

Deep Learning
Prof. Mithesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module - 6.5
Lecture – 06
PCA: Interpretation 2

So, that is what we look at in the second interpretation of PCA right.

(Refer Slide Time: 00:17)

Given n orthogonal linearly independent vectors $P = p_1, p_2, \dots, p_n$ we can represent x_i exactly as a linear combination of these vectors.

$$x_i = \sum_{j=1}^n \alpha_{ij} p_j \quad [\text{we know how to estimate } \alpha'_{ij}\text{'s but we will come back to that later}]$$

But we are interested only in the top-k dimensions (we want to get rid of noisy & redundant dimensions)

$$\hat{x}_i = \sum_{j=1}^k \alpha_{ik} p_k$$

We want to select p'_i 's such that we minimise the reconstructed error

$$e = \sum_{i=1}^m (x_i - \hat{x}_i)^2 \quad (x_i - \hat{x}_i)^T (x_i - \hat{x}_i)$$

45/71

Prof. Mithesh M. Khapra CS7015 (Deep Learning) : Lecture 6

So, again we have the same setup that given n are linearly independent for n orthogonal vectors. We can represent x_i exactly as a linear combination of these vectors, what do I mean by exactly? Perfect ok; if you actually describe the whole things in words ok. So, that is exactly what I mean right. So, you are going to write x_i as α_{1i} into p_1 plus α_{2i} into p_2 and so on. And when you do the summation on the LHS on the RHS, you just get back the x_i when you do the summation on the right hand side you get back the left hand side ok.

So; that means, it can exactly be represented, when you use all the n eigenvectors now, if I start chopping of stuff what will happen?

Student: (Refer Time: 01:04).

It will just be an approximation ok. Now we this is what I meant, and this is this the equation holds; that means, this is exact and we know how to find the alpha is, because p js are conveniently orthonormal. So, we know how to find that easily ok. Now what if we consider only the top k dimensions, what is going to happen? There is going to be some error in the reconstruction I am not capturing all the information in my original data, but there is some error which I am not being able to capture, and I made a conscious decision that that error is not important I am willing to let it go.

Hence I want to represent the data using fewer dimensions ok. So, this is exactly what you do in PCA when you take the top k dimensions is this fine ok. So now, we want to select p is such that; we minimize the reconstructed error ok. And this is again erratic actually we should try to write it as, $x_i - \hat{x}_i$, since these are vectors and the square of vectors would just meet this right.

(Refer Slide Time: 02:16)

The slide displays the equation for the reconstruction error e as a sum of squared differences between original data points x_i and their reconstructed counterparts \hat{x}_i . The equation is written as $e = \sum_{i=1}^m (x_i - \hat{x}_i)^2$. To the right of the equation, the vector difference $(x_i - \hat{x}_i)^T (x_i - \hat{x}_i)$ is written in red, indicating that the squared term represents the dot product of the error vector with itself. Below the equation, there is a video inset showing a man in a light blue shirt speaking. At the bottom of the slide, a blue footer bar contains the text "Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6".

So, but you get the point right were just trying to do the element wise squared error loss were trying to minimize that ok, we want to do this. So now let us try to see that if you are aiming to do this, what is the condition that, we arrive at ok. So, no I thought I would ask for some changes on this, let us see if you guys oh god ok; I will ask for some changes on this it is ok. I think they forgot let me just see how to deal with this ok.

For a minute all of you can you just bear with the fact that these are actually vectors and not scalars. So, this square actually does not mean anything it actually means $x_i - \hat{x}_i$

\hat{x}_i transpose, x_i minus \hat{x}_i . So, when I use square with vectors this is what I mean is that ok, everyone can work with that notation fine ok.

(Refer Slide Time: 03:09)

The slide contains the following content:

$$e = \sum_{i=1}^m (x_i - \hat{x}_i)^2$$

$$= \sum_{i=1}^m \left(\sum_{j=1}^n \alpha_{ij} p_j - \sum_{j=1}^k \alpha_{ij} p_j \right)^2$$

Handwritten diagram showing a vector x as a linear combination of vectors $p_1, p_2, \dots, p_k, p_{k+1}, \dots, p_n$ with coefficients $\alpha_1, \alpha_2, \dots, \alpha_k, \alpha_{k+1}, \dots, \alpha_n$. A bracket under the first k vectors is labeled k , and a bracket under the last $n-k$ vectors is labeled $n-k$.

Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6

So now what is x_i actually the real point right the correct point which can be obtained by the full reconstruction, if you consider all the n dimensions, what is \hat{x}_i just an approximation where you are considering only the k dimensions. Remember that each of these quantities is a vector fine ok. Now what is happening here? Let me just try to say this ok. So, let me just do this way. So, this is your original x and you are actually writing it as a linear combination of your p s somewhere you will have $\alpha_k p_k$ and then all the way up to p_n right.

So, this is $p_k \alpha_n$ ok. Now what is this full thing this is x and what is this \hat{x} ok. You see the picture what is the equation trying to tell you ok. Now what is the difference between these two then, these guys right if I want to take difference between x and \hat{x} everyone gets that it is the remaining term say; that means, $\alpha_{k+1} p_{k+1} + \dots + \alpha_n p_n$ is that clear.

(Refer Slide Time: 04:22)

$$\begin{aligned}
 e &= \sum_{i=1}^m (x_i - \hat{x}_i)^2 &= \sum_{i=1}^m \sum_{j=k+1}^n (p_j^T x_i)(x_i^T p_j) \\
 &= \sum_{i=1}^m \left(\sum_{j=1}^n \alpha_{ij} p_j - \sum_{j=1}^k \alpha_{ij} p_j \right)^2 &= \sum_{j=k+1}^n p_j^T \left(\sum_{i=1}^m x_i x_i^T \right) p_j \\
 &= \sum_{i=1}^m \left(\sum_{j=k+1}^n \alpha_{ij} p_j \right)^2 = \sum_{i=1}^m \left(\sum_{j=k+1}^n \alpha_{ij} p_j \right)^T \left(\sum_{j=k+1}^n \alpha_{ij} p_j \right) &= m X^T X \\
 &= \sum_{i=1}^m (\alpha_{ik+1} p_{k+1} + \alpha_{ik+2} p_{k+2} + \dots + \alpha_{in})^T (\alpha_{ik+1} p_{k+1} + \alpha_{ik+2} p_{k+2} + \dots + \alpha_{in}) &\downarrow [x_1, x_2, \dots, x_m]^{1 \times n} \\
 &= \sum_{i=1}^m \sum_{j=k+1}^n \alpha_{ij} p_j^T p_j \alpha_{ij} + \sum_{i=1}^m \sum_{j=k+1}^n \sum_{L=k+1, L \neq j}^n \alpha_{ij} p_j^T p_L \alpha_{iL} & \begin{matrix} x_{1,1}^2 & x_{1,2}^2 & \dots & x_{1,n}^2 \\ x_{2,1}^2 & x_{2,2}^2 & \dots & x_{2,n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1}^2 & x_{m,2}^2 & \dots & x_{m,n}^2 \end{matrix} \\
 &= \sum_{i=1}^m \sum_{j=k+1}^n \alpha_{ij}^2 & (\because p_j^T p_j = 1, p_i^T p_j = 0 \quad \forall i \neq j) & \sum_{i=1}^m \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix} \\
 &= \sum_{i=1}^m \sum_{j=k+1}^n (x_i^T p_j)^2 & & \begin{matrix} [x_{1,1} \ x_{1,2} \ x_{1,3}]^T \\ [x_{2,1} \ x_{2,2} \ x_{2,3}]^T \\ \vdots \\ [x_{m,1} \ x_{m,2} \ x_{m,3}]^T \end{matrix} \\
 & & & \begin{matrix} x_{1,1}^2 & x_{1,2}^2 & \dots & x_{1,n}^2 \\ x_{2,1}^2 & x_{2,2}^2 & \dots & x_{2,n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1}^2 & x_{m,2}^2 & \dots & x_{m,n}^2 \end{matrix} \\
 & & & \begin{matrix} 1,2 \\ x_{1,1} \ x_{1,2} \\ x_{2,1} \ x_{2,2} \\ \vdots \\ x_{m,1} \ x_{m,2} \end{matrix}
 \end{aligned}$$

So, can I write it as yeah can I write it as this ok. So, you get this right. So, I am only taking these guys because the rest will get subtracted. So, one is the full n dimensions the other is only k dimensions. So, if I take the difference between them what remains is k plus 1 to n dimensions, and that is exactly what I have written here ok.

