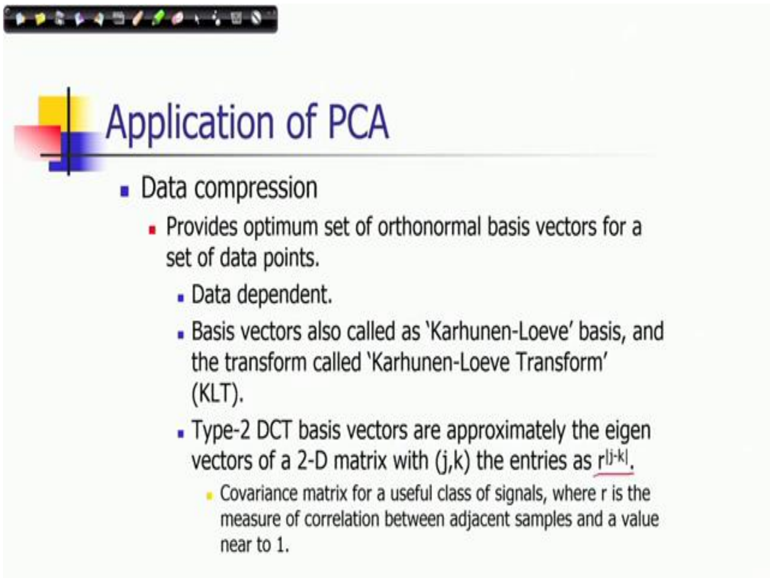
Computer Vision
Prof. Jayanta Mukhopadhyay
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 52
Dimension Reduction and Sparse Representation Part - II

(Refer Slide Time: 00:22)



In the previous lecture, we discussed about the computation of principle component analysis and how we can find out the lower dimensional representations of data points, which are represented in a particular space. Now, let us consider what are different kinds of applications those are possible using PCA. Now, I will consider three major applications here.

So, one of them is a data compression. As you have already seen directly that, actually you do not require that many dimensions to represent data, that itself gives a gives an efficiency of data representation, it requires less storage for representing data. So, it provides an optimum set of orthonormal basis vectors for a set of data points. Because, we have seen in the case of image transforms, that ortho normal basis vectors they are very convenient to transform any data point into another space, and if your basis vectors are properly chosen, then you can reduce the redundancy of that representations.

So in PCA you have that advantage actually it is optimum set of orthonormal basis vectors, which will give you that kind of transformation, but it is data dependent that is one issue here. So, for every set of data points you need to perform it is analysis, perform this analysis and then compute a new set of orthonormal basis vectors, which is not very convenient from the point of view of data compression. Because with the compressed data, then you have to also convey this information of orthonormal basis vectors which is an overview for this.

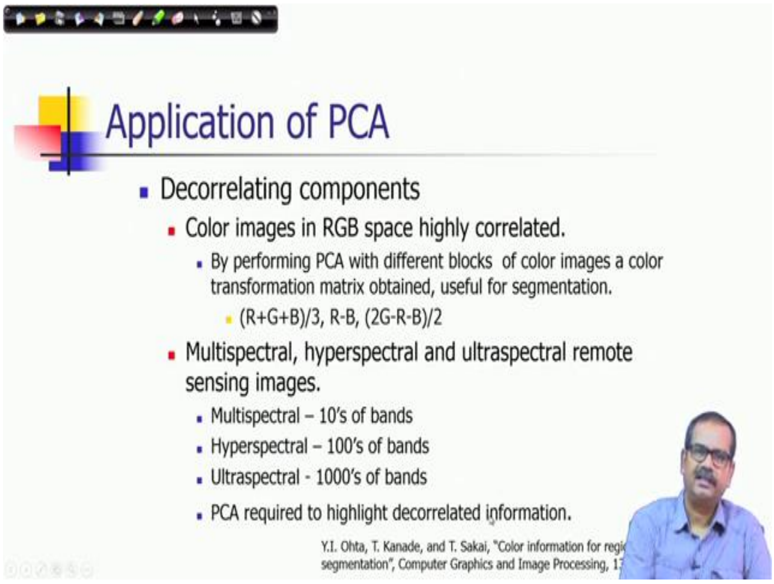
However, and these basis victors they are called ‘Karhunen-Loeve’ basis or the transform is called Karhunen-Loeve transforms which is an optimal representation of data as I mentioned. In fact, many standard basis vectors, they can be shown as eigenvectors of certain statistical representation of signals or images.

For example, type-2 DCT basis vectors are shown as it is approximately the eigenvictors of 2-D matrix with j, kth entries as $r^{|j-k|}$ as it is shown here. So, these are it represents the correlation between adjacent samples say j and k. So, if j and k it is deviates too much from that location it, then it is not it adjutancy is less and then the correlation should be less.

So, because r is a value which is less than 1, so as the j - k this magnitude increases it will be reducing this value will be decreasing. So, you have less correlations, but for high correlations it, so for adjacent samples it is expected they should be very highly correlated. Now, with this kind of data, it has been shown that the basis vectors are; that means, eigenvectors are similar to the almost similar to the type-2 DCT basis vectors. And, that is why type-2 o DCT is so, efficient for representing a large class of signals and images.

So, as I mentioned covariance matrix for a very useful class of signals, where r is major or of correlation between adjacent samples and it is the value with me at to 1. So, it is trying to represent the natural images and natural signals with these kinds of statistics it is trying to model that part.

(Refer Slide Time: 04:02)



The other advantage of PCA or other applicants which is used also different applications, that it de-correlates the components. We have already seen in the example that covariance matrix they show that there is high correlations between different components for the original data. But, after performing PCA actually those correlations they are those correlations are largely reduced. So, it will be almost 0 in that case ideally it would be 0.

