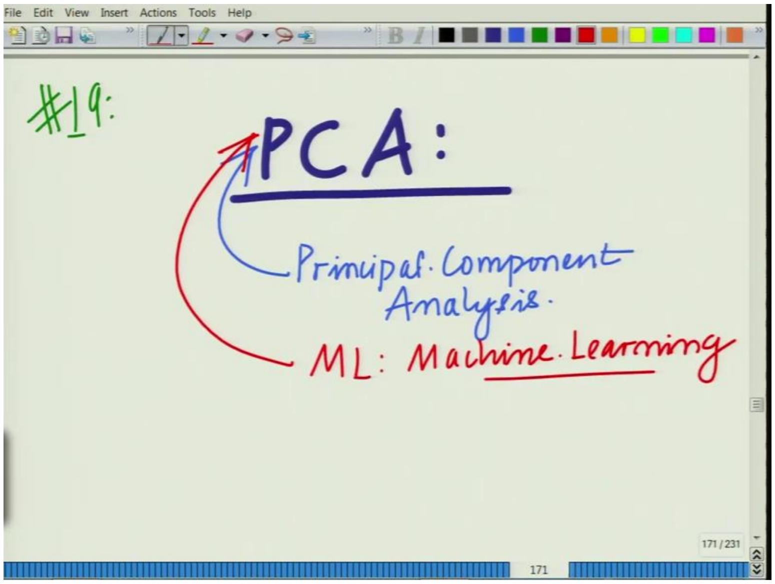
Applied Linear Algebra for Signal Processing, Data Analytics and Machine Learning
Professor. Aditya K Jagannatham
Department of Electrical Engineering
Indian Institute of Technology, Kanpur
Lecture No. 19
Machine Learning Application: Principal Component Analysis (PCA)

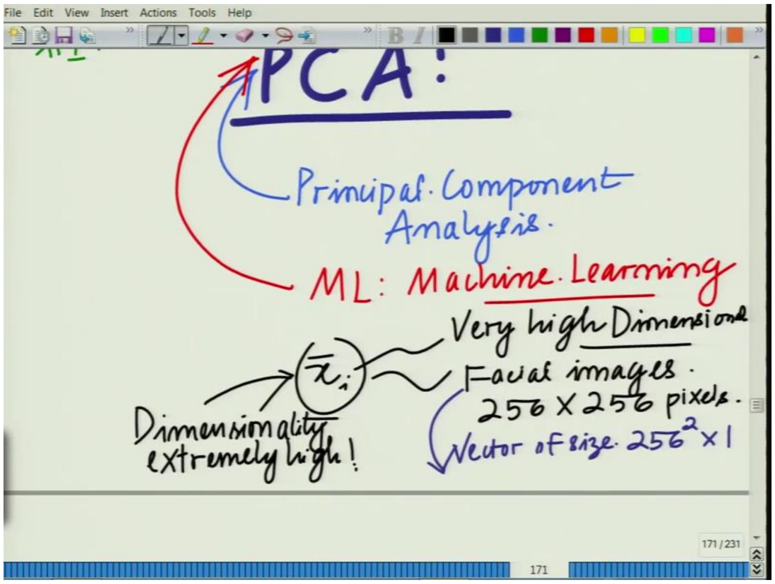
Hello, welcome to the another module in this massive online course. So, in this module let us look at a very important and interesting application of the concept of positive semi definite matrices and eigen values and that is in the context of PCA or Principal Component Analysis which is a very important concept in machine learning.

(Refer Slide Time: 0:45)



So, we want to look at the application of positive semi definite says and their Eigen vectors. This is termed as PCA which is Principal stands for Principal Component Analysis, stands for Principal Component Analysis. And this is one of the most important concepts in ML that is Machine Learning. We already seen an application in machine learning that is using the Gaussian classifier. This is another important application of linear algebra that is principal component analysis. And what this is roughly?

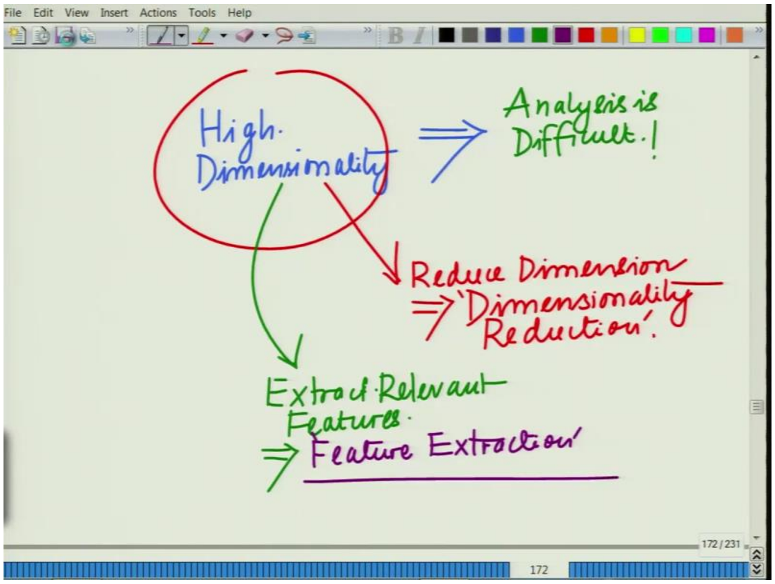
(Refer Slide Time: 1:51)



Essentially when you have the data vectors x bar i which are very high dimensional. For example, one can think of facial images. So, let us say you have facial images for instances 256 cross evens you can consider a typical image of size 256 cross 256 pixels and which is basically an image of size 256 cross 256. So, if you make this a vector, then it will be a vector of say, so from this if this facial image if you make this a vector of pixels, this will be vector of size 256 square cross 1.

So, if you take the pixels of the image put them column wise, this 256 cross 256 image it will be vector of size 256 square cross 1 that is 2 to the power of 8 square, 2 to the power of 16 cross 1. So, you can see it is a very large vector which means the data is going to be its multi-dimensional and dimensionality is very high. So, if you look at this machine learning application. So, the dimensionality can be very high, not just very high it is extremely high.

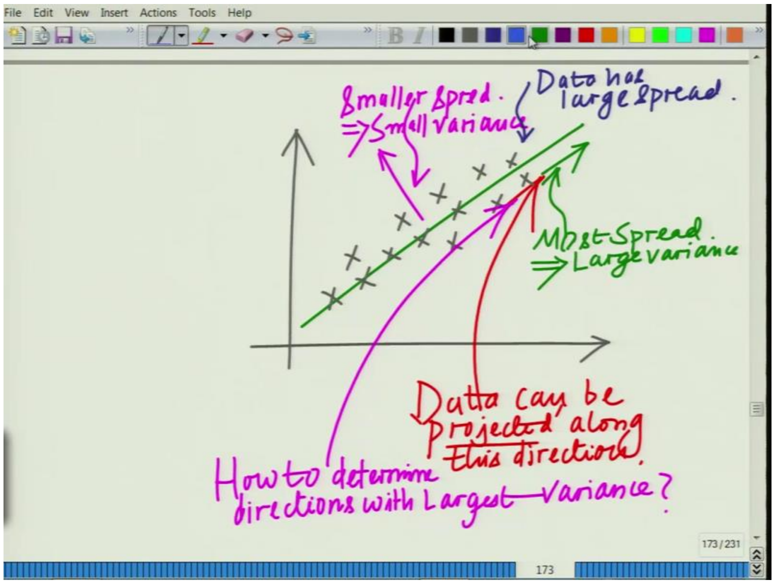
(Refer Slide Time: 3:42)

