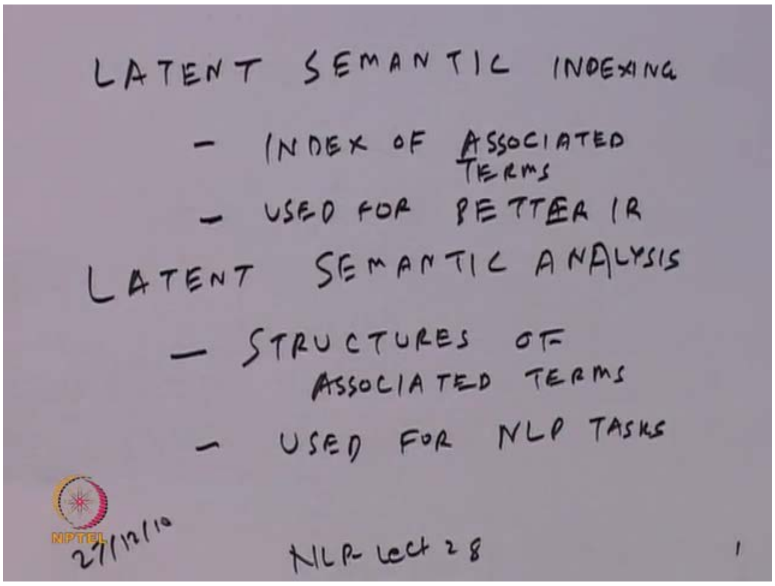
Natural Language Processing
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture - 28
PCA; SVD; Towards Latent Semantic Indexing (LSI)

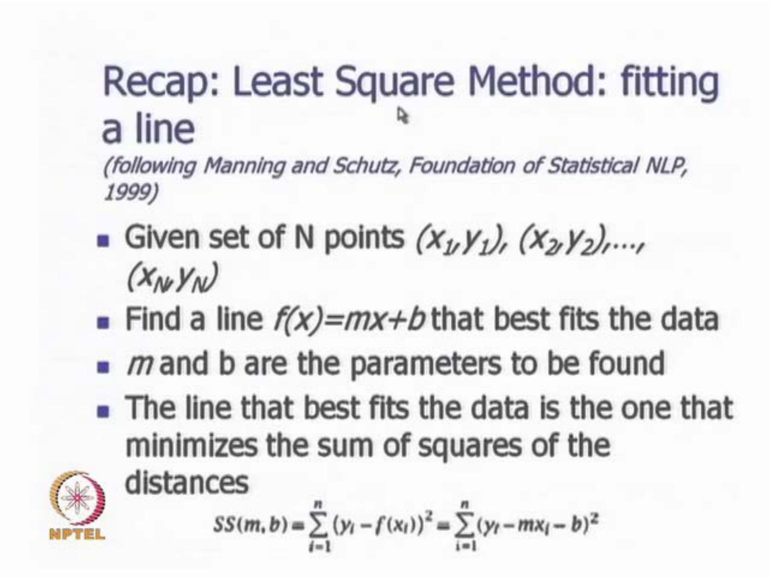
We continue discussing latent semantic analysis and indexing. So, we saw last time that latent semantic analysis, when apply to information retrieval refers to a list of index being built, which tries to capture the association between words. And when this world association is used for natural language processing tasks, then we have what is called latent semantic indexing, so it is better to put down these points.

(Refer Slide Time: 00:51)



So, latent semantic indexing, this is index of associated terms and latent semantic analysis, this is structures of associated terms and this is used for NLP tasks and here this indexive associated task is used for better IR. So, in this lecture, we will again look at principle component analysis, actually we started it last time, then singular value decomposition will be introduced and all this together final leading to our main topic which is latent semantic indexing.

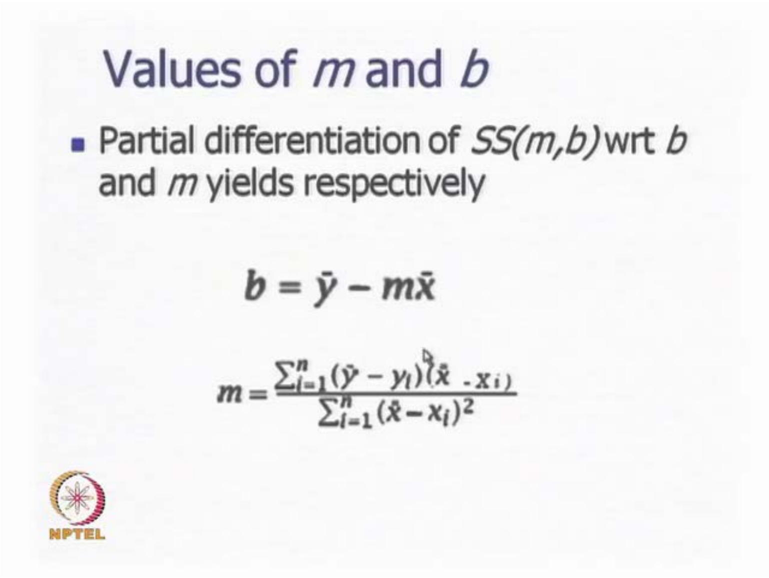
(Refer Slide Time: 02:02)



So, last time what did was that we did least square method of fitting a line and the example was chosen, so that we could go from a space of higher dimension to a lower dimension space. So, given this set of n points which are (x 1, x 2), (x 1, y 1), (x 2, y 2) up to (x n, y n), these points are defined by their 2 attributes x 1 and x and y. And from this we find a line f x equal to m x plus B that best fits the data, and m and B are the parameters to be found, so the line that best fits the data is the one that minimizes the sum of squares of the distances.

So, we take the distances of y i from if from f x i, f x i is the value that is written correspondent to particular x and the actual value is y, so y minus f x gives us the distance between the fitted line and the actual value. And the square of this distance eliminates any problem, it is sign and when we sum these square distances, we have a measure which is the measure of the fitness of the line for this data.

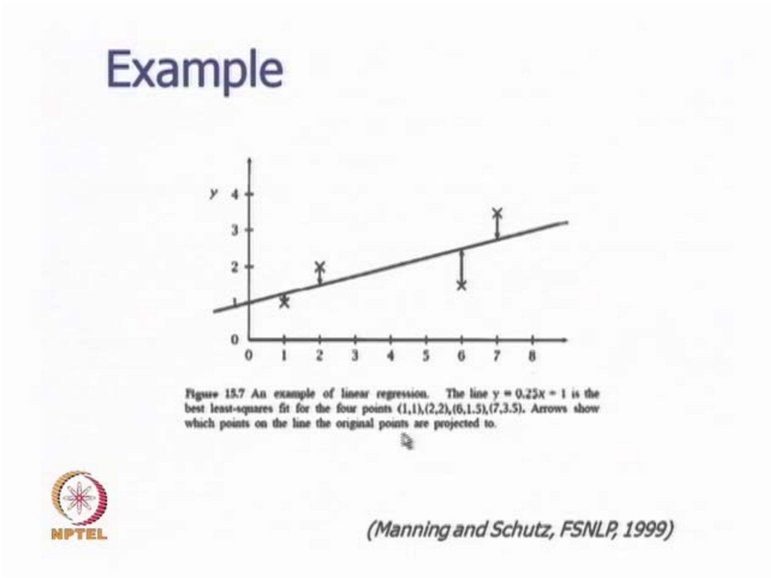
(Refer Slide Time: 03:31)



So, we have seen that we can partially differentiate this sum square expression with respect to m and b, and we get this expression for B which is mean of y minus m into x of y and m itself is found as a kind of covariance. Yes, covariance between the y attribute and x attribute divided by something like a variance of x, so one interesting question for you to think is why is it asymmetric with respect to x and y, why is it that m is found in terms of variance of x and not y.

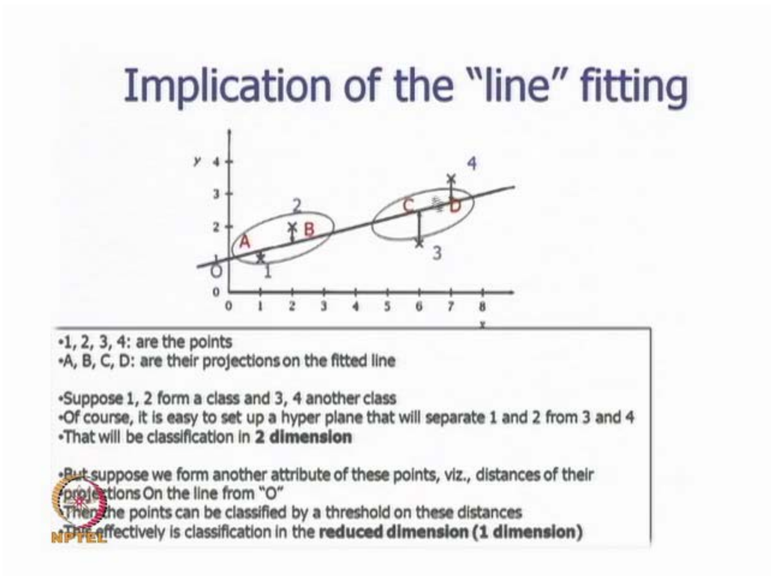
So, x is the controlling variable and that comes as the denominator in the form of a variance and this expression also is worth of wondering about because this is mean of the dependent attribute minus the slope of the line into the mean of the independent attribute. So, these expressions are worth reflecting on and have a developing an intuition about what they signify. So, this B and m they minimize the distance square distance between the line and the points and therefore, this line is the best fit according to this point of view.

(Refer Slide Time: 05:05)



So, there are these points here 1 1 2 2 6 1.5 7 and 3.5, this example is from Manning and Schutz foundation of statistical natural language processing published in 1999 by MIT press. So, these 4 points and the points are shown by crosses, now we have fitted a line for these points and this is the best line, where the slope is 0.25 1 by 4 and one is the intercept B on the y axis.

