Indian Institute of Technology Madras

NPTEL
National Programme on Technology Enhanced Learning

Pattern Recognition

Module 05

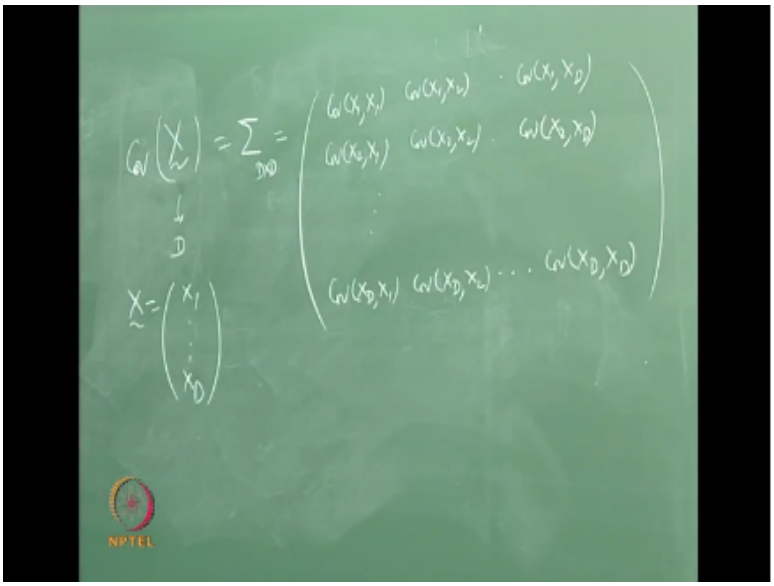
Lecture 01

Principal Components

Prof. C. A. Murthy
Machine Intelligence Unit,
Indian Statistical Institute, Kolkata

So I shall be discussing about principle components which probably many of you are aware of but in order to complete the things and as well as this probably I will try to give I mean a different way of looking at principal components. I think all of you may know what a covariance matrix is.

(Refer Slide Time: 00:39)



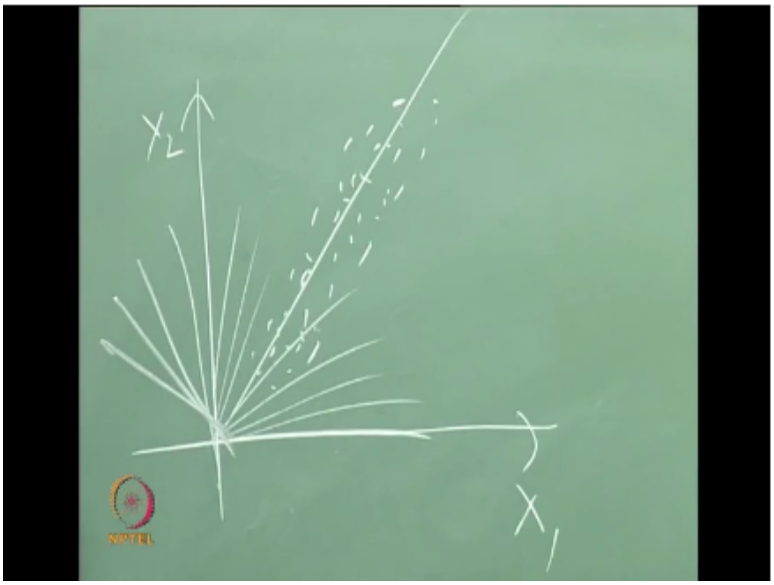
If you have a vector X then its covariance matrix, covariance matrix which are represented by σ and suppose this vector X has Here I am changing the notation slightly here say it has capital D dimensional vector, let us just say then this matrix is going to be basically covariance of the first

variable with itself covariance of X 1 with X 2 and covariance of X_1 with X_D then covariance of X_2 with X_1 covariance of X_2 with X 2 covariance of X 2 with X_D.

Then in the last row covariance of X_D with X 1 covariance of X_D with X 2 covariance of X_D with X_D where this vector X curve is actually X 1to X_D then this is the covariance matrix which is a capital D / capital D matrix which is a capital D / capital D matrix, still now when we are talking about feature selection we had a criterion function. Criterion function is defined based upon some I mean some particular characteristics which we believe that the features should possess.

