subject to the case that summation of delta 1 square plus delta 2 square plus delta 3 square is equal to 1. So, these delta 1, delta 2, delta 3 are corresponding to the second set which is Y 2.

So, once that is done then we go to the third stage of optimization which is again independent we will. Try to maximize the variability corresponding to  $Y_3$  such that  $\delta_{11}$  into  $X_1$  plus  $\delta_{21}$  into  $X_2$  plus  $\delta_{31}$  into  $X_3$  would be maximized the variability for that and we will also see that the subject to some conditions would be summation of  $\delta_{11}^2$  plus  $\delta_{21}^2$  plus  $\delta_{31}^2$  is equal to 1. Now, again this  $\delta_{11}$ ,  $\delta_{21}$ ,  $\delta_{31}$  are for  $Y_3$ . So, technically what I am trying to say is that the first time of optimization those were  $\delta_{11}$  plus then  $\delta_{12}$  and  $\delta_{13}$ .

So, the first one is corresponding to the first stage which is  $Y_1$  and the second suffix 1, 2, 3 are corresponding to  $X_1, X_2, X_3$ . Then we come to the maximization of the variability for the second which is for  $Y_2$ , those deltas would be  $\delta_{21}, \delta_{22}, \delta_{23}$ . So, the first suffix 2 would be corresponding to  $Y_2$  second suffix 1, 2, 3 would be corresponding to  $X_1, X_2, X_3$ . Then we go when we go to the third case it will be basically summation of as I mentioned summation of  $\delta_{11}^2$  plus  $\delta_{21}^2$  plus  $\delta_{31}^2$ . So, actually these  $\delta_{11}, \delta_{21}, \delta_{31}$  are actually  $\delta_{31}, \delta_{32}, \delta_{33}$ ; so, these suffixes which I mentioned 3 1, 3 2, 3 3 the first suffix would basically corresponding to  $Y_3$  and the second suffix 1, 2, 3 would basically corresponding to  $X_1, X_2, X_3$ . So, this is what we will have.

So, they would be  $\delta_{11}, \delta_{12}, \delta_{13}$  corresponding to  $Y_1$  the principal component 1, then we will have  $\delta_{21}, \delta_{22}, \delta_{23}$  which is corresponding to principal component 2. Then we will have  $\delta_{31}, \delta_{32}, \delta_{33}$  which will be corresponding to the principal component 3 which is  $Y_3$  and you can go on accordingly. So, again I will read a way of double checking whether the principal component transformations are correct at each stage is to calculate the sum of the squares of  $\delta_{11}^2$  plus  $\delta_{21}^2$  plus  $\delta_{31}^2$  at each stage, it is very intuitive to note that it will be true that they are 1.

So, let us consider. This is basically the values of  $\delta_{11}^2$ ,  $\delta_{21}^2$  corresponding to, let me highlight it in a different way, so, it will be easy for us to understand. So, would be this square we will find out 0.3697 will be this one, 0.32 will be this one. So, square of them would give up 1 which is true second case against square of them addition is one which is true.

Again, when we add up this square of the deltas add them up it is equal to 1. So, it proves that the principal component based on the fact that we have been able to break this principal component in the direction of the Eigenvalues are is the right method. So, that part is done second part would be obviously, the question if you remember I mentioned whether the variabilities are in descending order as it should be because the first case of the variability we want to take out the maximum variability corresponding to PC 1 then the next level of variability would be corresponding to PC 2 and so on and so forth. So, we will come to that also.

(Refer Slide Time: 18:01)

### Principal Component Analysis (PCA) (contd..) (Example # 06)

Another method to double check is to find the variances of Y.  $Y_1 = \delta_{11}X_1 + \delta_{12}X_2 + \delta_{13}X_3$

- $Var(Y_1) = 0.8719^2 \times Var(X_1) + 0.36976^2 Var(X_2) + (-0.3210)^2 Var(X_3) + 2 \times 0.8719 \times 0.36976 \times CoVar(X_1, X_2) + 2 \times 0.36976 \times (-0.3210) \times CoVar(X_2, X_3) + 2 \times 0.8719 \times (-0.3210) \times CoVar(X_1, X_3) = 1.6792$  which is the value of the first eigen vector as calculated before

DADM-I
R.N. Sengupta, IIM Dept., IIT Kanpur
401

Another method to double check is to find the variances of Y and we will check that the variability of Y 1, Y 2, Y 3 would be the in that order has already decided. So, variability of Y so, what was basically Y? So, Y 1 was equal to delta X 1. So, I should be using 1 1 so, 1 2 1 3. So, I want to find out the variability of that.

So, variability variance would be delta 1 1 square which is here good then variance of X 1 which is already known to us from the table if you remember. The variance covariance matrix of X then we go to delta 1 2 whole square so, which will be equal to this. So, I will just I am marking those values. So, I am going very little bit slow so, you can check as I do it. Variance of X 2 which would also be from that covariance matrix then delta 1 3 square would be this value multiplied by variability of X 3.