(Refer Slide Time: 05:55)

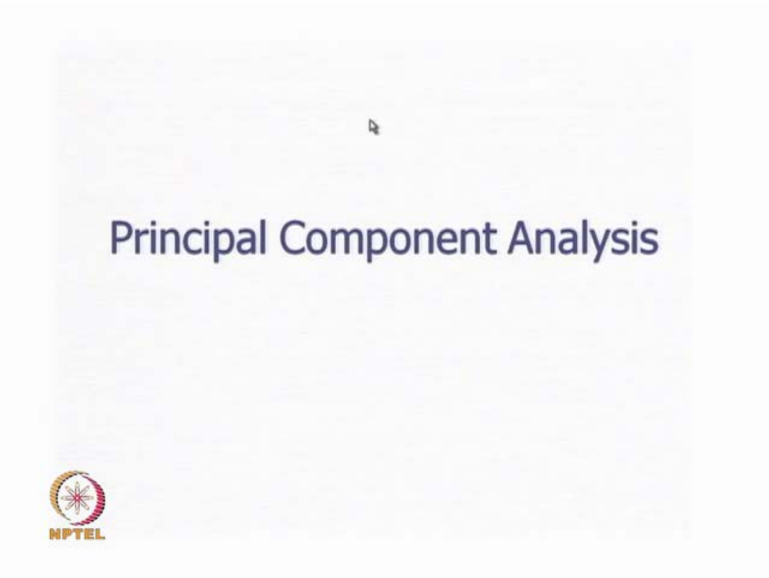


So, this is fine and many prediction problems who are fitting of line, but our interest is different what we are saying is that now assuming that, this 2 points C and D belong to a particular class and a and B belongs to another class. Then we can always draw a line this way, which is the hyper plane in 2 dimension this is a line which will separate C and D from A and B. So, this line can always you found out and the algorithm them exists for findings this, but now suppose we A classify these points in a reduce dimension, so the where to go about this would be for these points a and b, we get the values here.

We get the projected values on this line and for C and D we again get the projected values on this line, now you can see that separating the points through the projection is easy a task. So, we can fix a threshold value here and I classified is simply in terms of this threshold value it says that whenever the projection value is more than this threshold then the points belong to one class, and when this projected value is less than threshold belong to another class.

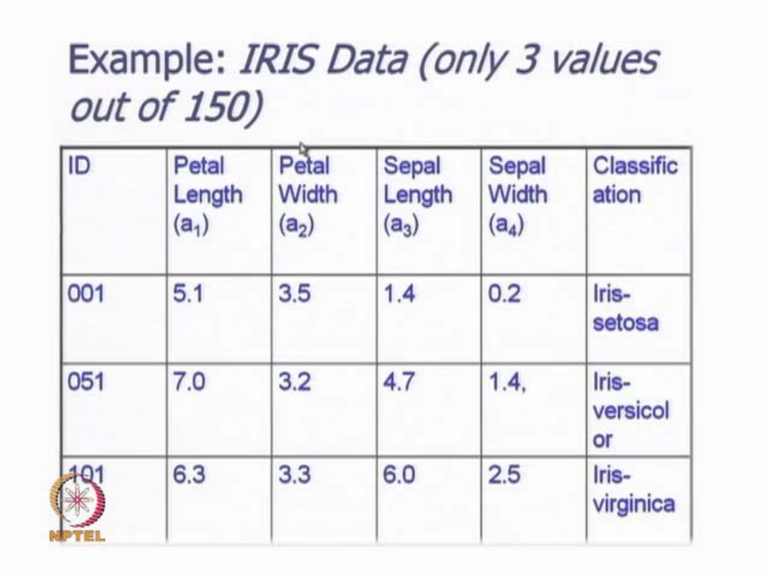
So, this classification according to ours is an easy a task then finding a line which separates C and D, from A and B get this a must simpler task, and what is happening is that we have reduced that dimension problem. So, the projection of A B C and D are now treated in terms of the distances from this special point O, and this distances are uni-dimension objects obtained from 2 dimensional representation of the points. And so we are solving the problem in reduce dimension, we have come down from a 2 dimension problem to a single dimension problem and this line is helping us to that. So, this is the whole point about dimension it reduction, and we have to take the problem into a different space with different parameters, so this point to was make last time and it will be to remember this point of view.

(Refer Slide Time: 08:21)



Now, we move on to principle component analysis and will do it slowly with some examples.

(Refer Slide Time: 08:28)



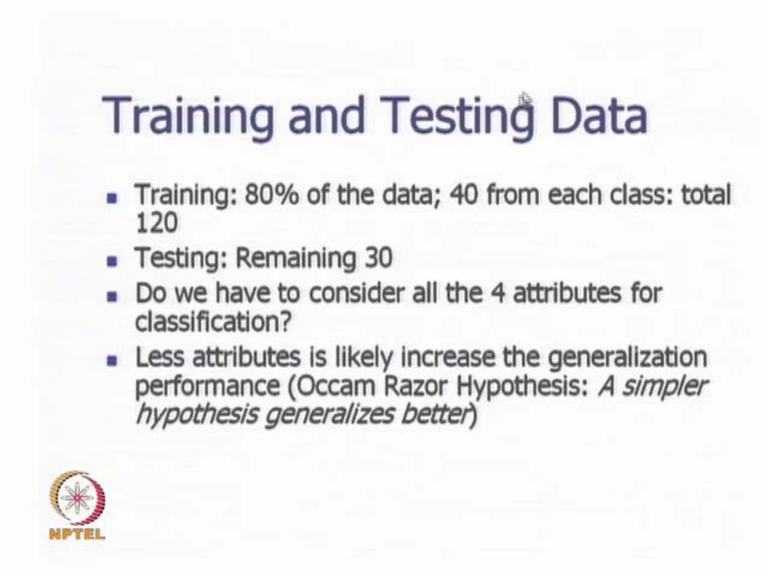
Example: *IRIS Data (only 3 values out of 150)*

ID	Petal Length (a_1)	Petal Width (a_2)	Sepal Length (a_3)	Sepal Width (a_4)	Classification
001	5.1	3.5	1.4	0.2	Iris-setosa
051	7.0	3.2	4.7	1.4,	Iris-versicolor
101	6.3	3.3	6.0	2.5	Iris-virginica

So, as was mention last time iris data is a very famous data for mission landing I will go to them, so there are 150 data points of this kind, these points are defined in terms of 4 attributes for flowers. So, petal length, petal width, sepal length, and sepal width, all the 4 attributes and the classification is in terms of this 3 class iris setosa, iris versicol and iris virginica.

So, this a 3 class problem, each class has 50 data points each point defined by 4 attributes, so now, the question that we ask is are all this attributes independent of each other, so that all of them are needed are is it that some attributes are highly correlated, may be petal length is correlated with sepal length. And then can we capture this correlation, create a function of petal length and sepal length and instead of 4 attributes were 3 attributes, there is a question, so then we have reduced the dimension of attributes by creating any function that is the whole idea.

(Refer Slide Time: 09:55)

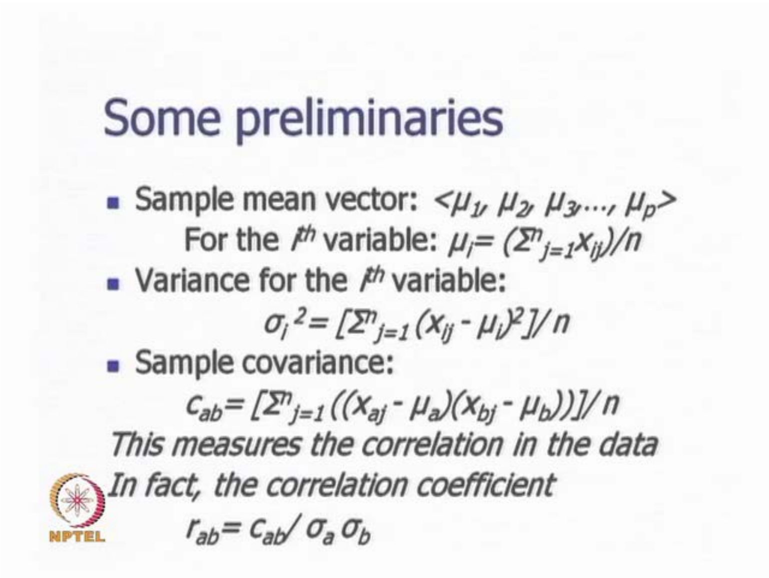


So, principal component analysis will help us here and the training and testing scenario is as follows 80 percent of the data is used for training. So, 40 from each class out of 50 and there are total 120 training example therefore, testing is on remaining 30 examples they form the 20 percent of data. So, as has been mention already do you have to consider all the 4 attributes for classification less attributes is likely to increase the generalization performance.

(Refer Slide Time: 10:29)

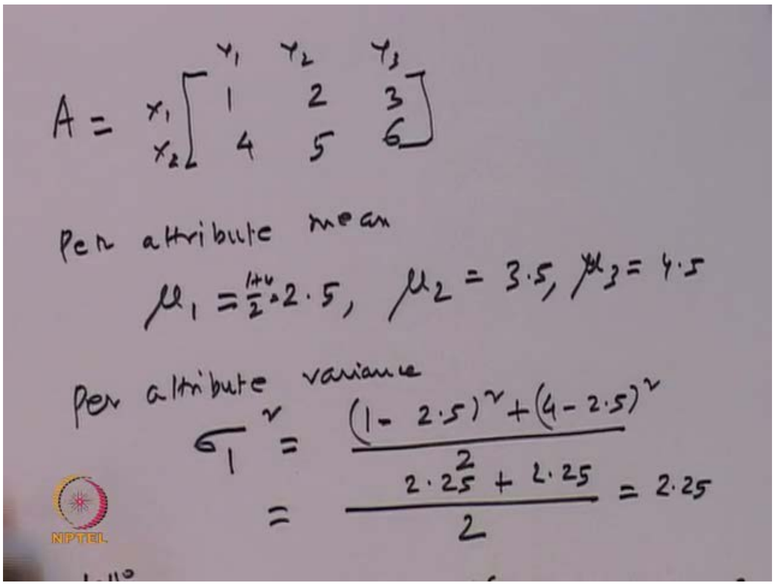
So, multivariate data typically is of this form we have this p attributes x 1 to x p, and there are n rows in this matrix corresponding to n data points, so this is a an example of multivariate data the structure of this.

