Stochastic Hydrology

Prof. P. P. Mujumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore

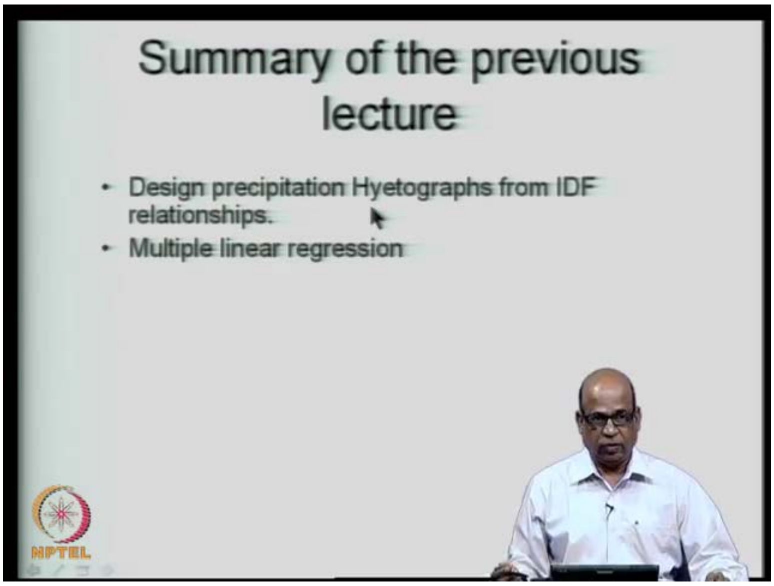
Lecture No. # 31

Principal Component Analysis

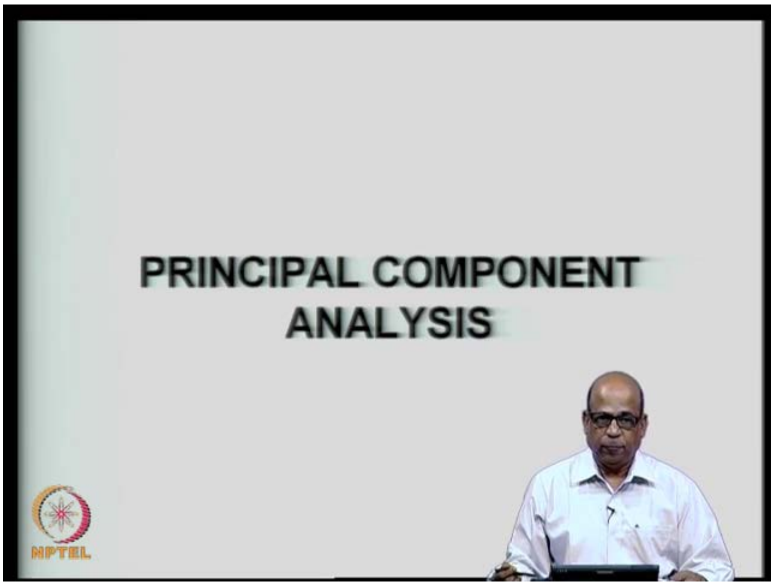
Good morning and welcome to this the lecture number thirty one of the course stochastic hydrology. In the last lecture what we essentially did is, to use the IDF relationships intensity duration frequency relationships, and convert it into the design hyetographs that was the first portion of the lecture, but in this subsequent part of the lecture we introduced a new topic this was to deal with multiple linear regression. So, starting with the simple linear regression where we are relating one dependent variable y with one independent variable x in a linear form y is equal to x plus v.

We extended that to relate one dependent variable y with several independent variables x 1, x 2, x 3, etcetera in a linear form. And, we saw how to obtain the coefficients beta 1, beta 2, etcetera when we are expressing y as a function of x 1, x 2, x 3 etcetera with the coefficients defined as beta 1, beta 2 and so on. Towards the end of the lecture, I mentioned that the independent variables x 1, x 2, x 3 and so on in the multiple linear regression may be correlate among themselves and also the number of variables that you may consider may be so large that the size of the problem itself becomes quite unmanageable in many realistic situations. And that is where we introduce the principal component analysis where we will deal with a set of uncorrelative variables and also (( )) only a few of the uncorrelated variables. So, that the size of the problems problem itself may reduce and that is what is called as the principal component analysis. So, in today is lecture we will introduce the principal component analysis.

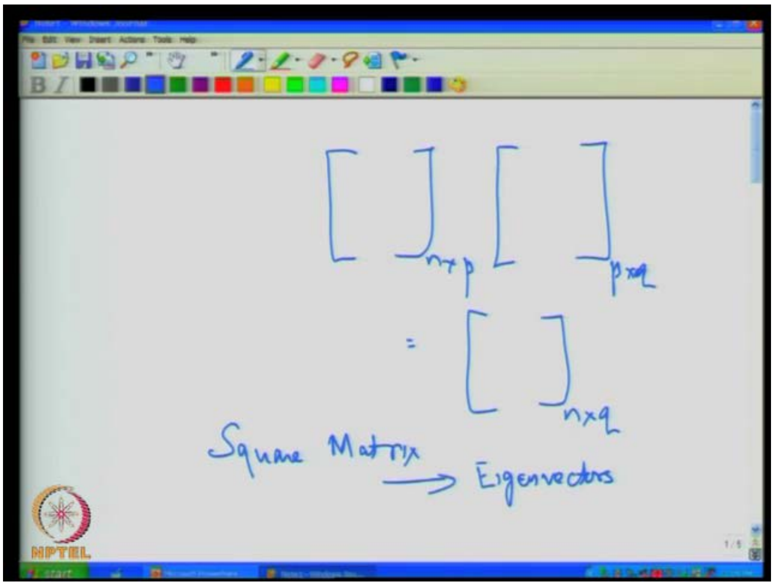
(Refer Slide Time: 02:35)



(Refer Slide Time: 02:49)



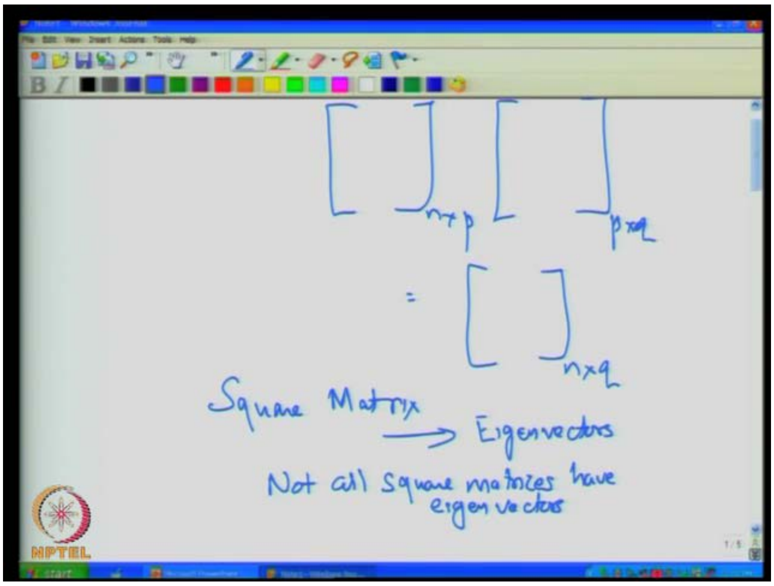
(Refer Slide Time: 04:16)



So, this is what we covered in the last lecture we discussed about the designed hyetographs from IDF relationships. And then, we introduced multiple linear regression we also discussed a problem related with multiple linear regression. So, today we will start with principal component analysis. Now, before going to the principal component analysis we need to revise a bit of matrix algebra. As you recall you know two matrices can be multiplied, if their sizes are compatible what do I mean by compatibility let us say you have a n by p matrix and another p q matrix. So, the number of columns of the first matrix equal to number of rows of the second matrix then, you can multiply two matrices with n by q in that particular example.

