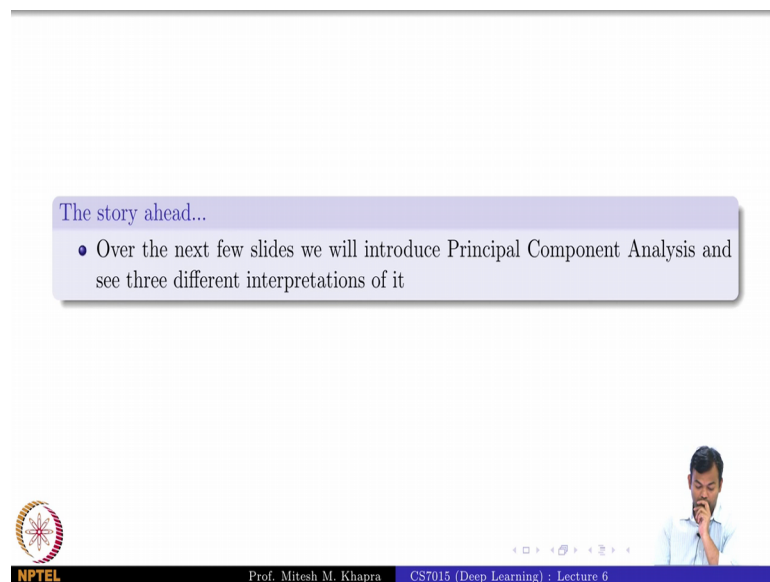


Deep Learning
Prof. Mithesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module - 6.4
Lecture – 06
Principal Component Analysis and its Interpretations

So, in this module we will talk about Principle Component Analysis and its different Interpretations. In this model we will look at one interpretation and then in the rest of the module some other interpretations.

(Refer Slide Time: 00:25)



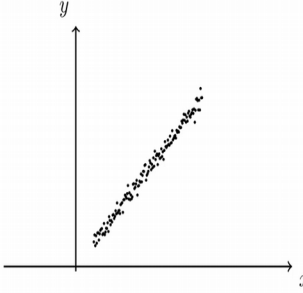
The story ahead...

- Over the next few slides we will introduce Principal Component Analysis and see three different interpretations of it



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

So, the story I add is going to be this we will talk about PCA and its Interpretations, ok.

(Refer Slide Time: 00:30)



- Consider the following data
- Each point (vector) here is represented using a linear combination of the x and y axes (i.e. using the point's x and y co-ordinates)
- In other words we are using x and y as the basis
- What if we choose a different basis?



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

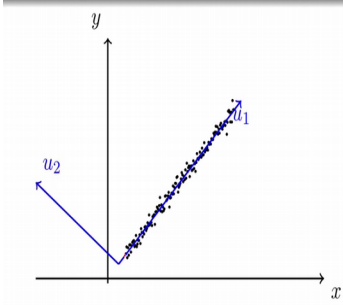
So now let us try to motivate PCA first consider the following data ok. In what dimension is this data?

Student: 2 dimension.

2 dimensions it is \mathbb{R}^2 ok. And each point here is represented as it is x coordinate and using it is x coordinate and it is y coordinate. Now it means that were using x and y as the basis right. That is clear that is the standard way that you would do any data point you will just represent using that basis.

Now, what if we choose a different basis? Let me give you 1 basis and then let me ask you some questions on this.

(Refer Slide Time: 01:05)



- For example, what if we use u_1 and u_2 as a basis instead of x and y .
- We observe that all the points have a very small component in the direction of u_2 (almost noise)
- It seems that the same data which was originally in $\mathbb{R}^2(x, y)$ can now be represented in $\mathbb{R}^1(u_1)$ by making a smarter choice for the basis

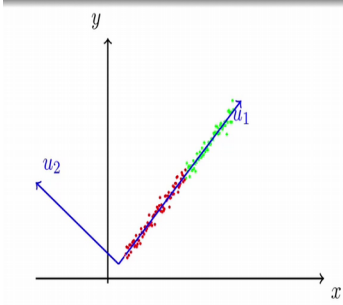
Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6

Suppose we chose this basis. So, in the previous modules we made a case for the x and y coordinate axis, there is nothing sacrosanct about it you could use any basis; The only condition on the basis that the vector should be linearly independent and in fact, if they are orthogonal it is even better right.