And now I am coming back to the proper notation where this is a vector right. So, I am writing the square as the dot product between the same vector is this ok. These are the m data point right; this sum this is overall the m data points you need to minimize that is that clear ok. So, this is fine, now beyond this is just some rearrangement. So, I have just expanded out that summation, this is what it would look like right. I have just expanded out these 2 summations.

Now let us try to do this in your head and see what are the kind of terms that you get there are 2 different types of terms that you will get. So, first of all let us understand that when you expand this you will end up with a lot of dot products, you will get a dot product between this and this and this and so on right. So, can you split those terms into two different types?

Student: (Refer Time: 05:37).

Square terms; So, one where i is equal to j and one where i is not equal to j is that clear fine. So, let me just write it as that. So, I will have k plus 1 to n right; that means, n

minus k terms, where i would be equal to j right so; that means, p_k plus 1 was getting multiplied by k p_k plus 1, p_k plus 2 was getting multiplied by p_k plus 2 and so on and then I will have these remaining terms where i is not equal to z right. So, these are the dot product between the other vectors is it fine. You see why I have split it this way, what will happen now? The second term will go to 0 ok; and what about the first term? α is a square ok, now what is α_{ij} actually how did you find α_{ij} .

Student: (Refer Time: 06:28).

It is a dot product between we did this right finding any of these components is just taking the dot product between x_i and that dimension. So, x_i transpose p_j is that fine ok. Is this fine and again this is slight abuse. So, this is actually, what no this is right, a this is sorry sorry, sorry sorry sorry I am just going to write it as this is this fine. I just written it twice and I can change the order. Since, it is a dot product ok.

Now, what I am going to do is, so this is actually summation over an index i and a summation over an index j . And I can change the 2 summations I can interchange them ok. So, that is what I am going to do now is this fine. I will push the summation all the way inside what is this actually this entire thing actually m times covariance of.

Student: (Refer Time: 07:25).

So, is this I is this what you are telling me that this is $m \times$ transpose x is this fine. How many if you do not get this I see a lot of blank faces how many if you do not get this quite a few; so, this is so i is equal to 1 to m right. So, you are going over the data points ok. So, this what is the dimension of this actually?

Student: n cross 1.

n cross 1 and this is 1 cross n , what does this product give you?

Student: (Refer Time: 07:49).

n cross n what are the entries in this matrix. So, this was say x_1 1 up to x_1 n . And this is again x_1 1 up to x_1 n ok. So, that is going to be x_1 1 square or rather let me just write it in the generic form right. So, it is going to be x_1 i into x_1 j right is that fine. And how many such matrices are you adding?

Student: (Refer Time: 17:07).

m of these. So, what would you get then? What would the first let us. So, ok so, let us do this. So, the first entry of this matrix is going to be x_{11} square, what about the first entry of the next matrix in this series?

Student: (Refer Time: 08:40).

x_{21} x_{21} square right ok. So, this is slightly tricky to demonstrate, let me just give me a minute I will just collect my thoughts and do it properly ok. Let us take a small example ok. So, x_{11} x_{12} x_{13} suppose we have a 3 dimensional matrix 3 dimensional data. So, I am taking a sum of m such matrices ok; i equal to 1 to m ; that means, this is going to vary this indexes the first index is going to vary from 1 to m . Now, let us see the first matrix and let us look at the first element of that matrix the first element of this matrix is going to be x_{11} square ok.

Now, let us look at the next matrix, what is the next matrix going to be? It would be x_{21} x_{22} x_{23} right and multiplied by x_{21} x_{22} x_{23} , what is the first element of this matrix going to be?