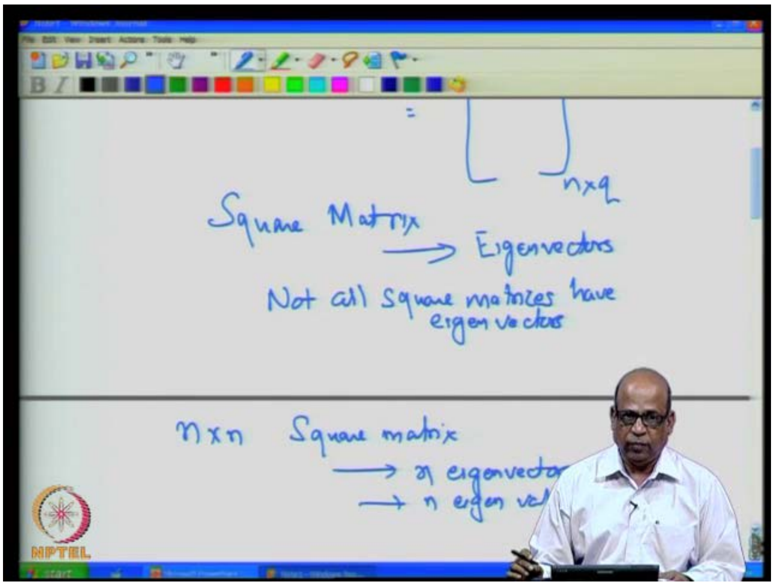
We also know from your from our earlier matrix algebra that a square matrix has associated with it eigenvalues and eigenvectors. So, we will start with a revision of the eigenvectors and eigenvalues which are necessary for discussing the principal component analysis. At least some of you would have gone through the matrix algebra earlier when we say two matrices are can be multiplied. As I mentioned let us say you have a n by p matrix and then you have a p by q matrix, and you multiply this you will get a n by q matrix. The concept of eigenvectors is closely associated with the matrix multiplication. We say for a square matrix that is, if you have a n by n matrix it can have eigenvectors. So, the eigenvectors are defined essentially for square matrix and not all square matrices can have eigenvectors.

(Refer Slide Time: 05:40)



So, some of the square matrices need not have eigenvectors that is the second one, second point not all square matrices have eigenvectors. Now what are these eigenvectors and how we determine etcetera we will see presently. Then if a n by n square matrix has a eigen vector let us say that you are considering n by n square matrix square matrix.

(Refer Slide Time: 06:30)



If it has eigenvectors then it has exactly n eigenvectors and associated with each of the eigen vector you have n eigenvalues. So, if you have a n by n square matrix you can you will have n eigenvectors and at most n distinct eigenvalues. So, how to determine these we will see today in todays class. It is also important for you to understand that the eigenvectors are orthogonal to each other in the, in the sense that they are perpendicular to each other. You remember that the vector indicates a direction say therefore, when you consider two eigenvectors the eigenvectors will be normal to each other or orthogonal to each other. We will now see how we determine the eigenvalues and eigenvectors which are necessary for carrying out the principal component analysis.

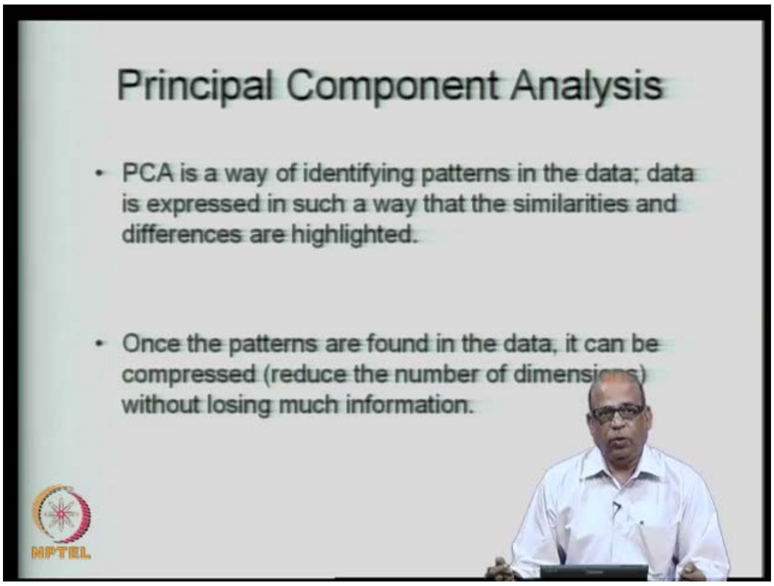
Just so, that you do not lose sight of what we are discussing the principal component analysis essentially we apply when you have to deal with multiple regression. Where you are dealing with a large number of independent variables. Typically let us say you are talking about hydrologic applications, where you want to estimate runoff at a particular location as a dependent variable. And it is dependent on several independent variables, let us say rainfall it is dependent and the area of catchment, it is dependent on the soil moisture; it may be dependent on the humidity and the temperature and So on. So, you want to relate the runoff at a particular location with all the independent variables.

Now these independent variables that you are talking about may be may have a correlation among themselves, in the sense that one variable may be dependent on the other. But we are still calling them as independent in as much as they are not dependent on y or the runoff in this particular case. For example your soil moisture may be related with rainfall and rainfall and soil moisture together are considered in the regression equation. So, to account for the correlations among the independent variables or actually to remove the correlations among the independent variables. And form another set of variables which are uncorrelated with each other and to possibly reduce the size of the problem itself.

What I mean by that is let us say you had ten variables y is equal to y is a function of x 1, x 2, x 3 etcetera ten variables, and each of the ten variables has 50 years of data and monthly data is what you are considering. So, each of the variables has six hundred values. So, the problem there mentioned is ten into six hundred now that may become slightly unwiedly. And therefore, to remove the correlations among the independent variables, and to possibly reduce the size of the problem we carry out the principal component analysis. And especially when you are dealing with the climate change impacts on hydrology towards the end of the lecture towards the end of this course. I will give one lecture on how do we handle the scaling issues.

When we are dealing with climate change impacts there it will become much more much more clear on where we use the principal component analysis. Especially when you are dealing with the scaling issues in the climate change impacts you deal with large number of climate variables. And therefore, it is important for us to reduce the size of the problem, and also to address the problem of correlations among the independent variables, and that is what we do in the principal component analysis. So, let us start with the basics of the principal component analysis.

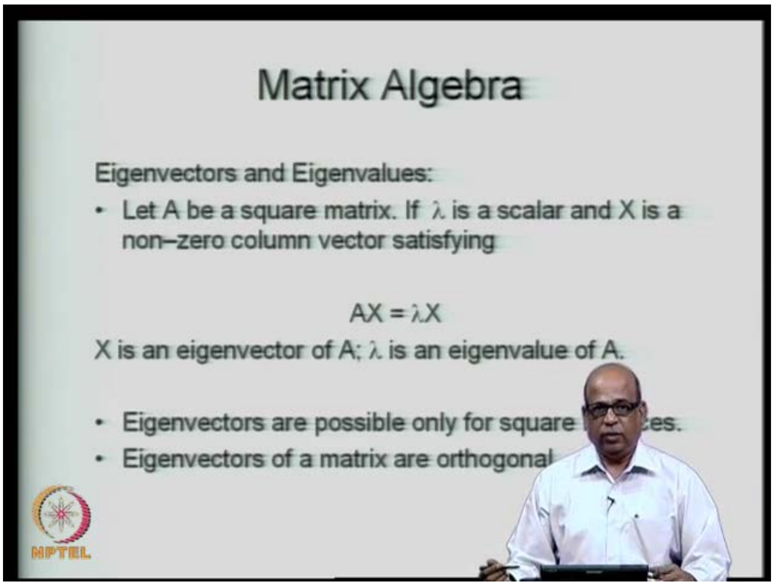
(Refer Slide Time: 10:26)



What does it do actually it is a way of identifying patterns in the data. You have ten variables all of the ten variables you have measured, and you have the observed data on all of the ten variables. So, there is a huge amount of data that is available to you the underlying pattern in this data set is what is captured by the principal component analysis. So, PCA is a way of identifying the underlying patterns in the data, and then once we identify the patterns, we express this data in such a way that the similarities among the data and the differences among them are highlighted in some senses.

And once the pattern is found in the data, the data can be compressed without losing too much of information. What do I mean by that, let us say that in the pattern of the data you see that some components. In some sense which we will see presently some components are much more predominant, in the in their information content compared to other components. Then you can usually only focus on these particular components which are much more predominant in their information content and that is what we mean by the data can be compressed.

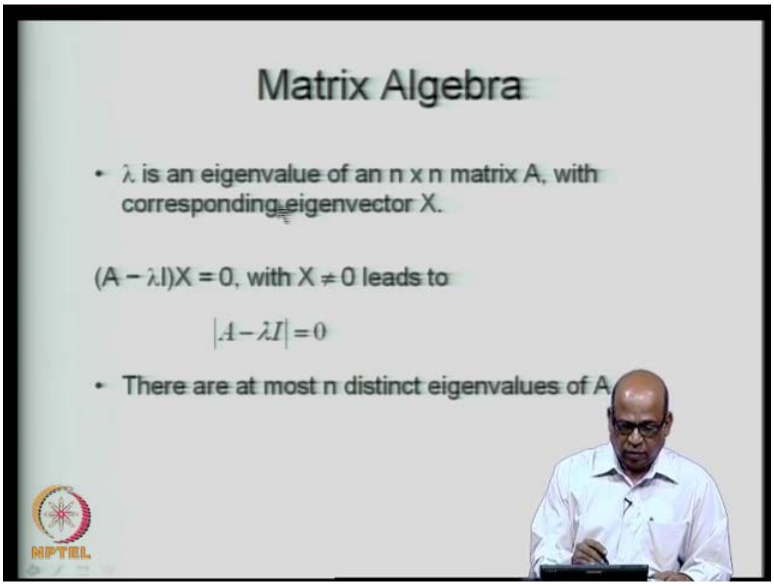
(Refer Slide Time: 12:11)



In fact, in image analysis etcetera the image compression uses principal component analysis in several situations. Now this will require a bit of background on matrix algebra. So, we will just quickly go through some preliminaries of the matrix algebra especially how we obtain the eigenvectors and the eigenvalues. So, first let us define what is a eigen vector and a eigen value. Let us say A is a square matrix and lambda is a scalar and X is a non zero column vector, satisfying A X is equal to lambda into X. So, let me demonstrate that that is we have A as a square matrix and this is a column vector lambda is a scalar and this is a column vector. Now this column vector X is a eigen vector of A and lambda is an eigen value of A.