So, that is how it de-correlates the components. So, one of the applications of this kind of using this property is that you can find out a new color space, that when color images they are represented in RGB color space they are highly correlated. So, there is a work I have given this reference here, it is the work done by Ohta Kanade and Sakai. So, what they did, that they took a large number of color images and, they performed PCA using the different blocks of the color images. And, then found out eigenvalues and eigenvectors and those eigenvectors that is the data that gives you the transformation new transformation space.

So, in fact, they found that if I transform the color component using those eigenvectors in this way, that (R+G+B)/3, R - B is a second component and (2G-R-B)/2, there is a third component. Now, these are the principle components from their principal component analysis.

And, as you can see this is nothing, but you are performing a color transformations where the fast one is indicating the intensity value and the other two they are the chromatic components. So, this can be, this has been obtained through this PCA itself. The major applications of PCA it is there when you have more number of components in images.

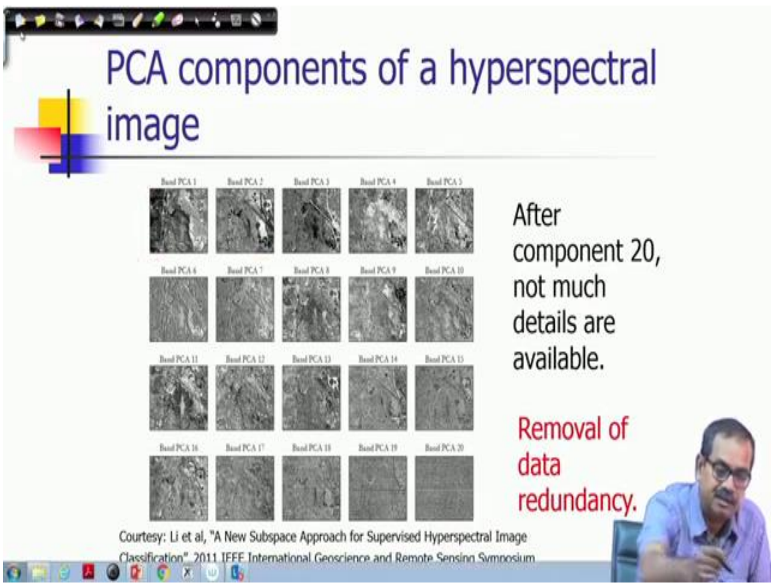
So, color images they have only three components, but if I consider remote sensing images, these components could be very large there are different kinds of remote sensing images depending upon the bands they are using, depending upon the electromagnetic wave length bands they are using. So, multispectral, hyperspectral, ultra spectral, remote sensing images. And, there could be many bands, like multi-spectral it could be 10’s of bands, hyperspectral it could be 100’s of bands, ultra-spectral it could be 1000’s of bands.

We have so many different bands so many numbers of bands and how to get an efficient data representation using those bands. So, what we can do? So, which means that at every pixel, it has dimension; data dimension is say if it is band number is N, data I mentioned is N.

So there are, these kinds of dimensional reduction becomes very useful we can find out only those components after deduction. We can reduce those components which are de-correlated and use those components to make the information analysis to analyze processing permission or to correlate with different ground truth or different ground level information.

So, PCA is required to highlight the correlated information.

(Refer Slide Time: 07:41)

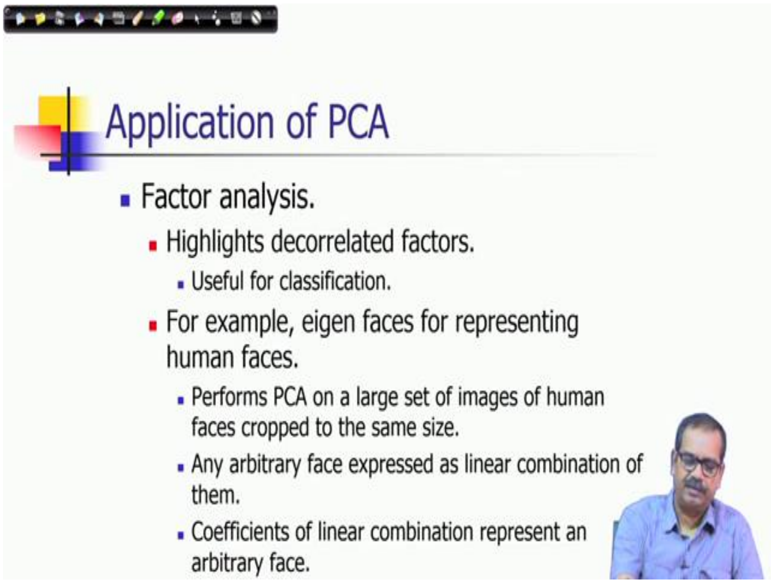


This is one example where I am showing that PCA component of a hyperspectral image and you can see that it starts from this, this is the PCA, which will corresponds to the maximum variance, maximum eigenvalue, you have lots of information there, a lot of details there and this is a minimum eigenvalue in this case it is actually 12th eigenvalue.

Now, if the number of band is in this case it is not specified, I suppose the number of bind is 20 in this case. So, this is a minimum one. In fact, there is no such information also almost no details are visible here. So, that is the advantage of PCA, you can prioritize or you can give references to those bands which are having more informations.

And, say this is a PCA 1, this is band PCA 2 band, second band, third band as you progressively go over the PCA components from say top to bottom and left to right by making these kind of scan. You will find finally, the details are slowly dying and it gets almost like a smooth region. And, so, you can use this components to find the, to analyze the information where if it is representing those information which right here. So this is what after component when (Refer Time: 09:22) much details are available and it removes that data redundancy in this representation.