Student: (Refer Time: 09:45).

x_{21} square what about the third one? x_{31} square this is fine so far now you are adding all these matrices. So, what is the first element of the resultant matrix going to be x_{11} square plus x_{21} square plus x_{31} square, what is this actually? This is the dot product of x_{11} with itself right, and what does that give you the variance if the data is 0 mean right ok. Now can you make a similar argument of the ij th entry is going to give you the covariance between the i th and the j th entry is that clear right. You could do a similar analysis you can actually work it out after going back, how many of you have found comfortable with this? There is still many who are not ok.

So, let us look at an ij th entry right. So, can someone help me with say that 1 comma 2 entry or the first matrix what is it going to be x_{11} into x_{12} right for the second matrix.

Student: (Refer Time: 10:52).

x no this is some ya correct and for the third matrix 3 2 ok. Now what is this sorry what is the summation of these? When you take the full sum you will get these 3 as as, what is this in this summation tell you?

Student: (Refer Time: 11:10),

Covariance between.

Student: First and second

The first column and the second column is that clear now, is it with everyone now, fine. So, what you have here is actually the covariance matrix you seems to be lost is it with you sure fine.

(Refer Slide Time: 11:32)

$$\begin{aligned}
 e &= \sum_{i=1}^m (x_i - \hat{x}_i)^2 &= \sum_{i=1}^m \sum_{j=k+1}^n (p_j^T x_i)(x_i^T p_j) \\
 &= \sum_{i=1}^m \left(\sum_{j=1}^n \alpha_{ij} p_j - \sum_{j=1}^k \alpha_{ij} p_j \right)^2 &= \sum_{j=k+1}^n p_j^T \left(\sum_{i=1}^m x_i x_i^T \right) p_j \\
 &= \sum_{i=1}^m \left(\sum_{j=k+1}^n \alpha_{ij} p_j \right)^2 = \sum_{i=1}^m \left(\sum_{j=k+1}^n \alpha_{ij} p_j \right)^T \left(\sum_{j=k+1}^n \alpha_{ij} p_j \right) &= \sum_{j=k+1}^n p_j^T m C p_j \quad \left[\because \frac{1}{m} \sum_{i=1}^m x_i x_i^T = \frac{X^T X}{m} = C \right] \\
 &= \sum_{i=1}^m (\alpha_{ik+1} p_{k+1} + \alpha_{ik+2} p_{k+2} + \dots + \alpha_{in})^T (\alpha_{ik+1} p_{k+1} + \alpha_{ik+2} p_{k+2} + \dots + \alpha_{in}) \\
 &= \sum_{i=1}^m \sum_{j=k+1}^n \alpha_{ij} p_j^T p_j \alpha_{ij} + \sum_{i=1}^m \sum_{j=k+1}^n \sum_{L=k+1, L \neq j}^n \alpha_{ij} p_j^T p_L \alpha_{iL} \\
 &= \sum_{i=1}^m \sum_{j=k+1}^n \alpha_{ij}^2 \quad (\because p_j^T p_j = 1, p_i^T p_j = 0 \quad \forall i \neq j) \\
 &= \sum_{i=1}^m \sum_{j=k+1}^n (x_i^T p_j)^2
 \end{aligned}$$

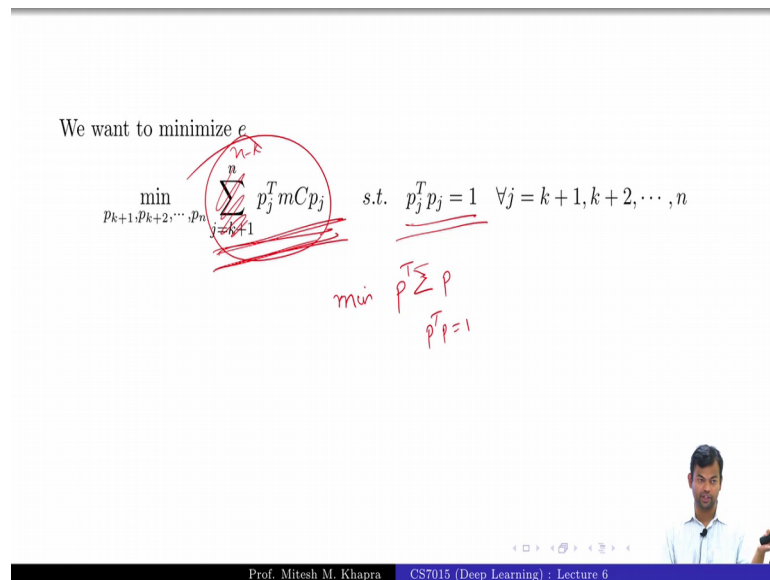
So, what we have here is something of this form ok.

