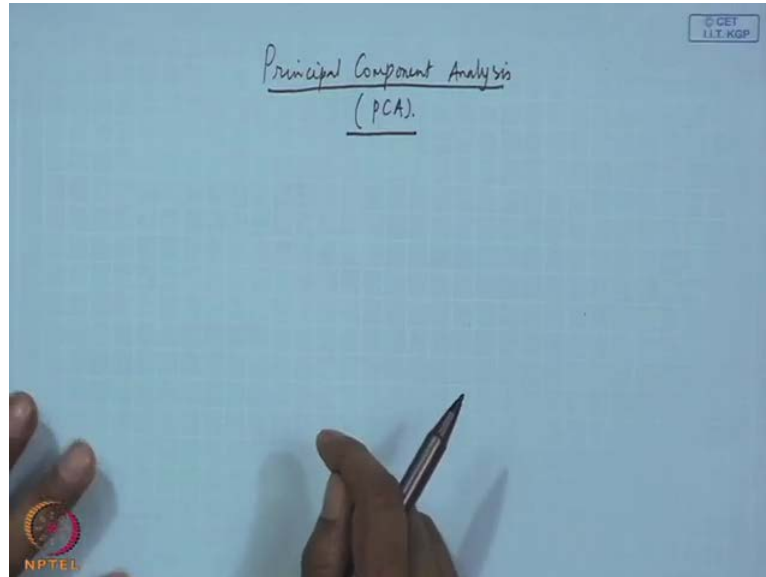


**Applied Multivariate Statistical Modeling**  
**Prof. J. Maiti**  
**Department of Industrial Engineering and Management**  
**Indian Institute of Technology, Kharagpur**


**Lecture - 30**  
**Principal Component Analysis (PCA)**

(Refer Slide Time: 00:28)



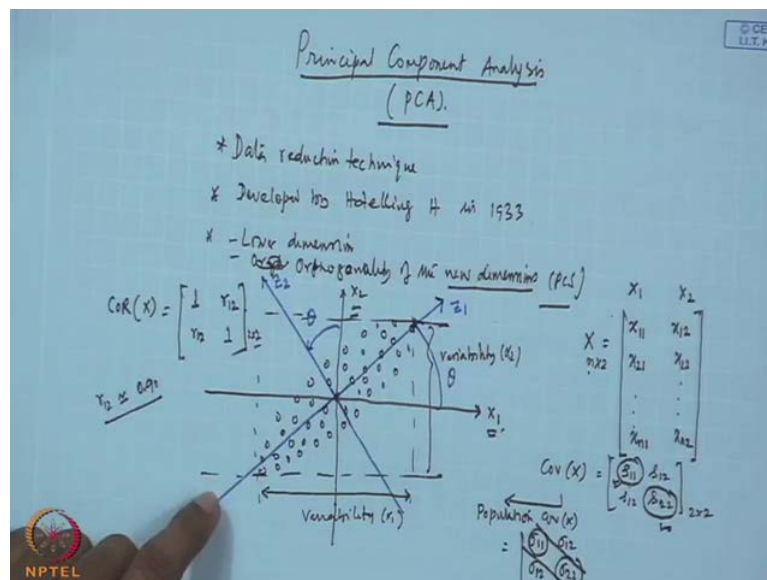
Good morning today we will discuss principal component analysis, principal component analysis PCA.

(Refer Slide Time: 00:44)

<h2>Contents</h2>	
<ul style="list-style-type: none"><li>■ Introduction</li><li>■ Conceptual model</li><li>■ Extraction of principal components</li><li>■ Sampling distribution of eigenvalues and eigenvectors</li><li>■ Model adequacy tests</li><li>■ A case study</li></ul>	 <span style="float: right; font-size: small;">Dr J Maiti, IEM, IIT Kharagpur</span>

So, let us see the content of today's presentation, first we will describe the basic concepts prevailing with principal component analysis. Then we will see that how principal component can be extracted from a given data set. Then we will go for sampling distribution of Eigen values and Eigen vectors. You will see that Eigen value, Eigen vector decomposition of the covariance or correlation matrix is the means for extracting principal component. Then followed by model adequacy test, then we will describe one case study. I think it requires around two hour of lecture first hour we will try to complete up to that extracting principal components. So, what is principal component analysis?

(Refer Slide Time: 01:53)



So, principal component analysis is a data reduction technique. Data reduction technique it is extensively used and developed by developed by Hotelling H in 1933. It is the data reduction is done with two prospective in mind, one is that lowered dimension, second one is orthogonality of the new dimensions. Orthogonality of the new dimensions new or transformed dimensions, which are basically P C S principal component. So, let us now go for the very simple two dimensional plot.

For example, let our dataset is X, which basically composed of n cross two that means m measurements on two variables. So, like this, so this one is variable X 1 and variable X 2 then this one X 1 1, X 2 1. So, like this X n 1, then X 1 2, X 2 2 like this, X n 2. If you plot the data is here suppose this axis X 1, this is X 2. Now, if you plot the data let us

assume that the data is like this data plot looks like this, now if you compute the covariance matrix of  $X$ . You will be getting as we are talking about data, so we have collected a sample in that sense. You have to write this will be  $s_{11}, s_{12}, s_{21}, s_{22}$ . So, two cross two matrix, so this scattered diagram shows that there is a relationship between  $X_1$  and  $X_2$

So, if you calculate correlation matrix between for  $X$ , what will happen? You will get certain correlation coefficient. This will be your correlation matrix two cross two, you will find out that  $r_1$  to that will be as you. Now, that correlation coefficient varies from minus 1 to plus 1 and here  $X_2$  and  $X_1$  the relationship is positive. So, you will get  $X$ , what large value may be it may be almost equal to 0.9 here 0.90 getting me. So, this is one aspect that we have taken two dimensional or regional data  $X_1, X_2$  their scattered plot looks like this. It shows there is linear relationship primary linear relationship.

If you go for correlation matrix you will be finding out that the relationship is strong enough. This is first, second issue here if I want to see the variability of  $X_1$  and variability of  $X_2$  then these are  $s_{11}$  and  $s_{22}$ , this is the variance component. Now, if we go for that this is sample, suppose population co variance population co variance, then what will happen? This will become  $\sigma_{11}, \sigma_{12}, \sigma_{21}, \sigma_{22}$ .