Now in this principal components here we are not going to talk about classification here what we are going to do is we are given some capital D number of features, we would like to see somehow where you have more I should say variance, the places where you have more variance you are seemingly going to get, I should say more variance provides there more of an idea about that particular variable or that particular I should say combination of variables.

(Refer Slide Time: 03:43)

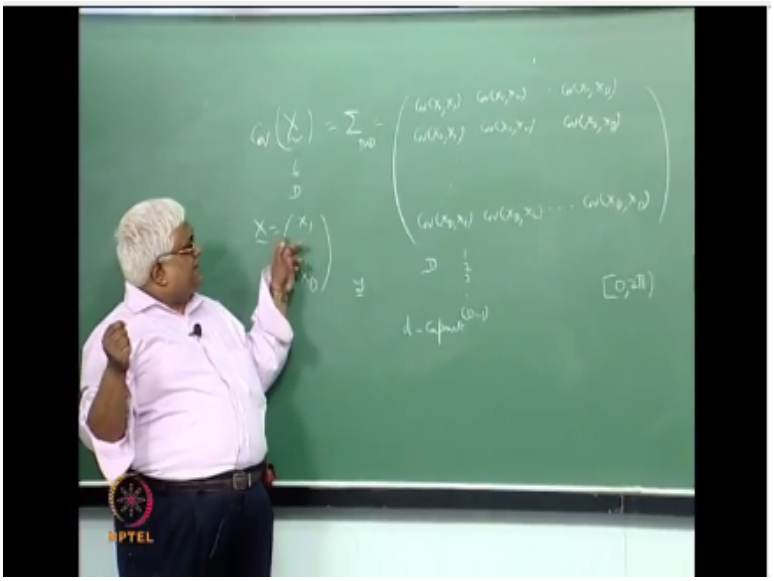


Basically you look at a data set like this say here you have two variables and you have a data set the variable X_1 has some variance and the variable X_2 also has some variance okay but probably if you take a linear combination and if the linear combination provides something like this may have more variance than this R than this am I correct? Basically in principal components this is what we are trying to get.

That is let me explain it you have totally capital D number of dimensions, there were number of dimensions is capital D then what we will see is that, in this capital D number of dimensions number of dimensions is finite, we are we would like to look at all possible directions. For example here how many directions are there you have this direction you have this you have this like this you have directions right.

So basically your directions are 0 to uncountable many directions, uncountable many directions you are going to have, then we would like to find out that direction where it provides maximum variance, what is the meaning of providing maximum variance? The meaning of providing maximum variance is the following.

(Refer Slide Time: 06:20)



So this is your data set my direction is let us just say this axis my first direction, then you take the projection of each point onto this you take the projection of each point onto this, say this point when it is projected it falls here, then you measure this length which is basically because it is already the x axis which is basically the x coordinate right. So if it is this axis you are basically going to get the x coordinate values of each one of the points but if it is not this axis.

Let us just say an axis likes this one then what are you going to do? You take their projections and each one you measure the distance from the origin you measure the distance from the origin okay, you measure the distance from the origin M they will be the projected points the when this point is projected onto this then the corresponding value is going to be the distance from the

origin to this one okay and similarly for this point the distance from the origin to this and something is positive another one is negative.

Here somehow this is negative and this is positive right, if you are taking this as your positive side of axis this as your negative side then from here to here you are going to get negative distance from here to here you are going to get positive distance, so the corresponding values are going to be that okay. So for any such direction say this is your direction then this point is projected here this is projected here then you take this thing.

So you take a direction and project the points onto that direction, so then you will get single dimensional values for each of these points then you can calculate the variance of this are you understanding what I am trying to say you can calculate variance. So for each direction you will get a value of the variance now you find out that direction for which the variance is maximum find out that direction where the variance is maximum okay.

How to find it out I will come to it later, say suppose you have found it then you store the direction, now look at all its perpendicular directions look at all its perpendicular directions, now among them find the direction with maximum variance. Then you have two directions now look at all of all the all the directions perpendicular to these two directions all the directions perpendicular to these two directions, then among them find the one which has the maximum variance.

Like that you just go on and on and on doing it when you come to the last one that means from D you will find 1,2, 3 up to say D – 1 you have found D - 1 directions always I mean if says you have found somehow D - 1 directions then the Dth direction is uniquely defined is it uniquely defined, there are totally capital D orthogonal directions sorry capital D orthogonal directions you have already found D - 1 orthogonal directions.