(Refer Slide Time: 10:50)



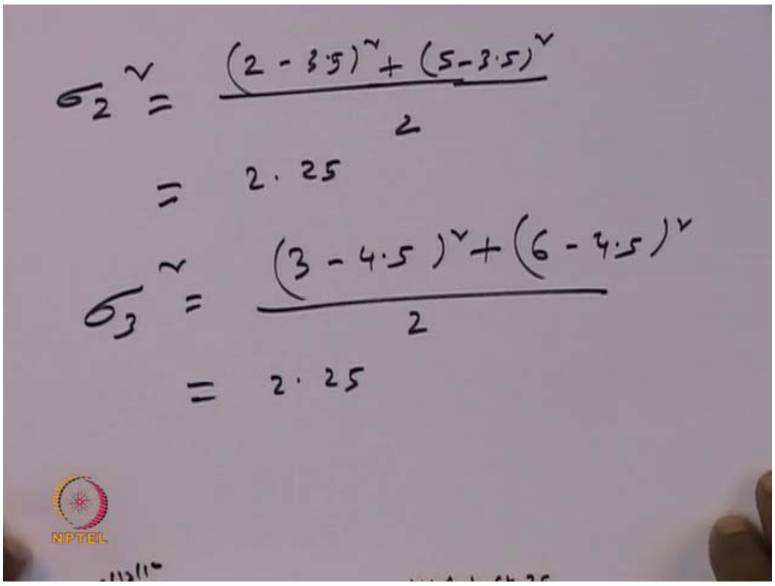
So, we do some preliminaries may be within example, so suppose we work with this matrix A which as 2 rows and 3 columns.

(Refer Slide Time: 10:59)



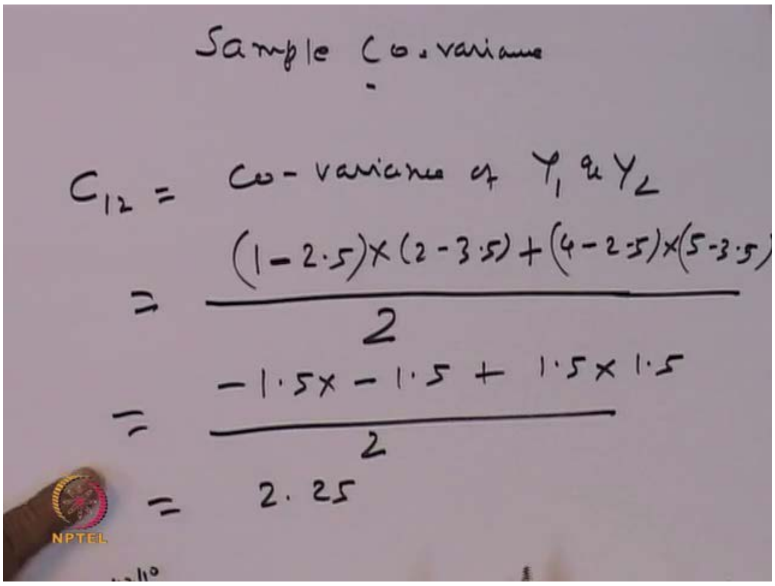
So, the values are lets a 1 2 3 and 4 5 6, this is a say x 1 x 2 y 1 y 2 y 3, so first thing to here is per attribute mean, so for y 1 we have mu one which is the mean of 4 and 1 that is 2.5 mu 2 is the mean of the attribute y 2 this is 3.5 and mu 3 is the attribute of 5 3, mean of 5 3 this is 4.5, this is the main. Now, we find out per attribute variance, so this mean it must be clear to you is nothing but the attribute values divided by the total number, so this actually is equal to 1 plus 4 by 2, which is 2.5. Now, per attribute variance, so for y 1 this is sigma 1 square, so we have to see the departure from the mean, now this is will be 1 minus 2.5 square plus 4 minus 2.5 square divided by 2. And that is equal to 1.5 square, which is 2.25 plus again this is 1.5 square which is 2.25 and this variance comes over to be equal to 2.25, so we found out the variance of attribute y 1.

(Refer Slide Time: 13:26)



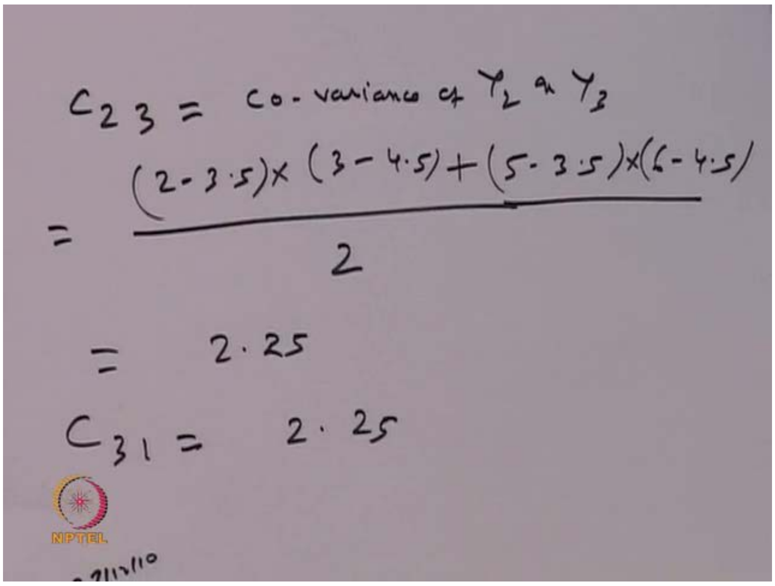
Similarly, sigma 2 square will be equal to 2 minus 3.5 square plus 5 minus 3.5 square by 2, this is also coming out to be 2.25 and sigma 3 square is 3 minus 4.5 square, so variance coming out to be same everywhere 6 minus 4.5 square by 2 which is equal to 2.25, so variance same everywhere.

(Refer Slide Time: 14:11)



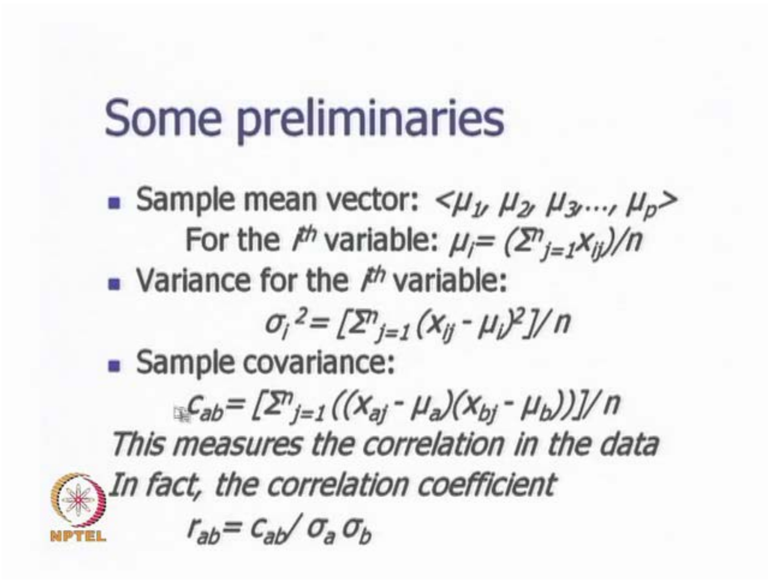
So, now, we also find out the sample covariance, so that mean we have find out C 1 2; that means, covariance of y 1 and y 2, so this will be nothing but it will be divided by 2 first of all. We have to take the difference from the mean of y 1 multiple it with difference of the value from y 2, so what a mean is you have to find out the sample covariance between y 1 and y 2. So, take the departure of one from the mean, take the departure of 2 from the mean multiply by the 2 here also departure from the mean and multiple. So, departure from mean will be 1 minus 2.5 into 2 minus 3.5 plus 4 minus 2.5 into 5 minus 3.5, so this comes at to be equal to minus 1.5 into minus 1.5 plus 1.5 into 1.5 by 2 which comes at to be equal to 0, so this again 2.25.

(Refer Slide Time: 15:49)



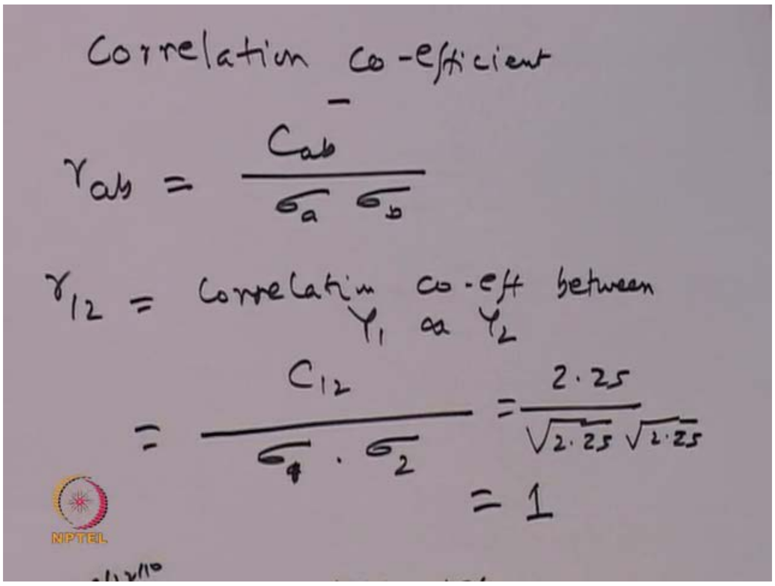
Then we have to find out, C 2 3 I think it will come out to be again 2.25, but let us check C 2 3 is covariance of y 2 and y 3 and this is 2 minus 3.5 into 3 minus 4.5 plus 5 minus 3.5 into 6 minus 4.5 again it is coming out to be 2.25, so we can possible say if you also say that C 3 1 also will be 2.25.

(Refer Slide Time: 16:52)



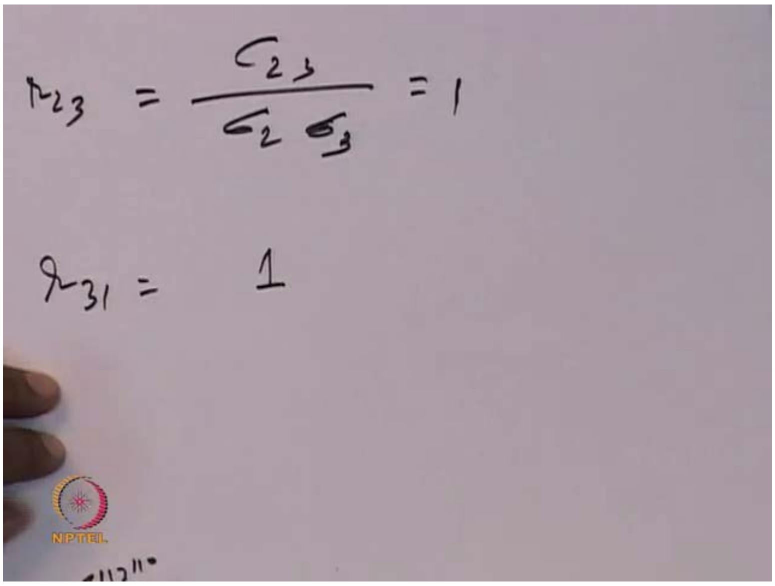
So, going to the slides now we have looking at the slides, way of the mean as way of found out the mean of each attribute, we have find out the variance of each attribute, we found out the sample covariance for pair wise attributes, and from this we get this very important quantity, namely the correlation coefficient.