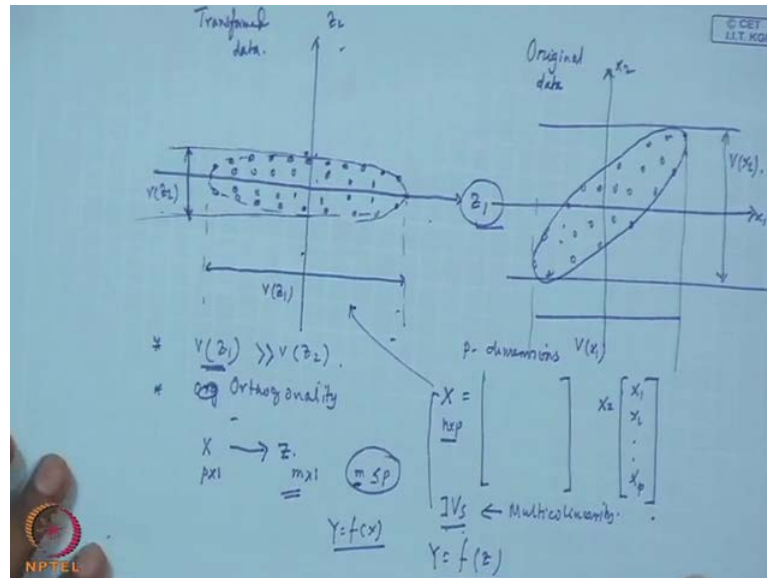
Now, this population variance component for  $X_1$  is this  $\sigma_{11}$  and for  $X_2$  is  $\sigma_{22}$ . If we see the scattered here if we say that  $n$  is representative enough for the population large value. If we see the scattered here then you see that the across this  $X_1$ , this is the range variability range. Similarly, if see across  $X_2$ , so you will be getting another range that is variability  $X_2$ .

So, you see that definitely the  $X_2$  and  $X_1$  variability are not same, but their substantial values that means if this is my diagram, this is my relationship, then definitely,  $s_{11}$  and  $s_{22}$  for  $X_1$  and  $X_2$  respectively, this value is high correct. Now, let us see that another the same diagram, let us see that you are rotating this axis  $X_1$  and  $X_2$  anticlockwise rotating the axis  $X_1$  and  $X_2$  rigid rotation keeping this origin.

This one visit by angle  $\theta$ , then what will happen you will get new direction for  $X_1$  and  $X_2$ ? Correct, if I rotate  $X_1$  by  $\theta$  then  $X_2$  is also will be rotated by  $\theta$ . Now, if I say in the new direction this one is  $z_1$  and this is  $z_2$ , this one is  $z_1$ , then  $z_2$ , what

can you say about the variability of the dataset along  $z_1$  and  $z_2$  getting me? So, if I draw one more figure along  $z_1$  and  $z_2$  here then...

(Refer Slide Time: 10:20)



Suppose, I am writing this one, basically in this, what I am doing? Now, this  $z_1$  is this along theta, now I am rot I am just making it like this, so that it will be we will be able to see it clearly in the sense. Correct? If you see the scattered plot, now this one is  $z_1$  and this one is other one is  $z_2$  I think it is like this only correct.

Now, if we see the variability across  $z_1$ , this is variability of  $z_1$ . Similarly, if you see this one across  $z_2$  you are getting this variability across  $z_2$ . What are you getting here? If I plot both the figures now here. So, basically you got something like this one is  $X_1$  and this is  $X_2$  and we have seen that variability of  $X_1$  and variability of  $X_2$ .

So, this is our original data dimensions and this is as we have rotated by theta degree anticlockwise. We are saying that transformed data getting me transformed data. Now, it is clear from this two diagram that although here from variability explain variability of  $X_1$  and  $X_2$  point of view. If I see that the both are large and may be  $X_1$  variability little more than  $X_2$ , if I go by this  $z_1$  and  $z_2$ , what I am finding out? Variability across  $z_1$  is much more then variability across  $z_2$ . So, what we have got then variability across  $z_1$  is greater than variability across  $z_2$ .

Second thing is that this is line second issue is here, that here in  $X_1$  and  $X_2$  both the variables are correlated. That is way you are getting an ellipse by variant 1 ellipse inclined one that mean this if I see that ellipse here the major and minor axis of the ellipse are not parallel to  $X_1$  and  $X_2$ . It shows the dependency between  $X_1$  and  $X_2$ . Now, here in the transform case, what is happening? Major axis of the ellipse is across  $z_1$  minor is across  $z_2$ . This shows  $z_1$  and  $z_2$  are independent. That is what I said the orthogonal dimensions orthogonal dimension or orthogonality is preserved.

Now, can we do this is there any mathematics by, which we can we can transform the correlated structure data structure into uncorrelated data structures well as from correlating. That means in the reduced dimension, while reduced dimension if the variability across  $z_2$  is much less compared to  $z_1$ , then what will happen? We can say that the information content across this  $z_2$  dimension is very less, we can ignore this information. We can just capture the information here in statistical sense information is the variability variance part. So,  $z_1$  alone is sufficient to give us the information, what is available in the original data set.

If this is the case, I will go for only one dimension than this  $z_1$  that this what this the reduction. So, principal component analysis we will do this not necessary for necessarily. For these two dimension it can be a then for  $p$  dimensions by  $p$  dimension, what we mean that your  $X$  data matrix is  $n$  cross  $p$ . That mean the variable you are considering  $X$  variable is  $X_1, X_2$ , there are  $p$  variables  $p$  can be per large, it can may be fifty. So, we are now converting what you are doing this  $X$  is converted to  $z$  if  $X$  is  $p$  variable case  $p$  cross one vector  $z$  can be  $p$  cross  $z$  also we can extract suppose  $z$  is  $m$  cross one vector

So,  $m$  can be less the equal to  $p$  depending on the correlation structure the matrix. We will be using the rank of the matrix and all those things, but essentially what do you want? Basically, we want that  $m$  should be less than equal to  $p$ . If  $p$  is very large  $m$  should be as low as possible compared to  $p$ . So, that is what is done in principal component analysis getting me. So, principal component analysis is a data reduction techniques. It transform original data matrix or data set into some other dimensions of reduced that components and it preserves the orthogonality of the components.