That is if you can form if you can find a vector X which satisfies this A X is equal to lambda X with lambda as a scalar, and X being a column vector if we can find such a vector then it is called as an eigen vector of A, and the associated value lambda is called as eigen value of A. Now as I mentioned earlier eigenvectors are possible only for square matrices, and the eigenvectors X are orthogonal to each other. Let us say you had two eigenvectors then both of them will be normal to each other. Vector denotes a direction and therefore, when you, you can talk about the direction of a vector, and the two eigenvectors will be orthogonal to each other or normal to each other.

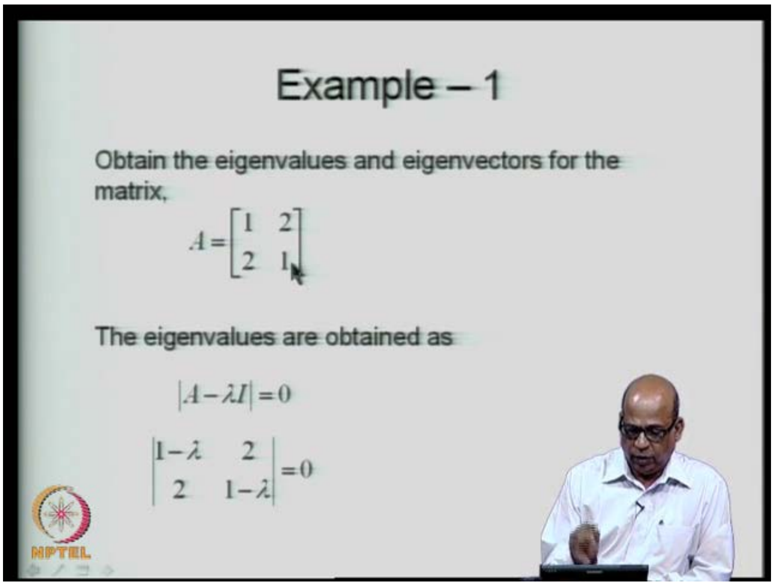
(Refer Slide Time: 14:15)



Then lambda is an eigen value of n by n matrix A, with corresponding eigenvector X. So, each of the eigen vector is associated with a eigen value lambda. Look at this now A X is equal to lambda X, this is how you define your eigen vector X. Now if I write A minus lambda I into X is equal to zero, with X not equal to zero this leads to the determinant A minus lambda I is equal to zero. And this is how we determine the values lambda that is, if A minus lambda into X is equal to zero you can either have X is equal to zero or A minus lambda A equal to zero. But we are saying that X is not equal to zero because you are stating that eigenvectors exist and therefore, you will get A minus lambda I the determinant of that must be equal to zero, this is how we determine the lambda or the eigenvalues.

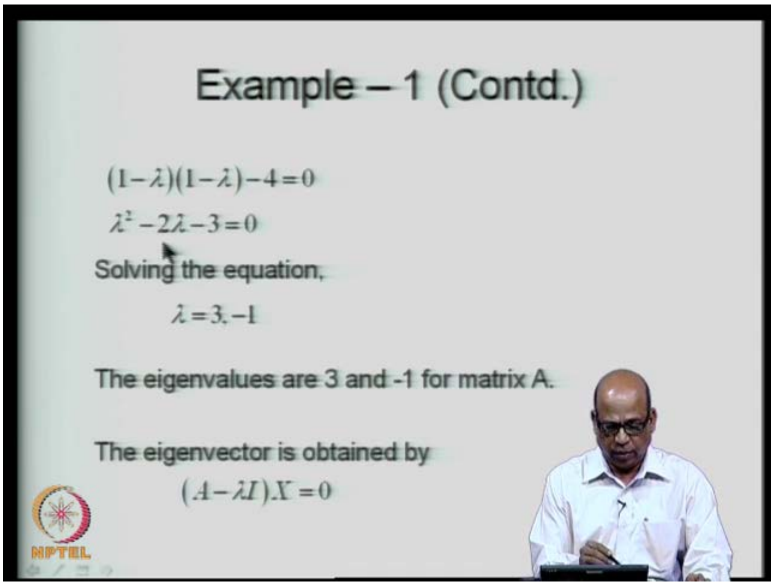
As you can see here this is the n by n matrix A is a n by n matrix and therefore, you may get maximum of n distinct eigenvalues of A you may have n value. But some of them may be equal to each other and therefore, you will get maximum of n distinct eigenvalues of A. So, these are the two important things that you must remember. One is you will determine the eigen vector X by this expression A X is equal to lambda X. And the other one is that you will determine the eigenvalues A minus lambda I is equal to zero determinant of A minus lambda I is equal to zero, and that is how you determine the eigenvalues lambda.

(Refer Slide Time: 16:33)



Any given square matrix first you determine the eigenvalues then use the eigenvalues in this expression to get the eigenvectors. So, let us look at one simple example now, let us say you take a square matrix A 1 2 2 1 very simple matrix, we will first obtain the eigenvalues. So, A minus lambda I this is a identity matrix I. So, A minus lambda I determinant of that is equal to zero. So, we will get 1 minus lambda I am picking up from here 1 minus lambda and 2 2 and 2 minus lambda that is how you form the A minus lambda I, I is a unit matrix remember 1 0 0 1 in this particular case. So, from here this is a determinant. So, I can form an expression for lambda this is equal to zero.

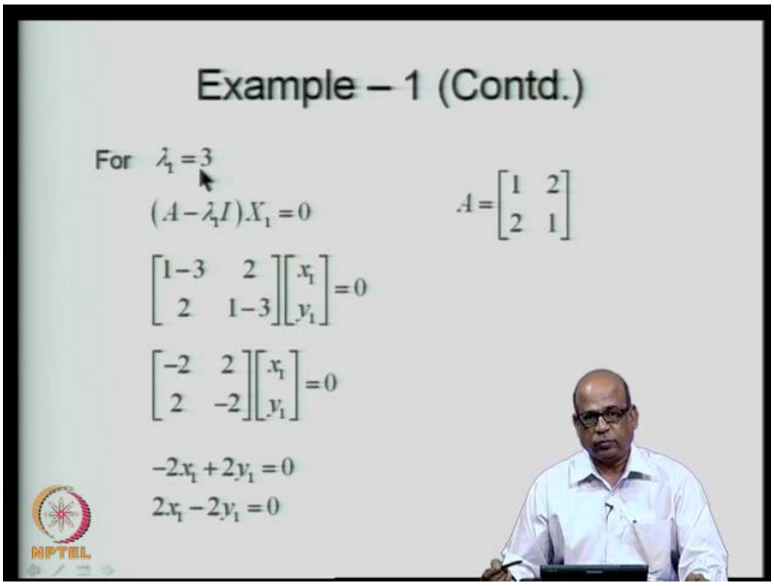
(Refer Slide Time: 17:31)



So, I will get this as 1 minus lambda into 1 minus lambda minus four is equal to zero. Are you get the expression lambda square minus 2 lambda minus three is equal to zero from here and when you solve this is a quadratic. So, when you solve this you get to solutions lambda is equal to three or lambda is equal to minus one, both of these satisfy this event expression and these are the lambda the eigenvalues for the matrix A. So, the eigenvalues are 3 and -1 for the matrix A. You had a 2 by 2 matrix therefore, maximum of two eigenvalues are possible and that these are the two values that you obtain.

In some situations when you have 2 by 2 matrix for example, you may get both of them equal to each other in which case there is only one eigen value. So, there are maximum of n distinct eigenvalues possible for a square matrix of the size n by n. And always you obtain the eigenvalues first and use the eigenvalues to obtain the eigenvectors. So, we will now obtain the eigenvectors once you have found the eigenvalues, remember corresponding to each of the eigenvectors there is a eigen value or they come in pairs eigenvalues and the eigenvectors they come in pairs.