(Refer Slide Time: 09:29)



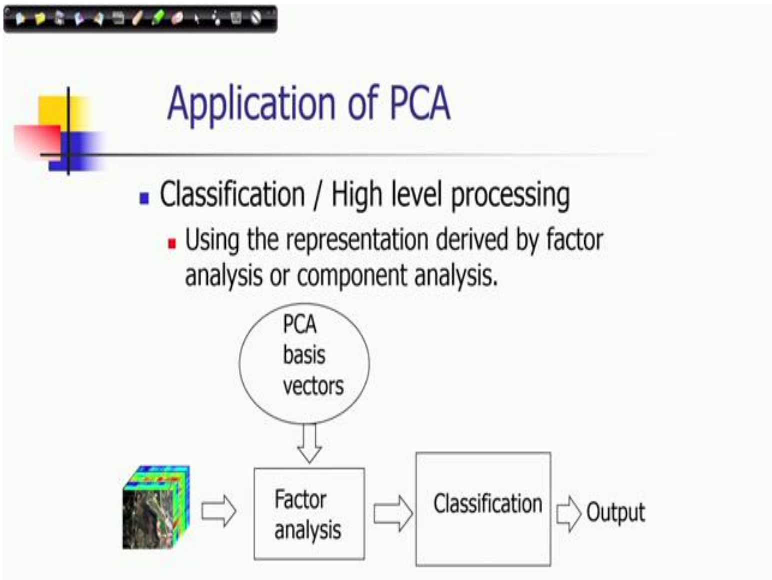
The third application of PCA could be factor analysis and it highlights this decorrelated factors. So, you can find out the factors related to that even the color component what I have shown, it is trying to find out those factors, intensity factors, coma factors and this factor analysis this is very much useful for classification. For example, you consider eigenvalues for representing human, human faces.

So, what we are doing here, we consider a large set of images of human faces crop to the same size and follow certain rules like it is not only just simple cropping, you are trying to maintain the different parts of the human face from the similar distances from the top. And, then you perform PCA and any arbitrary face you can express as a linear combination of those eigenfaces; that means, if you perform PCA you get eigenvectors those are called eigenfaces here. So, coefficients of linear combination it represent an arbitrary face.

(Refer Slide Time: 10:49)

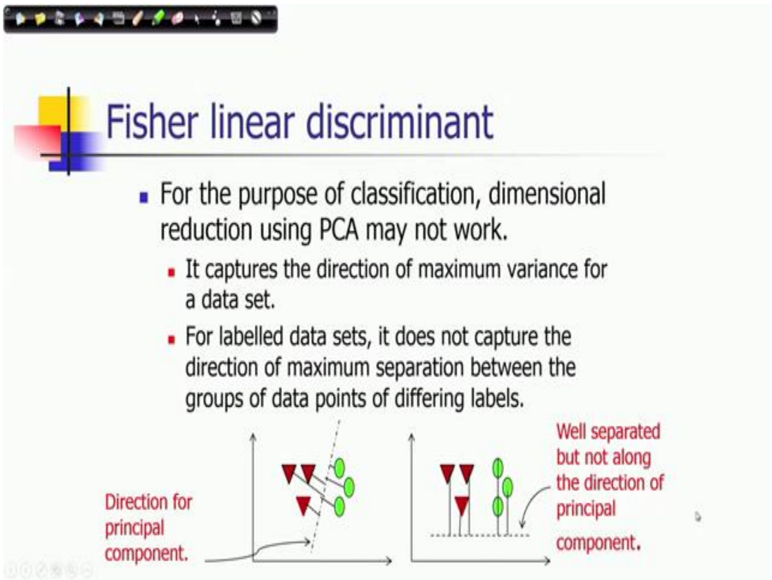
So, this is one example that it is showing four eigenfaces. So, if I consider any arbitrary face at least I can represent by four-dimensional vector using this PCA, using these factors. And, then subsequently we can use it for classification.

(Refer Slide Time: 11:08)



So, this is the another applications, that we use those factor analysis, use those representation or any classification or high level processing like, you can this is a pipeline. So, you can perform factor analysis and then those representation can we use for classification.

(Refer Slide Time: 11:31)



So, I will be now discussing another type of dimension reduction. In fact, the objective here is more for the purpose of classification. In the previous slide itself I have shown how PCAs could be used for factor analysis and those factors could be used for classification.

But, it is not necessarily that those representation of those factors are efficient for classifications particularly if you are considering linear discriminate functions.

So, we will be considering those kinds of dimensional reductions, where this linear discrimination becomes simple for in fact, this is a simple discrimination it is you are reducing it to just one dimensional component. So, the objective of this linear discriminates and which is known as special linear discriminate by the inventor Ronald Fisher is a very famous statistician and it captures the direction of maximum variance of a dataset.