So now I have given you a different basis now what do you make any observation here? So, they have all the points here have a very small component along the u_2 axis right. So now, this so far this point right, if I consider at this point then this is the component along the u_1 axis. So, that is it is u_1 coordinate as akin to the x coordinate and this is it is u_2 coordinate akin to the y coordinate ok, is a are the arrows clear here, ok.

So; that means, there u_2 coordinate is very small, and it is also very small for all the data points right. So, it is almost as if there is some noise there it is all within some epsilon. Now so it seems that the data which were actually represented in \mathbb{R}^2 can actually be represented in \mathbb{R}^1 by getting rid of this noisy dimension right. So, if you had chosen a different basis, you realize that with just one dimension you could have captured everything that was there in the data, and the other dimension was just adding noise it was redundant there is hardly any information there, ok.

(Refer Slide Time: 02:24)



- Let's try stating this more formally
- Why do we not care about u_2 ?
- Because the variance in the data in this direction is very small (all data points have almost the same value in the u_2 direction)
- If we were to build a classifier on top of this data then u_2 would not contribute to the classifier as the points are not distinguishable along this direction

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

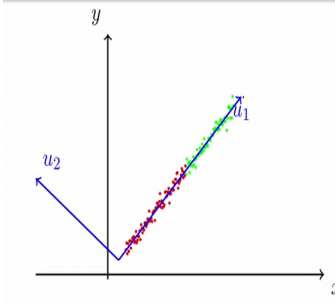
So now can you state this more formally because this is this intuition, but can you state more formally in terms of things that you have learned and say probability for it for example, what is wrong with the direction u_2 ? The spread of the data points along the u_2 direction is very small, what is the spread mean the variance right. So, we do not care about u_2 because the variance in the data along this direction is very, very small ok. And in particular right if I were to build a classifier, then would u_2 have any predictive power? Because along this dimension the points are indistinguishable ok; so, think of it that you are trying to find out whether you have. So, you have say 100 candidates and you want to decide whether they would be good basketball players or not right.

And quite naturally all the people that have shown up are say 6 foot 2 and 6 foot 3 inch and so on; and there is a very small height difference between them and all of them are 6.2 is the average and very close the spread is not much. So, this feature is not going to help you decide whether this person is going to be a good basketball player or not. You will have to rely on other features where the variance is more for example, how many teams has he participated in the past, how many matches as he won as a team as a member of some team and so on it.

So, those who expect some spread to be there all these 100 candidates might have different things right? But if the height is the same for all of them it is not going to be a

good predictor. And that is exactly what is happening along the u_2 direction. The points are almost indistinguishable there that is why; it does not matter.

(Refer Slide Time: 03:55)



- In general, we are interested in representing the data using fewer dimensions such that the data has high variance along these dimensions
- Is that all?
- No, there is something else that we desire. Let's see what.

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

So, in general given any data, now this was a simple case where the data was \mathbb{R}^2 I am talking about the general case where the data is \mathbb{R}^n right, and you will find this situation in higher dimensions also. So, you would not want to use that entire n dimensional data where you know that there are some columns, along with the variance is very small. So, you want to represent the data with fewer dimensions, such that; the data has high variance along those dimensions, ok.

Now, let me just clear a confusion here right. So, I am not saying that take your n dimensional data ok. Find the variance across each of these dimensions and then throw away the columns which have the lowest dimension, in this particular example if you had done this, what would happen? Could you have done that think of the original dimension's x and y . Along, these two dimensions there is enough variance in the data, right? The x coordinates vary from here to here. And the y coordinates also vary from this point right up to that point right? So, there is enough spread in the x and y coordinates.

So, in your original data I am not saying that pick look at each column, and see if there is no variance along that column then throw it away that would not work because you might end up with the situation that there is enough variance across each of these

dimensions. It is just that when you look at the data from a different angle; that means, you projected onto a different basis this becomes clear right?


So, you see the difference I that is not the same these two things are different operations. So, what I am looking at is projecting the data to a different basis, that is exactly what I did with u_1 and u_2 . And then some things became clear about the data. Now this projection along a different basis, I would be interested in doing that only if; I can get rid of the number of dimensions right? If now I had already had one basis where I had n dimensions. Now if the new basis is also going to be that all these new n dimensions that I have come up with are important then you are not gaining much, I do still have this high dimensional data, but you would like to project it in a way that you get rid of the lower variance dimensions.