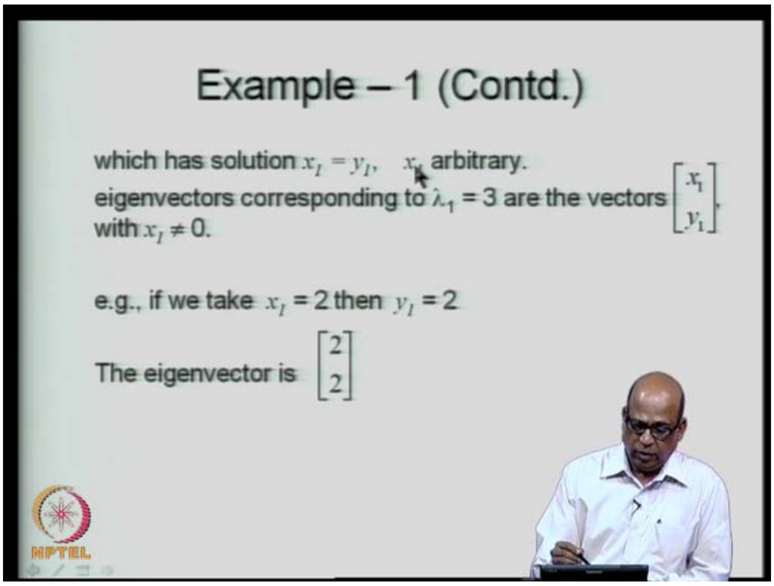
(Refer Slide Time: 19:16)



So, you have through 2 lambda values and corresponding to each of these lambda values we must find one eigen vector. The eigenvectors are obtained by A minus lambda I into X is equal to zero these are matrices A minus lambda I into X is equal to zero. And I specify the lambda value to obtain the value of X. So, I had 2 values of lambda. So, I will first start with lambda 1 is equal to 3. And substitute lambda 1 is equal to 3 in A minus lambda 1 I into X 1 is equal to zero, I is the unit vector and therefore, from this A here which is 1 2 2 1 I will get 1 minus 3 3 2 1 minus 3 because lambda is 3, and putting lambda 1 into I, I is the unit matrix, and I will write this X 1 as x 1 y 1 this is the eigen vector now.

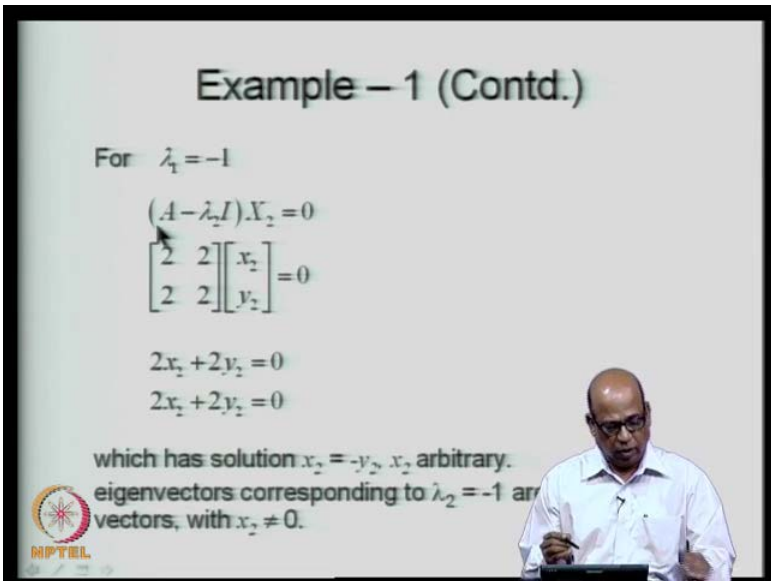
So, I will get this as minus 2 2 2 minus 2 and this leads to 2 equations minus 2 x 1 plus 2 y 1 is equal to 0, 2 x 1 minus 2 y 1 is equal to 0. The equations that you get out of solution of these are identical for example, you can multiply with this minus minus 1 so, you will get 2 x 1 minus 2 y 1 which is identical to this. So, from this I can write x 1 is equal to minus y 1.

(Refer Slide Time: 20:31)



You can write x 1 is equal to y 1 here and for any chosen x 1 arbitrary you can choose x 1 arbitrary you can get the eigen vector x 1 y 1. Your eigen vector was x 1 y 1 and the solution is x 1 is equal to y 1. So, any chosen x 1 you will get the corresponding y 1 as x 1 is equal to y 1 as y 1 is equal to x 1 and therefore, you will get the eigenvectors. So, the eigenvectors corresponding to the eigen value lambda 1 is equal to 3 are the vectors x 1 y 1 with x 1 not equal to 0 because that is what we have specified earlier. For example if you take x 1 is equal to 2 as I said any arbitrary value of x 1 satisfying x 1 is equal to y 1 we will define your eigenvectors. So, if you take x 1 is equal to 2 y 1 is equal to 2 then the eigen vector is 2 comma 2 corresponding to the eigen value lambda 1 is equal to 3.

(Refer Slide Time: 21:50)

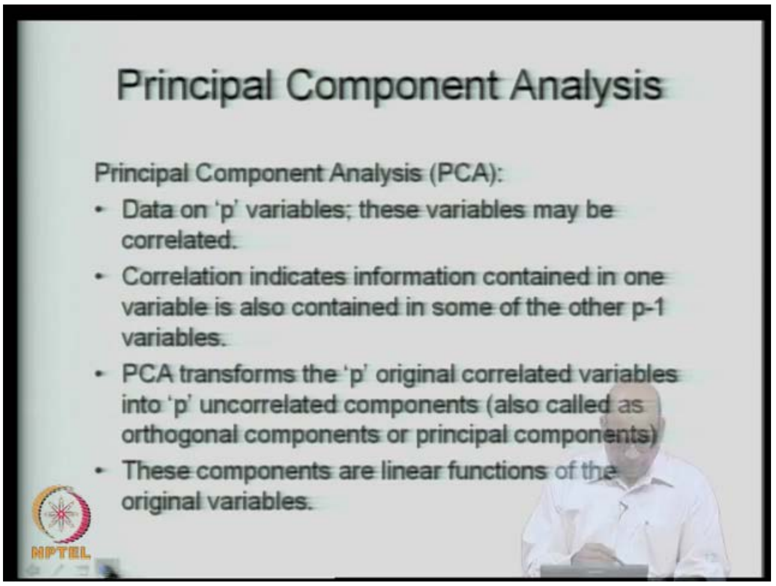


So, like this you can form several eigenvectors now we will take the other lambda 1 which is lambda 1 is equal to minus 1. If you put lambda 1 is equal to minus 1 I will get 2 and 2 remember your A is this matrix is 1 2 2 1. So, I am saying 1 minus 1 that is minus of minus 1 which will be plus 2 and here 1 minus minus 1 that will be again 2. So, this is how you get 2 2 2 2 x 2 y 2.This x 2 which is a vector I will write it as x 2 y 2 and you will get the equation 2 x 2 plus 2 y 2 is equal to 0, which is identical to the second one this has a solution x 2 is equal to minus y 2.

So, you can choose x 2 arbitrary and then you get y 2 and you will get the vector here x 2 minus y 2. So, the eigenvectors corresponding to lambda 2 is equal to minus 1 are these are the vectors. You can choose any value of x 2 and set x 2 is equal to minus y 2 and that is how you get with x 2 not equal to 0. So, this is how you obtain the eigenvectors and the eigenvalues. As I mentioned given any square matrix if you want to determine the eigenvectors first you find the eigenvalues by taking determinant A minus lambda I that determinant equal to 0.

A is a square matrix and I is a unit matrix. You will get at most n distinct lambda values for A square matrix of size n by n use these n distinct lambda values to obtain the corresponding eigenvectors how do you get this that will be A X is equal to lambda into X, and X is the eigen vector. So, corresponding to each of the lambda values you get one eigen vector and the eigenvectors and the eigenvalues come in pairs. So, we will use this method of obtaining the eigenvectors and the eigenvalues then we go to the principal component analysis. So, let us see what we do in the principal component analysis as I mentioned you have several variables and on each of the variables you have the data available to you.