(Refer Slide Time: 11:37)

We want to minimize e

$$\min_{p_{k+1}, p_{k+2}, \dots, p_n} \sum_{j=k+1}^n p_j^T m C p_j \quad \text{s.t.} \quad p_j^T p_j = 1 \quad \forall j = k+1, k+2, \dots, n$$

Handwritten notes:
 $\min p^T \Sigma p$
 $p^T p = 1$



So now what we want to do is we want to minimize this quantity subject to the following condition is that ok. What is the solution for this? If I did not have the summation ok; Suppose I just wanted one dimension. So, I want to minimize say $p^T \Sigma p$ such that $p^T p = 1$, what is the solution for this?

Student: (Refer Time: 12:08).

Smallest eigenvalue of Σ ; and you can show by induction that if you want k such things that here I am looking for $n - k$ such things right. Then these would be the $n - k$ smallest eigenvalues of Σ , but now I am talking about the smallest eigenvalues, but in the first solution I said we need to pick the largest eigenvalues. So, what is the difference?

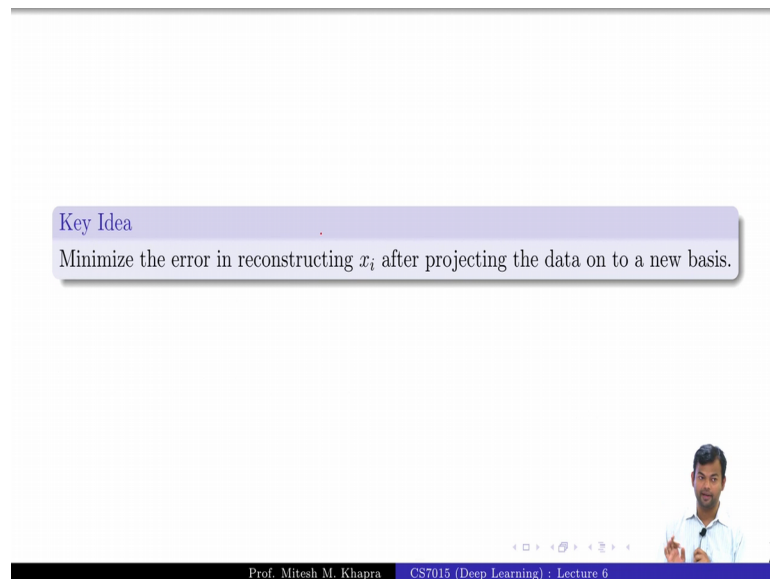
Student: (Refer Time: 12:35).

These are the ones we are throwing away; these are the ones along which the error is going to be minimum if we throw these away the error is going to be minimum. So, we will throw away the last $n - k$ dimensions which means we'll keep the first k dimensions is that clear. So, you arrived at the same solution is that right so; that means, in PCA you are actually trying to pick the dimensions in a way such that your reconstruction error is minimized, and this was exactly what our reconstruction error

was. So, do not worry about this math bit, just see that we started with this quantity this is what we wanted to minimize ok.

And we did some trickery and we came to this formula that minimizing that error is equivalent to minimizing this quantity. And for this we know the solution that the solution is the smallest eigenvalue and we want $n - k$ such things. That means, there would be the $n - k$ smallest eigenvectors is that clear; that means, we are going to keep only the k largest eigenvectors ok; that means, you are going to project your data on to k largest eigenvectors.