So capital D each one is uniquely defined okay and there also you project the points onto that then you get the variance okay. So now for each one of the points in this suppose one point is let us just say y it is a capital D dimensional point, then corresponding to the first direction that is the one maximal variance you will get the corresponding coordinate value to that corresponding to the second direction for this point you will get a value to that.

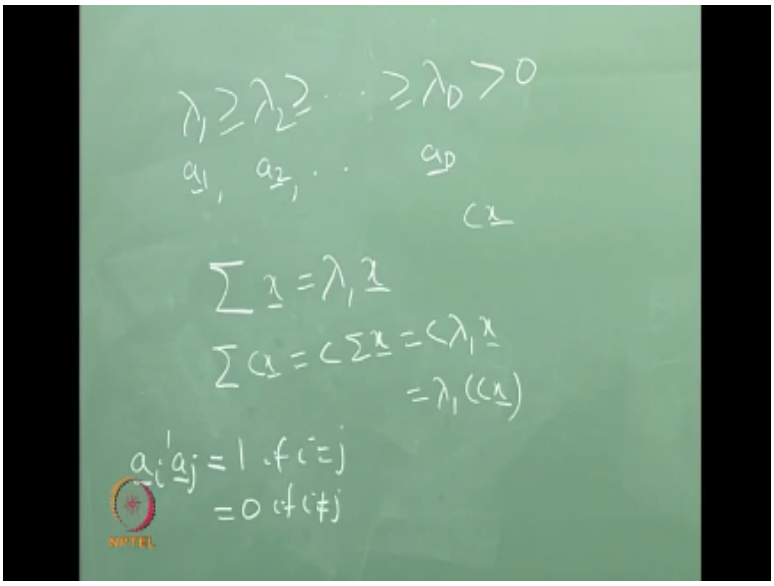
Like that corresponding to all these chosen D directions you will get a value, so that is the transformed value of this vector Y from the whole set of axis that you are given two new set of axis. We are basically going to get a new set of axis is it correct because all these are perpendicular directions and you have an origin and all these are perpendicular direction, so you are basically going to get a new set of axis right.

These directions are called the components, if you want to get small d number of components then 1 2 3 up to the small d you have to take they will be your small d components they are called principal components and there is also another term associated with this comes from electrical engineering, do you have any one of your background in electrical engineering? Honan low this is discrete carbon and low expansion this is discrete in low expansion.

So why the variance is given importance variance is given importance because variance will tell you where you have the more variation in the data set and you do not want to lose the information on variances, so if you have to represent the whole data set by a single component and you are choosing that principle that component with maximum variance that is why each time you are looking at the maximal variances okay. Now the next question is how do you calculate?

This is the basic principle but then it looks to be extremely complicated you have to take a direction for which you have the maximal variance and then you take the perpendicular direction off you consider all the perpendicular directions to this chosen direction, again find the one with maximal variance this seems to be a very highly cumbersome process. So there is a simple way of doing it that simple way is you take the covariance matrix of this vector from which you have got all these observations. And find it is Eigenvalues and eigenvectors find and write down the Eigenvectors eigenvalues in decreasing order that is.

(Refer Slide Time: 15:36)



There is a very basic question, here I just said write down the eigenvalues in decreasing order what happens if an Eigenvalue becomes a complex number? Can I write it in decreasing order? After all this is a matrix square matrix variance covariance matrix is a square matrix for every square matrix you can calculate Eigenvalues and eigenvectors and it is not necessarily true that Eigenvalues will be real, there can be complex numbers also, if they are complex numbers you cannot write down the eigenvalues in decreasing order or something like that right.

Now my question is that is it possible that for a covariance matrix the eigenvalues are complex numbers, the answer is no for a covariance matrix eigenvalues can never be complex why covariance matrix satisfies several properties one of the properties is covariance matrix is a positive semi definite matrix are non-negative definite matrix, positive semi definite are non-negative definite they mean the same thing it is if I write down the covariance matrix as σ then a ' σ a is ≥ 0 for all a.