(Refer Slide Time: 17:20)



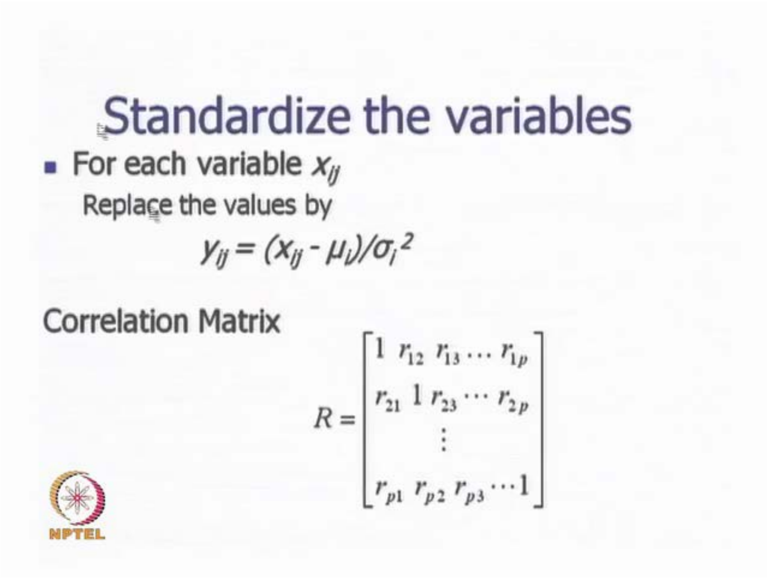
So, let us calculate the correlation coefficient for whatever we done so for, correlation coefficient for our matrix, which is this for this matrix now correlation coefficient r a b is defined has sample covariance divided by sigma a into sigma b. So, covariance divided by standard the variation of the 2 attributes, so now, what we have is lets first compute r 1 2, which is correlation coefficient between y 1 and y 2. So, C a b, we found to be equal to 2.25 everywhere C 1 2 by sigma 1 into sigma 2 and C 1 to was 2.25 sigma 1 is square root of 2.25, sigma 2 is also same 2.25, so there completely correlated.

(Refer Slide Time: 18:43)



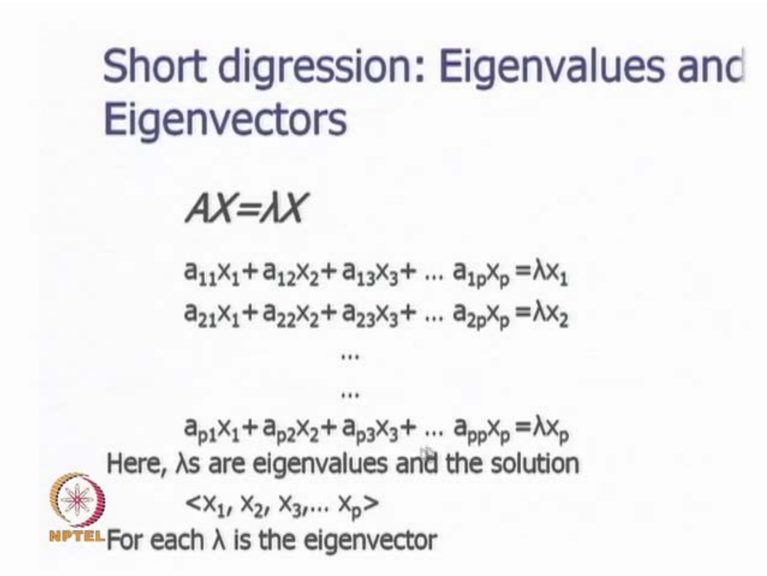
So, this is coming out to be equal to 1, so r 1 to is equal to 1, what about r 2 3? r 2 3 equal to C 2 3 by sigma 2 sigma 3, again this is coming out to be equal to be and I believe and r 3 1 is also equal to 1. So, now, going a slide, so found out the correlation coefficient to be equal to 1 everywhere.

(Refer Slide Time: 19:02)



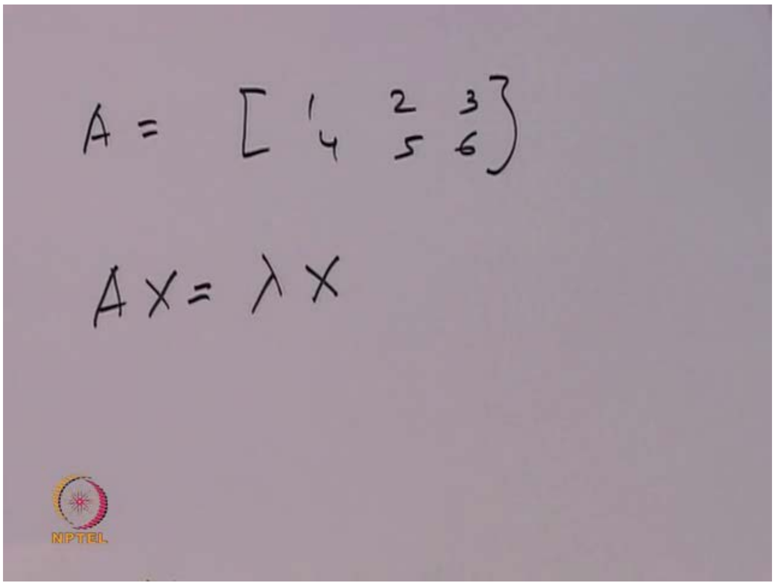
Now, for each variable x i j we replace the values by y i j equal to x i j minus mu i divided by sigma i square, we found out correlation coefficient with this values, actually what we have to do is that we first take the standardize the variance, and then find out the correlation coefficient anyway.

(Refer Slide Time: 19:31)



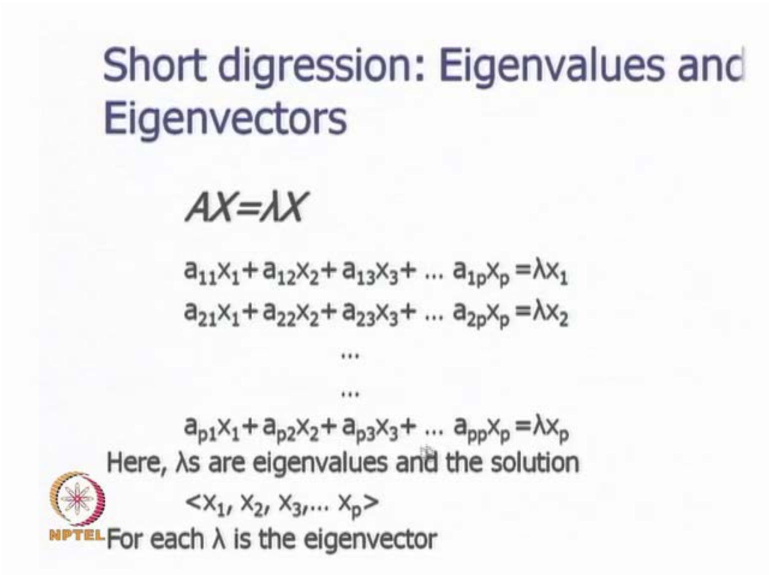
So, now, got the correlation coefficient and we do a principle component analysis you mean correlation coefficient matrix, so before that of course, we have to do this Eigen value and Eigen vector computation, again we do a beta basic task.

(Refer Slide Time: 19:46)



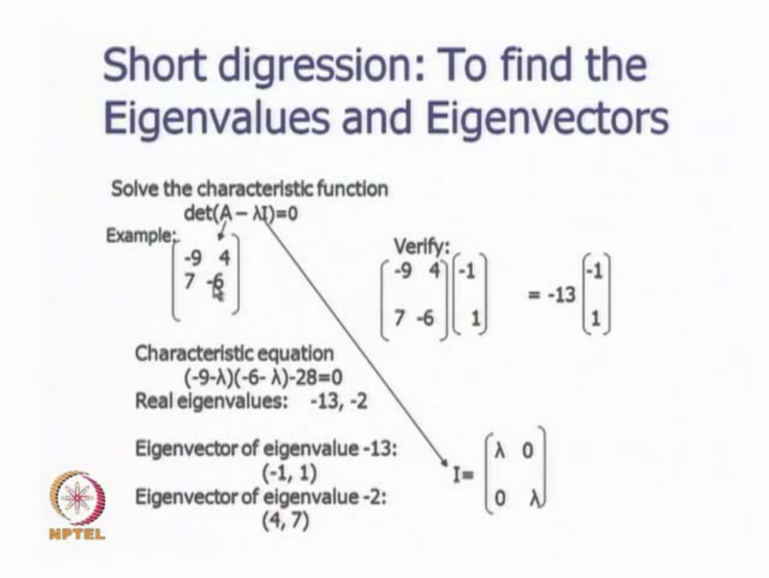
So, for this matrix A which is 1 2 3 4 5 6, the Eigen values are computed by this expression a x equal to lambda x, so going to the slide now.

(Refer Slide Time: 20:04)



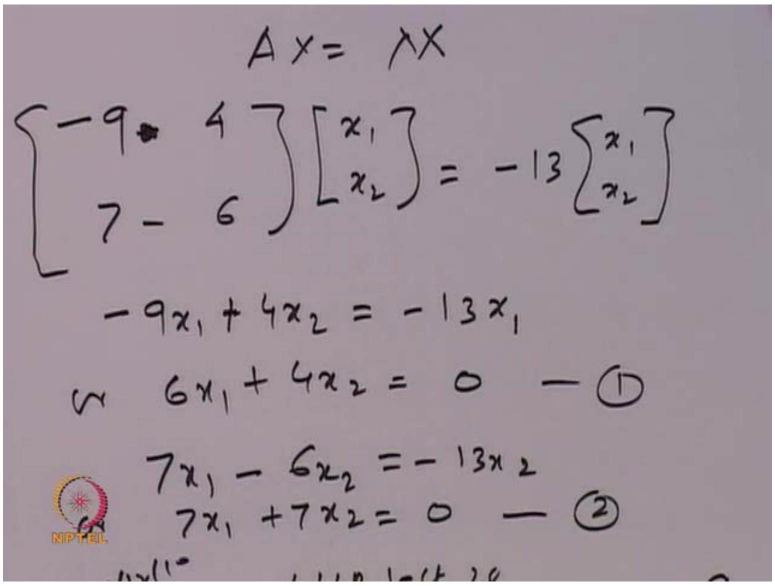
A x equal to lambda x for where a is a represented multivariate data we have x has the column vector, and this gives rights to the equation a 1 x 1 plus a 1 2 x 2 plus a 1 3 x 3 up to a 1 p x p equal to lambda x 1. And this way we set of this equation and find out different values of lambda from the determinate will see a movement, and lambda called the Eigen values and the solution x 1 x 2 up to x p for each lambda is the Eigen vector.