(Refer Slide Time: 13:37)



Key Idea

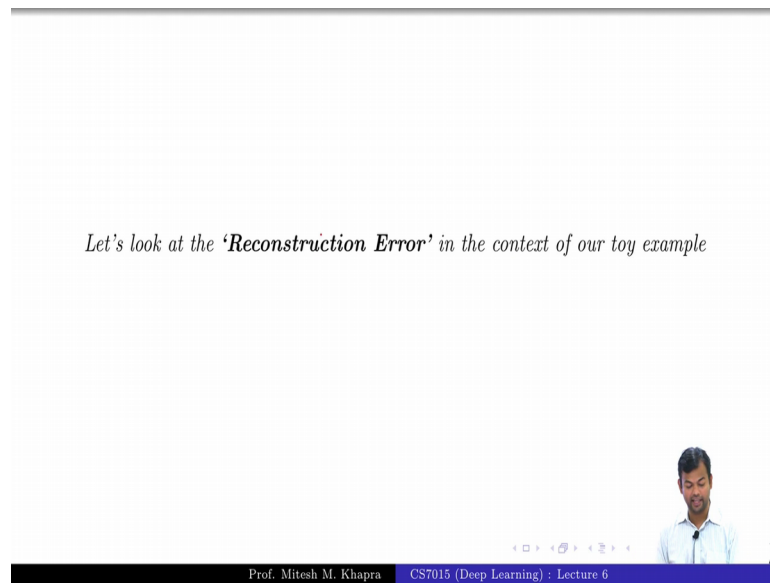
Minimize the error in reconstructing x_i after projecting the data on to a new basis.

Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6

Now, so the key idea here is this right minimize the error in reconstructing x_i after projecting the data onto the new basis.

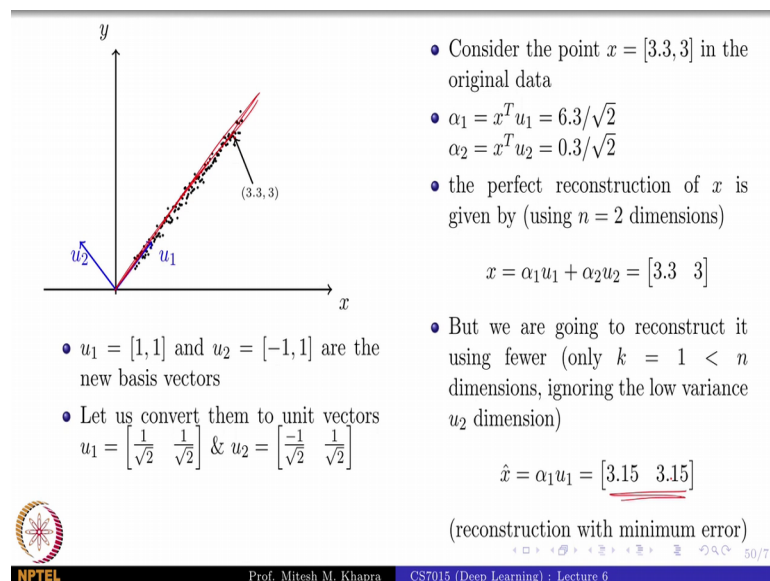
(Refer Slide Time: 13:43)

Let's look at the 'Reconstruction Error' in the context of our toy example



So, let us take an example and we will work with our toy example again.

(Refer Slide Time: 13:45)



- Consider the point $x = [3.3, 3]$ in the original data
- $\alpha_1 = x^T u_1 = 6.3/\sqrt{2}$
 $\alpha_2 = x^T u_2 = 0.3/\sqrt{2}$
- the perfect reconstruction of x is given by (using $n = 2$ dimensions)
$$x = \alpha_1 u_1 + \alpha_2 u_2 = [3.3 \quad 3]$$
- But we are going to reconstruct it using fewer (only $k = 1 < n$ dimensions, ignoring the low variance u_2 dimension)
$$\hat{x} = \alpha_1 u_1 = [3.15 \quad 3.15]$$