So, you might project it onto n dimensions, but you want to rank these dimensions according to variance and then throw away some of these dimensions, is that clear; is the objective clear? Ok fine. Is that all that we care about; n dimensions' project to a new basis and throw away the key dimensions which have less variance, is that all? What else would you want? People have done the MLPR course, no I would not. So, I am not going to classification or anything I just want a better representation of the data at this point. I am not really thinking about what I want to do with the data, maybe you are talking in terms of classification. And we have already seen even if the data is not linearly separable we have solutions for dealing with right. So, that is not a critical point ok. So, there is something else that very interested in and let us look at that ok.

(Refer Slide Time: 06:49)

x	y	z
1	1	1
0.5	0	0
0.25	1	1
0.35	1.5	1.5
0.45	1	1
0.57	2	2.1
0.62	1.1	1
0.73	0.75	0.76
0.72	0.86	0.87

$$\rho_{yz} = \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (z_i - \bar{z})^2}}$$



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

- Consider the following data
- Is z adding any new information beyond what is already contained in y ?
- The two columns are highly correlated (or they have a high covariance)
- In other words the column z is redundant since it is linearly dependent on y .

Now, consider this data I have 3 dimensional data ok. Do you find something odd about this data?

Student: (Refer Time: 06:56).

y and z are.

Student: (Refer Time: 07:00).

Are highly dash.

Student: Correlated.

Correlated right do you want these dimensions? Can you think of any practice such dimensions occurring? Height in centimeter and height in inches, someone would have just given you data right? Or if you if you take the credit card a credit card fraud detection case right? Someone would give you the salary and it would also give you the income tax now these 2 are highly correlated right?

So, then you do not really care if you have one you could probably almost with certainty predict the other right? Modulus some rules right because you get some tax exemptions and all that, but still. So, you can have this in practice, but even in our oil mining case your salinity pressure density those things could be related right? So, z is not adding any new information beyond what y is happening. So, the 2 columns are highly correlated.

So, actually yeah this is the formula for correlation, all of you know this anyone who does not know this formula good. So, not nothing is a stupid question right. So, you can always ask.

So, \hat{y} is the mean of this column a , sorry \bar{y} ; \bar{z} is the mean of this column and this is how you compute correlation this is just the formula ok. So, from every entry you subtract the mean ok. So, this is known as centering the data. So, if you do this what would the mean of the new data be?

Student: 0.

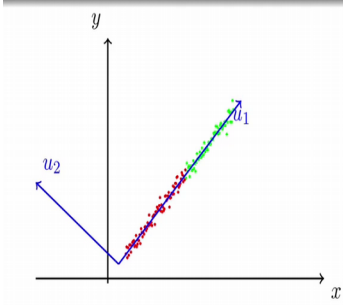
0 right? So, that is why it is called centering the data ok. So, I will have 0 mean 0 mean and you. So, what does this; what is the intuition behind this formula? Does anyone know? Can anyone tell me? So, this is a summation ok. So, this quantity is going to be high if the summation is high. It is a summation of some n terms now these terms could be positive or negative; If all the terms are positive what would we happen to the sum?

Student: (Refer Time: 08:50).

It would be high if there are some terms which are negative it would be low. Now when would all the terms be positive whenever y is above the mean, z is also above the mean right therefore, this quantity is positive this quantity is positive. Whenever both are below the mean again the product would be positive. When one is above the mean the other is below the mean, then there is something wrong happening right and in that case you will have a negative term right? So, you get the intuition fine

So, for more details of course, you can refer your other textbooks and so on, but this is just the intuition an important step here is to 0 mean the data right. We are computing the subtracting the mean of the data. Another way of saying this is that the column z is actually linearly dependent on y ok. It is almost linearly dependent I of course, have some noise 2.1, 0.76 and so on, but it is largely linearly dependent, I can get I can write z as some c times x fine.

(Refer Slide Time: 09:48)



In general, we are interested in representing the data using fewer dimensions such that

- the data has high variance along these dimensions
- the dimensions are linearly independent (uncorrelated)
- (even better if they are orthogonal because that is a very convenient basis)

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

So now can you tell me the refined goals that we have? We are interesting the representing data using fewer dimensions such that; remember that when I say fewer dimensions I mean a new set of dimensions right? It is not throwing away dimensions from the current data. We are looking for a new set of dimensions. What are the conditions that we want from these new set of dimensions?

Student: (Refer Time: 10:12).