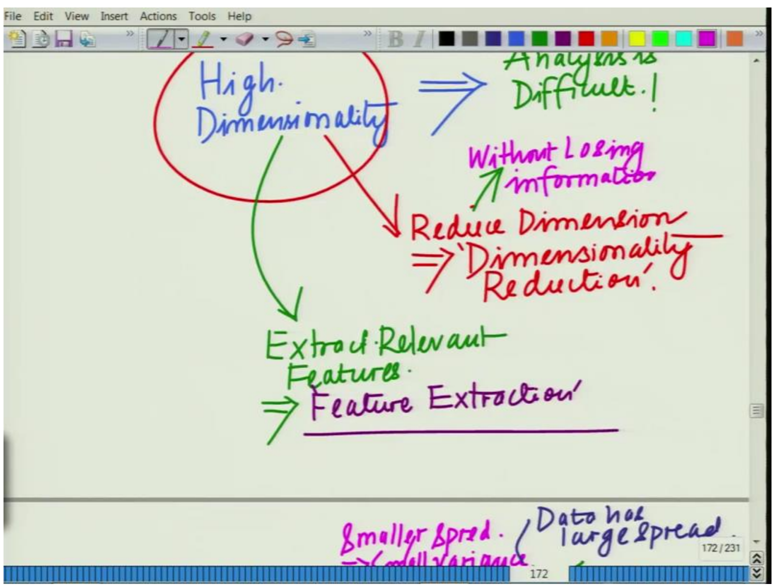


So, when the dimensionality is very high. So, high dimensionality this implies that the analysis or the analyzing the data is going to be highly difficult. So, when you have a data of very high dimensionality the analysis of data is going to be very difficult. And therefore, this high dimensional data has to be compressed. You have to reduce the dimension of this high dimensional data this is known as dimensionality reduction. Or you have to extracts certain most relevant features from this data which is known as feature extraction.

So, in this high dimensional data in machine learning applications what one has to do is, you have to take this high dimensional data and you have to reduce the dimension. Either you can think of it as reduction in dimension. This is termed as dimensionality. This is termed as dimensionality reduction or you have to extract the relevant features from this. This is termed as the feature extraction. This is termed the feature extraction step.

(Refer Slide Time: 5:44)





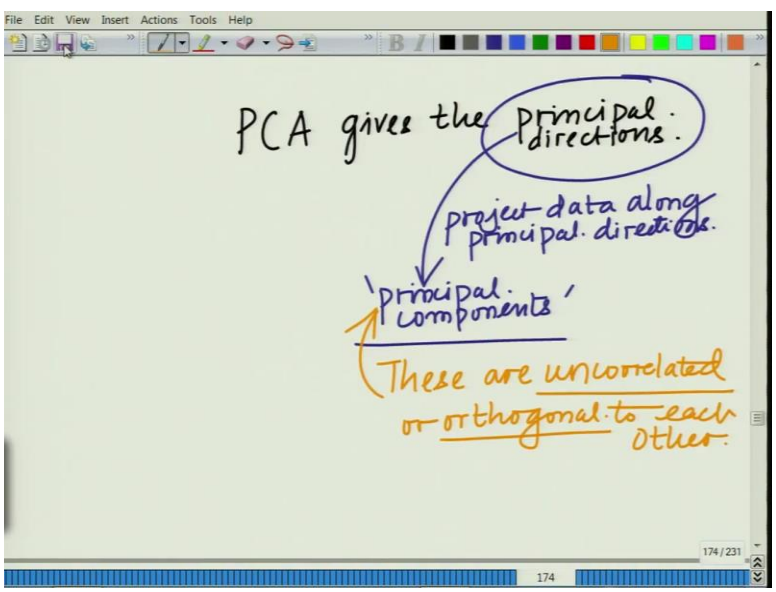
For instance, let us take a simple visual example of the I am going to illustrate a 2 dimensional data. So, let us say you have 2 dimensional data for instance. So, data has a large number spread, so you can see this is your data. Data has large spread but if you look at this data this data along this direction along this direction data has this has the most spread. Most spread along this direction. Most spread along this direction implies basically the largest variants implies large variants.

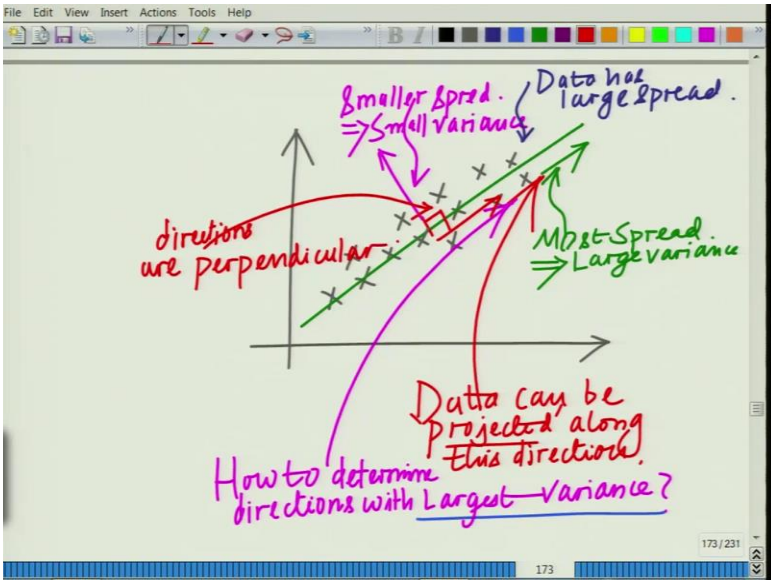
On the other hand, on this direction if you look at this direction perpendicular this has a smaller, this has a smaller spread implies, this implies this has small variants. So, essentially the data can be projected along this direction. The data can be projected along these directions. So, you can take, so what you can do is you can see your data. Your data has a large spread along a certain direction while along other the perpendicular direction the spread is smaller.

So, you can take this data projected along the direction which has, where it has a large spread without using much of the energy or without while capturing all the relevant information in the data. Projection along those directions which have a large spread captures most of the energy or most of the relevant information without losing. So, you can reduce the dimensionality without losing the relevant information. So, that is the essential idea.

So essentially, what we mean is this dimensionality reduction we reduce the dimension without losing. So, we want to reduce the dimension but in the same time losing I hope you appreciate this aspect without losing information. Or you cannot say without losing information while minimizing the loss of information. So, you still capturing the maximum moment of energy that is there in the data. Now the point is how do you determine these directions with the largest variants. Now the question that we want to ask is, how to determine the directions, how to determine this direction with the largest variants?

