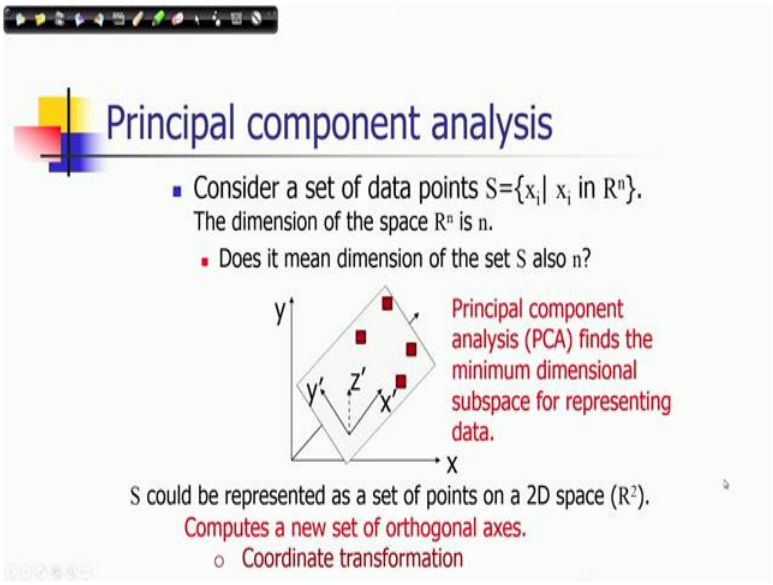
Computer Vision
Prof. Jayanta Mukhopadhyay
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 51
Dimension Reduction and Sparse Representation Part - 1

(Refer Slide Time: 00:24)



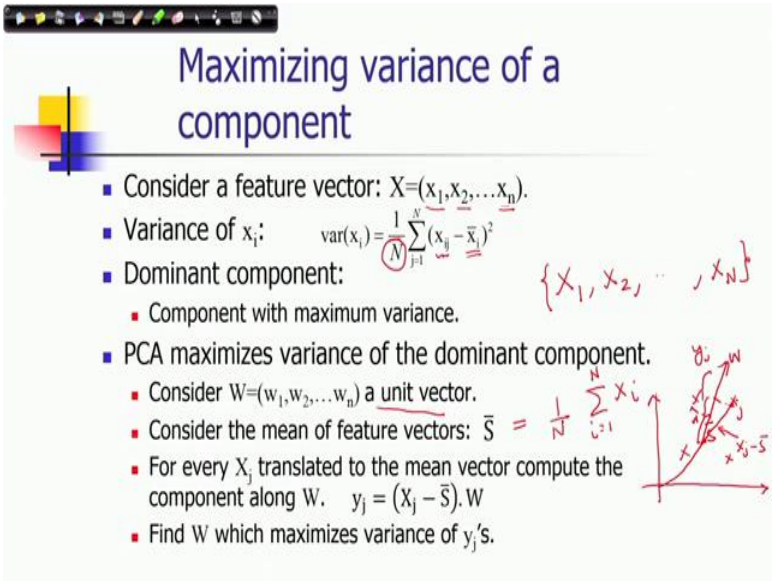
In this lecture, we will discuss about Dimension Reduction and Sparse Representation. So, let us first understand that, what is meant by dimension of a data? Consider a set of data points as shown in the slide that it is a set, where xi is a data point in the space of N dimensional real space. We can consider it is a vector in that real space so, the dimension of that space is naturally n.

Now does it mean the dimension of the set S is also n? So, let us take this example. So, it is a 3 dimensional space and there are 4 data points. Now, they may be arranged in such a way that they lie on a plane. So, when they lie on a plane we can always define a coordinate system within that plane and use that coordinate convention to represent every point.

So, in that case all the points could be represented as a set of points on it 2D space or which is a 2D real space. So, it is not necessary that dimension of data would be the same as a dimension of the space, it could be lower than that number, what the example has been shown here. So, the principal component analysis is a method by which it finds the minimum dimensional subspace for representing data. So, we will learn in this lecture how

this analysis could be done and the basic idea here is that, it computes a new set of orthogonal axis. And that and using that a new set of orthogonal axis you define new coordinates with respect to the representation.

(Refer Slide Time: 02:34)



So, it is a coordinate transformation in one sense. What principal component analysis does? It maximizes variance of a component let me explain that: you consider a feature vector representation of that data point X and since it is represented in n dimensional space. So, we have n components or n fields of the vector, which is shown here as {x1, x2,..,xn}; this is a convention we are using for representing a data.

Now variance of a particular component say ith component xi that is defined as using this mathematical expression, this should be clear that if there are N such data points; that means, N data points. So, for any xi's vector the corresponding jth data point we considered that is a value for the jth component of xi's vector.