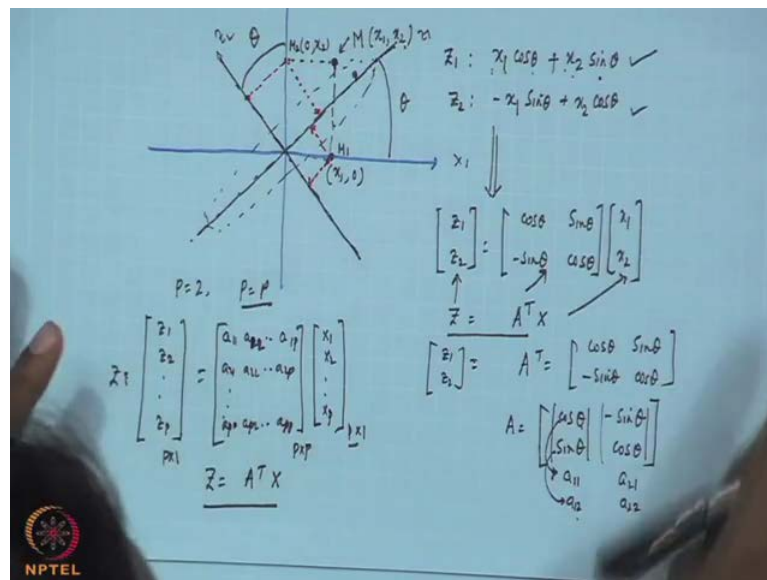
As a result, what will happen? What are the advantages? You will be getting from here advantages is, suppose we want to do a prediction model using multiple regression and

my X variable, these are all I V S independent variables. If I V S are correlate then that ultimately leads to multi co linearity problem multi co linearity are there. So, under multi co linearity condition the regression model what you want to fit that y equal to function of X linear model.

This model will not be a good one, because under multi co linearity it may not be possible to estimate the parameters. If you use to estimate, what will happen? It may be there may be distortion and many things will be there. There are different ways to do for example, is regression can be done in case of multi co linearity problem, but my question is, if I can make them independent, then this I V S be X can be transformed to this z scale. Like z here these are truly independent, now you can fit a model using f z getting me.

These are the advantage, now what is the principal? Basically, how this things that data reduction here or the orthogonal reduction? That is dimension orthogonality of dimension is maintained and data is reduced to lower dimension ,what is the method will go by that the two by two matrix case for s? So, the same two dimensional case.

(Refer Slide Time: 19:18)



So, if I see this, so you have seen that we got a ellipse like this in the original data set X 1 and X 2. Correct? Now, let us think that there is one point m in this for example, let this coordinate is X 1 and X 2, X 1 for X 1 variable X 2 for X 2 variable. Now, what you

have done? When you are converting to  $z$  that means  $z_1$  and  $z_2$  this is  $z_1$  and this one is  $z_2$ .

So, I am writing this point little high above for example, let it be here this point this is the point. So, you have rotated  $X_1$  by  $\theta$  anticlockwise and this also  $\theta$  I want to know first you say, what is this point? This value here, if  $m$  is  $X_1, X_2$  then this value is  $X_1$  0 for example, let this is  $m$  one point on the  $X_1$  axis similarly, if you project this  $m$  that projection on this point projection on  $X_2$ , what will get? You may get a point  $m_2$ , what is zero  $X_2$ ? This coordinate geometry.

Now, what we want we want to find out, what is the projection of this  $m_1$  on  $z_1$  as well as  $z_2$  getting me? So, you just let me draw this is axis, now clearly what you want here this is your  $\theta$ , our aim is we want to find out this is this one  $X_1$  0. I want to find projection of  $m_1$  on  $X_2$  at  $z_2$  and  $z_1$ . So, what you do? You will find draw a perpendicular here from this point you want to draw a perpendicular. So, from this point draw a perpendicular here similarly, you want also from this point you want to draw a perpendicular line here and perpendicular line here, what you are doing? Now, that  $X_1$  component is projected on  $z_1$  as well as  $z_2$ ,  $X_2$  component is projected on  $z_1$  and  $z_2$ .

Now, this angle is  $\theta$ , this is  $\theta$ , if this one is  $\theta$  and your this also  $\theta$ , so this will be  $\theta$ . So, ultimately, what you will find out? That the projection of  $m_1$  on  $X_1$  will be  $x_1 \cos \theta$ . So, if I write  $z_1$  we have there are two projection one is from  $X_1$  one is  $\theta$ . Similarly, there will be another projection that is a plus  $X_2 \sin \theta$   $X_1 \cos \theta$  plus  $x_2 \sin \theta$ . So, similarly in the  $z_2$  axis, what will happen? That minus  $x_1 \sin \theta$  plus  $x_2 \cos \theta$ , simple trigonometry. So, you will be able to find out only just know the  $\theta$  once you know the  $\theta$  you know the triangle and project it.

So, then this one if I write in matrix term I can write like this  $z_1, z_2$  equal to  $\cos \theta$   $\sin \theta$  minus  $\sin \theta$   $\cos \theta$  into  $X_1, X_2$  we can write like this. Because,  $z_1$  equal to  $\cos \theta$   $x_1$   $x_1 \cos \theta$  plus  $X_2 \sin \theta$   $z_2$  equal to minus  $x_1 \sin \theta$  plus  $x_2 \cos \theta$  this equation you are getting. So, then I can write this one equal to  $z$  equal to it transpose  $X$  where  $z$  is nothing but this one  $z_1$   $z_2$  and a transpose this one. We are saying a transpose this is a transpose and this is your  $X$  and this is your  $z$ . So, if I do like this then a transpose equal to our  $\cos \theta$   $\sin \theta$  minus  $\sin \theta$   $\cos \theta$ .

So, then what is the e a matrix transpose of transpose this will be cos theta minus sin theta sin theta cos theta this row transformation. So, instead of p equal to two if there is p equal to p? What will happen? You will get z 1 z 2 z p that will be your z and your this equal to A transpose. So, A transpose I am now writing that like this our A is a 1 here this is nothing but this column and a 2 is this column. So, if there are p variable there are two variable, so that why you are getting this one is equivalent to a 1 1 this is equivalent to what we are saying a 1 1, X 1 plus a 1 2, X 2.