(Refer Slide Time: 9:49)



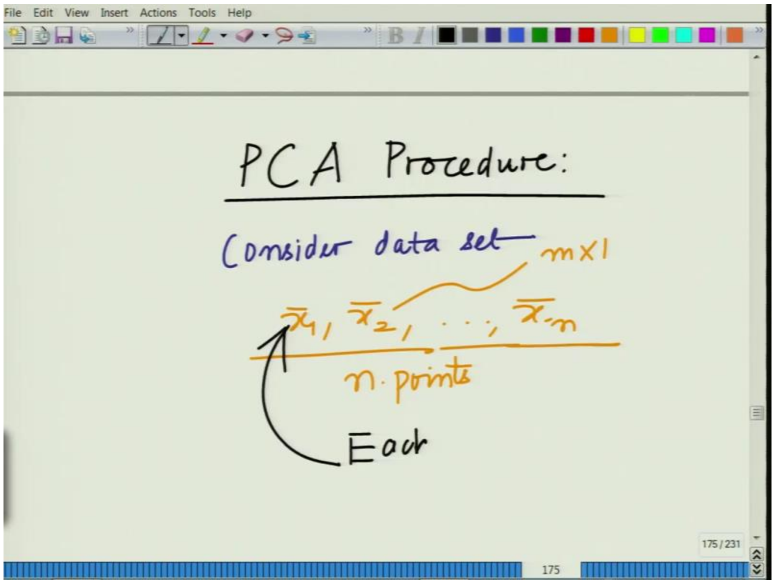


The answer to that how to determine the directions with these largest variants and this is essentially what is given as so the principal, that is basically how do you determine these principal directions along which you can project the data to obtain the principal components that is essentially what is given over there principal component analysis. So, that is the basis of the principal component analysis.

So, the principal PCA gives the principal directions. The principal directions remember and now project along these principal directions. Project our data along principal directions to obtain along these principal directions to obtain the principal components. So, you project the data. So, once you determine the principal directions, once you determine these principal directions, the next step is to project the data. Take the inner product along these principal directions to obtain the principal components of that data and that results in dimensionality reduction.

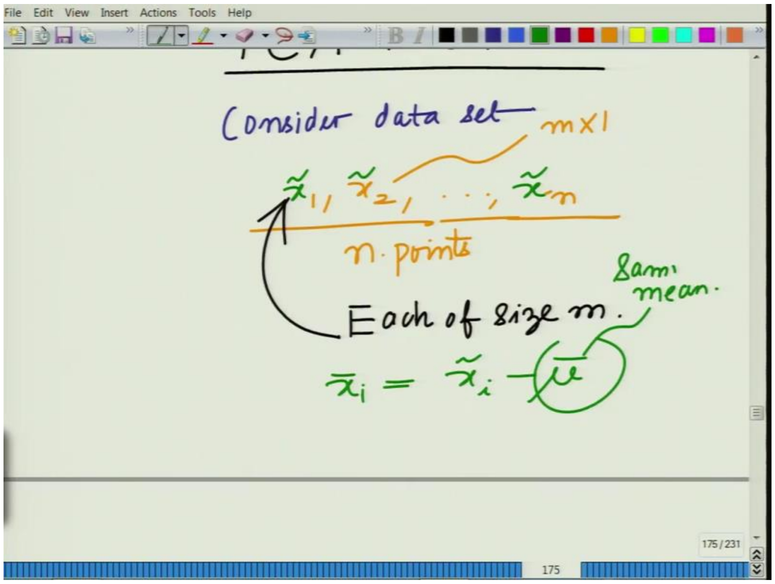
So let us, so this can be and the interesting so this are the largest variants and these are the other interesting these principal components are uncorrelated or orthogonal to each other. Like what you over here, these are perpendicular, these directions are perpendicular. So, you can see what you can see over here that these directions are perpendicular. So, these principal components these are the properties, these are uncorrelated. And we are going to see how that is going to be possible. These are uncorrelated or, these are uncorrelated or essentially orthogonal to each other. These principal components these are uncorrelated or orthogonal to each other.

(Refer Slide Time: 13:00)



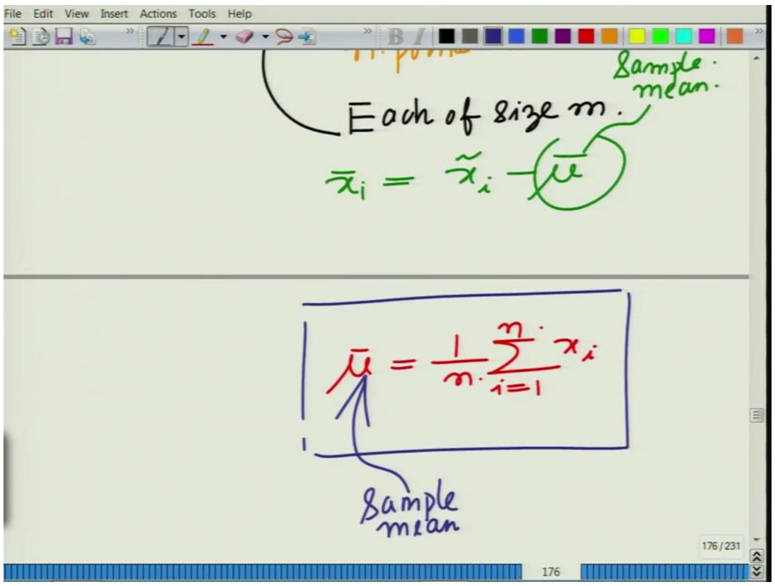
How do we do that? Now this can be so PCA can be, so what is the procedure for PCA? PCA technique what is the procedure. So, consider the data, consider the data set. Consider the data set, there are n points each of size m cross 1. Where m is of course very large that is why you want to compress it or we want to do the dimensionality reduction or feature extraction. Now so these are your n points or you can call this as n vectors of dimension m, correct? n points and data points. And each vector of size m, each is of size m.

(Refer Slide Time: 14:28)



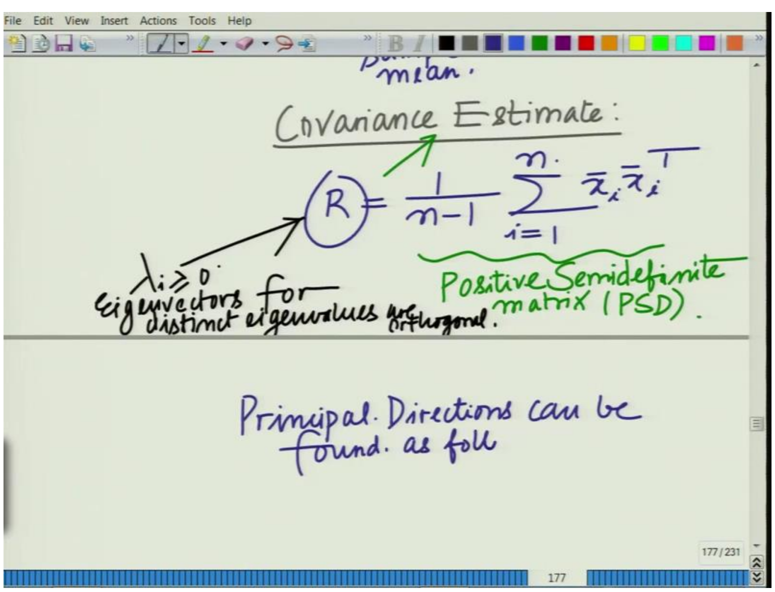
Now first what we do is we remove the mean. So, we obtain let us call this I think it would be better if I call this as x tildas. So, let me call this as x tilde 1, x tilde 2, x tilde n; then you remove the mean to obtain xi bar equals x tilde i minus mu bar. Where this is essentially what we call as the mean or in this case we have to estimate the mean.