(Refer Slide Time: 24:22)

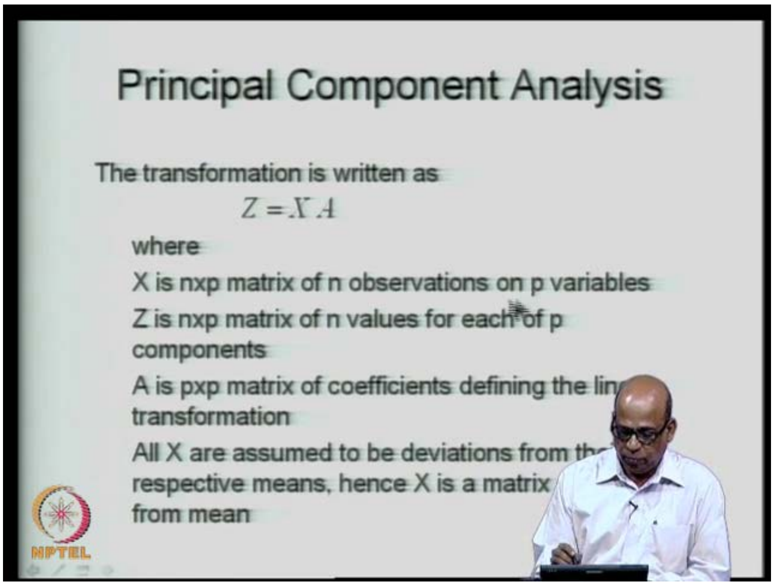


Let us say that you have a as I mentioned earlier you have variable such as rainfall, soil moisture, temperature, humidity and so on. And so, you have several such variables and your dependent variable may be runoff at a particular location and you have the data observed data on each of this variables. So, you have a data on p variables and these some among these may be correlated. Let us say rainfall may be related with temperature in some sense or the soil moisture is related with rainfall, and evapotranspiration is related with both soil moisture as well as temperature.

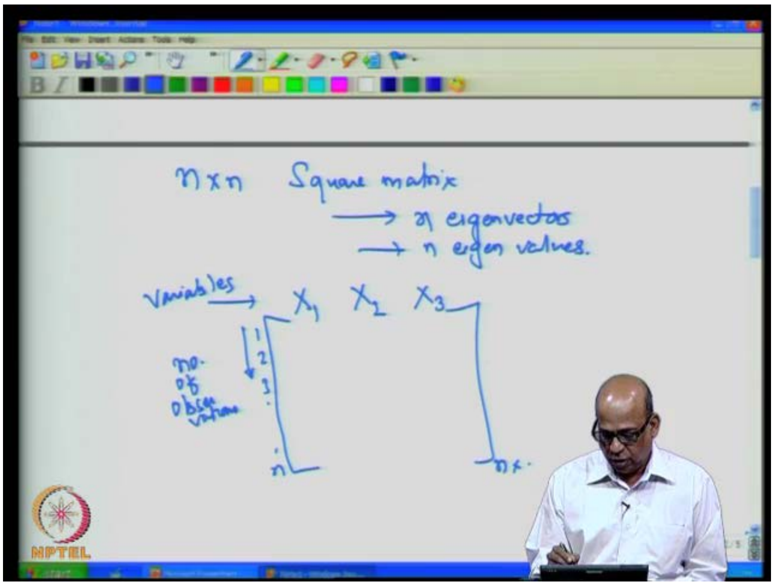
So, there may be significant correlations among the several variables. Now, what do we mean by correlation if you recall from your earlier lectures on this course correlation actually means that there is a common information, there is a information contained in one variable is also contained in some other variable, and this information we have to filter out. So, that we do not repeat what is contained in several variables. So, the PCA or the principal component analysis in some of the text books you will also see this is called as principal components analysis, but there are.

So, the PCA actually transforms the p original correlated variables into p uncorrelated components these are called as a principal components or they are also called as orthogonal components in as much they are orthogonal to each other. They are vectors and therefore, they are all orthogonal to each other. Now these components are actually linear functions of the original variables. So, you do a linear transformation of the original variables into some orthogonal axis. So, you are actually transforming the original data into some orthogonal components and these are linear transformations.

(Refer Slide Time: 26:47)



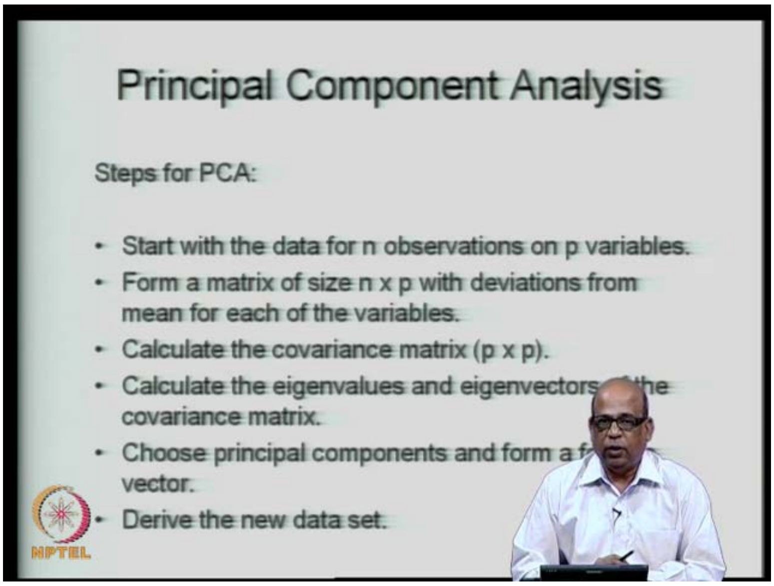
(Refer Slide Time: 27:22)



Let us say we write the transformation as Z is equal to X into A. Now X is your n by p matrix of n observations on p variables. We will not lose sight of what we are doing here. So, we will just go through our hydrology example let us say I have three variables one is rainfall another is runoff, and another is let us say soil moisture you have three such variables on each of these variables. So, these are the variables. On each of these variables you have number of observations let us say you have 1 2 3 etcetera you have n number of observations if you have 50 years of monthly data n will be six hundred. So, this is p and this is n. So, the X vector which is X matrix which is a matrix of the observed data will have a size of n by p and that is what I am writing here.

So, this is X is n by p matrix of n observations on p variables, and now Z is a n by p matrix of n values for each of the p components, now A is our matrix of components. First let us look at A you had p variables let us say you had three variables, then p is equal to three now A is a 3 by 3 matrix of coefficients defining the linear transformation this is in fact, the principal components. And Z is the transformed data. So, from the original data on the original data you apply the principal components to get the transformed data and that is a Z matrix. Now when we are doing the analysis we take the X matrix as A deviation matrix; that means, from the observed data you deduct the mean and many times you divide it by standard deviation also to form the X matrix, and A is the matrix of p coefficients p by p coefficients which we will see how to obtain this now.

(Refer Slide Time: 30:01)



And Z will give you the transformed data. So, this is what we essentially do in principal component analysis. So, how do we do the principal component analysis we have n observations on p variables we start with that data. In fact, you relate this with what we discussed in the multiple linear regression you have observations corresponding to each of the variables, and there is a dependent variable there are large number of independent variables. Now, from this matrix of n observations on p variables you form a matrix of size n by p with deviations from mean for each of the variables. Now typically we do this on the independent variables. So, you pick up all the independent variables and then deduct their mean and formulate a matrix of size n by p then this is n by p matrix. So, we can obtain the covariance matrix there are p variables.

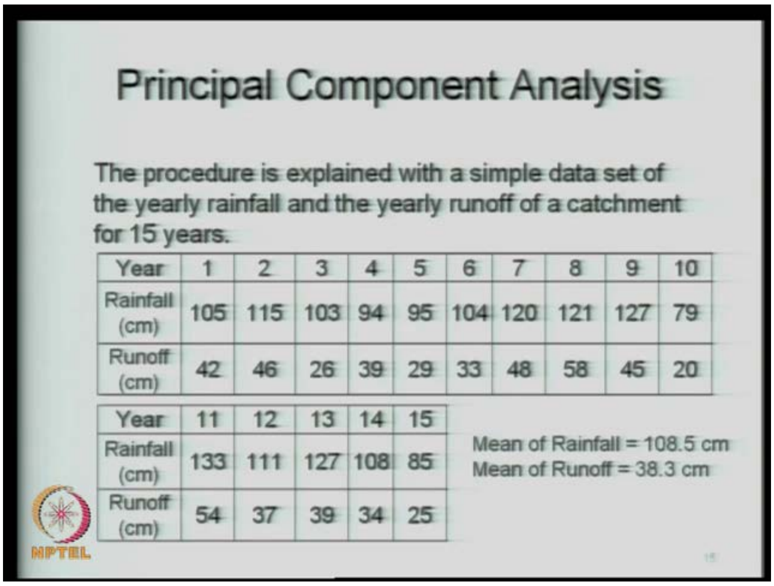
So, we also obtain a p by p covariance matrix that is correlation of variable x 1 with x 2 x 1 with x 3 and so on. x 1 with x 3 similarly x p with x 1 x p with x 2 and so on. So, you get a covariance matrix of p by p now this is a square matrix. So, we can obtain the eigenvectors and eigenvalues for this square matrix. So, we obtain the eigenvectors and eigenvalues for this covariance matrix. This is a p by p matrix therefore; we can get p eigenvectors and at most p distinct eigenvalues, from these p eigenvectors we can choose some eigenvectors for further analysis using regression depending on how important they are, in terms of the information contained on the dependent variable we will come to that slightly later.

