NPTEL

NPTEL ONLINE CERTIFICATION COURSE

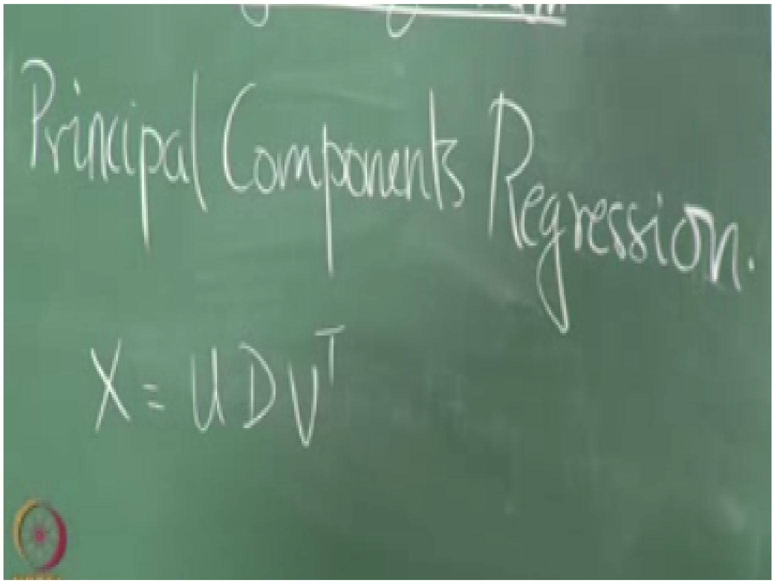
Introduction to Machine Learning

Lecture 17

Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

Principal Components Regression

(Refer Slide Time: 00:16)

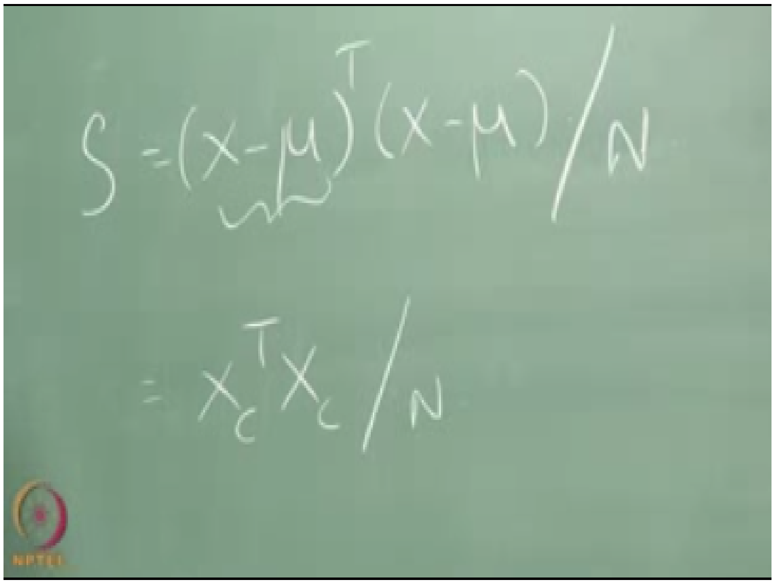


Principal Components Regression.

$X = UDV^T$

Right so D is a diagonal matrix right where the diagonal entries are your Eigen values if ideally or otherwise known as singular values right V is a V is at P x P matrix which has your eigenvectors and u the n x P matrix which typically spans your column space as x the same column space as x okay so this is essentially your singular value decomposition that we talked about so.