(Refer Slide Time: 15:03)



So, this will be the sample mean where mu bar equal 1 over n. summation I equal to 1 to n xi. So, this is essentially the sample mean of this point. So, this is essentially the sample mean this is essentially the sample mean of these points.

(Refer Slide Time: 15:38)

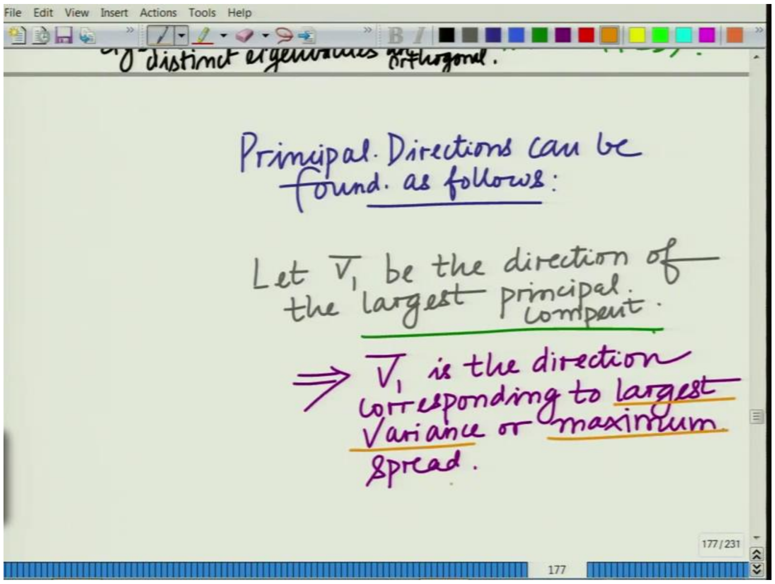


Now compute the covariance estimate the covariance. How do we estimate the covariance? Now you estimate the covariance as follows R equals 1 over n minus 1 summation i equal to 1 to n x bar i x bar i transpose this is the estimate of the covariance. This is your covariance estimate and now you can see this is a positive semi, this is a positive semi definite matrix remember any covariance matrix is positive semidefinite.

So, naturally their estimate also has to be positive semi definite. So, the covariance estimate is a positive semi definite which implies Eigen values are going to be greater than equal to 0. Eigen vectors corresponding to the distinct Eigen values are orthogonal these are the important properties of phd matrices remember. So, this implies Eigen values that is lambda i greater than and equal to 0 and Eigen vectors for distinct eigen values for distinct Eigen vector, Eigen vectors will distinct Eigen values these are orthogonal. And now we have to find the direction of the principal.

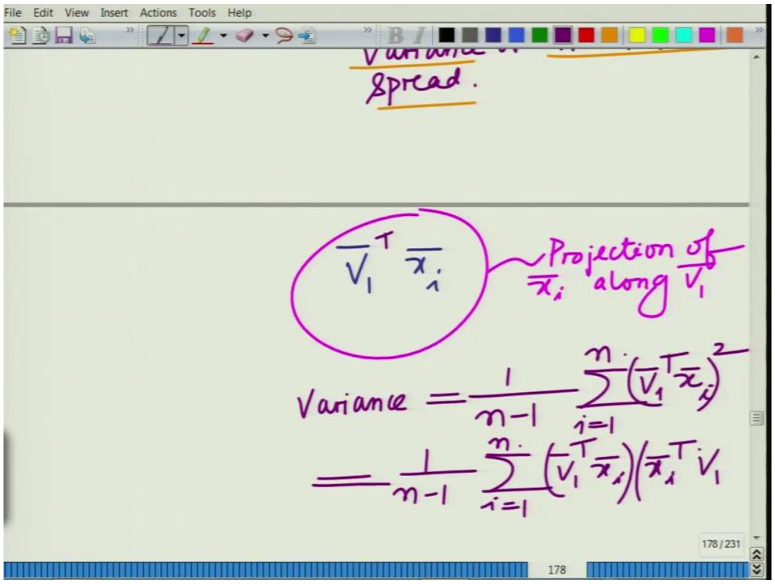
Now the principal components can be found as follows. The direction of the principal components.

(Refer Slide Time: 17:56)



So, the principal directions let us the principal directions can now be found as follows. Can be determined as follows let v1 bar be the direction of the largest principal component. Or let v1 bar, so v1 bar is the direction of the largest principal component which essentially implies this is the direction that has a largest variance, implies v1 bar is the direction corresponding to largest variance or what we call the maximum spread. So, this is the direction corresponding to largest variance or the maximum spread.

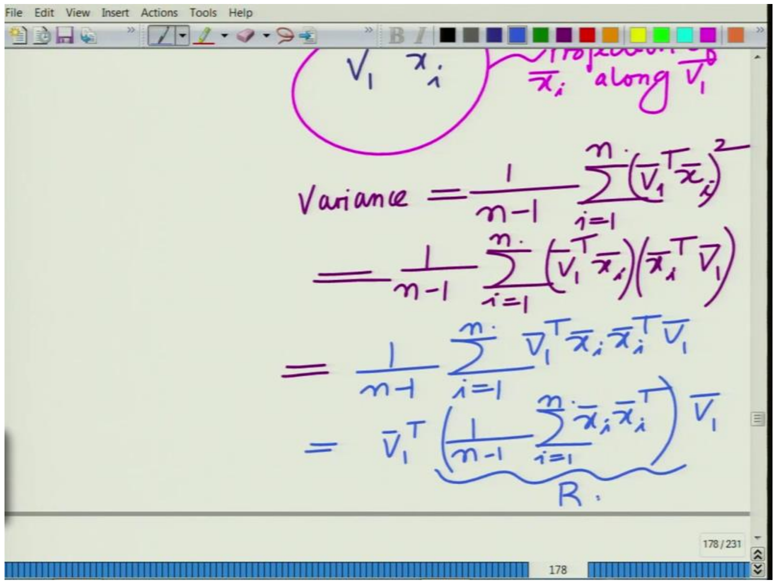
(Refer Slide Time: 19:46)



So, how do we obtain this first let us find the projection. So, let us find so we have v1 bar Hermitian x1 bar or xi bar. This basically gives me what does this give me what am I achieving by doing this. This give me the projection of xi bar along v1 bar. Now we compute the variance of this quantity.