But from these eigenvectors we pick up some of the eigenvectors we which we call it as principal components. And use this expression Z is equal to X into A to obtain the transformed data Z. So, this is what we do I will just quickly go through it we start with the data and then the data. we convert it into n by p matrix with deviations from the mean X minus X bar and in some cases we take X minus X bar by sigma or X minus X bar by S, where you are standardizing were you standardizing necessary because, you may have several variables all with different units. For example I may be dealing with different variables as rainfall in millimeters, area in hectares, then soil moisture in millimeters per centimeter or percentage and so on.

So, these units are all different and therefore, it is advantages to standardize them by standardizing I mean X minus X bar by S to deduct the mean and divide by the standard deviation. Then we get corresponding to each of the variables you take the covariance which respect to the other variables. Remember covariance you take with one variable with the other variable X into X into Y the covariance of X with Y covariance of X with Z and so on, covariance of X with itself is a variance. So, we get this covariance matrix of size p by p because it is square matrix.

We take the eigenvectors and eigenvalues of these and then define the principal components corresponding to the eigenvectors and the eigenvalues and obtain the new data set. So, from the original observed data set of n by p now we obtain another data set still of the same size n by p. And the new data set is in terms of the principal components and they are all uncorrelated, they are all the different principal components are orthogonal to each other. So, this is how we transform the original data set into a new data set. Where you are dealing with uncorrelated variables and you are dealing with orthogonal vectors.

(Refer Slide Time: 34:36)

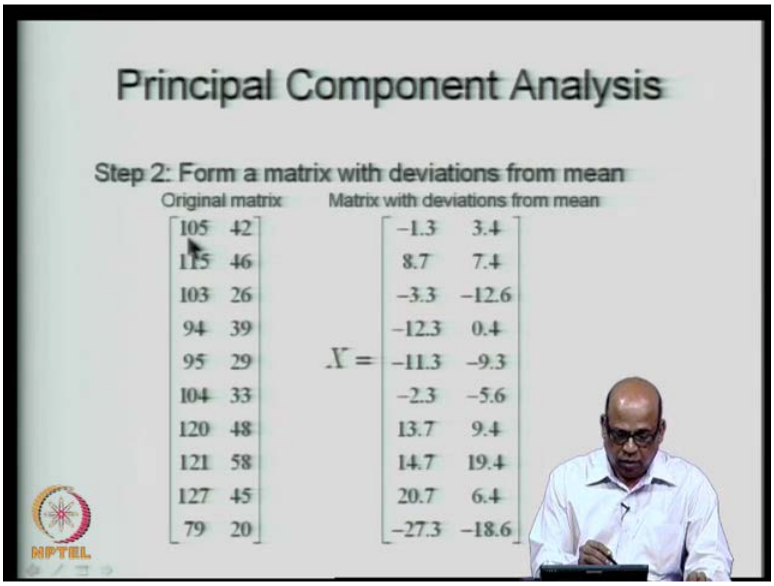


Year	1	2	3	4	5	6	7	8	9	10
Rainfall (cm)	105	115	103	94	95	104	120	121	127	79
Runoff (cm)	42	46	26	39	29	33	48	58	45	20

Year	11	12	13	14	15
Rainfall (cm)	133	111	127	108	85
Runoff (cm)	54	37	39	34	25

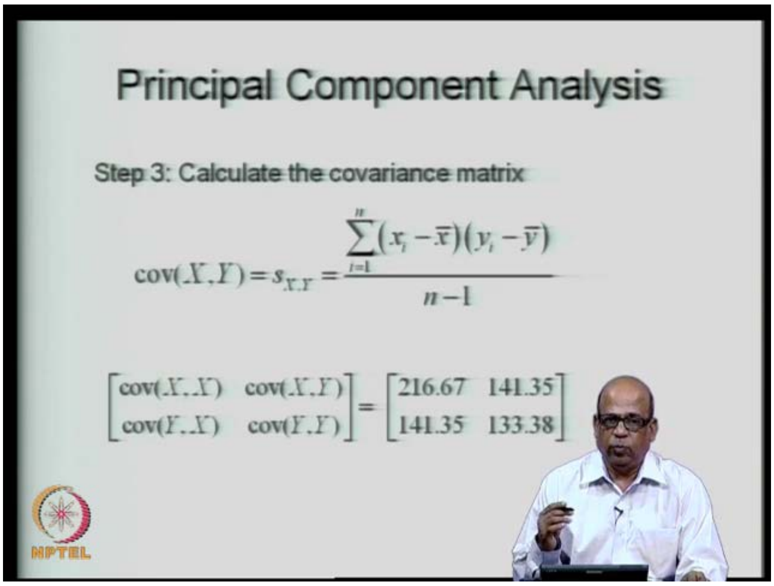
We will take a simple example now to demonstrate this procedure we are not right now worried about which is dependent, and which is independent variable for this particular exercise we are taking two variables rainfall and runoff. So, your p is equal to two and you have data for 15 years. So, these are fifteen values now.

(Refer Slide Time: 35:50)



The mean of rainfall as you obtain from here is 108.5 centimeters and the mean of runoff is 38.3 centimeters. So, with this now we will see how we obtain the principal components. First you look at this data 105 115 103 94 etcetera it goes up to 85 similarly for runoff it starts with 42 46 and So on. So, this is the original matrix of data. p is two this is rainfall this is runoff and n is fifteen, there are fifteen observations each available for the two variables. Now I form a matrix X by taking by deducting mean of rainfall from this column and the mean of runoff from this column, mean of rainfall is 108.5 mean of runoff is 38.3. So, 105 minus 108.5 I will round it off minus 1.3 similarly 3.4 and. So on.

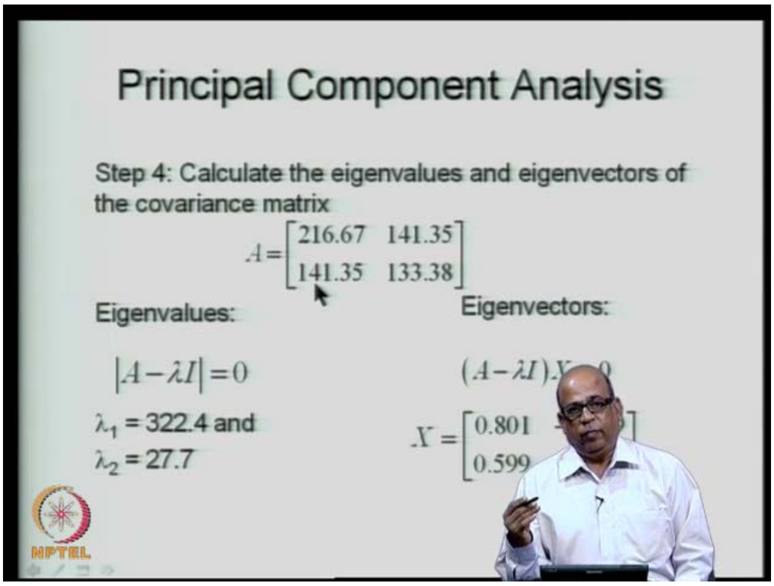
(Refer Slide Time: 37:02)



So, this is how I form the matrix X now X is simply X minus X bar that is here I am taking X minus X bar, where X is the original data and X bar is the mean both for your rainfall as well as your runoff, and that is how you obtain your values 1.3 3.4 and So on. So, once you formulate the matrix X we will also calculate the covariance matrix there are two variables X Y and X Y here. So, let us say if I write (X, Y) recall from your earlier lectures I am doing with random variables here. So, let us write capitals X and Y.

So, I will take a covariance and the covariance matrix is formed by covariance of (X, X) here and (X, Y) (Y, X) and (Y, Y). And the covariance between two variables is obtained as sigma X i minus X bar Y i minus Y bar divided by n minus 1, where n is the number of data fifteen in this particular case. And X i is from your table here this is X i this is Y i like this it goes. So, that is how you calculate the covariance matrix and these are the values as you obtain from them you can do it as a exercise you will get covariance is 216.67 141.35, and X Y is same as Y X and you will get this covariance matrix this is a square matrix. So, let us obtain the eigenvectors and eigenvalues corresponding to this this matrix now covariance matrix.