(Refer Slide Time: 01:07)

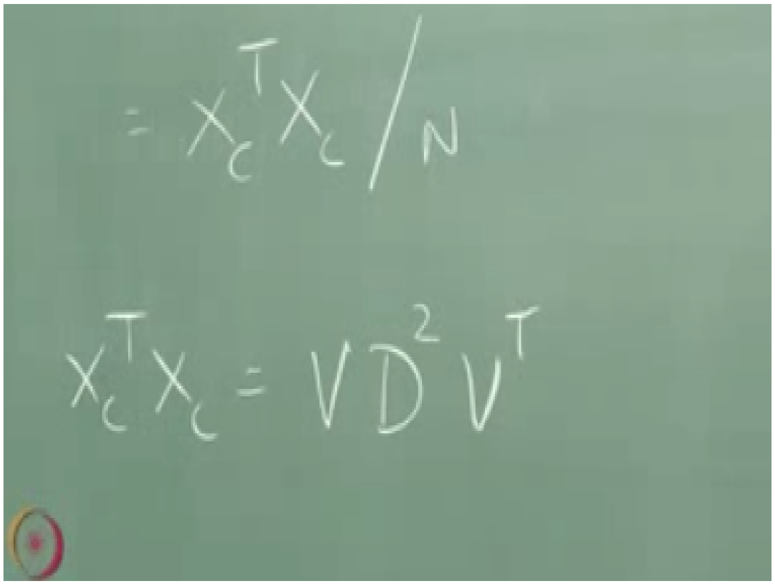


$$S = (\underbrace{x-\mu})^{T}(x-\mu)/N$$

$$= X_c^{T} X_c / N$$

So if you if you look at singular value decomposition or what is called the principal component analysis literature you will find the following you will find that they will talk about the covariance matrix yes okay what is the covariance matrix is a covariance matrix this is essentially if you think of whatever we have been doing, so far what would be this centered right it is centered so I take the centered data okay then this becomes this right so x tends to l then what I do is I find the Eigen decomposition of that. I find the Eigen decomposition of the covariance matrix right.

(Refer Slide Time: 02:16)

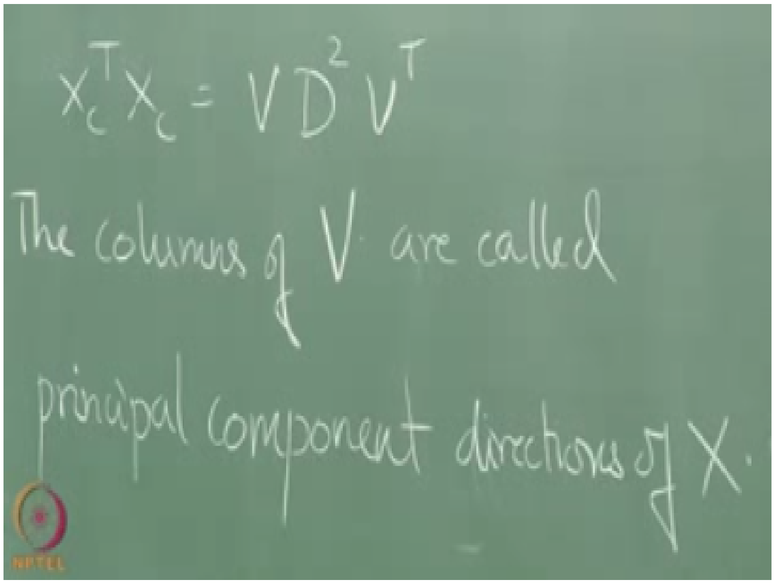


$$= X_c^T X_c / N$$

$$X_c^T X_c = V D^2 V^T$$

So I can essentially write this as so the same V and D that I wrote here assuming this was okay so if I take X_C so basically I am going to get the same thing right so it is essentially like doing singular value decomposition right and retrieving the V matrix right I am essentially taking the x^Tx which is the covariance matrix of the centered data okay and I am finding the Eigen valued composition of that so D^2 would be the Eigen vectors of x^Tx so this is standard self you should know okay.