(reconstruction with minimum error)

NPTEL Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6 50/71

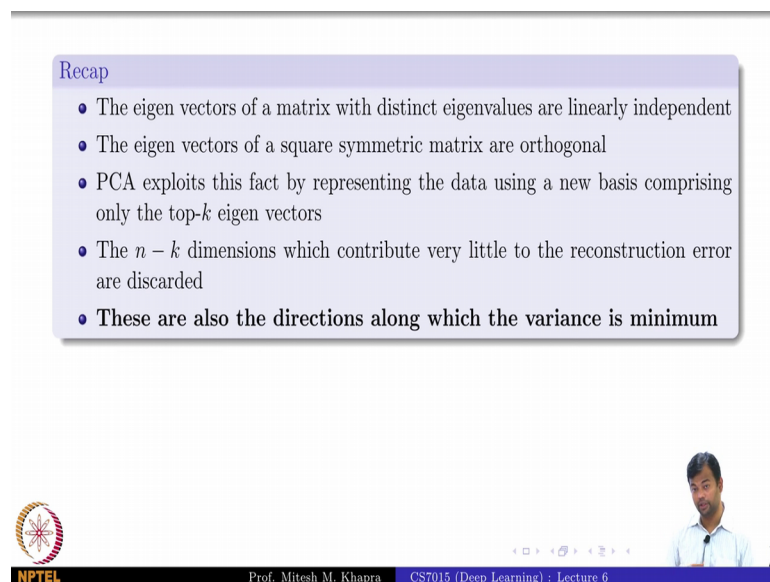
So, this was the data that we had and suppose I give you a new basis which is 1 comma 1 and minus 1 comma 1 ok. This is a new basis this is an orthonormal basis orthogonal basis you can see that $u_1^T u_2$ is equal to 0 ok.

Now, I need convert it to an orthonormal basis. So, I have just divided by the magnitude is it fine. Now consider the point 3.3 comma 3, this was our original point according to which coordinate axis x comma y ; that means, this was 3.3 and this was 3 ok. Now I can

find the alpha is right because this is an orthonormal basis I can directly find the alpha is, now the perfect reconstruction would be this. So, actually if I do this I get back the original point.

Now, what would happen if I throw away the second dimension, because the second dimension had corresponds to a smaller eigenvalue I will get this. So, you see that the point is still close to the original point I have not actually lost much right. What has happened is I have actually projected the boy lie point on this line right, the line x equal to y that is, why I get x equal to y . And in doing that I am not losing much information from the original data is this clear right. So, you understand what happens when you reconstruct the data fine.

(Refer Slide Time: 15:18)



Recap

- The eigen vectors of a matrix with distinct eigenvalues are linearly independent
- The eigen vectors of a square symmetric matrix are orthogonal
- PCA exploits this fact by representing the data using a new basis comprising only the top- k eigen vectors
- The $n - k$ dimensions which contribute very little to the reconstruction error are discarded
- These are also the directions along which the variance is minimum

NPTEL Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6

There is no end to this ok. So, just to recap the eigenvectors of a matrix with distinct eigenvalues are linearly independent. And we use this fact conveniently at least in the case of square matrix where the also happen to be orthogonal. So, we know that they can form a very convenient basis and PCA exploits this to find the top k eigenvectors which to be retained.

And while doing this they have seen that two things are at least ensured one the covariance between the dimensions is 0 because that is exactly how we formulated it and found the solution. We saw that it turns out that we need to diagonalize a certain matrix and the solution is the eigenvectors.

We also saw a different interpretation where we saw that it is the same as throwing away the dimensions along which the error would be minimum right. And both these interpretations led to the same solution which was project the data onto the eigenvectors of the covariance matrix of the original data. And this n minus k dimensions current contribute very little to the reconstruction, now what is the one thing which I have not proved yet? What was our wishlist?

Student: Variance and covariance.

Variance and covariance right high variance low covariance. I proved low covariance, I have also proved something with respect to reconstruction error because that is something I require for auto encoders. So, just remember this bit about reconstruction error.