(Refer Slide Time: 20:43)



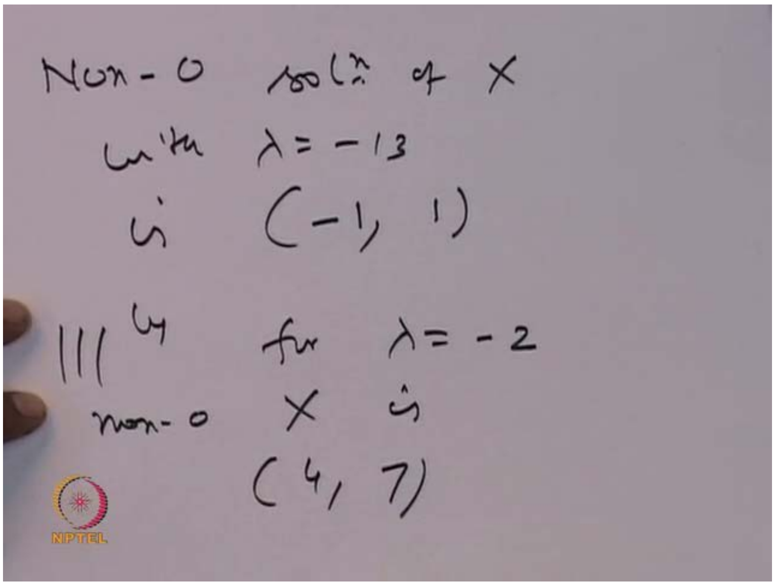
Here is an example, suppose the a matrix is minus 9 4 and 7 and minus 6, this is the A matrix, so first of all we get a minus lambda I, I is the identity matrix. So, lambda I gives as this matrix lambda 0 0 lambda, so A minus lambda I will be minus 9 minus lambda here, 4 minus 0 7 minus 0 minus 6 lambda, so this will be the A minus lambda I matrix. Now, we have to found what is call the characteristics equation so; that means, the determinate equal to 0, we have to found that equation, so minus 9 minus lambda into minus 6 minus lambda minus 7 into 428. So, this is equal to 0, now this left hand side is determinate equal to 0 is the characteristics equation, when we solve, then we find that lambda is equal to minus 13 or minus 2. So, this a quadratic equation lambda minus 13 minus 2 now when we put this value here that minus 9 4 7 6 into x equal to minus 13 x and solve for x let us do it once.

(Refer Slide Time: 22:15)



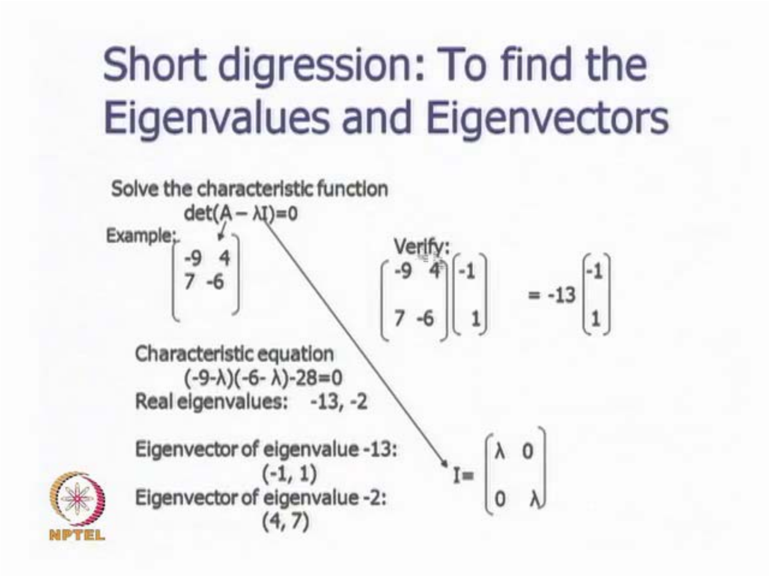
So, we have minus 9 minus 4 7 minus 6 is the matrix into x 1 x 2 is equal lambda which is minus 13 x 1 x 2 then we have minus 9 x 1 minus 4 x 2 equal to minus 13 x 1. There solving for a x equal to lambda x, for minus 9 4 plus 4 is equal to minus 13 x 1 or 6 x 1 plus 4 x 2 equal to 0, that is 1 and 7 x 1 minus 6 x 2 is equal to minus 13 x 2 or 7 x 1 plus 7 x 2 equal to 0. So, then so one solution is that x 1 and x 2 are all 0 vector, but others solution is that you can have x 1 minus 1 x 2 is plus 1 yes, yes 4 x 1 plus x 2 equal to 0, and this is 7 x 1 7 x 2 equal to 0.

(Refer Slide Time: 23:35)



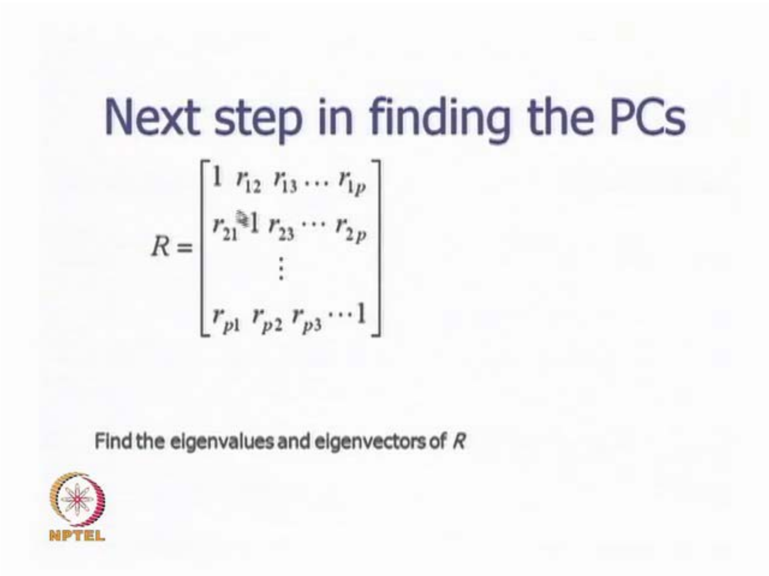
So, non zero solution for x 1 x 2 with lambda equal to minus 13 is minus 1 plus 1, similarly for lambda equal to minus 2 non zero x is 4 comma 7, so this the way this is calculated.

(Refer Slide Time: 24:04)



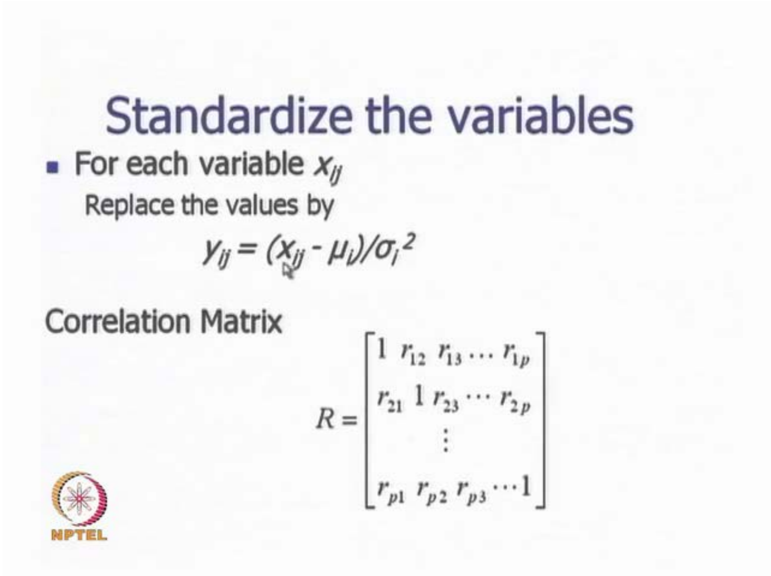
So, way it when come to the slides, we find that for this a matrix the Eigen values are minus 13 minus 2 the correspondent Eigen vectors are minus 1 plus 1 4 and 7.

(Refer Slide Time: 24:20)



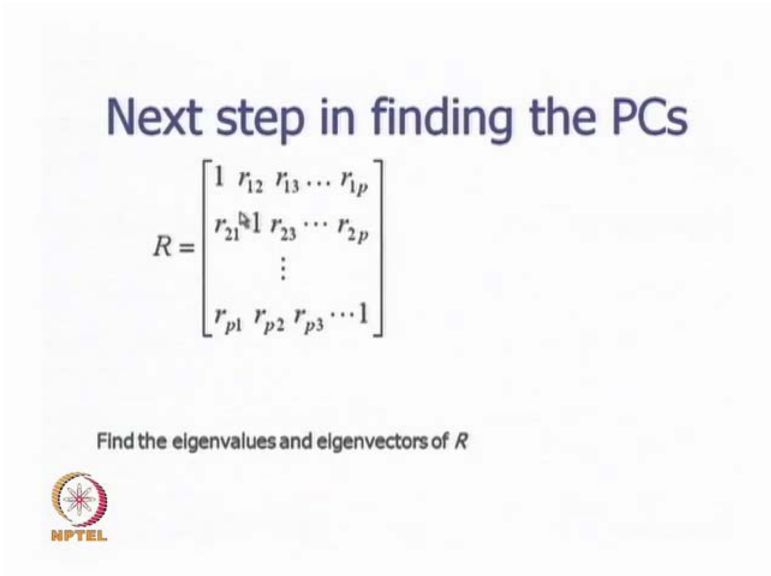
So, now, we proceed to finding the principle components, now from the given matrix we can find out the correlation coefficient matrix, just to remained our selves how do you find out the correlation coefficient matrix. We convert the given matrix into a matrix of standard variables, so this is one thing we had last time, so each value is subtract from the mean and is divided by the standard variance.