So, this is a 1 2 and this one is a 2 1 and a 2 2 you can do this in theta cos theta sin theta we have written here in place of cos theta sin theta I am writing that this is a 1 1, a 1 2, a 2 1, a 2 2 like this. If there is p variable case through, you can write down like this a 1 1, a 2 1 a this is transpose A transpose s a 1 on a 1 2 like this a 1 p a 2 one a 2 2 like a 2 p. Similarly, a p 1 a p 2 like a p p, this is your p cross 1, this one is your p cross p and then your X 1, X 2, X p, that means this two dimension this case. Now, we are making like this and we are writing in terms of s. So, this if I write further I can write this one is a transpose X, what we have written earlier also? Then what do you want to see that how this that we say that it will be orthogonal transformation. So, how this orthogonality is maintained here with this two dimension case we want to see.

(Refer Slide Time: 28:41)

$$A = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} = [a_1 \ a_2]$$

$$a_1 = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}, \quad a_1^T a_1 = \begin{bmatrix} \cos \theta & \sin \theta \end{bmatrix} \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} = \cos^2 \theta + \sin^2 \theta = 1.$$

$$a_2^T a_2 = 1.$$

$$A^T A = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = AA^T = A^{-1}A = I$$

orthogonal transformation

What we have written for the two dimension case A equal to our cos theta sin theta minus sin theta cos theta, that we are saying we can write like this a 1 and a 2 getting me.



Now, then what is you a 1 a 1 is cos theta sin theta. Now, if I want to know what is a 1 transpose a 1? What will be the value here? cos theta sin theta into cos theta sin theta will be cos square theta plus sin square theta will be one getting me. Similarly, you will be getting a 2 transpose a 2 equal to one. Correct? Now, then if I want to know that what is A T transpose a that will be cos theta sin theta minus sin theta cos theta into a is cos theta sin theta minus sin theta cos theta 2 cross 2 cross 2 cos cos sin sine plus cos square theta plus sin square theta that will be 1.

Now, then cos minus sinesine minus cos this will be 0,1. So, this will also be if you do a a transpose you will get the same thing. Correct? See this is s symmetric matrix in such a way that if you do like this A inverse, A is also be getting same thing even this is identity k matrix. That means what we are getting A inverse A is I when a inverse A is I this is what is orthogonal matrix that of diagonal elements are 0. So, you are getting orthogonal transformation and this transformation matrix, this is orthogonal transformation matrix A.


(Refer Slide Time: 32:04)

Conceptual model – p variables

$$\begin{aligned} z_1 &= a_1^T X = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p \\ z_2 &= a_2^T X = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p \\ &\vdots \\ z_j &= a_j^T X = a_{j1}X_1 + a_{j2}X_2 + \dots + a_{jp}X_p \\ &\vdots \\ z_p &= a_p^T X = a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p \end{aligned}$$

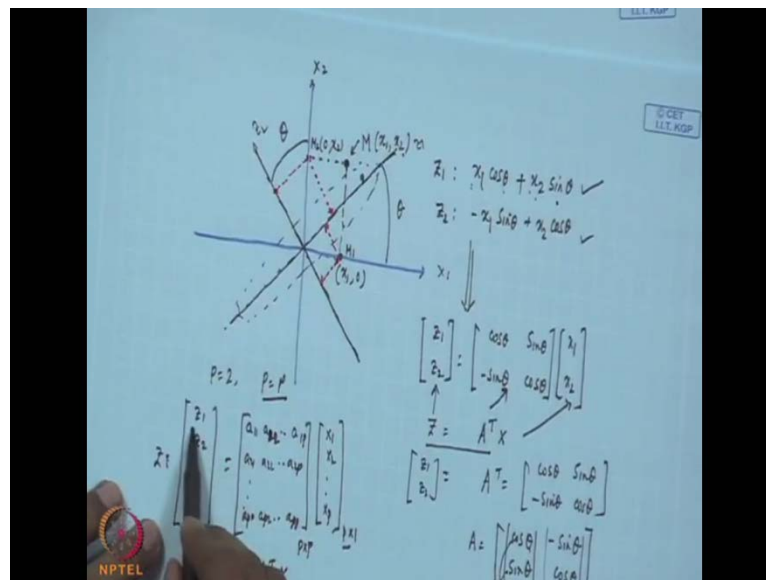
$a_j^T a_j = 1, j = 1, 2, \dots, p$

$Var(z_1) \geq Var(z_2) \geq \dots \geq Var(z_p)$


Dr J Maiti, IEM, IIT Kharagpur
7

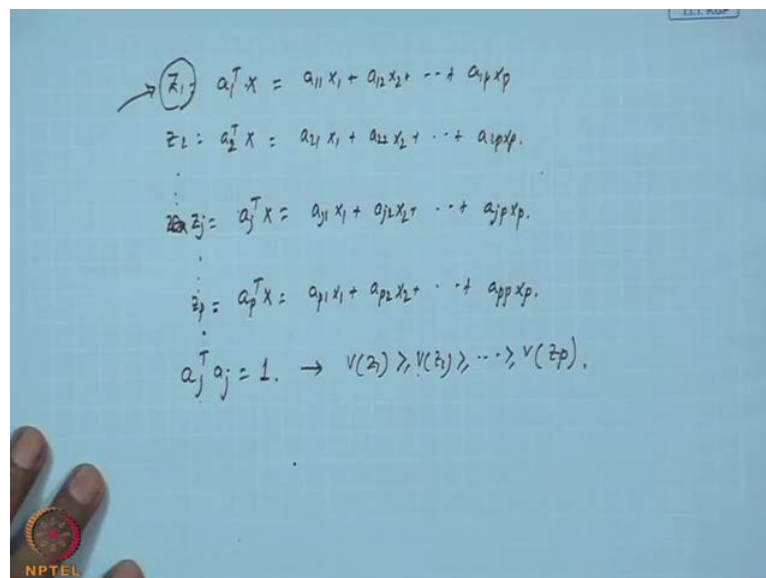
Now, what we will do, I will show you one slide here you see that there are p variable for X and you can extract p principal components z 1 z 2 to z p and their equation will be like this. What we have seen then what I mean to say if you just see this our earlier development that I think, I told you here.