If it is rather than or equal to zero for all a $\neq 0$ vector here equality is introduced then this matrix is said to be positive semi-definite are non-negative definite, non-negative means it is not negative that mean it can be zero or it can be > zero non negative is same as positive semi-definite positive means strictly greater than zero semi means you are including zero okay. So then σ is said to be non-negative definite matrix are a positive semi definite matrix and for a positive semi definite matrix the eigenvalues and this matrix is also symmetric σ is a symmetric matrix.

And positive semi definite matrix then the eigenvalues they are not only real they have to be also strictly ≥ 0 this is a I mean is a proven statement and from matrix algebra. Now for in for σ the eigenvalues are not only real but they are also strictly ≥ 0 that happens because for a positive semi definite matrix the determinant is can you tell me what the determinant will be? The determinant is actually for these matrices its product of eigenvalues the determinant is product of the eigenvalues.

So if the equality holds then there is at least one eigenvalue which = 0 then that means the determinant = 0 there is at least one eigenvalue which = 0 that means the determinant is also = 0, usually the covariance matrices are positive definite that is usually they satisfy this usually they satisfy this and if they satisfy this then this is true. Then that is true that means all the eigenvalues will be strictly > 0.

Now right let me ask you a question, corresponding to an eigenvalue how many different eigenvectors can you have? We generally write corresponding to this eigenvalue you have this eigenvector okay. My question to you is how many different eigenvectors you can have corresponding to a single eigenvalue? Do you have a unit eigenvector unique in the sense of the magnitude and the direction both have to be same or the direction is same the magnitudes are different can you say anything about it, Direction is same magnitude is different right.

That means suppose for the matrix σ suppose λ1 is an eigenvalue then σ X $_1$ s land okay so X is an eigenvector so σ X =λ $_1$ X and suppose I take some constant C times X then σ of constant C times X =C times σ X this is C times λ_1 X which is λ_1 times C $_X$ all right. So that means corresponding to an eigenvalue you are going to get vectors the same direction but different magnitude okay.

Now pose to eigenvalues are same okay before that let us just see suppose all the eigenvalues are different then can you say anything about the corresponding eigenvectors, suppose all the eigenvalues are different and can you say anything about the corresponding eigenvectors here what is the meaning of corresponding eigenvectors I take only those vectors with magnitude as one I take only those vectors with magnitude as one.

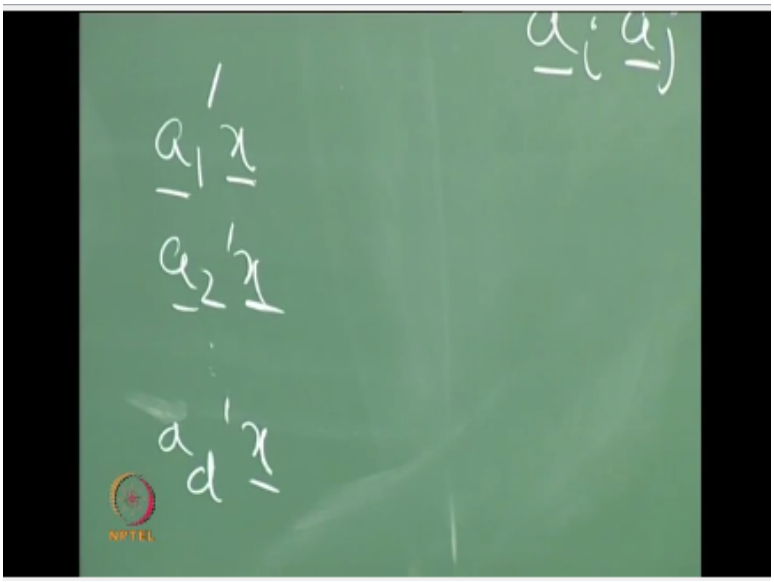
So corresponding to an eigenvalue you are going to get basically 2 eigenvectors since you are going to take √ okay - 1 ^{2} is 1 and 1 ^{2} is also 1 right so + running in two different directions you

are running it is the same thing in the same axis okay. So you might get two eigenvectors with magnitude as one but you take any one of them no problem, similarly for λ2 you take one such eigenvector so for λ B Capital D also you are going to take one such eigenvector.

My assumption is all these lambdas are different then what can you say about the corresponding eigenvectors we have an answer if eigenvectors are if i call them a_1 a_2 a_D then a_I 'A_J =1 if if i =j you are going to get a_I 'a_I which is actually the magnitude right square of the magnitude that =1 but if two eigenvalues are different here I am assuming all the eigenvalues are to be different then the corresponding eigenvectors this exercise this property that means they are orthogonal.