(Refer Slide Time: 24:53)



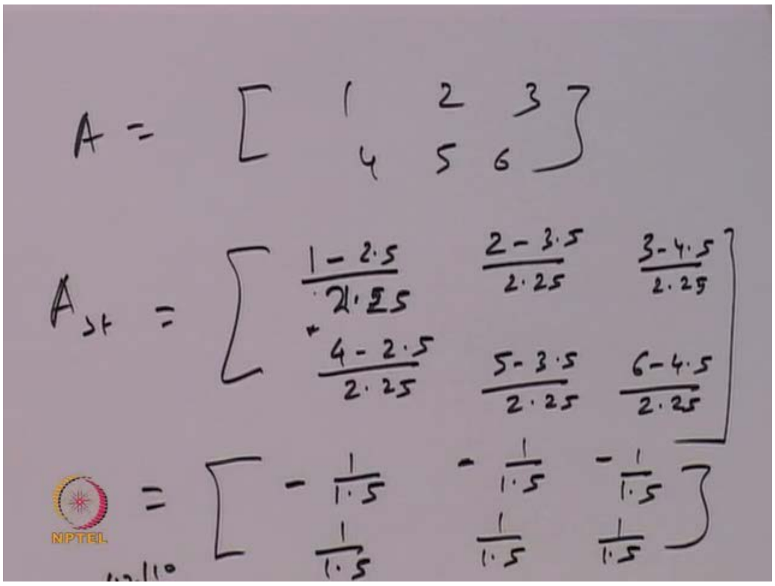
y i j is equal to x i j minus mu i dived by sigma square, so this is done for each value in the matrix.

(Refer Slide Time: 25:05)



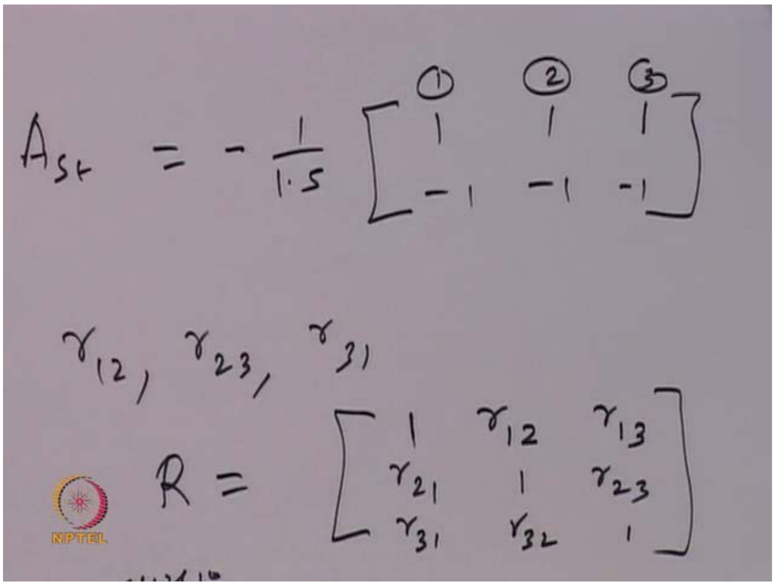
Then on this standard values we find out the sample covariance and sample covariance is dived by standard variation sigma a sigma b, and we get this correlation coefficient r 1 2 r 1 3 etcetera on the standardized matrix, so take this example of a matrix.

(Refer Slide Time: 25:29)



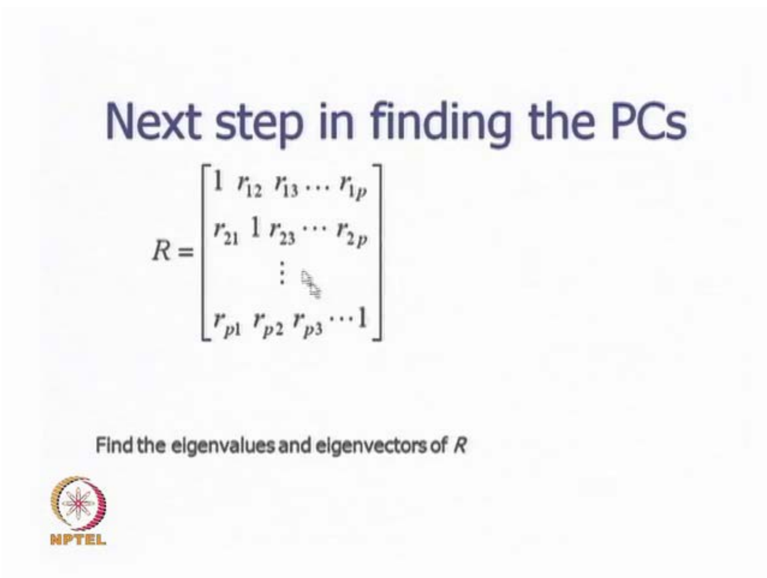
So, matrix A which is 1 2 3 4 5 6 is first converted to a standard which is nothing but 1 minus 2.5 divided by, so you have to divided by sigma square. So, 2.25 similarly 4 minus 2.25 divided by 2.25 2 minus 3.5 divided by 2.25 5 minus 3.5 divided by 2.25 3 minus 4.5 dived by 2.25 6 minus 4.5 divided by 2.25. This is known as standardizing the variables in the matrix, and this will come out to be equal to this is minus 1 by 1.5, this is also minus 1 by 1.5 this is 1 by 1.5, this is 1 by 1.5, this is 1 by 1.5.

(Refer Slide Time: 26:57)



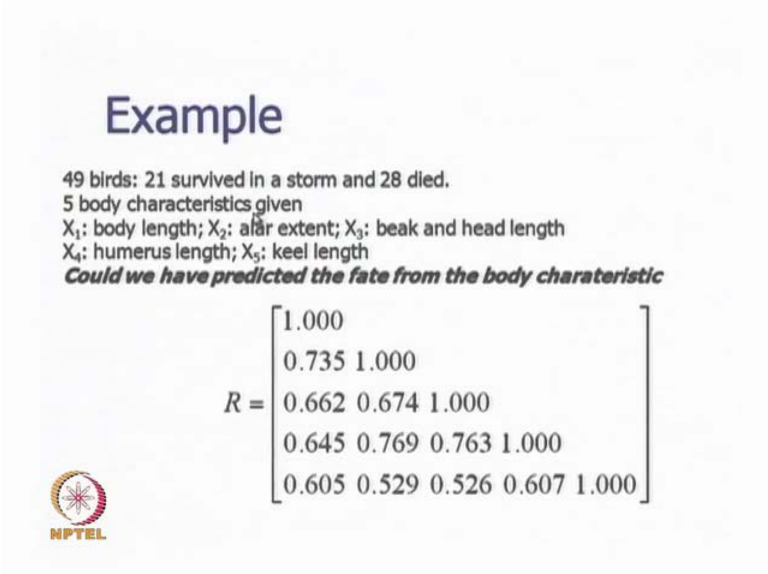
So, then A standard matrix comes out to be equal to minus 1.5 1 by 1.5 1 1 1 minus 1 minus 1 minus 1, this A matrix we get, from this we find out the r a b values. This is column number 1, column number 2 column number 3, and from this will have to find out r 1 2 r 2 3 r 3 1, so we going to through that computation now, but it is possible to get this value. And from this we will get correlation coefficient matrix which is equal to one r 1 2 r 1 3 then r 2 1 r 2 3 r 3 1 r 3 2 1, so whatever the matrix whether it is square or rectangle it has been convert to a square matrix and therefore, principle component this is can be done on this.

(Refer Slide Time: 28:05)



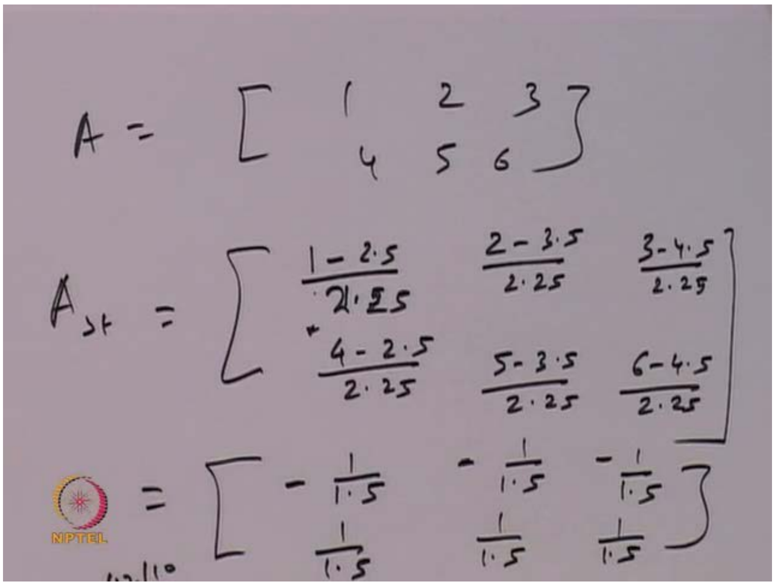
So, next step will be for this correlation coefficient matrix are we get that Eigen values and Eigen vectors.

(Refer Slide Time: 28:11)



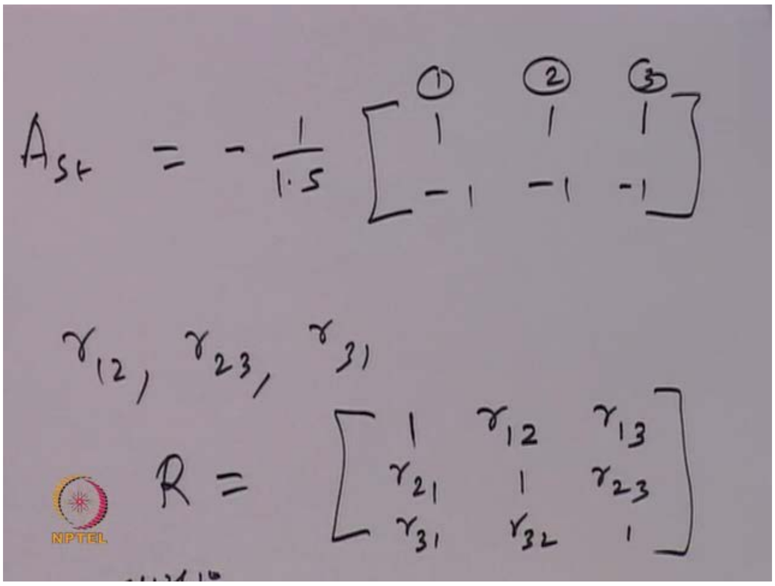
Now, we have taken example here, so for everything is clear we take the matrix A just look at the, whatever we wrote here.