And you consider the mean of that component. So, this is how the variance is defined, the standard definition of variance. Now we say a component is dominant out of all these n components whose variance is the maximum. So, that is then we say that component is a dominant component. Now, PCA it maximizes the variance of the dominant component. Let us understand what it means, consider a unit vector W and since as I mentioned that there is a coordinate transformation involved in PCA what we can do, that we can perform

certain type of coordinate combination, conventions. So, center of the coordinate our origin of the coordinate can be considered as a mean of these feature vectors.

So, let us represent that mean by this $\bar{S}$, which means you have N feature vector saved this is {X1, X2,..,XN} N feature vectors and $\bar{S}$ should be

$$\bar{S} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

So, this is how the mean could be computed. So, for every Xj translate it to the mean vector compute the component along W. Suppose you can consider these are the data points and this is $\bar{S}$. So, we are translating and this is the original coordinate of the sparse. So, first we transfer the center to $\bar{S}$ take any particular direction, say this is the direction W and take the component means the dot product of this W.

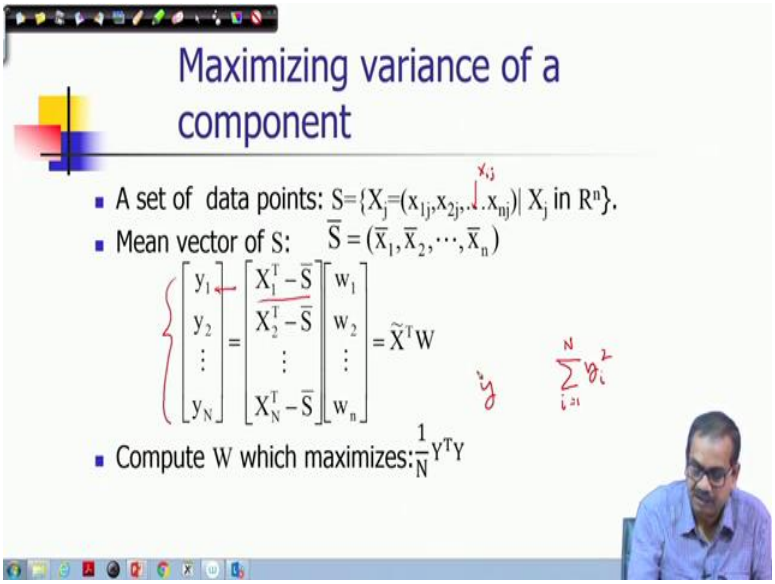
This is a unit vector, this should be a unit vector and so, the dot product of this, which would be the component that is how you will be computing it. So, this is what is yj.

$$y_j = (X_j - \bar{S}).W$$

So, you take which is a unit vector in this case, we consider this is a unit vector and then you take the dot product of this then you get this component as yj.

Now you consider the variance of this component. So, the problem of PCA is at least you have to find out one such direction, where it maximizes the variances of these projections of different vectors all the data a point centering at its mean.

(Refer Slide Time: 07:05)



So, let me continue that with that representation, we have a set of data points and now we are representing the data points in this way that, if I have Xj as a vector then, for the jth vector there are small n number of components. And, each one component is a variable which is indexed as the ith component should be Xij that is are how we are representing it. So, we consider ith component as Xij here.

So, the mean vector as we discussed is $\bar{S}$which means this is a mean of each component we have defined it earlier also. And, then we perform this transformation we are taking the component along a vector W which is a unit vector and you can see that every vector is translated towards mean. And, for all N vectors there are N numbers. So, for every vector translated it and we have the corresponding component along the unit vector.

So, finally, you get N observations or projections of all these data points along W centering at the mean of the vectors. So, now you consider the variances of these y's which can be represented as. So, we can write compute W which maximizes