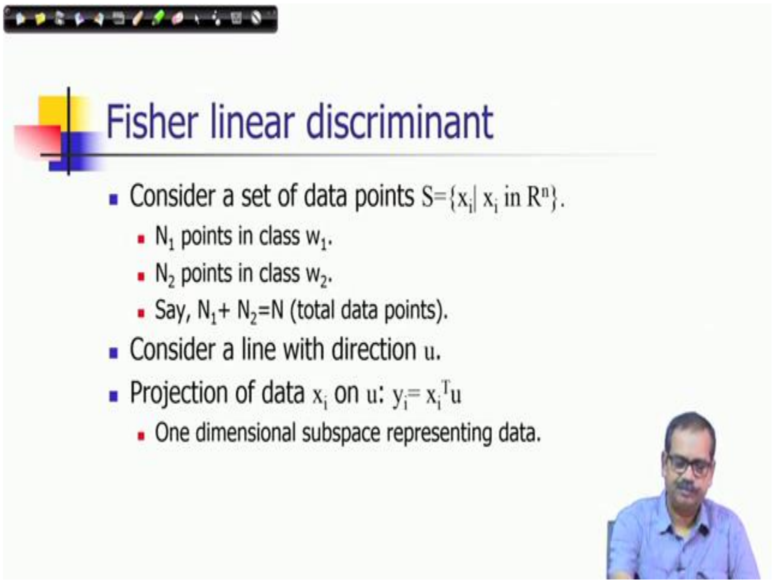
So, for label dataset it does not capture the direction of maximum separation between the groups of data points of different levels. So, here the point is that PCA, it captures the direction of maximum variance for the data set not the vicious linear discriminate. And, but if I consider the dataset is labeled then it may not capture the maximum separation. So, I will be showing you with I will be explaining it with respect to this diagram consider there are two groups of data points, these are the labeled data points; one is shown by the triangles, another other groups are shown by ellipses by different colors.

Now, the direction of maximum variance is a dotted line, that could be the direction, along that directions all the projections of these data points, they will have maximum variance. But, if I take the projections as you can see here that within these intervals the projected data points there they are intermingled. So, they are not well separated. So, they cannot be segregated using simple interval rules, they cannot be discriminated by that rules.

But whereas, if I consider another direction say this is a new direction another directions and if I take the projections. Now, you can see that all the projected points for this group they are lying in this within this interval and all the projected points from for these group they are lying within this interval along this direction of projection along that particular direction. So, they are well separated.

And, as it; so, this shows direction of the principal component is not really providing you the good separations between data points. So, as we mentioned it is well separated, but not along the direction of principal component.

(Refer Slide Time: 14:51)

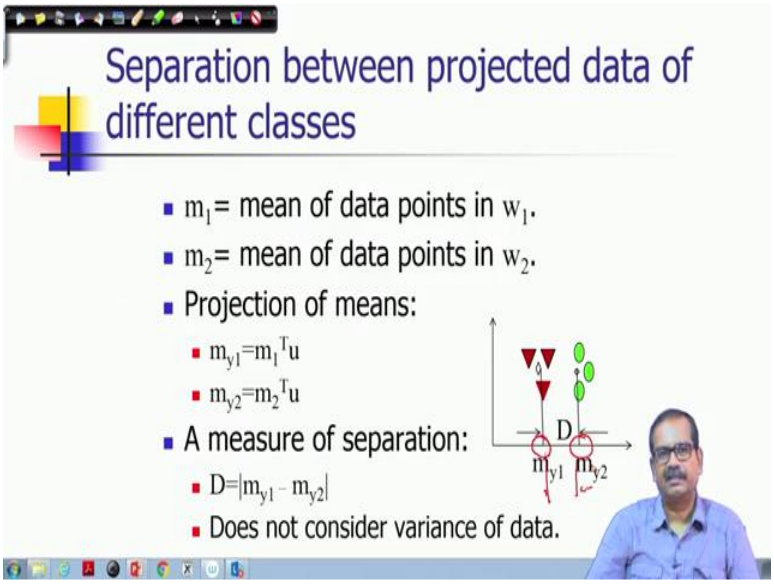


So, let me explain that what is the competition problem that involves in this particular analysis; Fishers linear discriminant analysis.

So, consider a set of data points as we have considered in the previous cases also. And, then out of those there are N1 data points which are in class w1. And, there are N2 data points which are in class w2. So, naturally we consider that N1+N2=N which is a total number of data points. And, now you consider a line with direction u, because our objective is to get a direction where this separation could be maximum. So, we should consider projection of data xi on u.

So, the projection can be operation can be expressed as a dot product of two vector xi and u or in the matrix representation, it is $xi^T.u$ matrix multiplication of xi^T and u. And that would give you the projection of data point yi, it is nothing, but a 1-dimensional sub space, which represents this data so, all the points around that lying on that line.

(Refer Slide Time: 16:18)

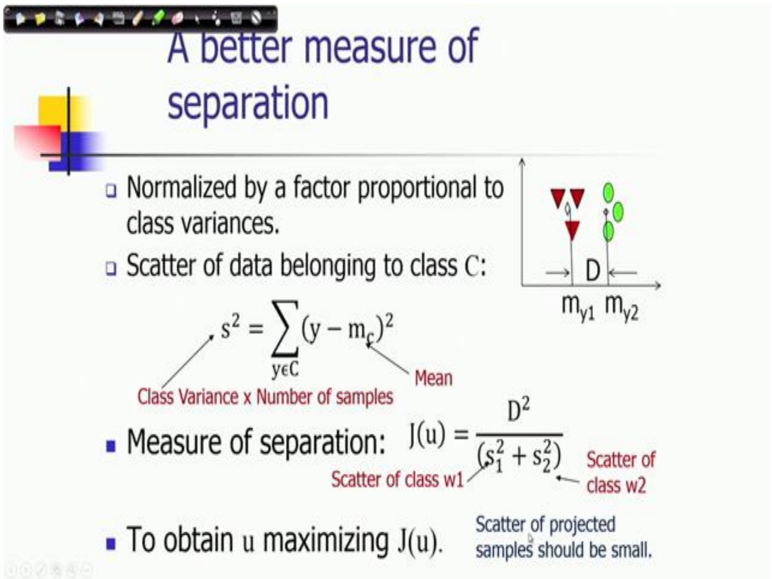


Now, we are trying to measure, what is how the projected data is separated? Because with that major then we can formulate a problem for maximizing that measure to get it well separated data. So, in this case let us consider mean of data points in class w1 is m1, and mean of data points in class w2 is m2. And, projection of this means along this direction u can be computed as projection of mean vector along u that $m_1^T u$ and $m_2^T u$

So, figuratively we can show here that m_{y1} and m_{y2} and one of measure of separation could be the separation of these two means; that means, how far there, we want this measure should be no large, this value should be large per well separated two groups.