(Refer Slide Time: 28:20)



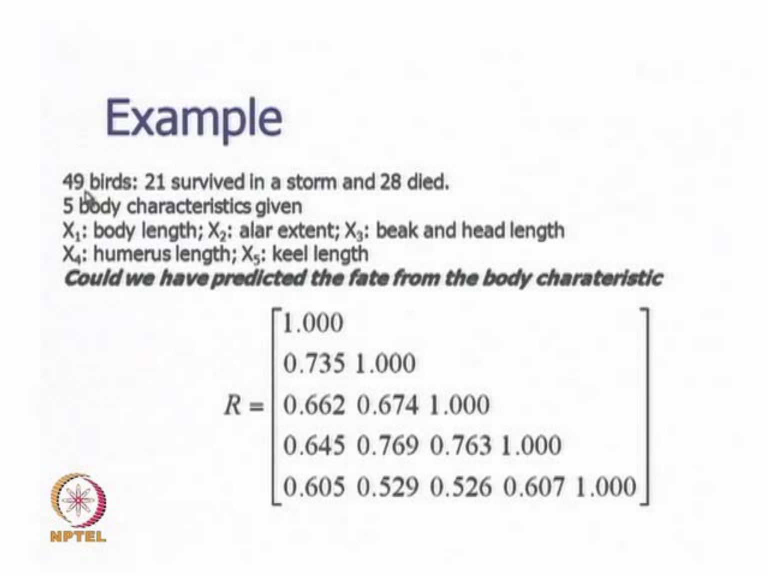
We take the matrix A first compute the column mean, everywhere compute the variance for each column standardize the variable, so take the variable value subtract the mean from it divide by variance everywhere, and that gives as the standardized A matrix.

(Refer Slide Time: 28:45)



From standardized A matrix we get the correlation coefficients, for each column and from this we found what is call the correlation coefficient matrix. Everything is done on the standardized view mean and variance. Yes, that is it, so this is done.

(Refer Slide Time: 29:12)



Now, here we taken example suppose there are 49 birds and 21 survived in a storm and 28 died 5 body characteristics given which are x 1 the body length x 2 the alar extent, x 3 the beak and head length, x 4 the humorous length, x 5 the keel length. So, some of this biological terms, so there are this 5 parameters which represented each part, so we have a

49 row by 5 column matrix describing the population of birds. Now, the task is to predict the fate of the bird from the body characteristic, so when the bird survive with the a bird with particular characteristic will about survive in the storm and this classification is done based on the body characteristic we try to learn their. So, we have a 49 by 5 multi wearer data we process the data through standardization compute the correlation coefficient, so for this data we have not shown the actual matrix, but what is interest to ask is the correlation coefficient. So, for this data we have found out the correlation coefficient and this is shown in terms of lower triangle matrix, because this a semantic matrix the values here will be same as the value on this side.

So, let us read one column which will be same as the first row, so correlation coefficient of first attribute x 1 with itself is one. The correlation coefficient between x 1 x 2 is 0.735, that between x 2 x 3 is 0.662 between x 1 x 4 is 0.645 between x 1 and x 5 is 0.605, so similarly all the correlation coefficient are given here. So, now from this correlation coefficient matrix we can find it is Eigen values and Eigen vectors by setting up a x equal to lambda x.

(Refer Slide Time: 31:23)

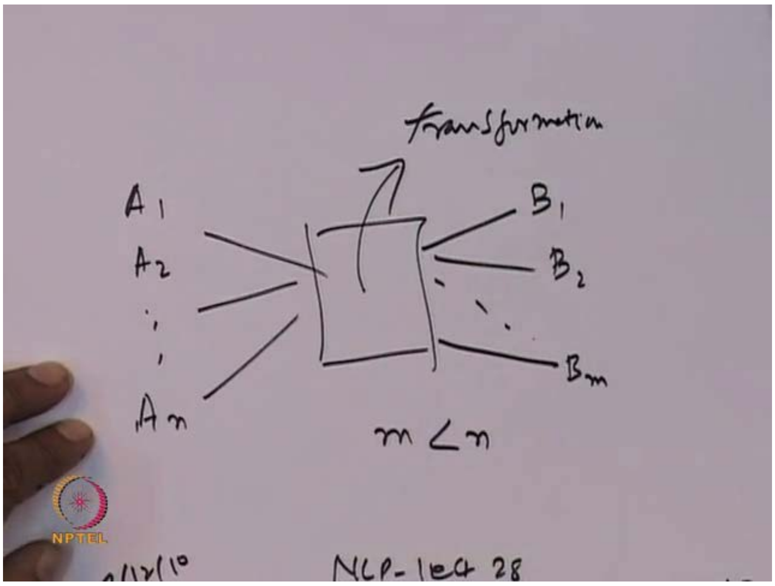
Eigenvalues and Eigenvectors of *R*

Component	Eigen value	First Eigen-vector: V_1	V_2	V_3	V_4	V_5
1	3.612	0.452	0.462	0.451	0.471	0.398
2	0.532	-0.051	0.300	0.325	0.185	-0.877
3	0.386	0.691	0.341	-0.455	-0.411	-0.179
4	0.302	-0.420	0.548	-0.606	0.388	0.069
[illegible]	0.165	0.374	-0.530	-0.343	0.652	-0.192

Now, this Eigen values an Eigen vectors computation is also not shown, but a method has been already descript a few mints back, so here we find that the Eigen values there will be 5 Eigen values at 3.612 0.532 0.386 0.302 and 0.165. This are the 5 Eigen values the first Eigen vector v 1 is 0.452 then the first component is this 0.452 then minus 0.051

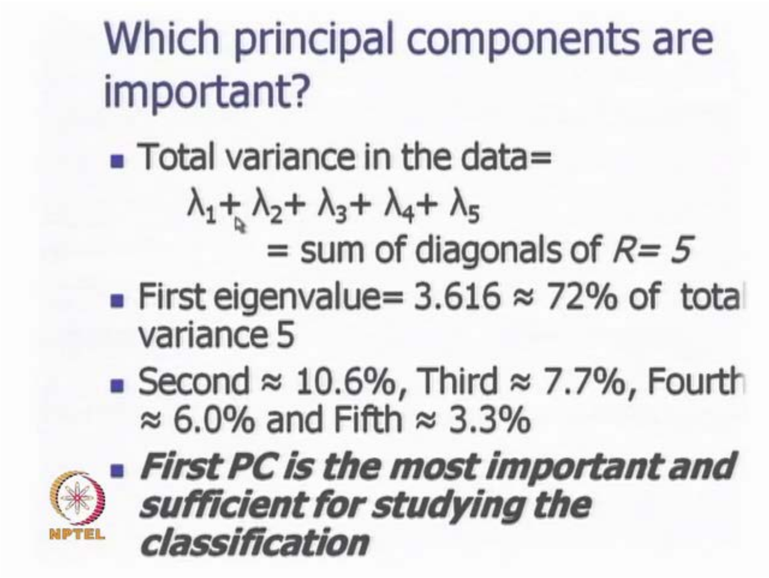
0.691 minus 4.420 minus 364 this the first Eigen vector. Then this is the second Eigen vector there is no correspondent between this and this are 5 Eigen values and corresponding to each the Eigen vector is shown, so may be this column being placed here is what misleading, so these are the 5 Eigen vector.

(Refer Slide Time: 32:36)



Now, the point is what we do with this data, we have already mention last time that the whole idea of dimensionality reduction is that we have attributes A 1 A 2 up to A n which when pass through A box gives rise to B 1 B 2 up to B m and m less than n. Now, what happens in this transformation is of interested, now from we have converted the a matrix into the correlation coefficient matrix, we have computed the Eigen vectors and Eigen values, so that is there in this box. Now, what we do with this is the question?

(Refer Slide Time: 33:13)

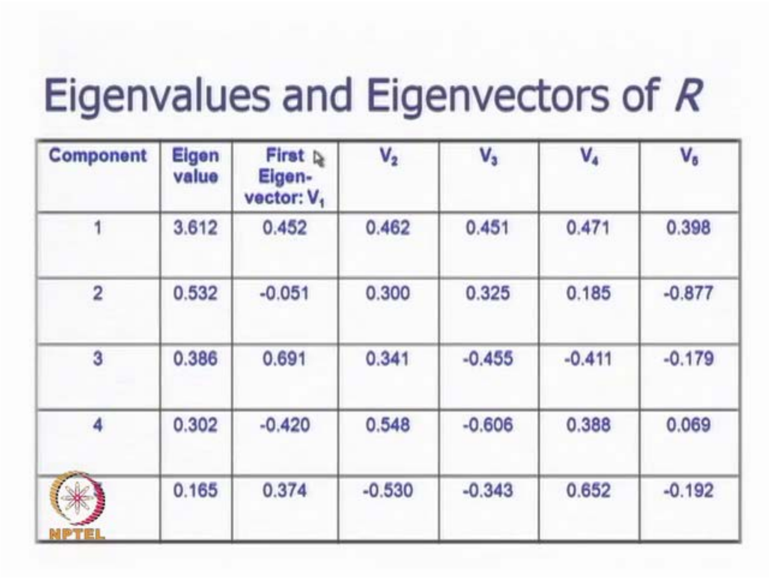


Now, without going into lot of theories of Eigen values in Eigen vectors we do say that the total variance in the data is measured by the sum of the Eigen values, we can discuss this point little later. So, this is equal to the sum of the diagonals of the correlation coefficient matrix, and this sum of lambda values Eigen values comes out to be equal to 5.

Now, first Eigen value is 3.616 that is 22 percent of 5 and this captures 72 percent of the total variance, the second Eigen value captures 10.6 percent of the variation third capture 7.7 percent fourth 6 percent and fifth 3.3 percent. So, first principle component is the most important and sufficient for studying the classification, so here some amount of insight into machine landing comes into picture.