One there should be high variance along these dimensions the new dimensions, and?

Student: (Refer Time: 10:15).

The dependence are linearly independent or uncorrelated fine.

And even better of course, if they are orthogonal, why?

Student: (Refer Time: 10:26).

Because we are looking for a new dash.

Student: Basis.

Basis and the most convenient basis is.

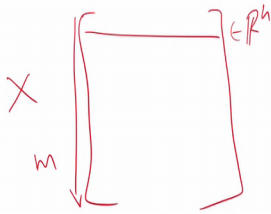
Student: Orthogonal basis.

Orthogonal basis, ok, fine.

(Refer Slide Time: 10:33)

Let p_1, p_2, \dots, p_n be a set of such n linearly independent orthonormal vectors. Let P be a $n \times n$ matrix such that p_1, p_2, \dots, p_n are the columns of P .

Let $x_1, x_2, \dots, x_m \in \mathbb{R}^n$ be m data points and let X be a matrix such that x_1, x_2, \dots, x_m are the rows of this matrix. Further let us assume that the data is 0-mean and unit variance.



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

So, now let us assume someone has given us this new basis ok. And let us call this p_1, p_2, \dots, p_n . So, instead of this x, y, z and so on, someone has given us this new basis eventually we will of course, figure out how to find the basis, but let us assume that someone has given this new basis right. And they are both linearly independent and actually it is redundant actually. So, yeah this example of a redundant feature such as, orthogonal vectors is sufficient, they are linearly independent.

Let P be an $n \times n$ matrix such that p_1, p_2, \dots, p_n are the columns of P right; Same thing as we had put the Eigen vectors in a column and probably I have unknowingly given out the solution, but ok. And let x_1 to x_m be the m data points given to us ok. So, we are given this data as usual we have this X matrix each one of them belongs to \mathbb{R}^n . And we have m such data points right that is the standard thing that we are operating, and you always write this as a matrix and we have already done the data is 0 mean and unit variance.

Actually unit variance is not required, but the data is 0 mean fine that we will sorry I am going to deal with covariance, as a unit variance is not required. So, the data is 0 mean is what I am going to assume, but what if the data is nonzero mean I can always make it 0, right.

So, if you have any basis any vector you can write it as a linear combination of that basis, is it fine? So far it is ok, ok. Now for an orthogonal basis we know that we can compute these alphas just by taking a dot product of the vector with the dimension ok, and just repeating some of the things right fine.

(Refer Slide Time: 13:25)

In general, the transformed data \hat{x}_i is given by

$$\hat{x}_i = \left[\leftarrow x_i^T \rightarrow \right] \begin{bmatrix} \uparrow \\ p_1 \\ \downarrow \end{bmatrix} \cdots \begin{bmatrix} \uparrow \\ p_n \\ \downarrow \end{bmatrix} = \underline{x_i^T P} \quad \text{1xn}$$

$\alpha_{i1} \dots \alpha_{in}$

So now let us see what this means; for one of the dimensions this is my data point x_i which I want to transform, for one of the dimensions I just had to take the dot product with that dimension and this will give me how many values; One value; that means, the coordinate along p_1 . I want to do it for all the n of them I can write it as this vector matrix multiplication right, what is the dimension of this?

N cross 1 , how many if you get that? Ok so this oh not many why?

Student: (Refer Time: 13:55)

1 cross n fine that is fine yeah how many of you get this; Ok fine yeah. So, this will give me all the n alphas is that clear for this data point, ok.

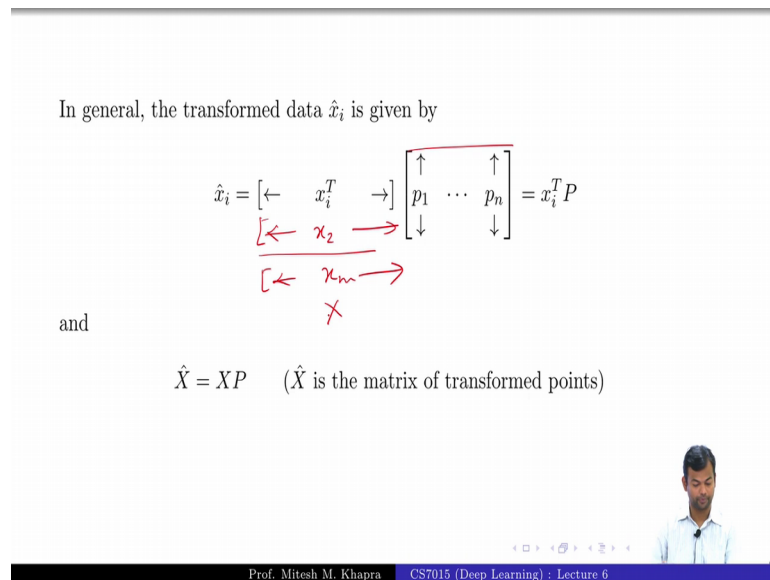
So, it will give me α_{i1} to α_{in} is it ok.