They are orthogonal am I right a 'a_J =0 when I is ≠ j and a 'a_J =1 if i=j, so if all the Eigen if no two eigenvalues are safe that means if all the eigenvalues are different, this property is satisfied but if two eigenvalues are same, can you say anything about eigenvectors. The eigenvectors first you are going to have several problems about eigenvectors.

(Refer Slide Time: 26:30)



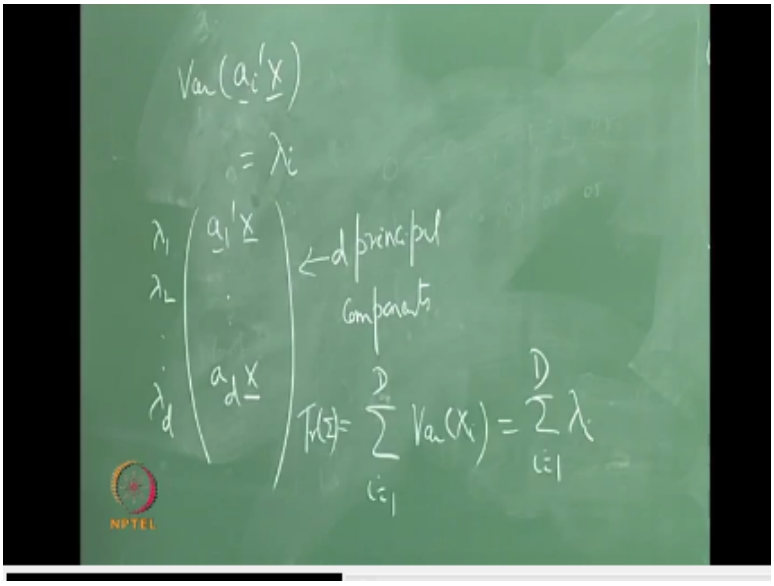
This is identity matrix what are the eigenvalues of this matrix this is identity matrix what the eigenvalues of this matrix are? They are 1 and 1 they are same every vector is an eigenvector right am I correct, so if I eigenvalues are same then this sort of property may not hold are you understanding, if eigenvalues are same this property may not hold but if Eigen if all the eigenvalues are different that means no two are same then eigenvectors are orthogonal.

And now you try to remember what I told you in the very beginning I said that you somehow find a direction now find all the directions perpendicular to that, so each vector is perpendicular to all the others am I correct, every if all the eigenvalues are different then eigenvector a with eigenvector is perpendicular to all the other vectors because of this okay. Now if you take the first eigenvector a one corresponding to this I said that you should get a real number.

What is that real number? That real number is a_1 ‘ for that particular X you remember this diagram, say this is the direction for this one you should take this for this you should take this again for this these are the values. So for a particular vector X the corresponding value is this corresponding to a_1 corresponding to a_2 the corresponding value is this corresponding to a $_d$ the corresponding to a $_D$ the value is this ad/‘X these are the product projected values.

And I was talking about variance the variance of all these values is actually for a $_1$ it is $\lambda_1$for a $_2$ it is λ_2 for a D it is λ_D, that is if I have to write it in mathematics.

(Refer Slide Time: 30:17)



Variance of a $_1$ ‘$_X$ = 1 it is λ_1 variance of a $_1$ ‘X =λ_1 so these are all the eigenvalues and eigenvectors of this variance covariance matrix. So the eigenvalues are going to give you the variances in those directions and eigenvectors will provide you the directions since we are trying to look at the one with maximal variance so you take the eigenvalue which is the maximum and corresponding to this you find eigenvector this is your first component.

Then I am assuming that the second one is strictly less than the first one the second eigenvalue is strictly less than the first, one then I get a $_2$ this is your second eigenvector, so this corresponds to the second component now you take this and up to you take D λ small d and corresponding to this you have the direction a $_d$ and you are small d these are your d principal components these are your small D principal components and the corresponding variances are λ_1 λ_2 λ_D the corresponding variances are λ_1 λ_2 λ_D.

In fact principal component analysis it is used extensively because of these properties that I told you and there is also another property, since I have been talking about variances is there any connection between these values and this diagonal note that every diagonal element is a variance term this is variance X $_1$ this is variance X $_2$ is variance X $_D$. So is there any connection between these diagonal elements and the λ_1 to λ D_S the answer is yes.