(Refer Slide Time: 03:09)

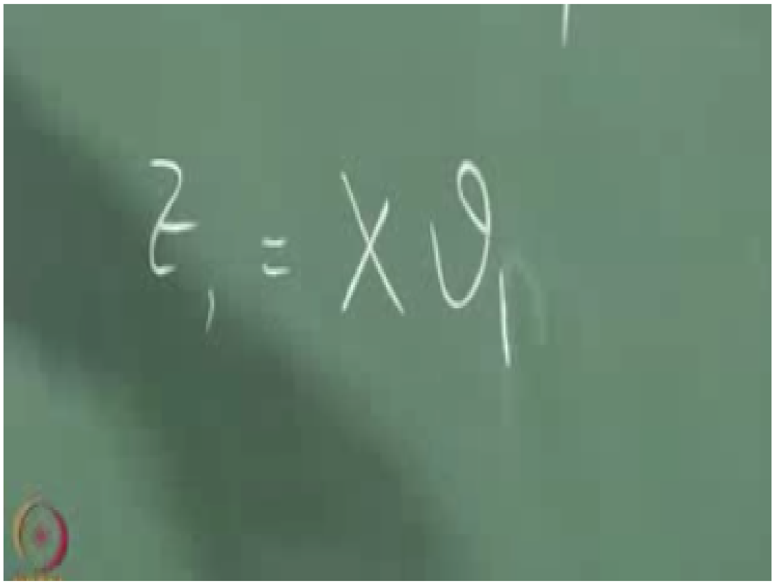


$$X_c^T X_c = V D^2 V^T$$

The columns of V are called principal component directions of X.

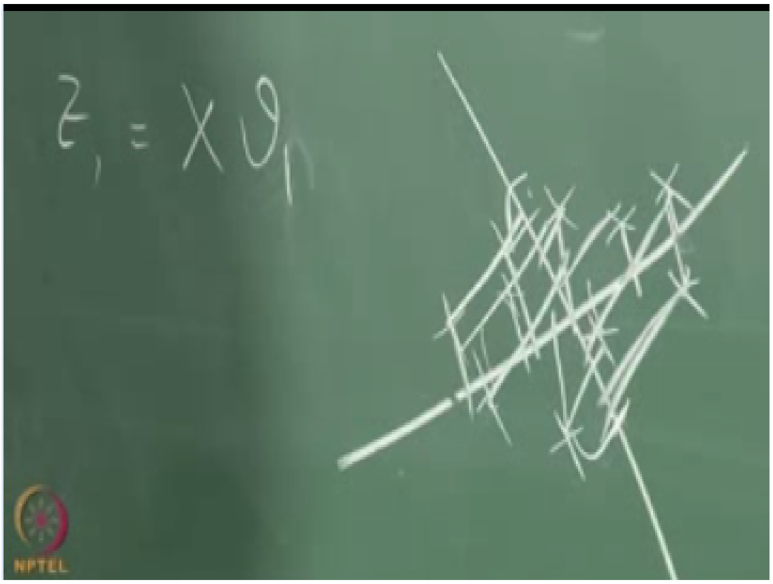
So the columns of so they are called the principal component directions of x. So there are a couple of nice things about the principal component directions, so we will talk about just one so I will actually come back to PCA slightly later right when I talk more about generally about feature selection not just in the context of regression but when I talk with generally about feature selection I will come back to PCA and tell you at least show you why PCA is good right now I will just tell you why PCA is good I will come back later and then I will show you why PCA is good right.

(Refer Slide Time: 04:33)



So suppose I take so where V_1 is the eigenvector corresponding to the first Eigen value right eigenvector corresponding to the first Eigen value so essentially what this means is I am projecting my data x on the first eigenvector direction okay so the resulting vector Z_1 okay will have the highest variance among all possible directions in which I can project x right.

(Refer Slide Time: 05:17)



So what does that mean right suppose this is x okay this is not x and y okay so it is a two-dimensional x this is x now I am claiming that V_1 will be such that when I project x onto V_1 I will have the maximum variance right, so in this case it will be some direction like this okay and projecting x onto this essentially means that right, so you can see that the data is pretty spread out it goes from here to here right on the other hand if I had taken a direction let us say that looks like that right.

So if I look at projection of the data right, so you can look at the spread it is a lot lesser in that direction than in the original direction I did the projection I know it looks pretty confusing to look at but the people can get my point right it is in the original direction that way the data was a lot more spread out as opposed to this direction where the data is lot more compact when I project it on to that direction.