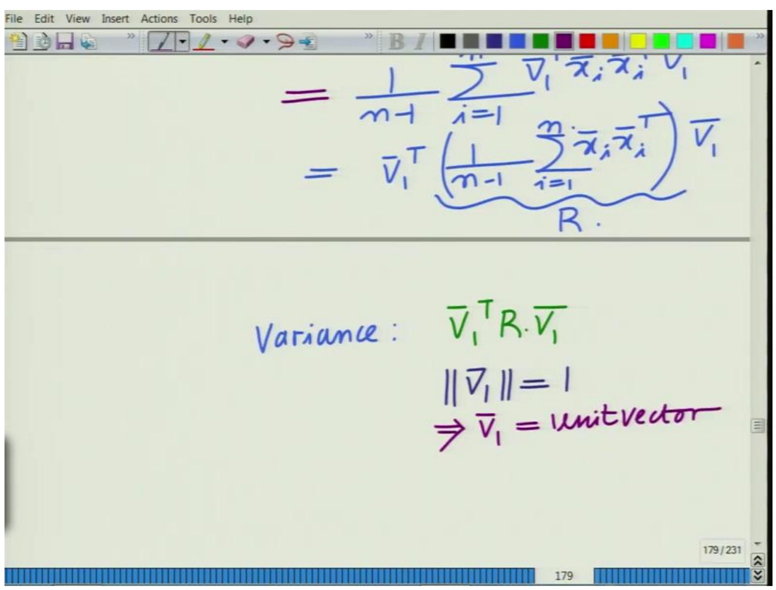
Now we have to evaluate the variance now the variance as you know can be evaluated as 1 over n summation i equal to 1 to n 1 over n minus 1 i equal to 1 to n vi bar Hermitian x bar i square which is 1 over n minus 1 summation i equal to 1 to n v1 bar Hermitian xi bar. Or we can call this as the transpose just to make simple let us make all this quantity is real. Vi bar transpose xi bar into the transpose of this quantity. This is scalar quantity so I can write this as xi bar transpose v1 bar.

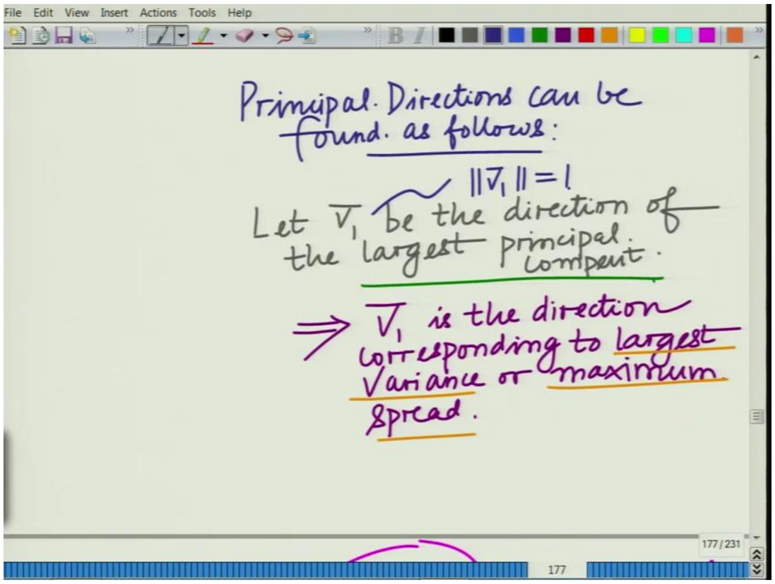
(Refer Slide Time: 21:32)



And now this is going to be 1 over n minus 1 summation i equal to 1 to n v1 bar transpose xi bar xi bar transpose v1 bar which is now take the v1 bar outside v1 bar transpose 1 over n minus 1 you can take this inside summation i equal to 1 to n xi bar xi bar transpose times v1 bar and now if you look at this this is nothing but the covariance matrix.

(Refer Slide Time: 22:23)



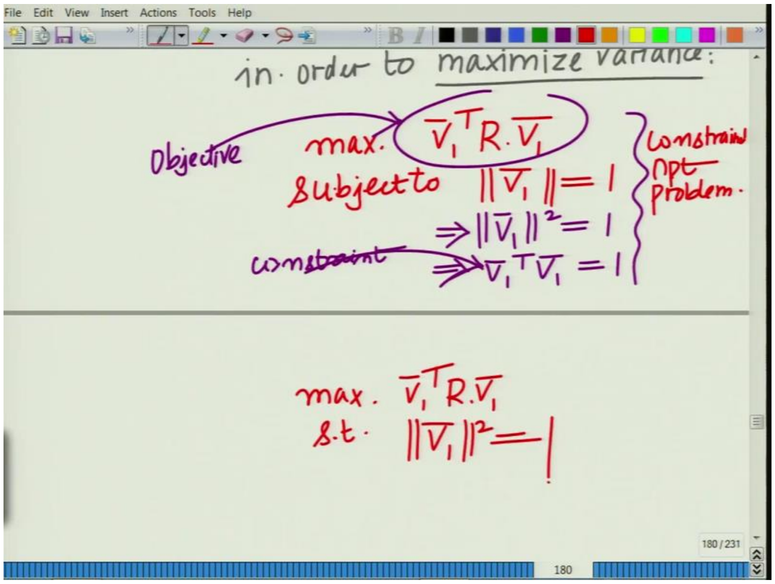


So, v1 bar transpose so variance corresponding to the principal component v1 bar that is the principal direction, direction of the principal component v1 bar is given by v1 bar transpose R v1 bar where R is the estimate of the covariance. So, this is given by v1 bar transpose R v1 bar. Now let us also normalize the energy of the direction. Since we are considering a direction we consider a unit non vector.

So, the normalize, so when we consider the direction of so let v1 bar denote the direction of the principal, direction of the largest principal component and which said non v1 bar equal to 1. So, this is a unit vector. And further so the direction is basically we have the property norm v1 bar equal to 1 basically this means that this is a unit vector. This implies v1 bar equals the unit vector.

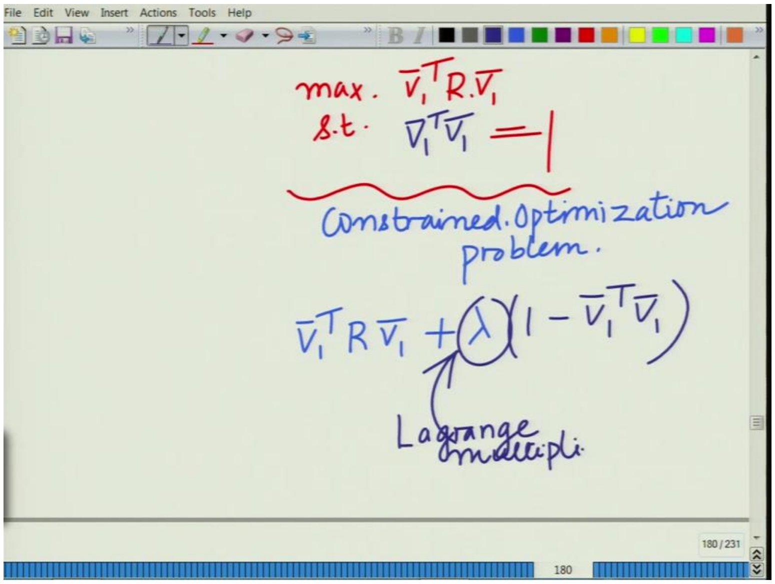
Because if you scale it alpha any direction if you scale it by alpha, then the direction variance increases by alpha square. So, to get a fair comparison of all the directions we simply consider a need may not be that is the particular principal component, the direction corresponding to the principal component but essentially it might be the case that it just has a with large magnitude.