There is a connection what is the correction? The connection is some of the variances this is the trace of the matrix σ trace of σ I hope you all remember the meaning of the word trace is the sum of the diagonal elements, the main diagonal elements this is nothing but I want you to check these things, I want you to check this I am not giving you any proofs for these things but please check it summation i =1 to capital D of λ_1 that means we are just summing up all the variances that we have got this is nothing but $\sum$ i =1 to D of variance of excise.

So basically what we are trying to do here this variance of X is the sum of variance of excise we are trying to make a partition, where somehow we are just trying to keep the information about the larger variances and we are removing those things smaller variances and now what is the meaning of this is a linear combination of the original variables right, a_1 'X it is a linear combination of the original variables original variables are X $_1$ capital X $_1$ capital X $_2$ capital X_D and this is their linear combination of that.

We have taken here small d such linear combinations originally what we have is capital D such linear combinations, that is the original set up where you have capital D such linear combinations and these are all orthogonal to each other, we have taken small d of them corresponding to the larger variances and the other capital D - small d they correspond to smaller variances and if the variance is small and if we remove those things, would it create a problem or I would like to ask the question in another way?

Variance is small how is it going to help you can you? Tell me if the variance is small we can replace all the values by the corresponding means are you understanding but the corresponding means because since the variance is small from the mean the distance will be very small so we can actually replace the values by the corresponding means, so in that way we are losing some information I am not saying that we are not going to lose any information but the information loss is small.

So when the variance is small you can replace the values for the corresponding means okay, so by keeping the larger variances and removing those things which smaller variances, yes we are losing some information I am not denying that since the variances are small, if we are place by the mean yes there will be some information loss but it is not really that much okay, it is not really that much and the information loss or the loss in this procedure is actually measured by this, the loss in this procedure is actually measured by this.

And some people may do this also that means maybe some people may take this ratio also you can measure the loss either by this quantity are this divided by this ratio okay and so that whole theory that I was mentioning that can be easily done by looking at the Eigenvalues and eigenvectors of the covariance matrix. There is a theorem and proof for this relationship between those directions and the covariance matrix Eigenvalues and eigenvectors of the covariance matrix.

That is generally available in many pattern recognition books it is also available in many electrical engineering books and I will not go into the detail so f that I will not go into the details of the proof of these statements, if the people who so ever is interested in these things they can always go through the corresponding proofs in the books and that they can find very easily okay and this is really a popular procedure.

Because of all these properties that I mentioned because of all these properties that I mentioned and it is used was that just too many, to many places it is used just to many places where principal components is used, this also resulted in I mean in fact usually computer scientists are statisticians they have to go through these Eigenvalues and eigenvectors because of principal components this is one of the reasons why he statisticians are computer scientists they have to go through the literature on Eigenvalues and eigenvectors because of this principal components.

And because of those wonderful properties of I mean the covariance matrix and when you see that there is a division that takes place the summation variance X is same as summation i=d to be λ_I this is a very strong property this is a very strong property and so just divides partitions that that is very nice you see that is very nice and there is a PCA LDA about which Dr. Sukhinder das goes anyway he will teach okay.

And that there are many other variations of PCA which are used at many places and about one of them I shall take a lecture probably tomorrow okay where that is PCS are used for feature clustering where principal components are used for feature clustering in fact PCS, have been used at I mean several places. One of the recent works is regarding principal components for sparse matrices, a sparse matrix is a matrix where you have more 0 elements than none zeros you have more 0 elements than non0 okay.

And then that means basically your data set is such that you have too many dimensions and in those dimensions too many of them, let us just say I was mentioning an example yesterday I will tell the same example today, it is you are your data set is something like a web mining data set that is say you have what a collection of web pages. Let us just say some documents let us just say you have 100 documents you have 100 documents okay.

So in each document you have some sentences some words and some sentences, let us just say the number of words per document is of the order of just give me some number let us just say 50 words are there of the order of 50 it may be 51, 52, 53 or it may be 47, 48, 49 or some of them may be even it may be much smaller let us just say 50. So for each document on an average you have let us just say 50 words now you have hundred documents.