So that is essentially what I am saying so z_1 right is essentially the projected data onto that direction onto x like z_1 actually has a highest variance among all the directions in which I can project the data right and consequently you can also show things like if I am looking to reconstruct the data original data and I say that you can only give me one coordinate right so you have to summarize the data in a single coordinate and now I am going to measure the data measure the error in reconstruction right. If you looked at it so the error in reconstruction would have been these bars that I did the projection over right.

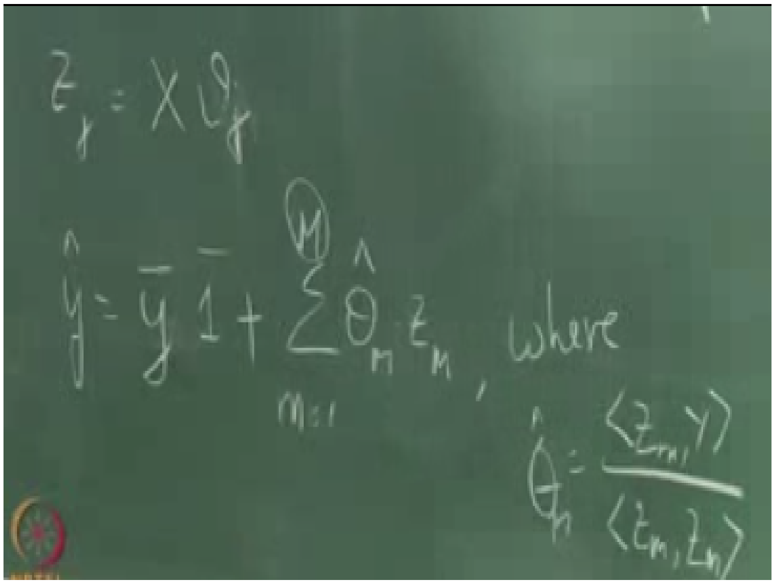
(Refer Slide Time: 07:40)

That would be the error in reconstruction, so I have the original data so that is the data so now I will give you this coordinates now I have to reconstruct the data right so essentially this will be the errors so the principal the first principal component direction the first principal component direction is the one that has the smallest reconstruction error first a principal comment direction will be the one that has the smallest reconstruction so we can show a lot of nice properties about this.

So I will actually come back and do this later when we talk about the general feature selection okay but here you can see the first thing you can see what each one write V_1 to V_P will be orthogonal right, so I have gotten my orthogonal directions right and the thing to notice is a lot of the variation in the data is explained by V_1 has the maximum variance likewise you take out V_1 right you take out V_1 so now what you have your data lies in some kind of a t - 1 dimensional space right and the direction in that the space which has the highest variance is P_2 it turns out that so V_1 has the highest variance over the data.

So in this space orthogonal to V_1 V_2 has the highest variance right in the space orthogonal to V_1 and V_2, V_3 will have the highest variance and so on so forth so essentially now what you can do is hey I am going to take all this directions one at a time right and I will do my regression right because each is orthogonal I can independently do the regression I can add the outputs and I can keep adding the dimensions until my residual becomes small enough that make sense so I will just keep adding this orthogonal dimensions until my residual becomes small enough at that point I stop.

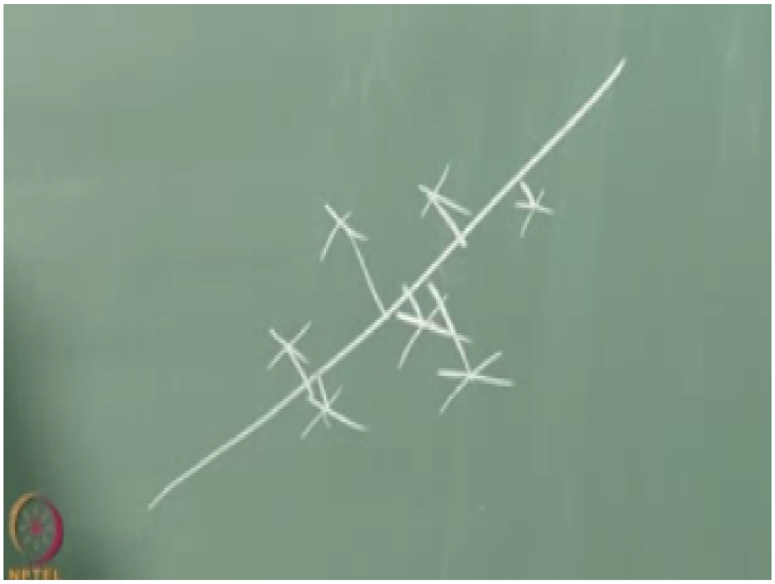
(Refer Slide Time: 09:44)