(Refer Slide Time: 14:11)

In general, the transformed data \hat{x}_i is given by

$$\hat{x}_i = \begin{bmatrix} \leftarrow & x_i^T & \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & & \uparrow \\ p_1 & \cdots & p_n \\ \downarrow & & \downarrow \end{bmatrix} = x_i^T P$$

and

$$\hat{X} = XP \quad (\hat{X} \text{ is the matrix of transformed points})$$


Now, I want to do this for the entire data right. So, I have done it for x_1 I also want it to be done for x_2 and all the way up to x_m , for each of these I would have such an operation where I have a vector multiplied by this matrix. If I just stack all these vectors I get back my matrix X . And the whole operation I can write as X into P is that clear to everyone ok, what is the dimension of X into P ?

Student: m cross n .

m cross

Student: n .

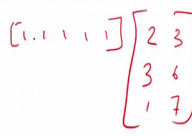
n right so, for all the m data points I have α_1 to α_n is that clear anyone who does not understand this? Ok fine.

So, \hat{X} hat is the matrix of the transformed points is that clear? I have now the new coordinates instead of the original coordinates according to the coordinate axis. I have the new coordinates in this matrix, ok.


(Refer Slide Time: 15:09)

Theorem:
If X is a matrix such that its columns have zero mean and if $\hat{X} = XP$ then the columns of \hat{X} will also have zero mean.

Proof: For any matrix A , $\mathbf{1}^T A$ gives us a row vector with the i^{th} element containing the sum of the i^{th} column of A . (this is easy to see using the row-column picture of matrix multiplication).



$$[1 \ 1 \ 1 \ 1] \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} = 6$$



Now I will just go through some very simple theorems or rather results. And I will not prove them you can prove them on your own, or other proof is there in the slides we can look at it later on right. So, if X is a matrix such that its columns have 0 mean and if \hat{X} is equal to XP , then the columns of \hat{X} will also have 0 mean. Is this obvious to most of you, not really is it, how many of you think it is obvious? Ok then let me just go over the proof.

So, for any matrix A , $\mathbf{1}^T A$ right so; that means, you have this vector this is a vector or a matrix; yeah this is a vector right. So, I have a vector of n 1s. So, $\mathbf{1}^T A$ this is nothing but a vector of n 1s. So, what is this product actually going to give me?

It will give me a vector containing n elements, what is each element?


Student: Sum of that column.

Sum of that column right, is this fine? Ok this is very obvious to see from if I have this suppose I have $\begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 3 \\ 6 \\ 7 \end{bmatrix}$ ok, and then of course, the corresponding.

(Refer Slide Time: 16:21)

Theorem:
If X is a matrix such that its columns have zero mean and if $\hat{X} = XP$ then the columns of \hat{X} will also have zero mean.

Proof: For any matrix A , $\mathbf{1}^T A$ gives us a row vector with the i^{th} element containing the sum of the i^{th} column of A . (this is easy to see using the row-column picture of matrix multiplication).

$$\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 3 & 6 \\ 1 & 7 \end{bmatrix} = \begin{bmatrix} 7 & 16 \end{bmatrix}$$


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

So, if I do this multiplication I will get a 2 dimensional output which would be just 7 and 16 right? So, that is just the sum of that column this is fine.

Student: (Refer Time: 16:33)

(Refer Slide Time: 16:36)

Theorem:
If X is a matrix such that its columns have zero mean and if $\hat{X} = XP$ then the columns of \hat{X} will also have zero mean.