Now, what we will have? We will have basically the covariance values. So, they would be they are a symmetric. So, hence they would be two values corresponding to 1 2, 1 3 then 2 3. So, 1 2 values would be covariance of  $X_1, X_2$  is given and we will basically have the corresponding values of  $\delta_{12}$  and  $\delta_{21}$  which we will have. This is the values, then when we go to covariance of 2 3 it will be multiplied by the corresponding values, then we go to covariance of 1 3 it will be again be monitored by the corresponding values.

So, technically,  $\delta_{11}$ ; so, that will come once here, second time when we are trying to find out corresponding to covariance of  $X_1, X_2$  and third time it will come when we are trying to find out the covariance of 1 and 3. So, let us check. Yes, it is there. So,  $\delta_{11}$  is taken care. Now, let us come to  $\delta_{12}$ ,  $\delta_{12}$  would be corresponding to this value. So, that is for the variance of 2 then it will come into the variance of 1 and 2 and it will also come into the variance of 2 and 3. So, let us check. Yes, it is there.

Now, we will come to  $\delta_{13}$ . So, that will again appear thrice. So, let us take the green color. So, that will be minus this is a minus sign if you remember for the Eigenvector. So, let us come there in the variance of 3, then it will come into the variance of 1 3. So, variance of 1 3 is here variance and 2 3 is also here. So, that takes care of all the delta values. So, we find out the variances. So, that would actually be equal to the first Eigenvectors because that is basically maximum amount of variance which we will trying to basically find out. So, that is actually equal to and that proves the second point corresponding to  $Y_1$ .

The first one was basically variability in the decreasing order second was orthogonality. So, we have taken care the first part and the second part; first part for all of them second part for the first  $Y_1$ .



(Refer Slide Time: 22:25)


**Principal Component Analysis (PCA) (contd..) (Example # 06)**

▪ Similarly 2<sup>nd</sup> d: Y<sub>2</sub>

▪  $Var(Y_2) = 9.4789$   $Y_2 = \delta_{11} X_1 + \delta_{22} X_2 + \delta_{33} X_3$

▪  $Var(Y_3) = 17.8418$

$\delta^2 Var(X_1) + \delta^2 Var(X_2)$   
 $+ \delta^2 Var(X_3) + 2 \otimes \otimes Cov(X_1, X_2)$   
 $+ 2 \square \square Cov(X_1, X_3) + 2 \circ \circ Cov(X_2, X_3)$



3x3

DADM-I R.N. Sengupta, IIM Dept., IIT Kanpur 402

So, let us come to Y 2 again, you will basically have I will write on the formula because the calculations would be simple. So, it will be delta for the first for variability of Y it will be. So, let me write down Y first, let me write down Y here. So, variabilities is coming 2 1 square. So, I will write down the principal values along the variance and covariance. So, this would be considered first. So, this is, so, the first value is taken care, second value is taken care, third value is taken care.

Now, I need to take care of the covariances. So, there is one value here and a one value here that will be twice into delta 1 2. So, the wait just. So, they would be basically I would not give the suffix now, wait. So, I should have basically written as. So, this will be delta should be delta 11 for the first case.

So, how should I, wait? Just wait because the nomenclature may get little bit confusing. So, I want to basically settle it out. So, the for this is for the second stage that is Y 2, I mention it 11 mention it. So, these would be for the first one, for the second one, the third one corresponding to the of the diagonal elements. So, you remove the suffixes. So, it will be easy for me to denote. So, it will be delta 1 which is here delta which is second case delta is the third case then I will have the values.

So, I will basically multiply these values. So, these are the value which I am multiplying by the covariance of X 1, X 2 then it will be twice into these values multiplied. So, it will basically one would be. So, I will take this here. So, one of so, this really, sorry, this

would not come, this would not come this is the third value taken care and the fourth value taken care. So, this would be twice of these are not symmetric remember.

So, once we have this for  $Y_2$  we find out the value of  $Y$  variance of  $Y_2$  the second case and then the variance of  $Y_3$ . So, based on that we will proceed step by step to find out the variability and the Eigenvectors which would basically come out from the corresponding case of the variance of  $Y_1, Y_2, Y_3$ .

So, this is a little bit involved problem I understand, solving it using the same problem using and then as simple Eigenvalues, Eigenvectors and then trying to find out the combinations to find out  $Y_1, Y_2, Y_3$  and then two things; find out the variability and the variance and the orthogonality. So, once you have done all these three things your rest assured the principal component method which you are try to followed gives you the actual result which is required, but remember again that the linearity of the axis is important and the standard form which is utilized where the units of the of  $X_1, X_2, X_3$ , till  $X_p$  are not affecting our calculation.

So, with this I will end the 39th lecture and continue discussing more about the multivariate statistics in the subsequent classes.

Have a nice day and thank you very much.