(Refer Slide Time: 24:10)



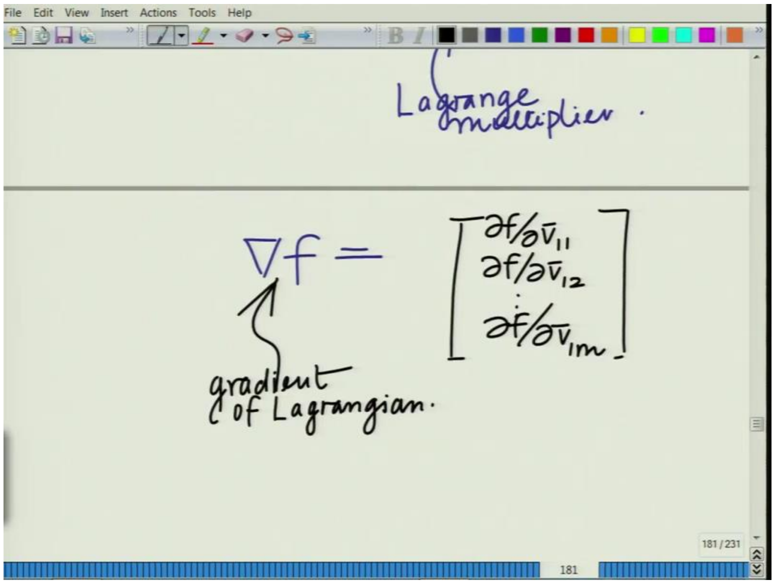
And therefore now to maximize the variance, in order to maximize the variance we have to have v1 bar transpose R v1 bar which is the variance subject to the condition norm v1 bar equal to 1. And this is an optimization problem or you can also say norm v1 bar square equal to 1 which implies v1 bar transpose v1 bar equal to 1. This is what we call as a constrained optimization problem this is the objective. This is the objective and this is the constrained. And therefore, this is what we call as a constrained optimization problem. So, this maximize v1 bar transpose R v1 bar subject to the constrained norm v1 square equal to 1.

(Refer Slide Time: 26:19)



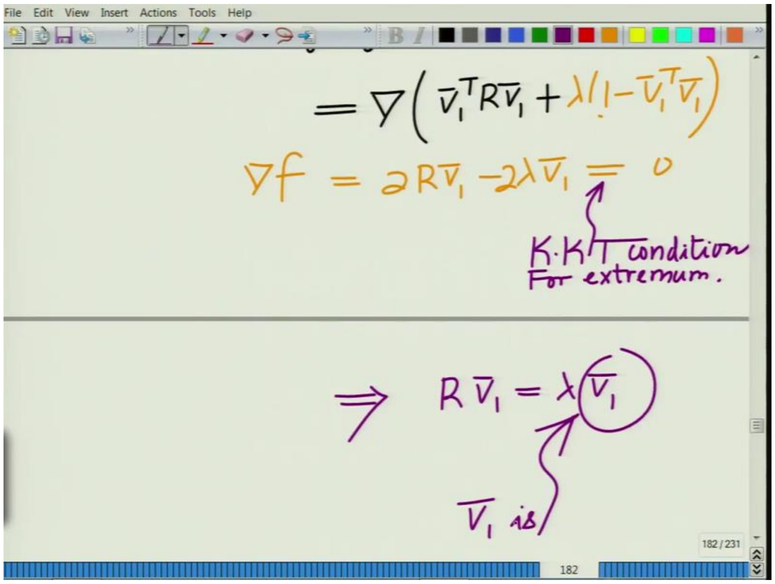
This is essentially what we are calling as a constrained, this is what we are calling and we can solve this using the KKT conditions. So, this first of all we have the objective v bar transpose lambda v bar transpose R v1 bar plus lambda times 1 minus v1 bar transpose v1 bar in fact you can write this constrained as v1 bar transpose v1 bar equal to 1. This is basically what we termed as the Lagrange multiplier this is the Lagrangian this is basically what we termed as the this is from the theory of convex optimization.

(Refer Slide Time: 27:28)



Now you find the gradients so this is your Lagrangian and now you find the gradient of the Lagrangian. So this is your gradient I am just giving you the key steps. Because this is a very interesting problem that arises frequently. Which basically simply partial of f with respect to every element of the vector. doh f by doh v bar 1 to doh f by doh v bar 1 doh f by doh v bar 1 I think this is m dimensional. So, doh f doh v bar 1m. So, we are taking the derivative with respect to each of these elements.

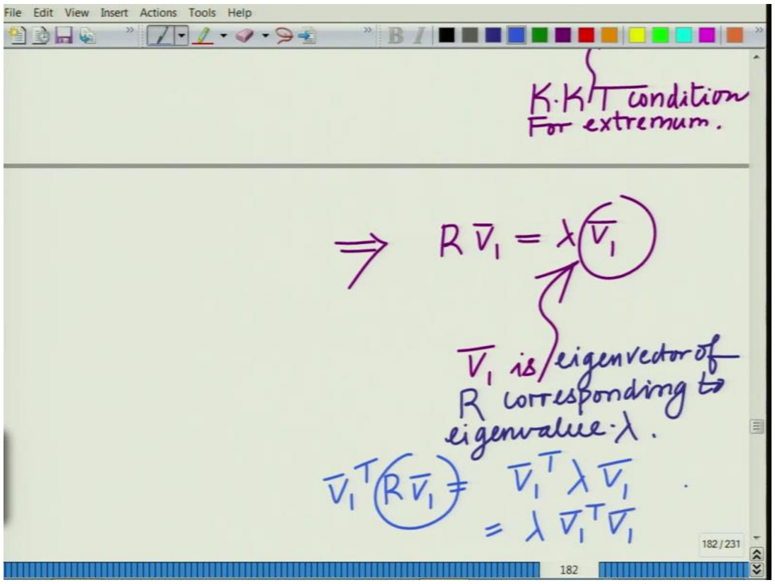
(Refer Slide Time: 28:23)



And this can be shown to be given by the gradient with respect to v1 bar of v1 bar transpose R v1 bar plus lambda 1 minus v1 bar transpose v1 bar which is essentially you can shown to which you can show to be 2 R v1 bar minus 2 lambda v1 bar. And now to maximizes to find the stationary point we set it equal to 0 we set to the gradient with respect to v bar equal to 0 just like you set that the derivative for a single dimensional function to find maxima or minima, you set the gradient equal to 0 so you set it equal to 0 to find the this is the KKT condition. Karush Kuhn tucker condition. So, this is the KKT condition for extremum.