Proof: For any matrix A , $\mathbf{1}^T A$ gives us a row vector with the i^{th} element containing the sum of the i^{th} column of A . (this is easy to see using the row-column picture of matrix multiplication).
Consider

$$\mathbf{1}^T \hat{X} = \mathbf{1}^T X P = (\mathbf{1}^T X) P$$

But $\mathbf{1}^T X$ is the row vector containing the sums of the columns of X . Thus $\mathbf{1}^T X = 0$. Therefore, $\mathbf{1}^T \hat{X} = 0$.
Hence the transformed matrix also has columns with sum = 0.

Theorem:
 $X^T X$ is a symmetric matrix.

Proof: We can write $(X^T X)^T = X^T (X^T)^T = X^T X$



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


So now, we have this \hat{X} that is the transform matrix. Now let us see if I do this operation $\mathbf{1}^T \hat{X}$ what happens. I can write it as this I can club it as this, what is this? It will be all 0s because the original matrix was mean 0; that means, the sum of the elements of all the columns each column independently was 0; that is what this is going to be a 0

vector. So, 0 multiplied by any matrix is going to be 0. Now is it obvious I hope this is obvious $X^T X$ is a symmetric matrix, I still have the proof for that, ok.

(Refer Slide Time: 17:06)

Definition:
 If X is a matrix whose columns are zero mean then $\Sigma = \frac{1}{m} X^T X$ is the covariance matrix. In other words each entry Σ_{ij} stores the covariance between columns i and j of X .

Explanation: Let C be the covariance matrix of X . Let μ_i, μ_j denote the means of the i^{th} and j^{th} column of X respectively. Then by definition of covariance, we can write :

$$\begin{aligned}
 C_{ij} &= \frac{1}{m} \sum_{k=1}^m (X_{ki} - \mu_i)(X_{kj} - \mu_j) \\
 &= \frac{1}{m} \sum_{k=1}^m X_{ki} X_{kj} \quad (\because \mu_i = \mu_j = 0) \\
 &= \frac{1}{m} X_i^T X_j = \frac{1}{m} (X^T X)_{ij}
 \end{aligned}$$


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

Now, if x is a matrix whose columns are 0 mean. Then a matrix sigma which I am going to call as a covariance matrix, which is given by this is actually the covariance matrix. How many of you agree with this; how many of you have seen the covariance matrix before? Ok good. So, all of you agree that this is the covariance matrix if you do not please raise your hands; If you do not you will not understand the rest of the stuff now you have to be given the right in center, ok.

So, let us see be the covariance matrix of X . Now what is the covariance matrix actually first of all tell me that? If I say that I have an n cross n matrix x

(Refer Slide Time: 17:45)

Definition:
If X is a matrix whose columns are zero mean then $\Sigma = \frac{1}{m} X^T X$ is the covariance matrix. In other words each entry Σ_{ij} stores the covariance between columns i and j of X .

The slide contains handwritten notes in red ink. At the top, it says 'm x n'. Below that, it shows the formula for the covariance matrix entry: $\frac{1}{m} \sum_{i=1}^m (x_{i1} - \mu_1)(x_{i2} - \mu_2)$. To the right of this formula is a matrix representation of X with columns labeled $x_1, x_2, x_3, \dots, x_m$. The entries are $x_{11}, x_{21}, x_{31}, \dots, x_{12}, x_{22}, x_{32}, \dots, x_{13}, x_{23}, x_{33}, \dots, x_{2m}$. Below the formula, there is a boxed formula: $\frac{1}{m} \sum_{i=1}^m x_{i1} x_{i2}$. At the bottom of the slide, there is a video player interface showing a person's head and shoulders.

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

Let me not make it any cross n , let me make it m cross n ok. What does the covariance matrix actually capture; what is the dimension of the covariance matrix first of all?

Student: n cross n .

n cross n ok, and what does each entry of the covariance matrix capture the covariance between the i th column and the j th column.

Student: (Refer Time: 18:09).

So, the entry ij of the covariance matrix captures, the covariance between the i th column and the j th column is that fine. Now what is the formula for covariance suppose I give you 2 columns right let us see I have give you x_1 1×1 2×1 3 and x_2 1×2 2×2 3 , can you give me a formula and of course, I will go up to k or rather m , right.

So, what is the formula summation?

Student: (Refer Time: 18:52).

i equal to 1 to.

Student: m .

m .

Student: (Refer Time: 18:52).

Mu 1.

Student: (Refer Time: 18:53).

Mu 2 anything missing?

Student: By m.

By m anything else in the denominator? No, no is it fine ok. So, an what is mu 1? Mu 1 is just an average of this ok. So, this is the covariance formula now if the mus are 0, then what does this boil down to?