$z_j = X v_j$
$\hat{y} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m z_m$, where
$\hat{\theta}_m = \frac{\langle z_m, y \rangle}{\langle z_m, z_m \rangle}$

So this is essentially the idea behind so remember we are working with the center data right, so you automatically add in your intercept which is y bar the coefficient is y bar right and then your if you if you choose to take the first M principal components your thing will be θm ZM where ZM is given by this right and θ m is essentially regressing Y on ZM right so that is a univariate regression expression we know that well now so this gives you the principal component regression fit so one of the drawbacks of doing principal component regression is that I am only looking at the data the input right I am not looking at the output.

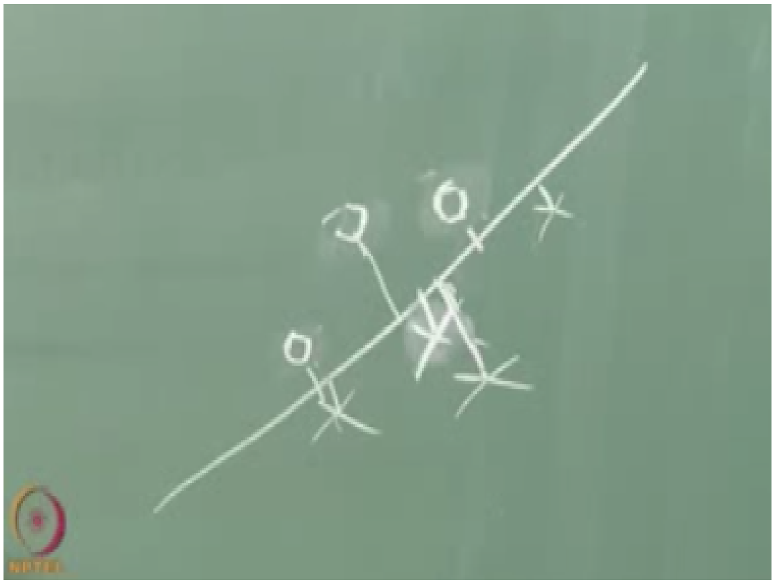
So it could very well be that once I consider what the output is right I might want to change the directions a little bit right, so I can give you an example is easier for me to draw if I think of classification.

(Refer Slide Time: 12:00)



Let us say this is the data and what would be the principal component direction you want to choose something like this right so that would be the ideal direction that you would want to choose okay so now what will happen the data will get we get projected like this right but suppose I tell you that.

(Refer Slide Time: 12:32)



Suppose I tell you that that is fine that these three were in a different class and if you want to think of it in terms of regression let us assume that these three have an output of -1 and these 4 have an output of +1 okay now if you think of this direction so the +1 and -1 are hopelessly mixed up right the +1 and -1 are hopelessly mixed up and I cannot I cannot draw is give a smooth prediction of which will be +1 which will be -1 on the other hand if you project onto a direction like this right the variance is small right I agree the variance is much smaller but if you think about it.

So all the -1 go to one side right all the +1 go to one side, so now if I want to do a prediction on this so it will be like okay this is this side is -1 and that side is +1 I can essentially do a fit like this which will give me a lot lesser error than the other case right so in cases where you are having an output that is specified for you already it might be beneficial to look at the output also when trying to delay directions as opposed to just looking at the input data so in classification you can see right in classification this will be say class 1 this will be class 2 and having this direction allows you to have a separating surface somewhere here right we talked about classification in the first class right.

So you just having a separating surface here will be great but in this case if I am projecting on this direction coming up with a linear separating surface is going to be hard everything gets completely mixed up.

IIT Madras Production

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

www.nptel.ac.in