So100 x 50 let us just say 5000 words and for the sake of convenience let us assume that all these five thousand words are different even if some of them are same the number of words will be quite a bit okay. So let us just say you have 5000 words and all these words are different now what we will do is that were present a web page by a 5000 dimensional vector a word one is present you write one otherwise 0 if word 2 is present and that location write one otherwise 0.

So your vector is going to be a 5000 and dimensional vector where you have 0 or 1 like that you have 100 documents, that means 100 vectors but your number of dimensions is 5000 number of dimensions is 5000 but the number of such vectors is 100. Now if you have to look at the

corresponding variance covariance matrix to find principal components you are going to have some problems, if it is instead of 5000 in fact if you look at WebPages you have too many words.

The number of words may be in lakhs then your variance matrix will be 1 lakhs x 1 lakhs such a big matrix right, so but then most of the elements are 0 if you have to describe one web page where you have hardly 100 or 250 words 100, 150, 200, 250 or just say 500 words but then your number of words that you have taken is 1 lakhs you know one lakhs words among which 500 are here so the rest all of them are 0, so your data matrix is basically a sparse matrix.

And nowadays many data sets are like this and for these sort of data sets if you have to do all these things then you may have to develop some new methods, the reason is that whenever we do this sort of thing we assume there is an inherent assumption, that the number of points is much more than the number of dimensions. The number of data points is much more than the number of dimensions as for many real-life problems that may not be true for many real-life problems the number of data points may be much less than the number of dimensions.

I have mentioned for a remaining data set okay and you have many other data sets many data sets involving bioinformatics, you have cancer patients, you have many gene expression data sets where again the number of dimensions is after order of two three four thousand but the number of points may be of the order of 100 or 150 or 200. So these are some of the latest problems where when you are trying to apply principal components you may face some problems because of since the number of dimensions is much more than the number of data points.

Then and your computer may not be able to support finding a finding eigenvectors for a letter just a 5,000 by 5,000 matrix your computer may not be able to support it but for the same thing for a 100/100 matrix probably our computer can support, it so sparse data and sparse matrices are occurring many times and in many applications in real life. So there are some papers where somehow people are trying to find the principal components for when you have sparse matrices.

There is one paper by tipsy Ronnie on this Riga in this regard I think that they appeared in one of the statistics journals, tipsy Ronnie is a famous person working in machine learning and I hope by now you know that many of these things are we are calling it pattern recognition some people are calling the data mining, some people are calling it machine learning and some people are calling it artificial intelligence.

So many of these things are actually occur in too many disciplines okay and Tipsy Ronnie and a few such others statisticians they call themselves as machine learning people, so they are working on this thing some papers are already published and there are several problems related to principle components in very high dimensional data sets because your computer may not be able to support such high dimensional. I mean finding eigenvalues and Eigen value vectors for such high dimensional matrices, I am stopping it here if you have any questions please ask me, no more questions okay.

End of
Module 04 – Lecture 02

Online Video Editing / Post Production
M. Karthikeyan
M. V. Ramachandran
P. Baskar

Camera
G. Ramesh
K. Athaullah
K. R. Mahendrababu
K. Vidhya
S. Pradeepa
D. Sabapathi
Soju Francis
S. Subash
Selvam
Sridharan

Studio Assistants
Linuselvan
Krishnakumar
A. Saravanan

Additional Post – Production
Kannan Krishnamurty & Team

Animations
Dvijavanthi

NPTEL Web & Faculty Assistance Team
Allen Jacob Dinesh
Ashok Kumar
Banu. P

Deepa Venkatraman
Dinesh Babu. K.M
Karthick. B
Karthikeyan. A
Lavanya. K
Manikandan. A
Manikandasivam. G
Nandakumar. L
Prasanna Kumar. G
Pradeep Valan. G
Rekha. C
Salomi. J
Santosh Kumar Singh. P
Saravanakumar. P
Saravanakumar. R
Satishkumar. G
Senthilmurugan. K
Shobana. S
Sivakumar. S
Soundhar Raja Pandian. R
Suman Dominic. J
Udayakumar. C
Vijaya. K.R
Vijayalakshmi
Vinolin Antony Joans

Administrative Assistant
K.S. Janakiraman

Principal Project Officer
Usha Nagarajan

Video Producers
K.R. Ravindranath
Kannan Krishnamurty

IIT Madras Production

Funded By
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in