(Refer Slide Time: 32:35)



So, like this you will be getting, so what will be getting?  $z_1$  equal to a  $1_1$ ,  $X_1$  plus a  $1_2$   $X_2$  plus a  $1_3$   $X_3$ .

(Refer Slide Time: 32:48)



$Z_1$  equal to a  $1$  transpose  $X$ , which is a  $1_1$   $X_1$  + a  $1_2$   $X_2$  dot dot a  $1_p$   $X_p$ . Similarly,  $z_2$  a  $2$  transpose a  $X_2$   $1_1$   $X_1$  + a  $2_2$   $X_2$  like a  $2_p$   $X_p$ . So, if you write down here  $z_k$  or  $z_j$  you write let it be  $z_j$  then a  $j$  transpose  $X$  a  $j_1$   $X_1$  + a  $j_2$   $X_2$  dot dot a  $j_p$   $X_p$ . So, following if you go up to the last principal component that may be possible to get that is  $f$   $p$  transpose a  $X$ , which is a  $p_1$   $X_1$  + a  $p_2$   $X_2$  like a  $p_p$   $X_p$ .

We said that we want orthogonal transformation, so when you are reducing the dimension the principal components are, we want this type of a trans a j transpose a equal to 1. That is one thing second one, what do you want? We want that when we are extracting the principal component z 1, which is the first principal component the variability in z 1 must be the maximum must be the maximum.

What we have discussed earlier? In this explanation we say we want the first principal component in such a manner that, it will explain the maximum variance possible. Second principal component will explain the next maximum variance followed. Similarly, third one like this with this is the case, then I can say here variance of z 1 greater than equal to variance of z 2 greater than equal to variance of z p. So, what we will do then? We will, just one minute, so I have shown you this thing and that, this is possible. Now, we want to find out what will be the principal components given a data matrix X? How to extract the components? So, here before going to this extraction part I have to tell little bit of that we are saying that.

(Refer Slide Time: 36:28)

$$\rightarrow z_j = a_j^T X \quad \underline{V(z_j)}$$

$$\text{jth pc} \quad \uparrow$$

$$V(z_j) = V(a_j^T X) = a_j^T V(X) a_j$$

$$= a_j^T \Sigma a_j$$

$$Cov(X) = \Sigma$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma_{pp} \end{bmatrix}$$

$$E(z_j) = E(a_j^T X)$$

$$= a_j^T E(X)$$

$$= a_j^T H \leftarrow 1 \times p, 1 \times 1$$

$$a_j^T X \sim N(a_j^T H, a_j^T \Sigma a_j)$$

In general, we are saying z j is a j transpose X were z j is the jth principal component. This is jth p c and this a j and X, what are those a j and a x? You will be X you know what is this a j, we will be defining later on. It is if you see the equation and similar to regression coefficients a 1 1 X 1 plus a 1 2 X 2 plus a 1 p X plus. Similar, to regression coefficients, but it is not.

So, now I want to know I say that variance of  $z_j$  you require to compute, because variance of  $z_1$  greater than variance of  $z_2$  like this, what will be the variance of  $z_j$ ? It will be variance of  $a_j^T X$ . Then it will be  $a_j$  is the constant here, so  $a_j^T$  transpose variance of  $X$   $a_j$ . As we know this is a vector basically  $a_j$  is a vector later on you will be you know that  $a_j^T$   $a_j$   $1$  to  $a_j^T$   $p$  that is a vector. Now, what is the variance of  $X$ ,  $x$  is what is the variance of  $X$  means, that will be the covariance of  $X$ . So, covariance of  $X$  will be  $\sigma$  you have seen earlier in different class  $\sigma_1$  to  $\sigma_1$   $p$   $\sigma_1$  to  $\sigma_2$   $\sigma_2$   $p$  that is  $\sigma_1$   $p$   $\sigma_2$   $p$   $\sigma_p$   $p$ .

So, that means this will be  $a_j^T \sigma a_j$  then what will be the expected value of  $z_j$  expected value of  $z_j$  will be expected values of  $a_j^T X$ , which is  $a_j^T$  expected value of  $X$ , which is  $a_j^T \mu$ . Because, we said that  $X$  is  $p$  variable vector and it is it has  $p$  mean vector  $\mu_1$  to  $\mu_p$ . Correct? So, that means essentially that  $a_j^T X$ , which is something, which is having  $a_j^T \mu$   $a_j^T \sigma a_j$  that is the mean and variance part. So, if there is normally distributed then you will put normal  $n$ , so  $a_j^T$  is one cross  $p$ .

Now, what will be your  $\mu$  part? Suppose, this one if you  $1$  cross  $p$  into  $p$  cross  $1$ . So, one into one into one and this one also will be one into  $p$  into  $p$  into  $1$  that is one into  $1$ . So, this is  $a_j^T X$  this one will give you a univariate case this is linear transformation of variable  $X$  all the principal components are linear transformation, that  $a_j$ ,  $a_j^T X$  that is the linear transformation. Now, if you do not get the values of  $\sigma$  capital  $\sigma$  as well as the  $\mu$ , because these are population these are population parameter that is  $\mu$  and  $\sigma$ .

If you know  $\mu$  and  $\sigma$  and then giving a data matrix you go for this type of extraction of  $z_j$ . Then this will be known as population PCA getting me, what I mean to say if you know  $\mu$  and  $\sigma$  of the population that  $X$  variable case. Then you are doing population principal component analysis, but it is seldom known you will not get this two values. So, what is the then the rescue is you will use  $\bar{X}$  as estimate of  $\mu$  and  $s$  as estimate of  $\sigma$ . So, when we use  $s$  actually here, we will be we will be playing with the covariance matrix.

So, if you use  $s$  capital  $S$ , which is the covariance sample covariance matrix then your principal component analysis will be known as sample principal component analyse. So, we will be writing it as sample PCA and essentially population PCA is not possible.