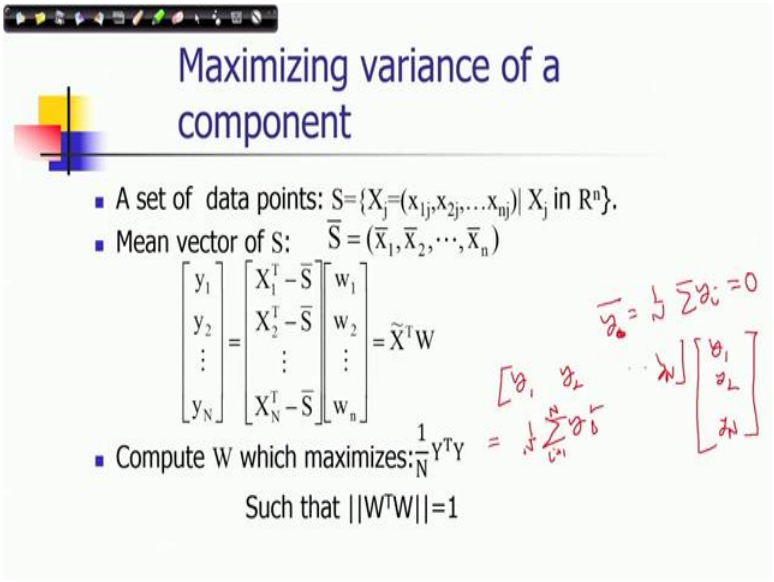
$$\frac{1}{N} Y^T Y$$

Actually you can say that since we have translate towards mean. So, the mean of the y's would be 0.

So, it is sufficient if I simply maximize the square of the magnitudes of that vector.

$$||W^T W|| = 1$$

(Refer Slide Time: 09:47)



So, these are magnitude of this vector that would be

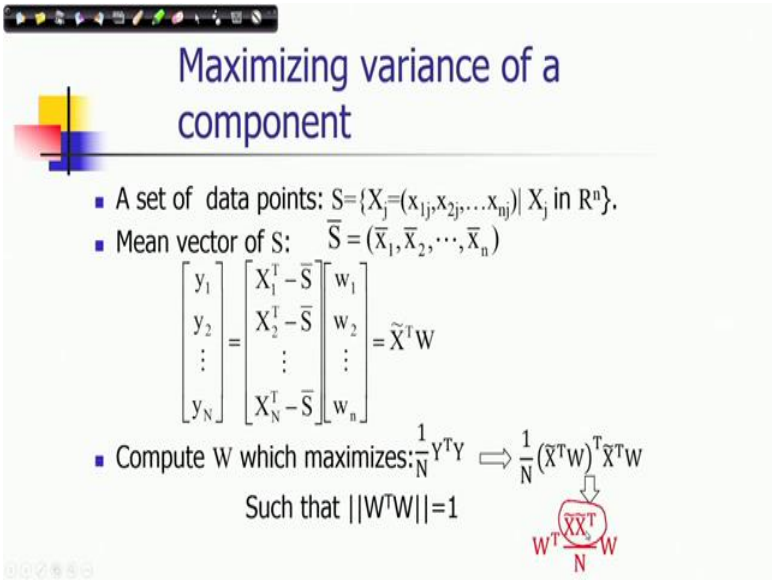
$$\frac{1}{N}\sum_{i=1}^{N} y_i^2$$

And we are taking the mean of it, which is the variance of this value. And, if I consider the mean of these mean of the y 1's at all we can write

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad = 0$$

And effectively we are finding out the variance of the components and we would like to get W which maximizes this particular factor. There is a constraint on W you have already mentioned that it has to a unit vector so we should have the norm of the W that should be equal to 1.

(Refer Slide Time: 10:50)

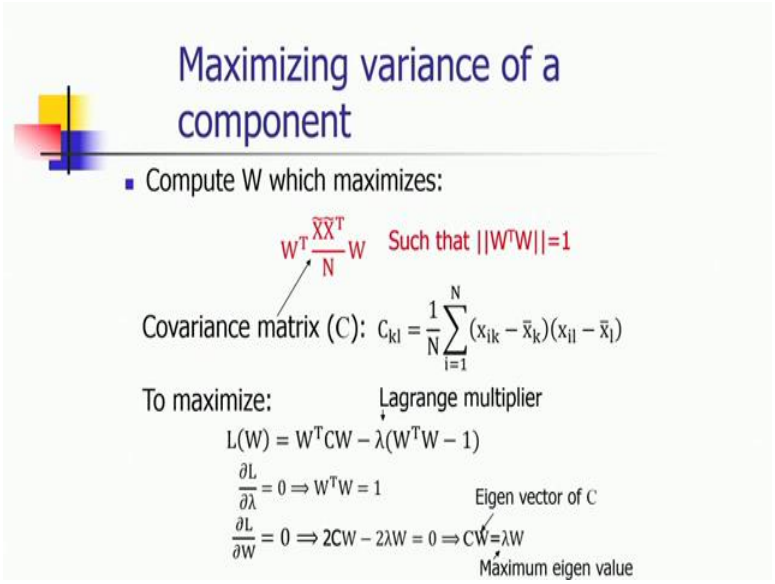


And, effectively if I consider expand Y^T or Y, y is it has been shown it has been X^TW. So, this can be written in this form and which is giving the expression as

$$\frac{1}{N}(\widetilde{X^T}W)^T\widetilde{X^T}W$$

This quantity is interesting because, what it is measuring? It is measuring the covariance of X of course, you have to consider the averaging of that those values. So, this is the averaging there is a term which is divided by N which will average it out.