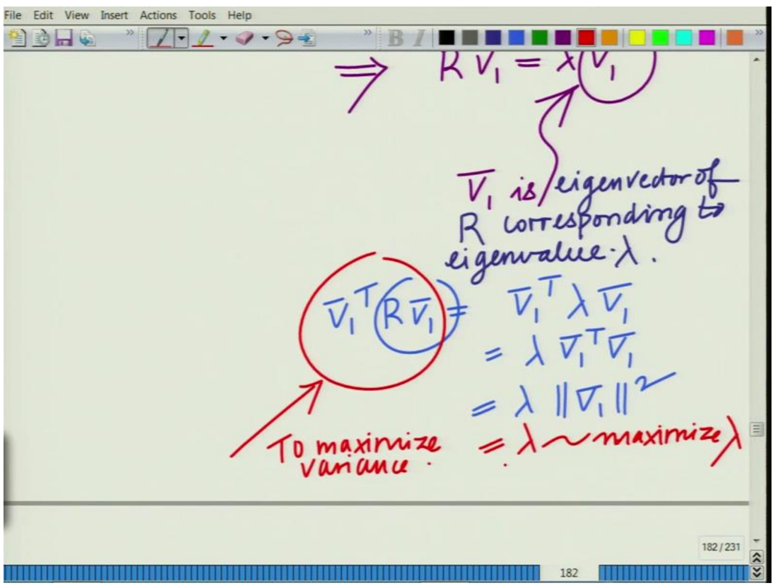
Of course this is a convex of course you can show that this will have a maximum and therefore this is v1 bar transpose or v1 and this essentially implies that now what this implies is that R v1 bar equal to lambda v1 bar which essentially implies that v1 bar is if you can see this is nothing but exactly the definition of an Eigen value that is R v1 bar equal to lambda times v1 bar. So, v1 bar is an Eigen vector of R corresponding to the Eigen value lambda. So, v1 bar is an Eigen vector of R.

(Refer Slide Time: 30:24)



Corresponding to Eigen value lambda. And further v1 bar transpose R v1 which is v1 bar which is a variance is given us now R v1 bar equal to lambda v1 bar so this is v1 bar transpose lambda v1 bar which is lambda v1 bar transpose v1 bar.

(Refer Slide Time: 31:02)

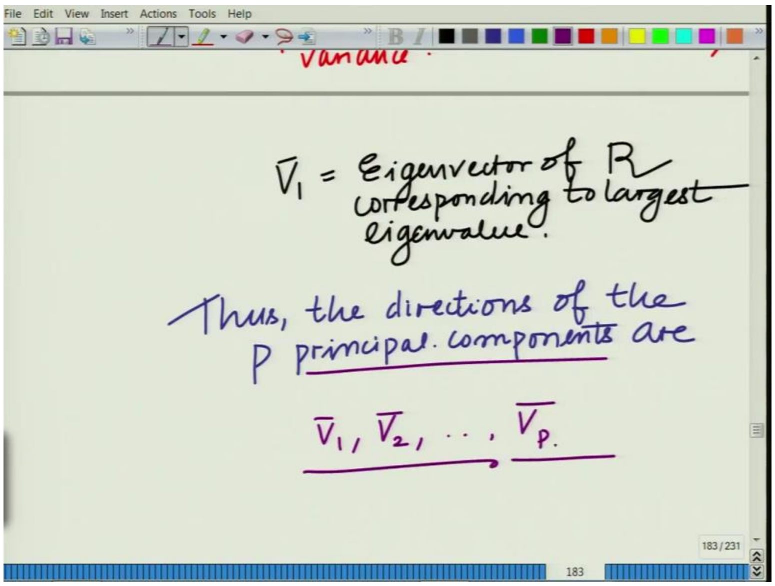


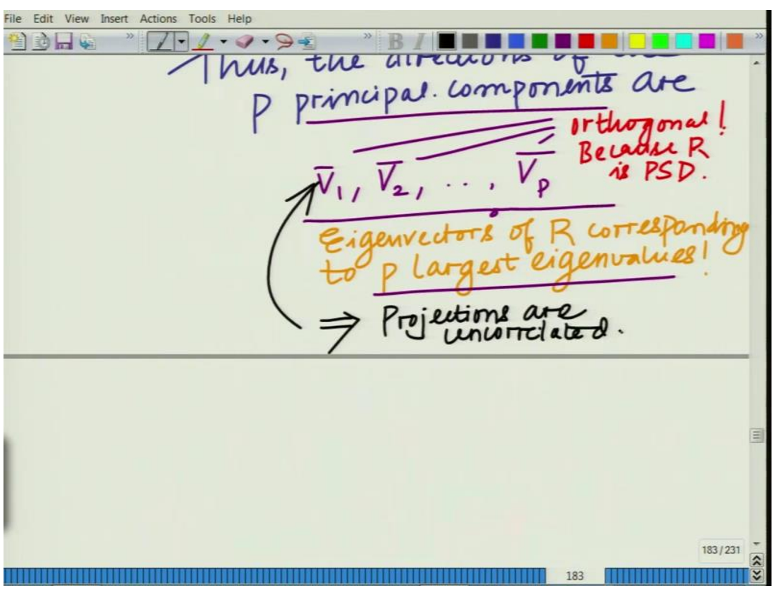
Which is equal to lambda times norm v1 bar square. So, this is lambda times norm v1 bar square and norm v1 bar square is 1 so, this is lambda. So, to maximize variance this very obvious that we need to maximize lambda that is to maximize the variance of this to maximize the spread along this principal component you have to maximize the lambda. And therefore, v1 bar is the

Eigen vector of the covariance matrix corresponding to the largest Eigen value lambda that is the interesting result.

So, the direction of the principal component the largest principal component is basically the Eigen vector of the covariance matrix corresponding to the largest Eigen value that is the very interesting result.

(Refer Slide Time: 32:09)



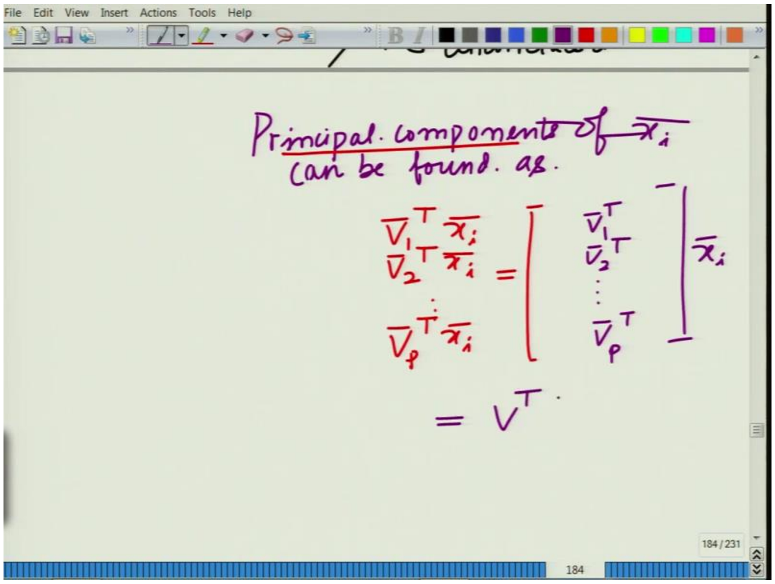


So, v1 bar equal to Eigen vector of R corresponding Eigen vector of R corresponding to largest thus the direction of the so if you want to have P principal largest components thus, if you want to extract P features directions of the P principal components. Direction of the P principal components are essentially call them as V1 bar, v2 bar, vp bar which are nothing but the Eigen vectors of R corresponding to the P largest Eigen value.

So, these are the Eigen vectors of R corresponding to the P these are the Eigen vectors of R corresponding to the P largest Eigen values that is we call them lambda 1, lambda 2, lambda P the corresponding Eigen vectors v1 bar, v2 bar, vp bar. These are the direction of the principal components. And remember the Eigen vectors corresponding to distinct Eigen values orthogonal. So, these are you can find these vectors that these are orthogonal.