So, if you consider the difference absolute difference between these two values, that would give you a separation measure, but the problem here is that it is not capturing the variance of data. So, some data could be very well spreaded, very largely spread it, in an area and some could be very closely groups, closely spaced and after projection. So, ideally what we would like to have, ideally you should have the data points they should be closely spaced around mean. And, they should be largely separated. Then only those kinds of measures I mean those are the desirable situations desirable cases for this projection, but that is not captured by simply by the value D.

(Refer Slide Time: 18:13)



So, what we do in that case we normalized this D by a factor proportional to class variance. So, in fact, we call it this factor scattered. So, scatter of data belonging to class C is defined in this way it is on the projected space.

$$s^2 = \sum_{y \in C} (y - m_c)^2$$

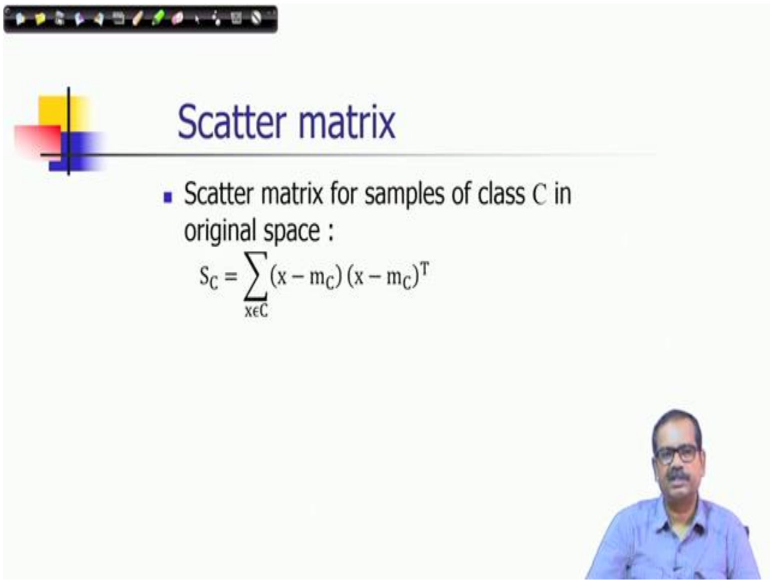
So, it is nothing, but as you can see, it is a sum of mean square of mean deviations from the mean from the mean of the data points. And, it is a proportional factor with variance, because it is nothing, but class variance product of class variance and number of samples. And, as I mentioned the m c here represents of the mean of class C and small s square is a scatter.

So, the measure of separation a good measure of separations could be that

$$J(u) = \frac{D^2}{(S_1^2 + S_2^2)}$$

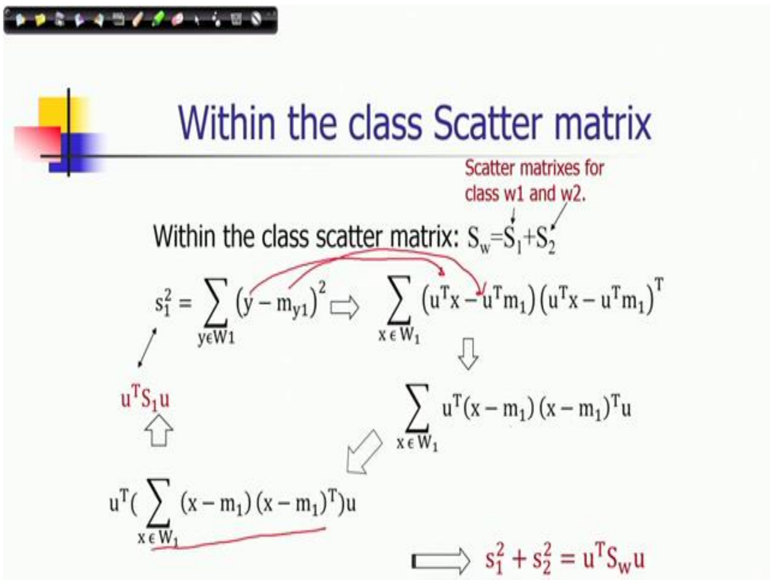
And, our objective is to maximize this value J(u) with respect to get a direction and so, scatter of projected samples should be small.

(Refer Slide Time: 19:44)



Let us define also scatter matrix of the original data. So, scatter matrix of sample of class C in the original space that is also defined in the same way. So as you can see it is the multidimensional space. So, earlier we have only one-dimensional space sub space for projections. So, we have used standard definitions of variances kind of definitions; that means, square deviations, but now you have to use the corresponding outer product of this matrix to denote a scatter matrix. So, this is represented as, as you can see

(Refer Slide Time: 20:33)



So, within the class scatter matrix, that is defined as scatter matrix for S1 and S2; sum of S1 and S2 that is within the class scatter matrix. So, each one for S 1 you have to consider once again that definition; that means, suppose for S1, we had m1 as the mean. So the vector $(x - m_1)(x - m_1)^T$ and sum of if I consider there are ith samples. So, these ith component and if there are N1 samples in S1.