(Refer Slide Time: 11:36)



So, we have to compute W which maximizes this particular factor and as I mentioned that this is nothing, but the covariance matrix C, where the element of say kth lth element is covariance between the kth component and lth component. Covariance between kth component and lth component of the vectors, of the data points and that is how the covariance matrix is defined.

So, the objective function you can consider for maximizing the variance is that to maximize a function, which is the function of weight vector W or the unit directions W. And, which will maximizing this quantity; this is the quantity which is same as this one which has to be maximized, but there is a constraint so, this constraint in the objective function can be incorporated using Lagrange multiplier. And this is that particular term. So, this lambda is the Lagrange multiplier.

So, if I take the derivative with respect to lambda it will be $W^TW=1$ which is actually enforcing the constraint what we want to for while getting a solution. And, then if I take the partial derivative with respect to the weight direction unit vector W, then we get this system of equations as that I earlier also mentioned that the similar properties or similar rules of differential geometry for one dimensional variable can be extended for multi dimensional space also when you are using matrix operations, linear operations.

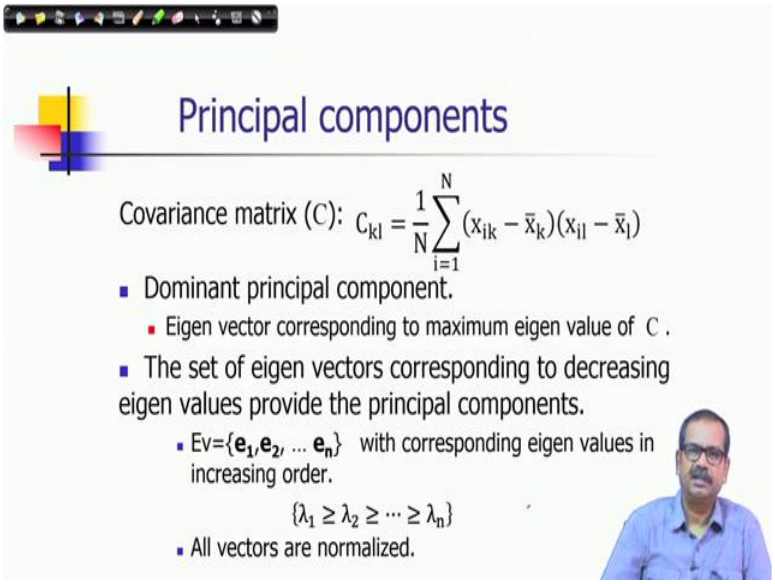
So, you can consider W^TCW it is a quadratic product.

$$L(W) = W^T CW - \lambda(W^T W - 1)$$

$$\frac{\partial L}{\partial W} = 0 \rightarrow 2CW - 2\lambda W = 0 \rightarrow CW = \lambda W$$

So, λ is an eigen value and since we are considering W has to be an unit vector. So, you will consider the unit eigenvector in this case. And, since we would like to maximize the variances and we will be considering the maximum eigen value for the vector corresponding to the maximum eigen value.

(Refer Slide Time: 14:47)

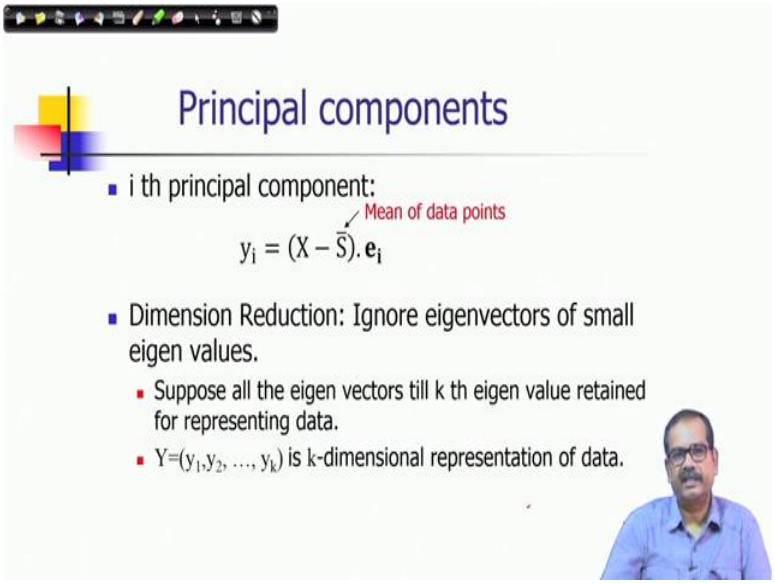


So, what you get actually dominant principal component that is eigenvector corresponding to maximum eigen value of C. Now, what about other eigenvectors because as covariance matrix it is a symmetric matrix so, you will get N eigenvectors if they because a dimension is small N here in this case and n X n matrix.