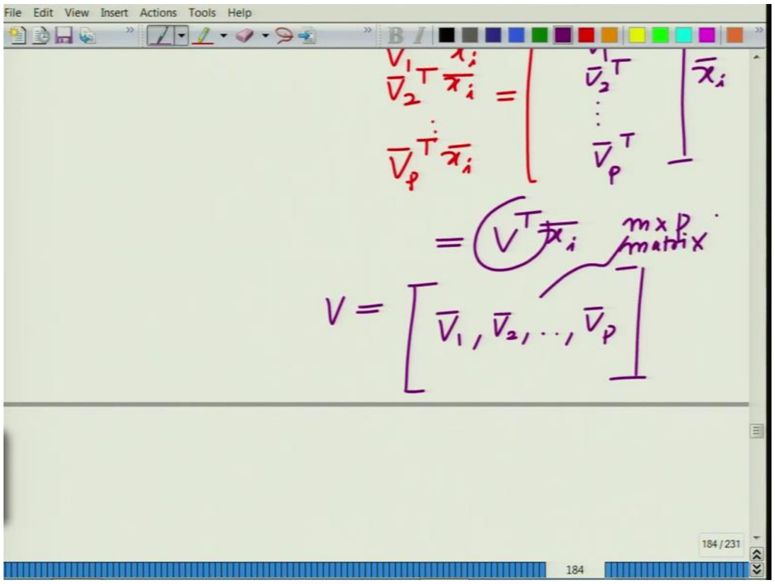
So, these are orthogonal because we call R is PSD. So, this implies the projections are going to be uncorrelated when you take the projections along this orthogonal. So, the projections are going to be uncorrelated. So, this implies that the projections are going to be uncorrelated.

(Refer Slide Time: 35:27)



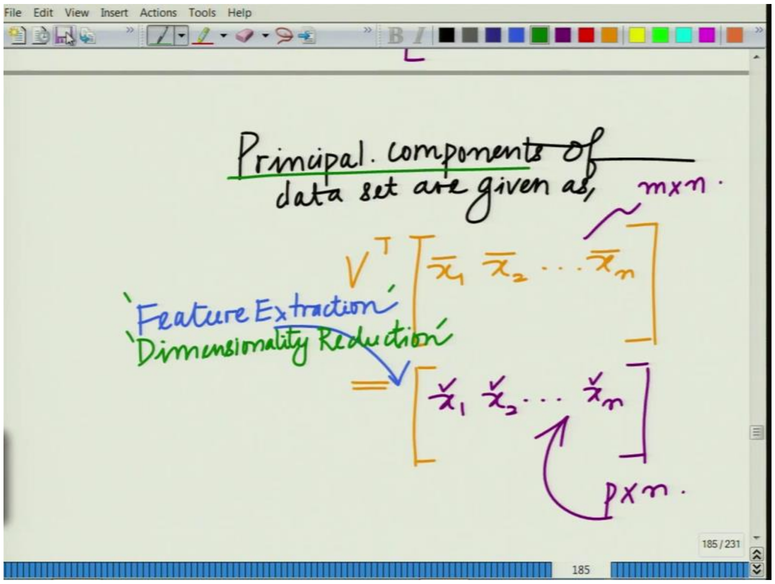
And now the principal components of the vector xi bar can be found as so principal components of xi bar can be found as v1 bar transpose xi bar, v2 bar transpose xi bar, and so on and so forth vp bar transpose xi bar which you can write as the in the compact form we can write this as the matrix v1 bar transpose, v2 bar transpose, vp bar transpose times xi bar which can now write as the matrix v transpose xi bar.

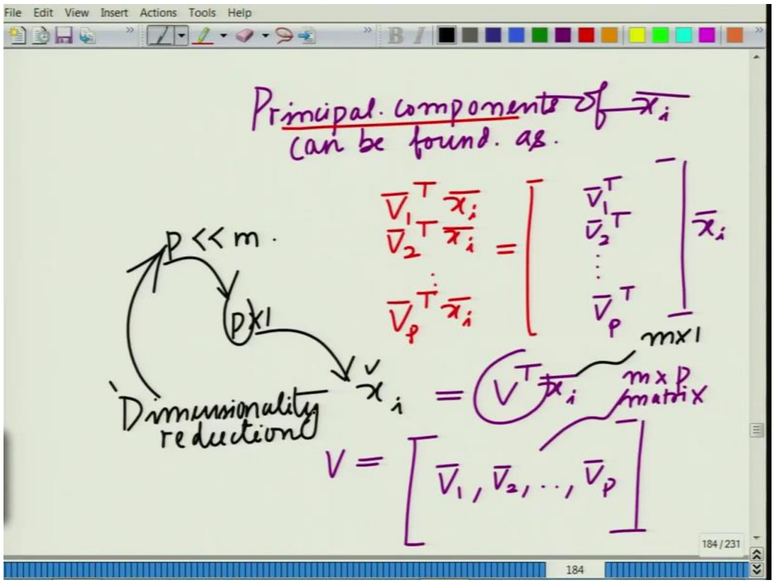
(Refer Slide Time: 36:43)



Where v is the matrix containing the Eigen vectors v bar v equal to v1 bar, v2 bar, vp bar. So, this will be your each of the vector of size m. So, this will be your, I would like to say each is a vector of size m. So, this would be vector of this would be m cross P matrix. So, this would be an m cross P matrix.

(Refer Slide Time: 37:29)





And now the principal component of the entire data set. So, the principal components of the data set so this is basically your dimensionality reduced vector. So, if you look at this v transpose xi you can call that as xi check. So, this will be P cross 1 vector this will be P cross 1 vector because this is your m cross 1 and P can be much smaller that is whole point of dimensionality reduction P can be much smaller of reduction. So, this is basically the dimensionality reduction.

This is basically where the dimensionality reduction is arising because you are choosing much smaller number of principal components that is Eigen vectors v1 bar, v2 bar, vp bar. So, I said they capture a significant fraction of the energy that is there in the data set x1 bar, x2 bar up to xn bar. So, and the principal components of the entire data set.

So, principal components of the data set are obtained as so principal components of the data set these are given as that is you have v bar or v transpose into x1 bar, x2 bar, so on up to xn bar which essentially gives you the matrix which gives you the matrix this gives you the matrix. What does this give you? This gives you the matrix x1 check, x2 check, xn check. So, this is we started with this m cross m matrix and this is the extracted features. These are the P cross n.

This is the reduced dimension or this is basically the dimensionality reduction. So, this is given by the so these are the what we are calling as the principal components of the data set. So, this is basically your dimensionality reduction or feature extraction. This process is basically your feature extraction. Or basically your dimensionality reduction. This is basically you feature extraction or dimensionality reduction.

So, this is essentially the concept of principal component analysis that is used extensively in machine learning in ML as we already said to take large dimensional data and compress it into data of much smaller dimension that is to reduce the dimensionality of large dimensional data. And this step can also be termed as feature extraction. So, we will stop here and I hope you enjoyed this and application this explains in every practical way I mean this explains a very interesting practical applications some of the concepts that we have learned so far.

That is positive semidefinite meter is covariance matrix is Eigen values, Eigen vectors, and so on. These I have already said this arise the concept of I mean positive semidefinite matrices Eigen values, Eigen vectors this arise very frequently in practice. So, let us stop here and continue in subsequent lectures. Thank you very much.