(Refer Slide Time: 42:03)

Handwritten notes on a blue background:

$\rightarrow Z_j = a_j^T X$        $V(Z_j)$   
 jth PC  $\uparrow$   
 $V(Z_j) = V(a_j^T X) = a_j^T V(X) a_j$   
 $= a_j^T \Sigma a_j$   
 $E(Z_j) = E(a_j^T X) = a_j^T E(X) = a_j^T H$   
 $a_j^T X \sim N(a_j^T H, a_j^T \Sigma a_j)$

Population:  $\mu, \Sigma$   
 Population PCA  
 $\bar{X} = \frac{1}{n} \sum X_i$   
 $S = \frac{1}{n} \sum (X_i - \bar{X})(X_i - \bar{X})^T$   
 Sample PCA

$Cov(X) = \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma_{pp} \end{bmatrix}$

So, we will go for sample PCA, because when we are talking about applied principal component analysis, then we have to rely on the data and will go by sample PCA.

(Refer Slide Time: 42:34)

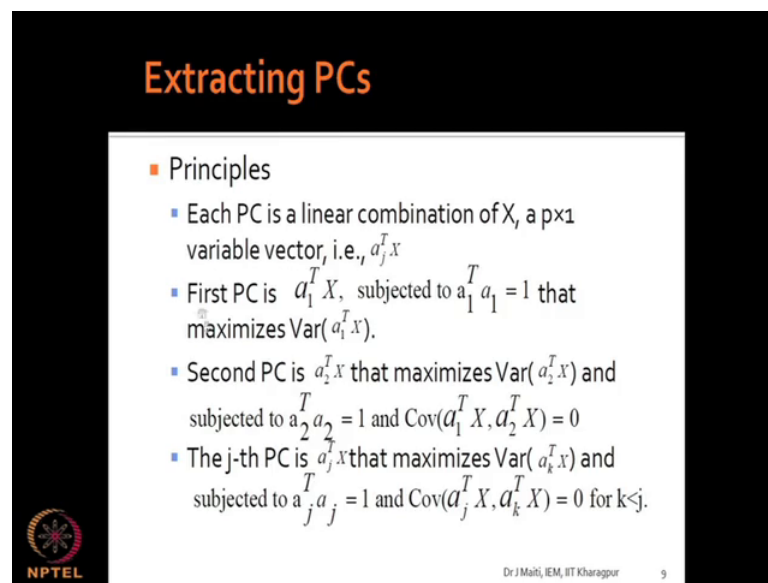
Handwritten notes on a blue background:

$E(Z_j) = E(a_j^T X) = a_j^T E(X) = a_j^T \bar{X}$   
 $V(Z_j) = V(a_j^T X) = a_j^T Cov(X) a_j = a_j^T S a_j$   
 $a_j^T X \sim N(a_j^T \bar{X}, a_j^T S a_j)$

So, next we will discuss sample PCA that means how we will extract all those thing. When we will use sample PCA. Then your extracted value of  $z_j$  will be expected value

of a  $j$  transpose  $X$ . So, this will be a  $j$  transpose expected value of  $X$ , which will be a  $j$  transpose  $\bar{X}$  from sample point of view. Similarly, your variance of  $z_j$  this will be your variance of a  $j$  transpose  $X$ , which will be nothing but your  $j$  transpose covariance of  $x$  a  $j$ . This is a  $j$  transpose, now covariance of  $s$  will be replaced by capital  $S$  a  $j$ . Correct? If I say that this can be normally distributed then a  $j$   $x$  it is normally distributed with a  $j$  transpose  $\bar{X}$  a  $j$  transpose  $s$  a  $j$ .

(Refer Slide Time: 44:08)



**Extracting PCs**

- Principles
  - Each PC is a linear combination of  $X$ , a  $p \times 1$  variable vector, i.e.,  $a_j^T x$
  - First PC is  $a_1^T X$ , subjected to  $a_1^T a_1 = 1$  that maximizes  $\text{Var}(a_1^T x)$ .
  - Second PC is  $a_2^T x$  that maximizes  $\text{Var}(a_2^T x)$  and subjected to  $a_2^T a_2 = 1$  and  $\text{Cov}(a_1^T X, a_2^T X) = 0$
  - The  $j$ -th PC is  $a_j^T x$  that maximizes  $\text{Var}(a_j^T x)$  and subjected to  $a_j^T a_j = 1$  and  $\text{Cov}(a_j^T X, a_k^T X) = 0$  for  $k < j$ .

NPTEL  
Dr J Maiti, IEM, IIT Kharagpur 9

Now, I will explain you when we extract PC? What are the principals? We follow getting me, we will we will discuss the principal. Now, let us see the slide each PC is a linear combination of  $X$  1 linear combination of a  $X$ , if  $p$  cos one variable vector. That is a  $j$  transpose  $X$  you have seen this one already first PC is a 1 transpose  $X$  subjected to a 1 transpose  $a_1$  equal to 1. That also you have describes that maximises variability of this variance of a 1 transpose  $X$ . Second PC is a 2 transpose  $X$  that maximizes variability of a 2 transpose  $X$  and subjected to a 2 transpose  $a_2$  equal to 1. Covariance of a 1 transpose  $X$  a 2 transpose  $X$  equal to 0, keep in mind this one, what is happening.

(Refer Slide Time: 45:18)

$$E(z_j) = E(a_j^T X) = a_j^T E(X) = a_j^T \bar{X} \leftarrow$$

$$V(z_j) = V(a_j^T X) = a_j^T \text{Cov}(X) a_j$$

$$= a_j^T S a_j$$

$$a_j^T X \sim N(a_j^T \bar{X}, a_j^T S a_j)$$

$$z_1 = a_1^T X \leftarrow V(z_1) = \text{the max}^m$$

$$a_1^T a_1 = 1$$

$$z_2 = a_2^T X \leftarrow V(z_2) = \text{the max}^m \text{ of the remain}$$

$$a_2^T a_2 = 1$$

$$\vdots$$

$$\text{Cov}(a_1 X, a_2 X) = 0$$

$$z_j = a_j^T X \leftarrow V(z_j) = \text{the max}^m \quad a_j^T a_j = 1$$

$$\text{Cov}(a_j X, a_k X) = 0 \quad \text{for } k < j$$

We are saying first component you are extracting that is our  $z_1$  is the first component, which is a  $1^T X$  you will extract in such a manner that the variability of  $z_1$  will be the maximum. Also you a  $1^T a_1 = 1$  that is the normalisation case. So, I have a data set, so and in this suppose this is my total variability portion. So,  $z_1$  will extract as max as possible by  $z_1$ . So, let  $z_1$  variability extracted is this much then I am coming to the second p c, which is my  $a_2^T X$  here. What happen the remaining variance? I want to extract maximum of the remaining variance. So, variance of  $z_2$ , that is the maximum of the remaining.