Student: (Refer Time: 19:14).

x_{1i} into x_{2i} , what is this quantity actually?

Student: (Refer Time: 19:22).

This is the dot product between the i th column and the j th column fine ok. Now that is pretty much the explanation right. So now, the C_{ij} th entry is supposed to be given by this formula. If the means are 0 you are just left with this formula. And this is nothing but the dot product between the i th row and the j th, I mean the i th column and the j th column is that fine ok.

And now if you write it as a matrix then you can just say that it is the ij th entry of the X transpose X matrix everyone gets this; no one has any confusion the people who raised their hands fine good.

(Refer Slide Time: 20:09)

$$\hat{X} = XP$$

- Using the previous theorem & definition, we get $\frac{1}{m}\hat{X}^T\hat{X}$ is the covariance matrix of the transformed data. We can write :

$$\frac{1}{m}\hat{X}^T\hat{X} = \frac{1}{m}(XP)^T XP = \frac{1}{m}P^T X^T X P = P^T \left(\frac{1}{m} X^T X \right) P = P^T \Sigma P$$

- Each cell i, j of the covariance matrix $\frac{1}{m}\hat{X}^T\hat{X}$ stores the covariance between columns i and j of \hat{X} .
- Ideally we want,

$$\begin{aligned} \left(\frac{1}{m}\hat{X}^T\hat{X} \right)_{ij} &= 0 & i \neq j \text{ (covariance = 0)} \\ \left(\frac{1}{m}\hat{X}^T\hat{X} \right)_{ij} &\neq 0 & i = j \text{ (variance } \neq 0 \text{)} \end{aligned}$$

In other words, we want

$$\frac{1}{m}\hat{X}^T\hat{X} = P^T \Sigma P = D \quad \text{[where D is a diagonal matrix]}$$

Prof. Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 6 41/71

So now ok. So, we now this is where the we are so far, that we have assumed that someone has given us these dimensions' p_1 to p_n , which we have put in the matrix p right. And we have also made a case that X into P which is what I have written here actually is just a projection of the original data onto this new basis right. Everyone gets that ok. And I am calling that new projection or the new result that I get as X hat. So, that is, what my transform data is.

What is missing here?

Student: (Refer Time: 20:42).

We do not know what p is that I am assuming someone has given me that P , now I need to figure out what is the P here. Now using the previous definition, we get that this is the covariance matrix of the transform data ok. So, let us just write that this is fine this is fine, what is this?

Student: (Refer Time: 21:03).

Covariance matrix of the original data ok; So, I will just write it as Σ fine ok. Now each cell ij of the covariance matrix towards the covariance between columns i and j of X hat, where X hat is the transformed data, what is the property that you want to hold. I give you 2 conditions or I will give you only one condition for now when i is not equal to j .

Student: (Refer Time: 21:28).

0 ok, so, what should the covariance matrix look like?

Student: (Refer Time: 21:37).

Remember that this is, what is this? This is the covariance matrix of the transformed data right that is what I started with right. This is the covariance matrix of this transformed data, what do I want this covariance matrix to look like?

Student: Diagonal matrix.

A diagonal matrix because I want every non diagonal element to be 0 right; And this point I am not telling you what I want the diagonal elements to be I am just telling you I do not want them to be 0.

Well if it is 0 what would that mean?

Student: (Refer Time: 22:05).

That is the variance right if you take the along a diagonal, what you get is the variance it is if it is not clear right now well return back to that. Right now we just know that the off diagonal elements are the covariance between the i th and j th column and we want that to be 0. So, we want this condition to hold. This is something very new that you have never seen in this course before they have actually not seen in this course before have you seen this or not?

Student: (Refer Time: 22:3).

Thank god fine.

So, what is this?

Student: Diagonalization.

The diagonalization of which matrix? This matrix right and what was this matrix it was X transpose X this is clear. So, what is the solution? All rows always lead to.

Student: Eigenvectors.

Eigenvectors, right.

(Refer Slide Time: 22:51)

• We want,

$$P^T \Sigma P = D$$

• But Σ is a square matrix and P is an orthogonal matrix



• Which orthogonal matrix satisfies the following condition?

$$P^T \Sigma P = D$$

• In other words, which orthogonal matrix P diagonalizes Σ ?