(Refer Slide Time: 38:34)



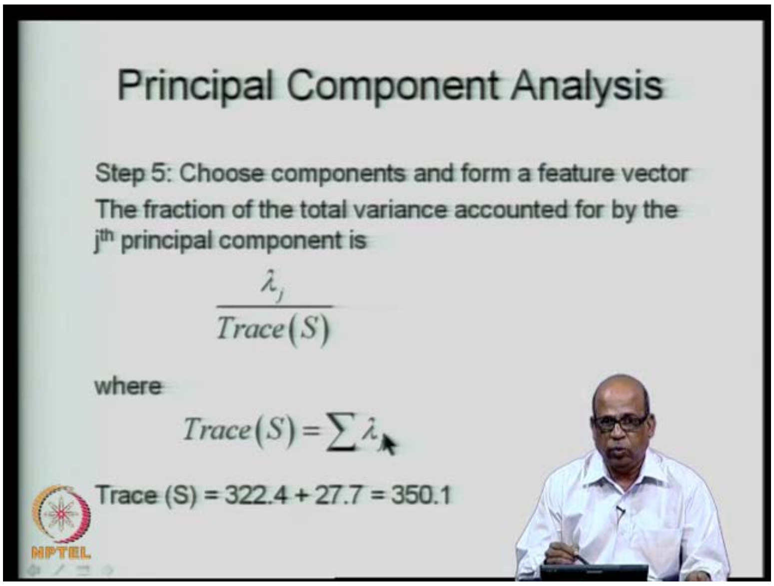
So, how do I get the eigenvectors remember whenever you have a square matrix to get the eigenvectors first you obtain the eigenvalues. So, this is the A matrix let us say which is a covariance matrix as we obtain just now we use determinant a minus lambda I is equal to zero. So, A minus lambda into this is unit matrix is equal to zero you will get lambda 1 is equal to 322.4 and lambda 2 is equal to 27.7. One of the lambda values is significantly higher compared to the other lambda value, we will see what is a significance of this, what is a implication of this then we do the regression using the principal components. We use these lambda values and like I did in the example corresponding to each of the lambda values you get a eigen vector.

So, I will get the matrix of eigenvectors 0.801 0.599 is 1 eigen vector and minus 0.599 and 0.801 is another eigen vector remember these eigenvectors are unit vectors. In the sense that let us take the distance of the eigen vector indicated by the eigen vector that is this is one eigen vector. So, I will take square root of 0.801 square plus 0.599 square this is equal to 1. So, generally any program if you use mat lab etcetera and ask for eigenvectors it returns unit vectors unit eigenvectors, and by unit you recall that we mean the distance given by the vector is one similarly you can see that the distance given by this is also equal to one. So, starting with your original data first you transform the original data into a data consisting of deviations from the mean, and then you also formulated the covariance matrix and this is the covariance matrix here. Covariance matrix is a square matrix of size p by p. p is equal to two here because you are dealing with two variables. Once you get the covariance matrix you get the eigenvalues of covariance matrix, because it is a size 2 by 2 you will get two eigenvalues in general. So, lambda 1 is equal to 322.4 and lambda 2 is equal to 27.7.

You obtain the eigenvalues go to the eigenvectors equation a minus lambda X is equal to zero from this you get X because lambda is given. So, first you put lambda 1 is equal to 322.4 you will get one eigen vector, and lambda 2 is equal to 27.7 you get and you get the other eigen vector and you form the matrix of eigenvectors. And these eigenvectors are unit vectors. Now I just mentioned about the relative magnitudes of the eigenvalues one is 322.4 another is 27.7. These have an implication on how we choose the principal components how many of the principal components.

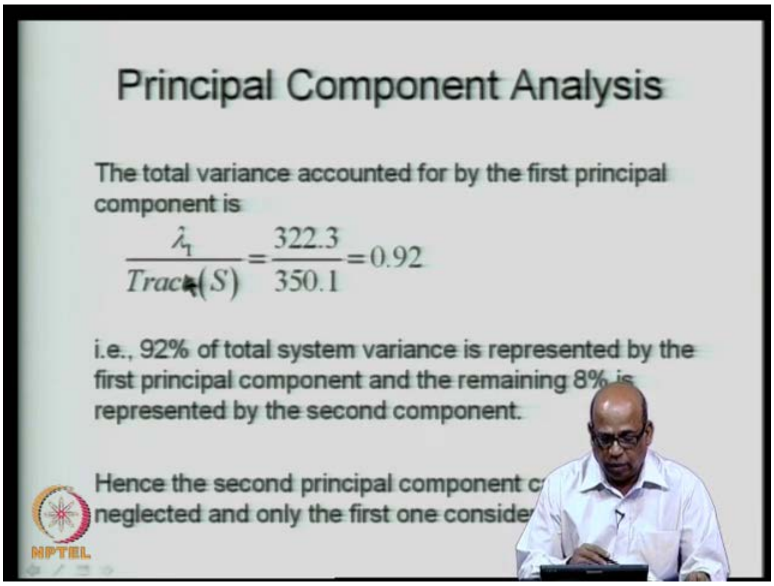
We choose the eigenvectors corresponding to each of the lambda values they explain the original variance to certain degree, which is related to the eigenvalues what do I mean by explain that is. So, much of variance has been contributed by this particular eigen vector is what I mean by explaining the variance. Let us say you are talking about runoff as a function of rainfall, and then you obtain as a function of rainfall, and another variable soil moisture let us say. And you obtain the eigenvectors corresponding to each of these that is, you get one eigenvectors the first eigen vector may explain 95 percent of the variance in the runoff whereas, the second one may explain only 5 percent which means you can afford to neglect ignore the second one.

(Refer Slide Time: 43:26)



And therefore, you look at how much of the variance in the dependent variable is. In fact, explained by each of these different eigenvectors and how do we do that we compute or calculate the trace of S, here trace of S is defined by simply summation of lambda js. So, you got the eigenvalues. So, add all the eigenvalues and that will define your trace of S. And the fraction of the total variance accounted for by the jth principal component is lambda j by trace j simply you get lambda j by trace j, that will give you the amount of variance that is explained by that particular principal component. So, in this particular case let us say lambda 1 is you have 322.3 and this trace S comes out to be 350.1.

(Refer Slide Time: 44:20)



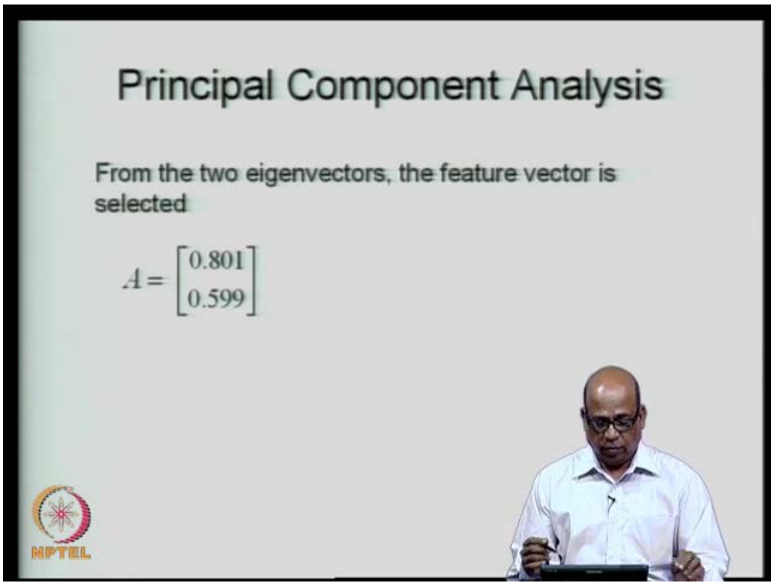
So, the first component first principal component lambda 1 explains 92 percent of the variance 0.92 fraction, which is 92 percent of the total variance is explained by the first principal component. And the remaining 8 percent is represented by the second component. Remember the first the identity of the original variables namely runoff rainfall, and runoff is gone now we are only dealing with the principal components. And principal components are a linear combination of both the variables. So, you cannot relate lambda 1 to be runoff lambda 2 to be or lambda 1 to be rainfall and lambda 2 to be runoff and so on.

The original identity is gone we are simply dealing with some transformed variables lambda 1 lambda 2 which are the eigenvalues and the eigenvectors. In fact, which define the principal components we are dealing with the principal components. And therefore, in this particular case you can afford to ignore the second principal component. If you are dealing with a regression relationship which will come to presently. So, this is how we carry out the principal component analysis first transform the data into a data consisting of deviations. So, that you are centering the data in some sense and if you have variables with different units you also standardize them, and then formulate the matrix matrix X consisting of the transformed data then you form the covariance matrix.

Corresponding to the covariance matrix you obtain the eigenvectors and eigenvalues. And look at how much of variance has been explained or how much of these or how much of the variance is contributed by each of the eigenvectors, and that is the objective of the principal component analysis. So, you found the principal components in terms of the eigenvectors, and look at how much of this information is coming from each of the principal components. What is the purpose of doing all this; we had a set of large number of variables. And then we want to fit a regression relationship between a dependent variable and a number of independent variables.