So, let I am able to extract this one  $z_2$  see, ultimately if this is my total variability  $z_1$  and  $z_2$  has already may be around 70 percent is extracted by the two components. There are  $p$  components this is also here  $a_2^T a_2 = 1$  and covariance of  $a_1 X$   $a_2 X$  this equal to 0. Because, orthogonal component now in this manner if you go for  $z_j$  then you will write  $a_j^T X$ . Here, you will definitely maximise the remaining variance already  $j - 1$  components are extracted the remaining maximise this one, the remaining variance. Here, also you will write  $a_j^T a_j = 1$  and covariance of  $a_j X$   $a_k X$  this equal to 0 for  $k < j$ . Why  $k < j$ ? What I am saying you are means  $k$  up to  $k$  can be 1 to  $j - 1$ .

So, when you are extracting this, please keep in mind this is the principals what I am saying that when you are extracting the first component that basically, takes care of the

maximum variance of the data set subject to this normalization. Then a 2, a x, z 2 second principal component goes for the second max of variability of the remaining variance and subject to this as well as covariance component between the first two. When you go for the third, there also you will you will maximize the variability of that component subject to the remaining variability. You also make a normalize it that a three transpose a three equal to one and covariance between a 3, a 1 as well as a 3, a 2. That X is common that will be 0.

(Refer Slide Time: 48:58)

Handwritten mathematical derivation on a whiteboard:

$$z_j = a_j^T x$$

$$v(z_j) = a_j^T S a_j \quad a_j^T a_j = 1$$

$$a_j^T a_j - 1 = 0$$

$$L = a_j^T S a_j - \lambda(a_j^T a_j - 1) \quad \lambda = \text{Lagrange multiplier}$$

$$\frac{\partial L}{\partial a_j} = 0$$

$$2S a_j - 2\lambda a_j = 0$$

$$2(S - \lambda I) a_j = 0$$

$$(S - \lambda I) a_j = 0$$

Annotations on the whiteboard include "P.P.P." and "P.P.P." near the final equation, and a small box in the top right corner containing "© CRET I.I.T. KGP".

So, in the same manner you will extract. So, then essentially what we are doing now? Actually, if I go by the general component that is  $z_j$ , this is my  $a_j$  transpose  $X$  we are we want to maximize the variability of this, which is our  $a_j$  transpose  $s$   $a_j$ . Now, I am using sample covariance matrix and subject to what we are saying that  $a_j$  transpose  $a_j$  this equal to 1. Correct? So, we have two things first one, we want to maximize this second we have this concept. So, this second one I can write like this  $a_j$  transpose  $a_j$  minus 1 equal to 0.

As we want to maximize this with respect to this condition, so what we can do we can create a function. Suppose, this function is  $l$ , which is  $a_j$  transpose  $s$   $a_j$  minus  $\lambda$  into  $a_j$  transpose  $a_j$  minus 1. Correct? So, we are using a function this is the  $l$   $\lambda$  is the Lagrange multiplier  $\lambda$  is Lagrange multiplier. So, you know that this is the general way of maximisation process using langrage multiplier. So, what you require to



do now? You require to find out the a j value in such a manner that this function will be maximized. That means what I want dell a by dell a j, that is what I want getting me, so this you want to put it 0 provided dell 2 l by dell a j that will be that matrix. You will also get this should be that maximisation that positive definite and negative definite case is there that condition must satisfy.

So, this ultimately results into this equation it will be like this s minus lambda I into a j equal to 0. See this is if you take derivative with respect to a j it will be 2 s a j. If you take derivate with respect to a j it will be 2 lambda a j this derivative. So, ultimately it will be 2 s a j minus 2 lambda a j that equal to 0. So, 2 will be cancelled out that 0 is there, so s minus lambda, but s is a matrix lambda is scalar this into a j equal to 0.

So, that is the equation, now s is your p cross p matrix. Because, p variables are there lambda is a scalar it all depends on that what will be there number of lambda, but I definitely p cross p matrix. Then this equation is a famous equation in matrix algebra, anyways you will find out that set of linear equations case suppose and more famous one is this one, suppose I want to know the lambda values, then what you require to do? You have to find out the characteristics equation this equal to 0.

(Refer Slide Time: 53:15)

Handwritten notes on a blue background:

- Characteristic eqn has  $p$  roots.  $\lambda_1, \lambda_2, \dots, \lambda_p$
- $S_{p \times p}$
- $|S - \lambda I| = 0$
- $|S - \lambda I| = 0$
- $\textcircled{S} \begin{bmatrix} 1.15 & 5.76 \\ 5.76 & 29.54 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$
- $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- $\begin{vmatrix} 1.15 - \lambda & 5.76 \\ 5.76 & 29.54 - \lambda \end{vmatrix} = 0$

So, determinant of this equal to s minus lambda I equal to 0 determinant of this equal to 0. This is a characteristics equation getting me, so if s is p cross p then this equation has p roots that means you will be getting lambda 1 with the condition greater then equal to

lambda 2, like this lambda p. So, let us see now some more slide here I said that this will be the that p the order polynomial is there p root you are getting. You have to this is the if you know lambda, basically from this equation you will be able to find out lambda values. Now, take 1 lambda put into this equation and then you find out the a values. I will show you one data set.

Now, this is the dataset what is the profit and loss? For example we have started with and if you of you plot the scattered, what you are getting? You are getting that it is almost linear relationship. So, that means one here data reduction is very much possible, so we tried with two variable case and what we found out that these are the things. So, first is the covariance matrix. Covariance matrix is s, so with this example s is 1.15, 5.76 then your 5.76, 29.54.

Now, what we want we want s minus lambda I determinant equal to 0. So, your I will be 2 by 1001 I is identity matrix, so if I do like this minus lambda I. So, that means instead of lambda I am writing 1001 this total determinant if I write and if I put into 0, then what is the resultant? This one 1.15 minus lambda 5.76 then 5.76, 29.54 minus lambda this equal to 0, determinant of this equal to 0.