So, after all our learning see depends on how representative the data is for the concept, so here we are talking about let us say birds which can survive the storms and birds which cannot, so as many variations of birds which survive and which do not survive. We see the better is the expose of the landing algorithm, so landing algorithm is better exposed to the situation, and therefore, is reach experience should enable to learn better. So, now the Eigen values will discussed these theories more in more detail and with more intuition that this Eigen value capture the variation in the data.

(Refer Slide Time: 35:13)

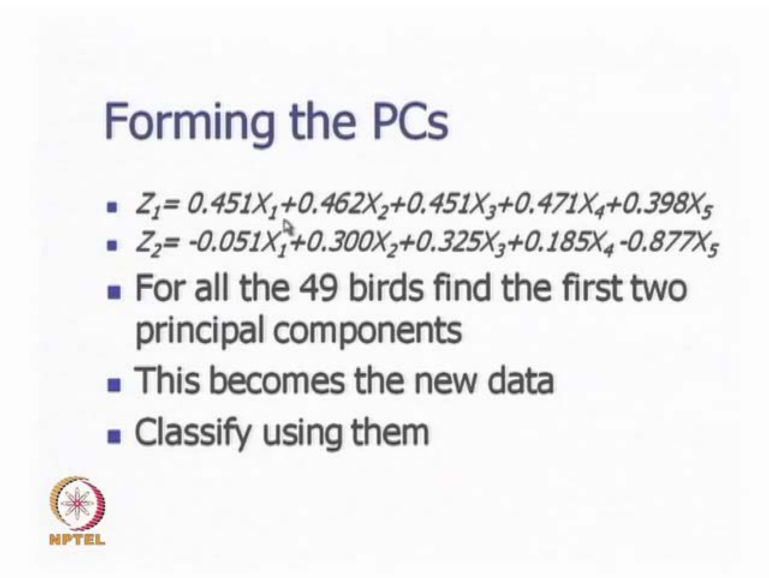


Eigenvalues and Eigenvectors of R

Component	Eigen value	First Eigen-vector: V_1	V_2	V_3	V_4	V_5
1	3.612	0.452	0.462	0.451	0.471	0.398
2	0.532	-0.051	0.300	0.325	0.185	-0.877
3	0.386	0.691	0.341	-0.455	-0.411	-0.179
4	0.302	-0.420	0.548	-0.606	0.388	0.069
5	0.165	0.374	-0.530	-0.343	0.652	-0.192

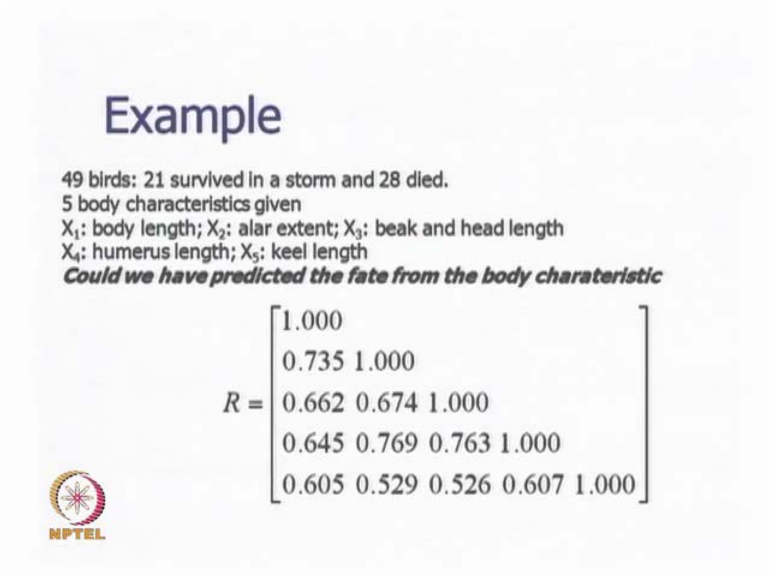
And the first Eigen value, which is shown here 3.612 and rest of them and you can see are miniscule compare to this value, this sum of 2 1, so there sum of lambda values computed from the correlation coefficient matrix on standard matrix values always comes out to be equal to the dimension of the correlation coefficient matrix. So, some of the Eigen values is equal to dimension of the correlation coefficient matrix, now this particular Eigen value captures 72 percent of the variance, how it is variation will see later, so now, we perform an operation the first component of each Eigen vector is taken this is 0.452 0.462 0.451 0.471 and 0.398.

(Refer Slide Time: 36:01)



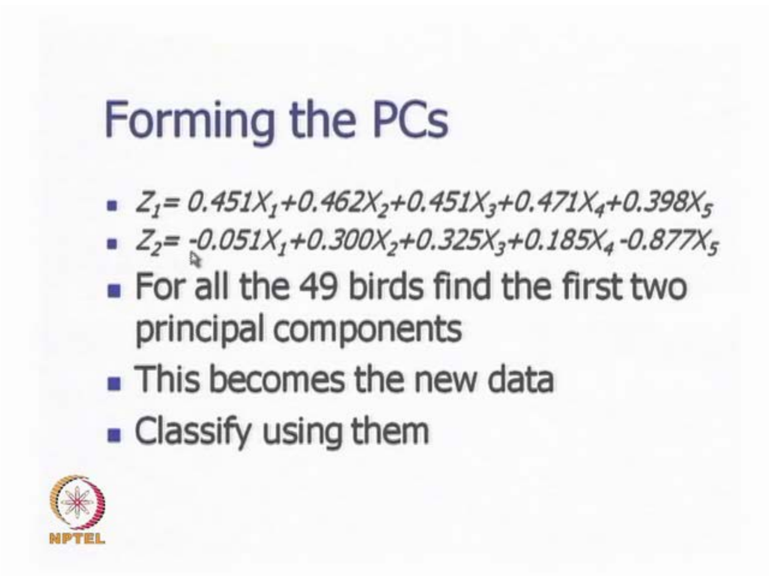
So, what are x 1 x 2 x 3 x 4 and x 5 these are the 5 characteristic values, so x 1 if I take one particular bird lets I take first bird then x 1 is it is body length.

(Refer Slide Time: 36:22)



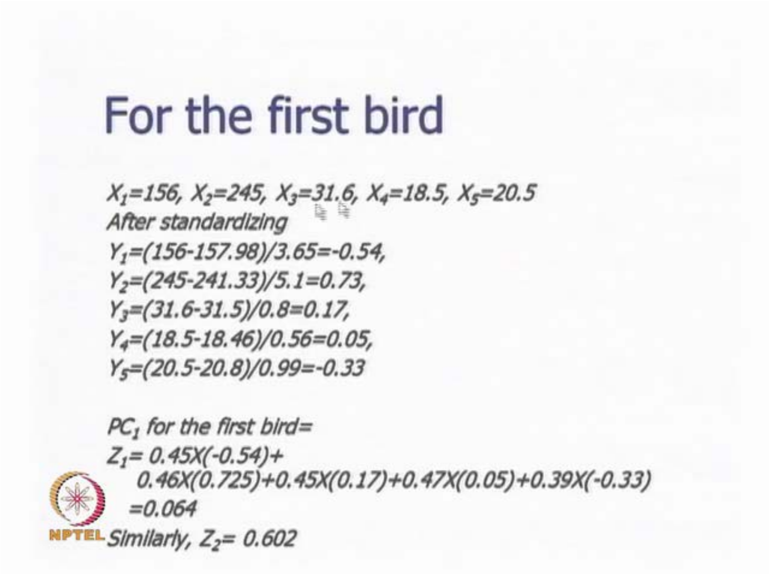
So, x 1 is body length, x 2 is alar extend, x 3 is beak and head length, x 4 is humerus length, x 5 is keel length, there are 5 different characteristics and this have 5 different values.

(Refer Slide Time: 36:36)



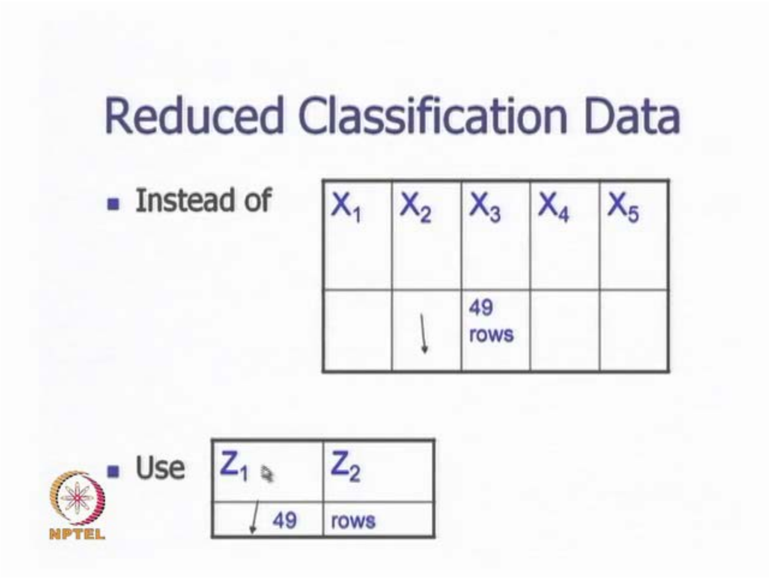
Now, they are multiplied by the first component of the Eigen vectors the first component of the Eigen vectors and then we obtain the values z 1, then this x 1, x 2, x 3, x 4, x 5 are multiplied by the second components of the Eigen vector. And we obtained z 2 so for all the 49 birds find the first 2 principal components, so these are called the principal components, and this becomes the new data and we classify using.

(Refer Slide Time: 37:13)



So, if I take this example now, the for the first bird x 1 is 156, x 2 is 245, x 3 is 31.6, x 4 is 18.5, x 5 is 20.5, now after standardizing again we have to do standardizing, we get these values y 1 y 2 up to y 5. So, the principal component one for the first bird will be 0.45 into minus 0.54, the standardized first attribute value, plus 0.46 which is the first component of the second Eigen vector into standardized second attribute value. So, each attribute value is multiplied by the first component of the Eigen vector and this gives me z one similarly z 2 comes this way.

(Refer Slide Time: 38:01)



So, now, instead of a 49 by 5 matrix we have a 49 by 2 matrix, so there are 49 data points, yes each data point instead of being represented by 5 attributes r represented by 2 attributes. And these attributes are functions of the previous attribute x 1, x 2 up to x 5, z 1 is the functional x 1, x 2 up to x 5, the function is obtained from the Eigen first components in the Eigen vector and now we have a reduced dimension data and we classify based on that, so will discuss this in more detail tomorrow.