So, we can show how it is related with this within the class scatter matrix. So, this can be written as in this form. So, you can see that this is what is the projected sample of x along the direction u. So, we can write $u^T x$; similarly, this is the definition of this is a projected mean of the samples along y along direction u. So, it is $u^T m_1$. So, this itself can be written

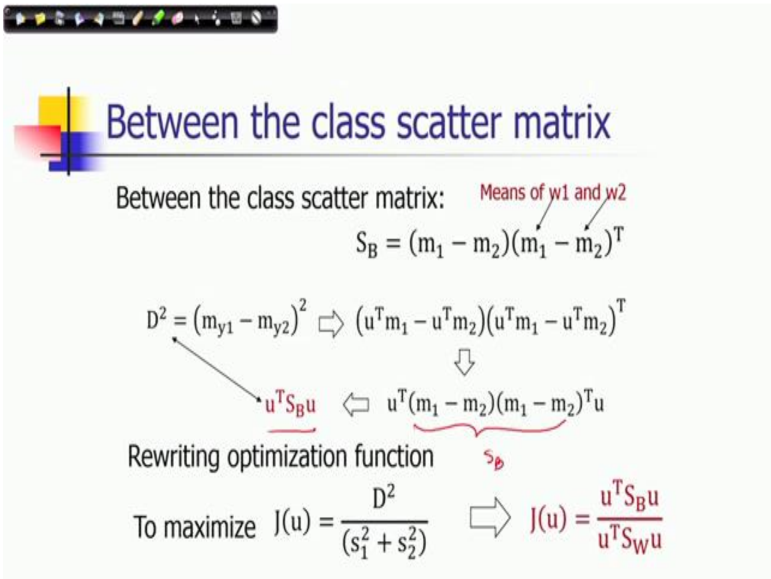
$$s_1^2 = \sum_{x \in W_1} u^T (x - m_1)(x - m_1)^T u$$

So, we can write in this way you can simplify this algebraic form and you can write in this form. You can note here actually since it is one-dimension. So, whether I put the transpose here or here does not matter. So, we have taken the convenient form for the sake of derivations for the sake of the final derivations. So, we can see that from here you can take out u transpose u outside and this will give you the scatter of the original sample for class W1.

And, this is the reason y^T here instead of taking u transpose a minus this could be as transposed and you could have taken this also. On the other hand we may and (Refer Time: 23:17) this is also correct. So but, the advantage of this notation is that, then we can bring the scattered matrix nicely within this u^T and u in this form, we can get it in this form.

And so, from here we can see that a small s_1^2 is related with the scatter matrix of that class by this explanation. So, s_1^2 is nothing, but $u^T s_1 u$. Similarly, s_1^2 would be $u^T s_2 u$ and if I add them finally, the both S_1^2+S_2 square can be written as $u^T S_w u$, where Sw is the within the class scatter matrix as we have defined.

(Refer Slide Time: 24:08)



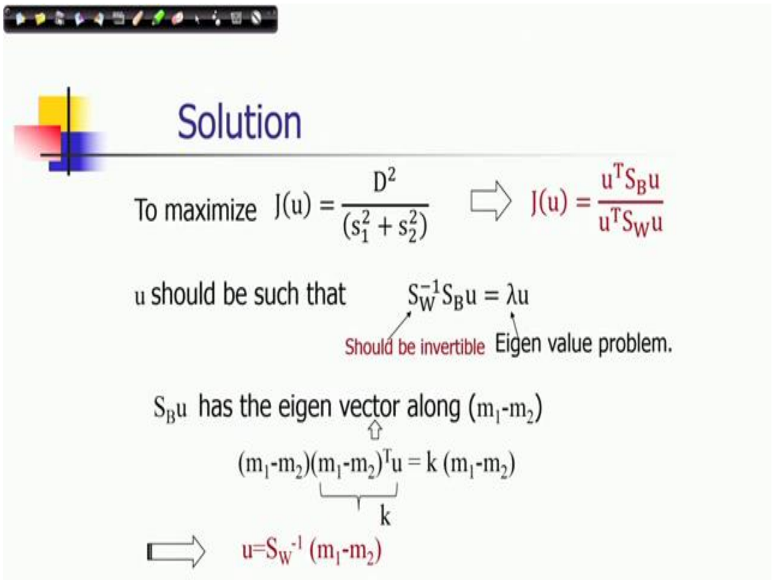
Between the class scatter Matrix; that is another definition. Now, this is the scatter matrix formed by the means of those classes. So, this is how the definition is there. So, $S_B = (m_1 - m_2)(m_1 - m_2)^T$, it is the difference between two mean vectors. So, difference vector of means that is defining this scatter matrix. And, here also we can show that the separation is related with this.

So the D^2 which is nothing, but the square deviation of two projected means that could be written in this form. Once again conveniently we have used the outer product from 2 to outer product matrix representation for showing the square of a one- dimensional term and this could be represented as

$$J(u) = \frac{u^T S_B u}{u^T S_W u}$$

Now, this part is nothing, but this is what is S_B and so, you can write it in this way and that is related with D^2. So, finally, we can rewrite this optimization function as it is to maximize these factor J(u) which is $\frac{D^2}{(S_1^2 + S_2^2)}$ this could be re written as $\frac{u^T S_B u}{u^T S_W u}$.

(Refer Slide Time: 25:51)



So, this is the problem of optimization, that you have to maximize it, once again u should be unit victor that is a constraint.