And so, all the eigenvectors in fact, it can be shown that they are providing the maximum variances alone the residuals one after another. And, so the solution of this particular analysis or what you can say that set of eigenvectors corresponding to decreasing eigen values they provide the principal components.

Suppose we represent it set in this form that there are small in number of eigenvectors as I say {e1, e2,..,en} such that there are corresponding eigen values are also in the decreasing order. So, e1 corresponds to the maximum eigen value, e2 corresponds to the second maximum and the minimum eigen value λn that corresponds to the eigenvector en. You should note all vectors are normalized here, we are only considering unit vector eigen vector.

(Refer Slide Time: 16:09)



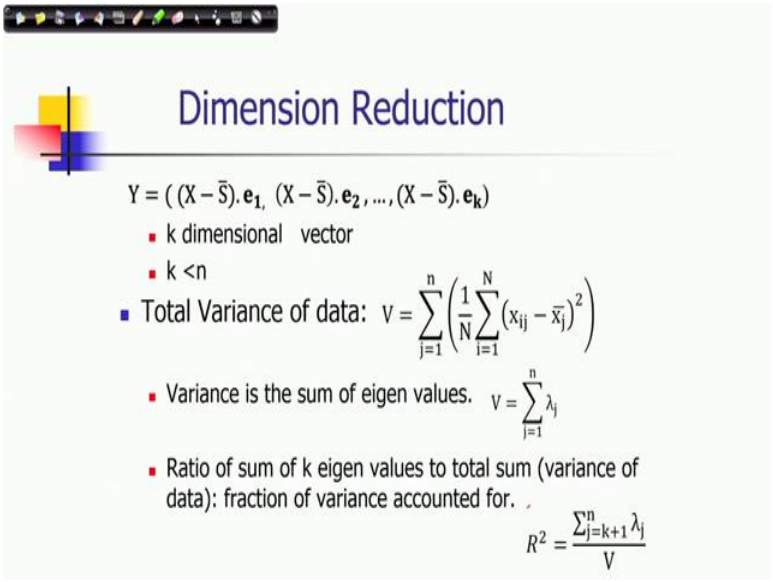
So, the ith principal component is defined as the projection along the eigen vector of the centering at $\bar{S}$ centering at the mean of the data points. So, we can mathematically write it as

$$y_j = (X - \bar{S}).e_i$$

The dimension reduction could be in this way, that we can ignore eigenvectors of small eigen values. That means, for a data point we can reduce the dimension, we can ignore those components whose eigen values are very less.

So, there is an interpretation of eigen values, they are representing the variances of the residuals for at that point. So, suppose all the eigen vector still kth eigen value retained for representing data, then the data can be approximately represented by k dimensional, it kind of a k dimensional representation of data.

(Refer Slide Time: 17:09)



So, we had an N dimensional data, but as I mentioned the dimension of data is not necessarily the dimension of the space. So, they can lie on a subspace which is a k dimensional subspace in this case. And using the principal component analysis we have performed that coordinate transformations. So now, your coordinate axis are given by these eigen vectors and you are considering and your center of your coordinates becomes the center of the data point. And, you consider the projections of data points along each eigen vectors that would give you the components.

So, first k components in decreasing order which are sufficient which may be sufficient to represent the data, which may be sufficient to capture the variances of the data. So, this is a thing that we have k dimensional vector and as you understand k has to be less than n or it could be equal n also. So, total variance of data that can be you know, that can be represented as variances of each component sum of variances of each component.

There are n components so, you consider variance of each component and that would give you the total variance of data. So, variance is the sum of eigen values. So, this can be shown, this can be proved also, that variance is nothing, but sum of eigen values. So, that is why the eigen value which are very small which is negligible it is not contributing to the data variance part and we can ignore those components. So, ratio of sum of k eigen values to total sum that is a variance of the data that is the fraction of variance accounted for. So,