• **Answer:** A matrix P whose columns are the eigen vectors of $\Sigma = X^T X$ [By Eigen Value Decomposition]

• Thus, the new basis P used to transform X is the basis consisting of the eigen vectors of $X^T X$

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

So, we want $P^T \Sigma P$ to be a diagonal matrix and we know which are the set of vectors which I put in; P such that they will diagonalize Σ .

Student: Eigenvectors.

Eigenvectors of.

Student: (Refer Time: 23:06)

$X^T X$ right ok, why did I put this; it is the matrix of the eigenvectors right. So, it is a matrix of the eigenvectors of $X^T X$.

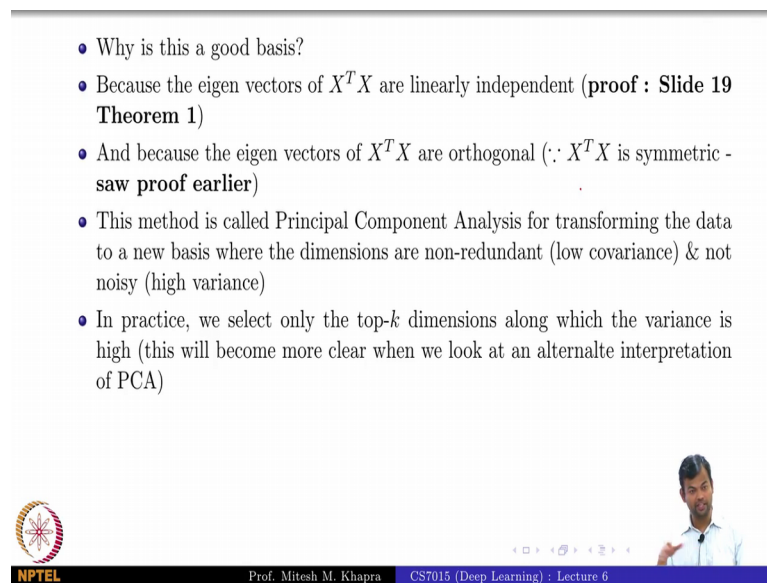
So now have we finished it, do we know principal component analysis now. So, we started with the intuition that we wanted to transform the data ok. I cannot stress enough that we want to transform the data not chopped off dimensions from the existing data ok; that means, we need to project the data to a new basis and we had a couple of conditions the variance should be high, and the covariance should be 0 we have satisfied one condition which is the covariance is 0. And we arrive at a solution which says that the eigenvectors forms the basis that you should project on; so, that the covariance would be 0 ok.

So, we have a solution, we know exactly which basis to use to represent the data ok; so, that the covariance condition is satisfied, what about the variance; did we do anything about the variance?

Student: (Refer Time: 24:10).

So we will come back to that, fine.

(Refer Slide Time: 24:08).



The slide contains the following text:

- Why is this a good basis?
- Because the eigen vectors of $X^T X$ are linearly independent (**proof : Slide 19 Theorem 1**)
- And because the eigen vectors of $X^T X$ are orthogonal ($\because X^T X$ is symmetric - saw **proof earlier**)
- This method is called Principal Component Analysis for transforming the data to a new basis where the dimensions are non-redundant (low covariance) & not noisy (high variance)
- In practice, we select only the top- k dimensions along which the variance is high (this will become more clear when we look at an alternate interpretation of PCA)

The slide also features the NPTEL logo, the name Prof. Mitesh M. Khapra, and the course information CS7015 (Deep Learning) : Lecture 6. A small video inset shows a man speaking.

Why is this a good basis; what does the what is a good basis, the best basis?

Student: Orthogonal.

Orthogonal right because the eigenvalues of X transpose X are linearly independent that is and they are also orthogonal because X transpose X is a dash matrix.

Student: (Refer Time: 24:24).

Ok.

Good real symmetric ok, good, ok.

So, this method is called the Principle Component Analysis for transforming a data to a new basis. And that where the dimensions are non-redundant because they have low k covariance and not noisy, because they have high variance. The second part I have not

proved right and I will get to that at some point fine. Ah No that is what we saw right no, what is I did not get that now in practice how many eigenvectors would you have.

Student: n eigenvector.

N eigenvectors, do you want to keep all of them? Which ones will you throw away.

Student: (Refer Time: 24:58)

The low variance ok.

And now in the next interpretation actually we will try to see, what is the; what happens when you throw away the least important dimensions right? What do you mean by the least important dimensions?