So, u should be such that it can be shown that

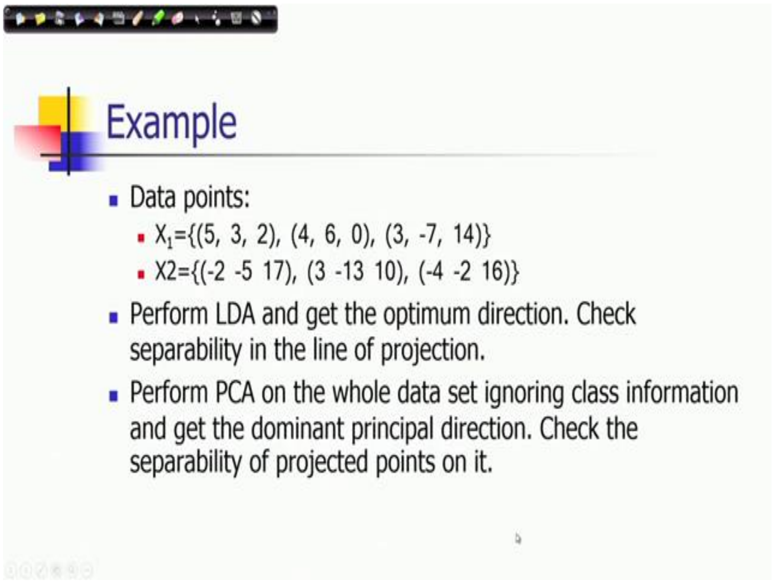
$$S_W^{-1} S_B u = \lambda u, S_W \textit{ is invertible}$$

So, I am not giving you the divisions I am providing you the final solutions that u is the eigenvector of these particular matrix. It is like an eigenvector solution $S_W^{-1} S_B$ and if we compute the eigenvector there. And, since now you want maximization which means you have to consider the maximum eigenvector corresponds to the maximum eigenvalue.

So, and one of the thing is that interestingly that S_Bu it has eigenvector along already (m1-m2), this can be shown easily, if I consider this particular expansion of S_B. So, this is nothing, but S_B. So, this part as you can see that this is the dot product of the difference vector with respect to u and we should be some scalar value. So, this is some scalar value k.

So, finally, this expression is nothing, but it is a vector this is some scalar term this is nothing, but a vector m1 - m2. So, the eigenvector of SB is also m1-m2. So, in that case the solution of $u = S_W^{-1}(m_1 - m_2)$, that is what is your solution and, in this way you can get the direction where you can get the maximum separation of projected samples; this is how the Fisher's linear discriminant works.

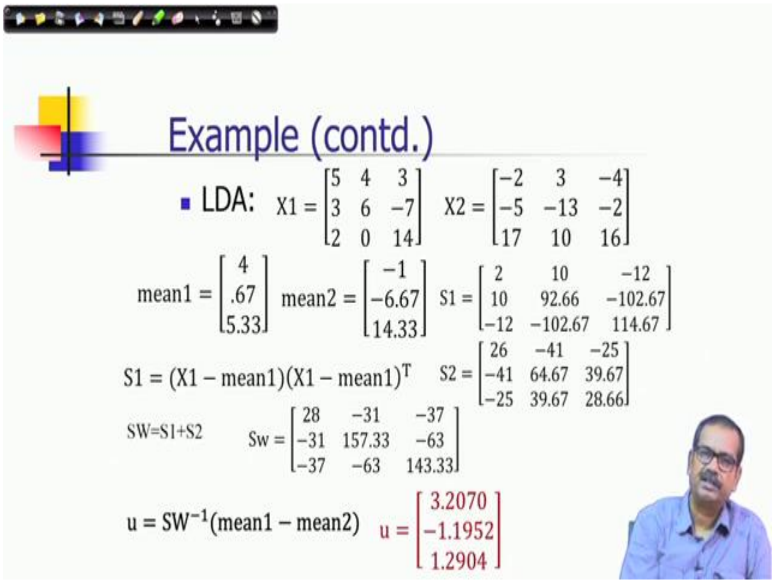
(Refer Slide Time: 27:47)



So, let me explain it with respect to an example consider a set of data points and now we are considering labeled set of data points. So, we have I show given you 2 such sets X1 and X2 and we perform linear. So, we have to perform linear discriminant analysis or to get the optimum direction or FLD actually Fisher's linear discriminant analysis to get the optimum direction.

And, we should check also the separability in the line of projections. As an alternative also we will perform PCA on the whole data set, because PCA does not consider class information, it considers the whole data set and get the dominant principle direction and check the separability of the projected points on it.

(Refer Slide Time: 28:38)

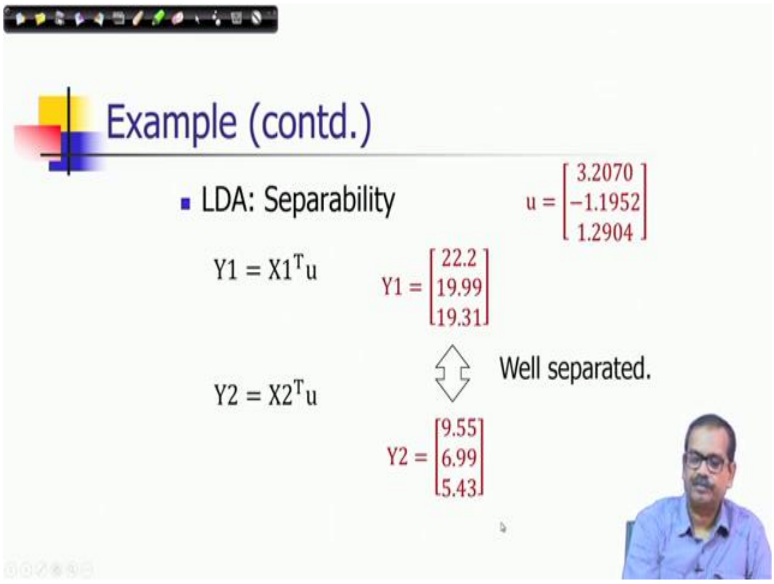


So, for a linear discriminant analysis or performing here Fisher linear discriminant, we consider this representation of data into two matrixes X1, 1 group of data all column vectors are data points X2, other column vector are here also column vector data points per group 2.