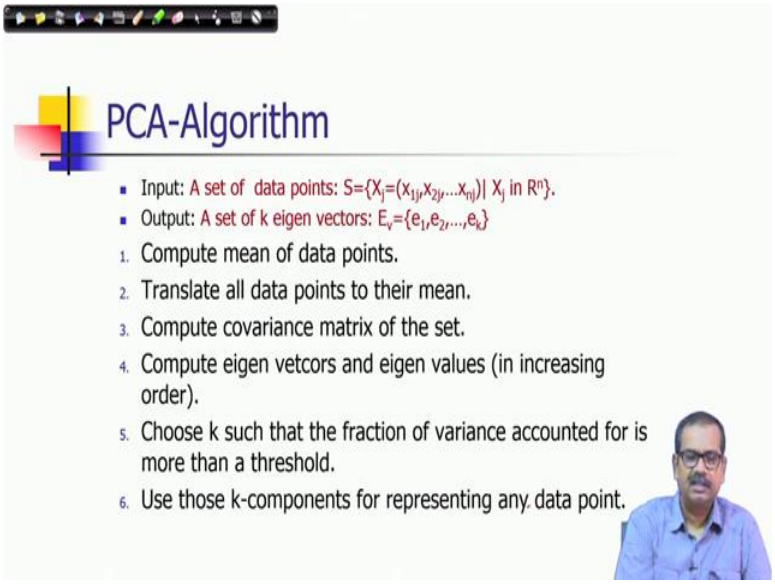
in dimension reduction this is what we would be considering that as high as this fraction is better is the information content of the data retained.

So, this fraction is a fraction as you understand it varies from 0 to 1. So, we will be considering a very high value of this fraction nearly 1 for representing data. So, mathematically we represent this statistics as

$$R^2 = \frac{\sum_{j=k+1}^{n} \lambda_j}{V}$$

As you can see that this is not fraction of (Refer Time: 19:36), this is a fraction of variances which are rejected.

(Refer Slide Time: 19:50)

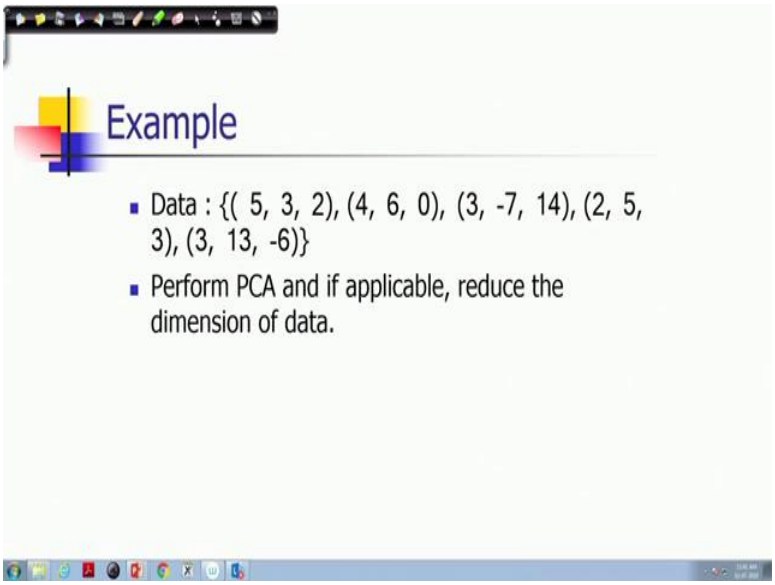


So, it is a sum of those components from variances of those components which are not accounted in your representation. So, this R^2 should be as small as possible. So, to summarize the PCA algorithm is that you have an input, it has a set of data points as we mentioned that Xj and each data point is a point in the n dimensional space. And, then the output should be a set of k eigen vectors, now k will be determined by the threshold fraction threshold that fraction of variances that we are accounting for with that particular factor.

So, the algorithm goes like this you have to compute the mean of data points, and then translate all data points to their mean. Compute covariance matrix of the set, then compute

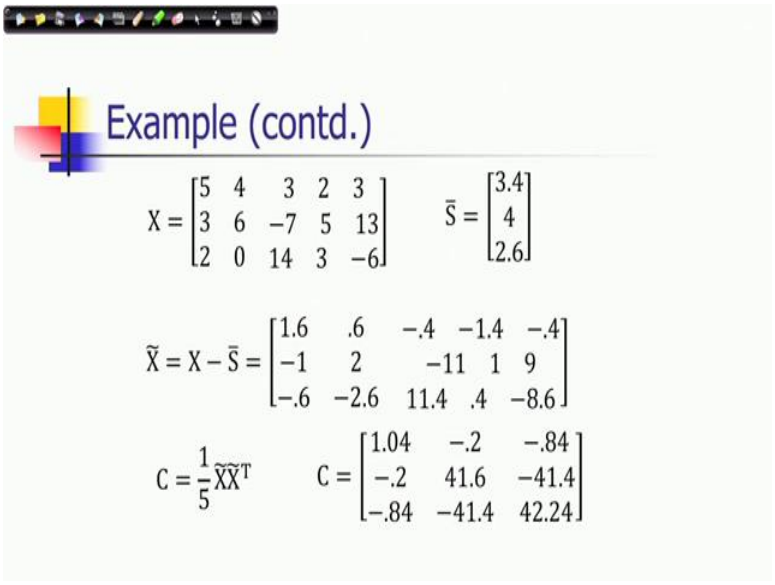
eigenvectors and eigenvalues that is an increasing order then choose k such that the fraction of variance accounted for is more than a threshold so, that threshold is also a parameter to this algorithm. Usually the typical value could be say 0.95; that means, 95 percent of the variance you are taking care of by this representation, and use those k components for representing any data point.