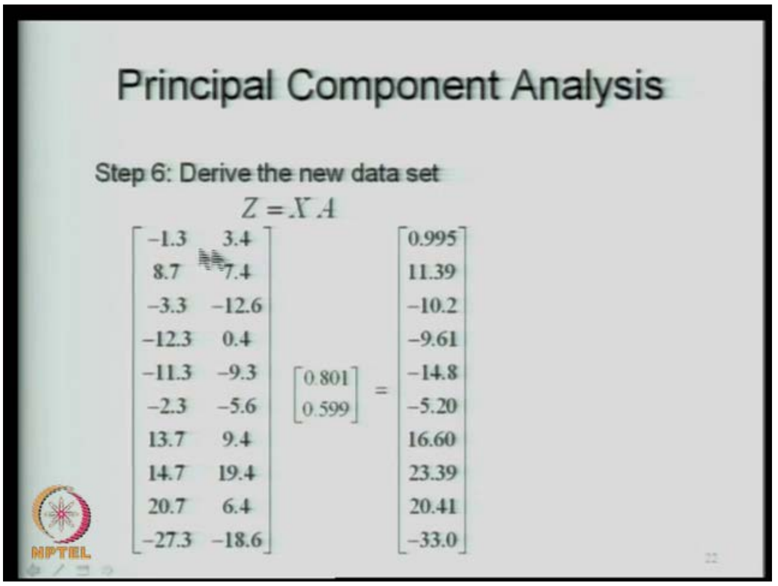
And we carry out the principal component analysis on the independent variables x 1 x 2 x 3 etcetera up to x p, there are p independent variables, we carry out the carry out the principal component analysis. We know let us say you had p variables you get the p principal components or p eigenvectors. And we know each of these eigenvectors are the principal components their contribution to the variance in the dependent variable. And we choose only those particular components only those principal components, which are contributing significantly to the variance in the dependent variable, and that is what we do in regression using the principal components.

(Refer Slide Time: 48:22)



So, we know now how to carry out the principal components we use this principal component analysis to do the regression. So, we had the original regression from the original regression, we are going on to regression with the principal components let us see how we do this. So, to continue the example now as I mentioned we said 92 percent of the variance is explained by the first component. So, we will neglect for the time being we will neglect the second component. So, we will take only the first eigen vector and call it as the feature vector. So, we call it as the feature vector we are choosing only the first one you can choose both of them as feature vectors. So, feature vectors are those eigenvectors which are chosen for the further analysis. So, this feature vector now is 0.801 0.599 corresponding to your first eigen vector.

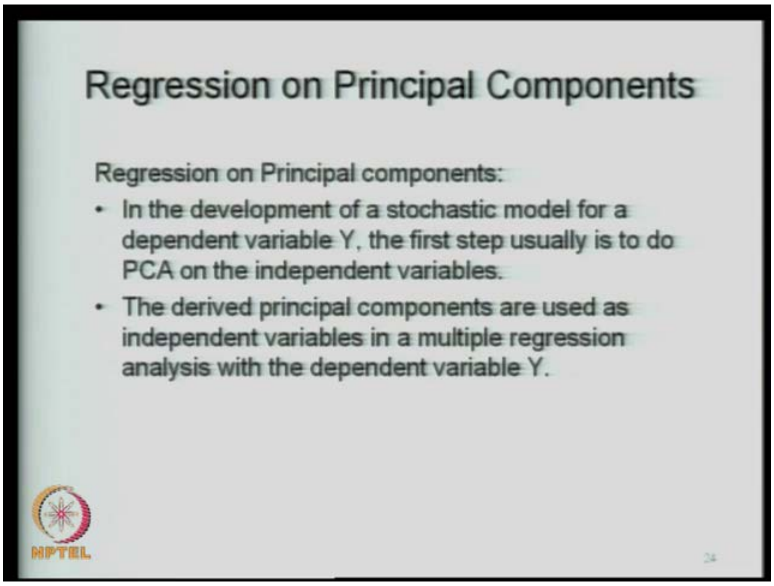
(Refer Slide Time: 49:12)



Then with this now we obtain the new data set this was your original data set minus 1 .3 3.4 that is by original, I mean these are the data sets that you use for the analysis this is X and this is your A. These are the data sets formed with the deviations from the mean, and this is A and you will get Z as the new data set. What is happened now the original p by n matrix has been converted into 1 by n matrix. So, this is the transformed data. So, if you had ten variables you would have had ten by n matrix. And if you choose only one eigen vector then you would get you will still get 1 by n, which means the remaining nine you would have discarded in the sense that they are not contributing.

So, much to the variance or the first component alone is able to explain most of the variance, or the first component alone is contributing to most of the information contained in the data set therefore, you discard the other principal components and then deal with only this data set. Let us say we did not ignore the second one, and still we want to use the second eigen vector, also in which case we can choose this is 2 by n matrix and n is fifteen year, I have taken ten values here. So, n is ten here and then this is the 2 by 2 eigen matrix this is one eigen vector this is another eigen vector and. So, you will get 2 by 10.

(Refer Slide Time: 51:22)



So, this is the transformed data. So, like this from the original data you use the eigenvectors to get the transformed data. Then we come to the question of regression using principal components. So, let us quickly recall what we did in this example now in this example what I have done is I have chosen only the ten values here and then carried out the exercise I gone up to 79 and 20. So, remember I have only used ten values here. So, the same thing can be done with the remaining with the complete data set fifteen values. So, these are ten values and then I get a transformed data of remaining ten values.

So, from the two independent variables let us say that your independent variables where rainfall and runoff for some purpose from the two independent variables. You have got one data set now which are all principal components these are called as the principal components. This is a eigen vector and in fact, this is called as a principal component which has a size of one by ten. So, this is a principal component that we are talking about. We can use this information because, we know now that this principal component contributes to 92 percent of the variance in the data set. So, we can discard the other and then simply use this data set.

In the multiple linear regression expression we will come back to that topic now in a multiple linear regression you had several variables x 1 x 2 x 3 etcetera there are p variables, you get p principal components. And p principal components you know the contribution of each these principal components to the variance of the dependent variable

Y. Let me quickly tell what I mean by that let us say you have Y is equal to you want to write it as a function of x 1 x 2 etcetera x p p variables. And therefore, you get p principal components and each of these p principal components the contribution to the variance in Y, or the variance in this particular data set can be obtained like we did in the example here we got 92 percent.

So, each of the principal components you can obtain what is the contribution to the variance then depending on that, you can choose q of them less than p to be included in the regression equation that is why I choose only 2. If two of them can together explain about 95 percent of the variance in why and so on. And then using only those two we fit the regression relationship not on the original data, but now on the principal components. So, we have obtained the principal components now you regress the dependent variable y with respect to the principal components and that is what we do in the next lecture.

So, we will see how we take this principal component analysis forward and then apply them in regression where, we are relating the dependent variable not with respect to the original variables, but with respect to the principal components. And remember principal components are a linear combination of the original variables themselves and therefore, individual principal component does not indicate any particular physical variable that we considered earlier. For example principal component number one does not indicate rainfall, it is some linear combination of rainfall and runoff water together in that particular example that I talked about.

So, the regression that we developed now Y as a function of principal component number one principal component number two and so on. We may have chosen q number of principal components with q being less than equal to p we will be a new regression relationship now. And what is the advantage that this has this is now dealing with all uncorrelated variables, because all the principal components are uncorrelated with each other. And it has also compressed the original data set in the sense that we are not choosing all the principal components, but we are choosing some lesser number of principal components compared the original number of variables.

So, we will summarize now. So, we started with how to compute the eigenvectors and eigenvalues for a matrix. The eigenvectors are computed only for square matrices, eigenvectors are defined for square matrices and to determine the eigenvectors. You first determine eigenvalues, eigenvalues are determined by determinant A minus lambda I is equal to 0. Where A is the square matrix, and I is the unit matrix. And then once you get the lambda associated with each of the lambda values, you have an eigen vector you will get the eigen vector as A X is equal to lambda X, where X is the eigen vector. Then we went on to principal component analysis where we defined the principal components as eigenvectors.

In fact, of the covariance matrices covariance matrices of p variables. So, you have a size of p by p, and that is how we will get the eigenvectors these eigenvectors are in fact, the principal components. So, we also know how much of the variance is contributed by each of the eigenvectors depending on the contribution of to the variance, we choose a few of them few eigenvectors to use in regression. So, in the next lecture then we will start with regression using the principal components. So, we have arrived at the principal components starting with the original data we will use the principal components in the regression, and see what is the information that we can derive out of that, and what are the advantages arising there of thank you for your attention we will continue the discussion next time.