(Refer Slide Time: 53:12)

Characteristic eqn has p roots.  $\lambda_1, \lambda_2, \dots, \lambda_p$

$S_{p \times p}$

$$\begin{vmatrix} 1.15 & 5.76 \\ 5.76 & 29.54 \end{vmatrix} - \lambda \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 0$$

$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$$\begin{vmatrix} 1.15 - \lambda & 5.76 \\ 5.76 & 29.54 - \lambda \end{vmatrix} = 0$$

The two roots of  $\lambda$ ;  $\lambda_1$  and  $\lambda_2$

$(1.15 - \lambda)(29.54 - \lambda) - 5.76^2 = 0$   
Quadratic equation of  $\lambda$ .

NPTEL

So, ultimately you will be finding out 1.15 minus lambda into 29.54 minus lambda minus 5.76 square equal to 0. It will be it is a quadratic equation of lambda because our s is two cross two. So, there will be 2 roots, so lambda 1 and lambda 2, so that roots will be the

two roots of two roots of lambda. Let it be lambda 1 and lambda 2 using this equation you can find out.

(Refer Slide Time: 57:11)

$$1.15 \times 29.54 - 1.15\lambda - 29.54\lambda + \lambda^2 - 5.76^2 = 0$$

$$\lambda^2 - 30.69\lambda + (1.15 \times 29.54 - 5.76^2) = 0$$

$$\lambda = \frac{-(-30.69) \pm \sqrt{(-30.69)^2 - 4(1)c}}{2 \times 1}$$

$$\lambda_1 = 30.66$$

$$\lambda_2 = 0.03$$

$$(S - \lambda_j I) a_j = 0$$

$$\lambda_1 = 30.66$$

$$\begin{bmatrix} 1.15 & 5.76 \\ 5.76 & 29.54 \end{bmatrix} - \begin{bmatrix} 30.66 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = 0$$

So, what I means to say the resultant equation is 1.15 into 29.554 minus 1.15 lambda minus 29.54 lambda plus lambda square minus 5.76 square equal to 0. So, lambda square minus you add this 30 I think 30.this lambda, this one then this two is to be added that you will be getting some value. This is the sum positive value you will be getting if I consider this as c then you lambda is minus b minus b is 30.69 minus b plus minus root over b square, b square minus four a c, a is 1, c by 2 into this will give you the Eigen values two values lambda one will be your 30.66 and lambda 2 will be 0.03. Correct?

So, you are getting your lambda 1 and lambda 2, then what you require to find out you require to find out the a values. So, what you will do? Now, we know s minus lambda I h j that will be equal to 0, so here will be using lambda j. Now, so as lambda 1 equal to 30.66, so you put s is 1.15, 5.76 then 5.76, 29.54. This minus lambda is your 30.66 and into this matrix 1001. This matrix this into a 1 equal to 0, now you are a 1 will be what a 1 1, a 1 2, because there are two variable.

(Refer Slide Time: 01:00:53)

$$\lambda^2 - 30.69\lambda + (1.15 \times 29.54 - 5.76^2) = 0$$

$$\lambda = \frac{-(-30.69) \pm \sqrt{(-30.69)^2 - 4 \cdot 1 \cdot c}}{2 \cdot 1}$$

$$\lambda_1 = 30.66$$

$$\lambda_2 = 0.03$$

$$(S - \lambda I) a_j = 0$$

$$\lambda_1 = 30.66$$

$$\begin{bmatrix} 1.15 & 5.76 \\ 5.76 & 29.54 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = 0$$

$$a_1^T a_1 = 1$$

So, I can write a 1 1, a 1 2, a 1 1, a 1 2 that will be 0 this value will be 0, but what will happen? Ultimately, you will be getting two equation and you will not get unique solution. You will get infinite solution under this conditions, because this is a equation is 0 and that means there will be infinite number of this a 1. So, in order to restrict that what we will do? a 1 transpose a 1 equal to 1 so if a 1 transpose a 1 equal to 1. Then what will happen? a 1 1, a 1 2, a 1 1, a 1 2 that is one. So, a 1 1 square plus a 1 2 square equal to 1. So, whatever equation you are getting here that earlier equation you put a 1 1 equal to from this equation. What is the value you are getting? You will be getting positive and negative values.

(Refer Slide Time: 01:01:07)

The image shows a whiteboard with handwritten mathematical work. At the top left, the equation  $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = 1$  is written. Below it, the equation  $a_{11} + a_{12} = 1$  is circled. To the right, the equation  $a_{11} + a_{12} = 0$  is written and underlined. Below that, the equation  $a_{11} = -a_{12}$  is written and underlined. At the bottom left, the equation  $\left(\frac{-a_{12}}{1}\right)^2 + a_{12}^2 = 1$  is written and underlined. In the bottom left corner of the whiteboard, there is a small logo for NPTEL.

So, you put this one once you put into this equation or a 1 1, now a 1 2 will be replaced by a 1 1 from this first equation a 1 is already there. Our a 1 relation you will be getting from here from this you will be getting, because this is two by two this into this plus this into this. So, it is a two variable equation will be like this some value into a 1 1 plus some value into a 1 2 that will be 0 like this. So, similarly now you are getting one relation from here suppose this is your p and this is your q. Then you are getting a 1 1 equal to q by p minus a 1 2. Here, you put instead of a 1 1 you put minus q by p square a 1 to plus a 1 2 square plus a 1 to square plus a 1 to square equal to 1. You will be getting a 1 to and putting here you will be getting a 1 1.


(Refer Slide Time: 01:02:52)

### Extracting PCs - example

Covariance matrix	Profit	Sales	Eigen-values	Proportion	Cumulative
Profit	1.15	5.76	30.66	0.999	99.90
Sales	5.76	29.54	0.03	0.001	100.00

Loading	PC1 (Z1)	PC2 (Z2)
Profit (X1)	0.98	0.19
Sales (X2)	0.19	-0.98

$$Z_1 = 0.98X_1 + 0.19X_2$$
$$Z_2 = 0.19X_1 - 0.98X_2$$


Dr. J. Malvi, IEM, IIT Kharagpur

13

Similarly, a 2 1 and a 2 2 you will be getting. So, the final one this is the that why this is your a 1 1, a 1 2 this is a 2 1, a 2 2. You are getting third principal component z 1 and z 2, this is the manner of extraction of principal component.