(Refer Slide Time: 21:05)



So, let me explain elaborate these computations using an example, consider this is a data point and we want to do PCA on this data point for reducing the dimension of the data.

(Refer Slide Time: 21:23)

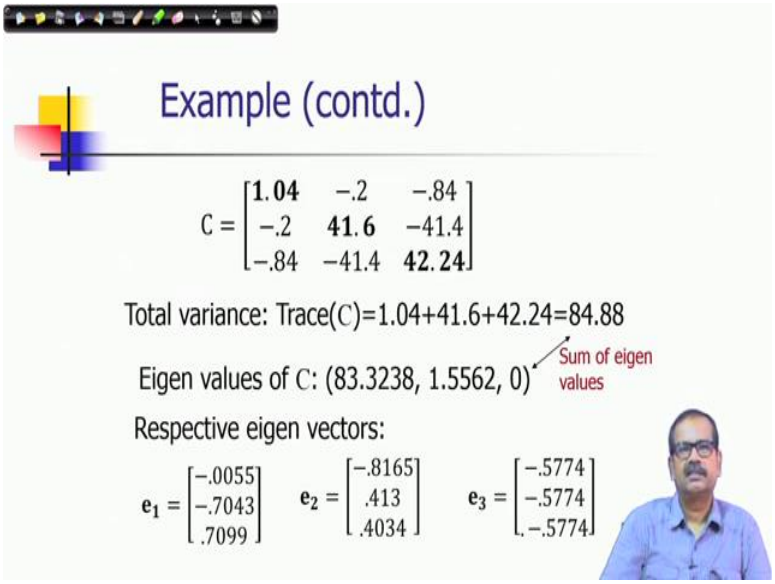


So, all these data points now they are represented as a column vector of a matrix X, there are 5 data points as you see they are all data points are in three dimensional spaces. So, each column vector is a data point so, these are data points in that three dimensional space those are the coordinates.

So, if I take the mean of those data points it is at all these three dimensional points so; that means, mean of those columns column vectors your mean is computed in this form. Now, you compute the vectors translated towards the mean. So, you simply subtract mean from the from each column vector, you will get this particular matrix X tilde, which is representing the translated vectors to around in the mean. And, then you compute this particular covariance matrix; which is $\frac{1}{5}\tilde{X}\tilde{X}^T$.

So, if I perform these matrix computations this is a covariance matrix you get. So, for principal component analysis what you need to do? You need to find out the eigenvectors and eigen values of this covariance matrices. So, since this is a 3 X 3 covariance matrix and it is a symmetric covariance matrix; you have 3 eigen values. It could be distinct, it could be non-distinct, but there were 3 eigen values and correspondingly there will be 3 eigen vectors.

(Refer Slide Time: 22:57)



In this particular example we also note the diagonal elements of this covariance matrix, actually diagonals they represent variances of components. Say first component variance

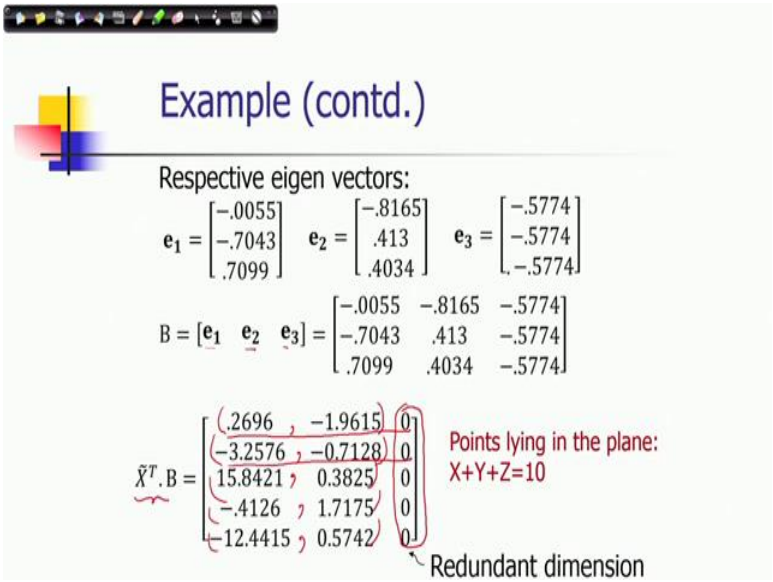
is 1.04 that is a translated, but around the mean so, variance of first company is 1.04, second component is 41.6 and third component is 42.24.