So, mean of X1 is given this that is what is m1 and mean 2 is given this, that is what is m 2 in our notations we have discussed earlier and this is what is your scatter matrix for the X 1. So, S1 is a scatter matrix for S1; similarly, you computes scatter matrix for X2 S2 and we within class scatter matrix can be represented. Now, this is a definition of S1, which we already we have discussed. So, within class scatter matrix is S1+S2, which is given this. And, now the solution you have to take the S_W^{-1} and then multiply that with m1-m2 that would give you the direction.

So this is what is your u, S_W^{-1}(m1-m2) and it could be found that u value is given this, it could be unit vector also it could be any vector, because it is just projection. So, magnitude of the vector does not.

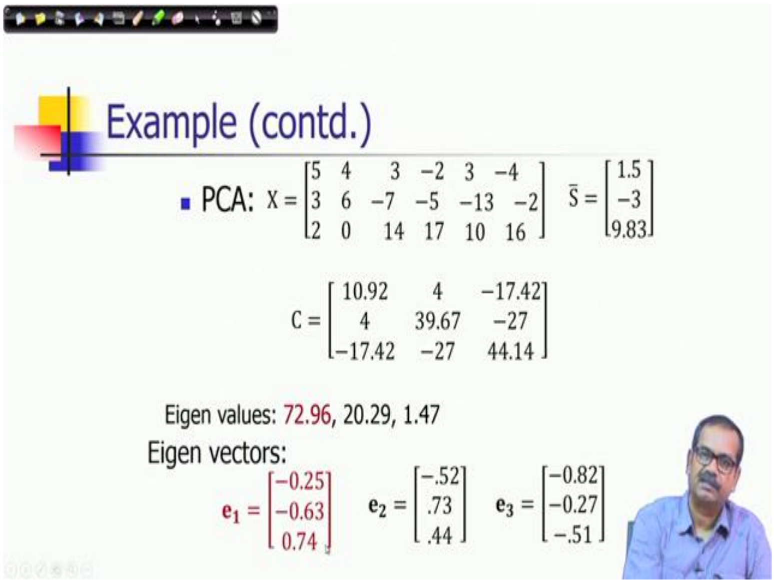
(Refer Slide Time: 30:08)



So, the separability if you would like to study you take $Y_1 = X_1^T u$. So, you are taking the projections of all data points in X1 with respect to u. So, you can perform all these operations simultaneously using these matrix multiplications. Similarly, you consider projections of all data points of X2 with respect u.

So, you will get them as column vectors there are three data points here. So, these are the projections for Y1, these are the projections for Y2. And, you can see that their intervals are well separated, because the range of Y1 is 19.31 to 22.2 whereas, range of Y2 is 5.43 to 9.55 and they are well separated.

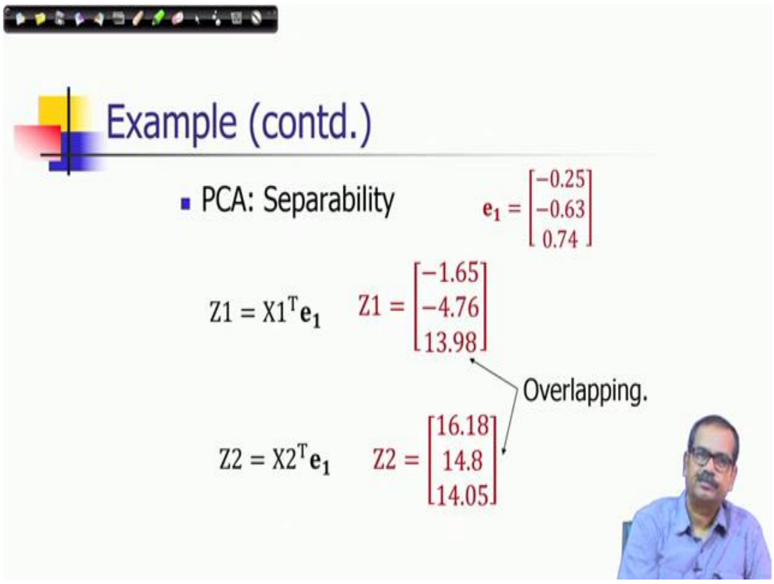
(Refer Slide Time: 30:55)



Whereas, if we take principle, if you perform principle component analysis on the whole data set which is represented here by the matrix X all column vectors are data points and you have considered union of data points of X1 and X2. So, part principle component analysis as we did earlier, first we compute the mean vector S bar, then you compute the covariance matrix, the variance of data.

So, sum of diagonal elements that would give you the total variance of the data. And, if I perform the Eigen; compute the eigenvalues and eigenvectors in order of there, values I am showing here. So, the maximum eigenvalue is 72.96 and eigenvectors corresponding eigenvectors are this e1, e2, e3 and we are interested on finding out the directions of maximum eigenvalue or maximum variance. So, this is your principle component direction of the principle component is given by e1.

(Refer Slide Time: 31:58)



So, if I take that direction and check the separability of projected samples so, you perform the same projections on that direction and these are the values we will get. We will see that at least there is one sample this is the (Refer Time: 32:14) example. So, I mean it approximately it is good well separated, but there is at least compared to the previous one, one sample is here 13.98, which lies in the interval or very close to the interval of the data points of Z 2.

So, the separation is not that much, it is not really (Refer Time: 32:38) overlapping in this sense here, but it is very close to that integral. So, that is it shows that the utility of fishers discriminate analysis. So, with that let me stop here and I will continue this topic in the next lectures.

Thank you very much for listening to my talk.

Keywords: Fischer linear discriminant, between class variance, scatter matrix, eigen faces.