If you note that the maximum component actually maximum variance is actually in the third component in this particular data set which has been given in its original form, and also there are high correlations between the factors. Because, if you note the corresponding off diagonal terms, you will find that these terms are quite significant. So, the total variance could be, total variance is the sum of all these diagonal terms.

So, if I take the diagonal terms you can see the total variance is 84.88. And eigen values of the covariance matrix they can be computed and they are computed as; you can see that one of the eigen value becomes 0 that is the minimum, but the maximum is 83.3238 in this case which is quite high and which is capturing almost all the variance of the data, but the second eigen value is also 1.5562 which is also; it is much less than the first component, but still it has some significant component.

So, what we can do? We can represent this data because the third is 0. So, we can represent this data only using these two components itself and let us see what are the eigen. And, you should note that some of these eigen values they are the same as the total variance of the data. So, the respective eigenvectors with respect to the eigen values; that means, e1; e1 corresponds to the eigen vector 83.3238; e2 and e3 similarly they correspond to the eigenvectors 1.5562 and the other one is 0. So, we can perform the dimension reduction by considering projections alone e1 and e2 only.

(Refer Slide Time: 25:12)

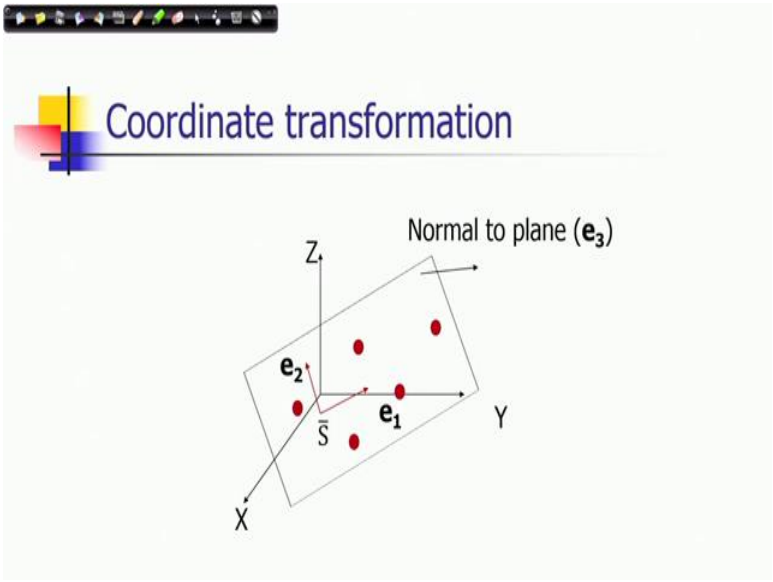


So, what we will do? We will consider as if they are the basis vectors now, and each basis vectors as the columns e 1 e 2 e 3 is a similarity there is an, it is almost it is the same as what we learned for image transform. So, you have a new basis vectors and the original data point can be you know their components with respect to this new basis vectors can be computed of course, you have to translate towards mean for principal component analysis.

So, if I consider the components of translated data points at mean of those original data points then with these computations we are computing each one of them. So, it is taking care of for each data point, it is computing the translation, it is computing the corresponding product, corresponding dot product along e 1 e 2 and e 3. So, now this is your new data point, translator and you can see that one of the component is 0 which means: now I can represent it in a two dimensional space.

So, I can represent it this is a data point, this is another data point, this is another data point like this ignoring the third dimensions. So, this is a redundant dimension. In fact, if you note the point set what I have given they are lying in the plane X+Y+Z=10. And, that is why since they are lying in a two dimensional plane; through principle component analysis what you could find out the plane of why they are lying. In fact, the third eigen value that would give you the normal to that plane.

(Refer Slide Time: 27:11)



And, that is why any projection along that direction it is 0 and e1 and e2 they are giving two directions which those are the vectors which are lying on that plane, and this is a new axis and in that plane you can once again express the coordinates of each data points. So, this is how using principal component analysis you can find out the lower dimensions on which data points lie.

So, let me stop here and we will continue this topic in the next lectures.

Thank you very much for listening to my talk.

Keywords: Principal component analysis, dimensionality reduction, maximizing variance.