Applied Multivariate Analysis

Prof. Amit Mitra
Prof. Sharmishtha Mitra
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

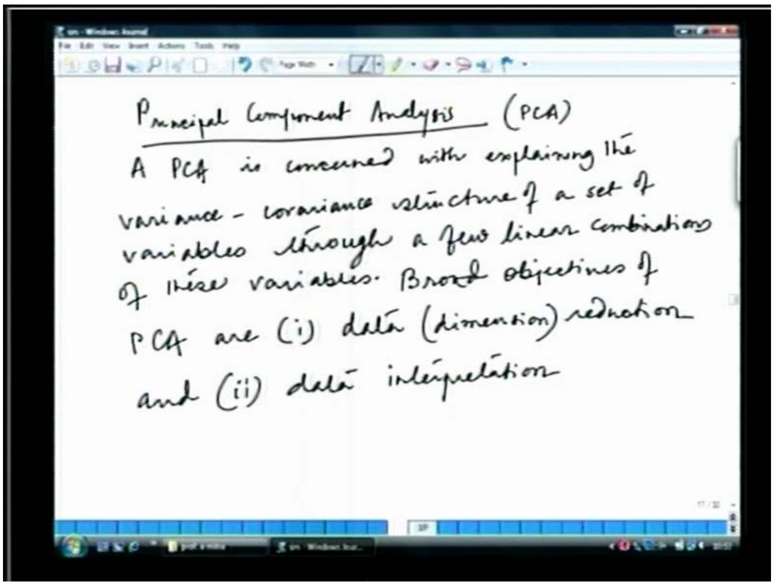
Lecture No. # 23
Principal Component Analysis

We are going to start this discussion with the topic of principle component analysis. If you recall manova - manova was all about partitioning the total variability in the data into components, which word you to the difference sources of variation. A principle component analysis, it also tries to explain the total data variability present with the help of a fewer number of linear combinations of the original data, meaning if I have a P dimensional random data vector say which means that I have P random vectors X 1, X 2 to X P say… Now, I am going to explain the total variability present in the data with the help of k new variables say Y 1, Y 2 to Y k, where k is the number which is much less than P, and these Y, i's are actually linear combinations of the original variables X 1, X 2 to X p.

So, so basically you can see the broad objective of principle component analysis is reduction in the data dimension; now once the reduction, and data dimension is achieved we achieve many more things. And one of the most important of which is interpretation, data interpretation, data projection, etcetera.

Now, strictly speaking the all the P variables are required, if I want to explain the the variability - the total variability present in the data, (( )) but in most situations we will see that our fewer number of the linear combinations of these variables will be good enough to explain the total variability. I mean coming parlance this is said that the information content of the variables X that is X 1, X 2, X P is as much as the information content of the or conversely we should say that information content of the new variables Y's are as much as the information content of the original variables, but we should take this with the bit of cushion and we must remember that this is with respect to the total variation in the data.

(Refer Slide Time: 02:35)

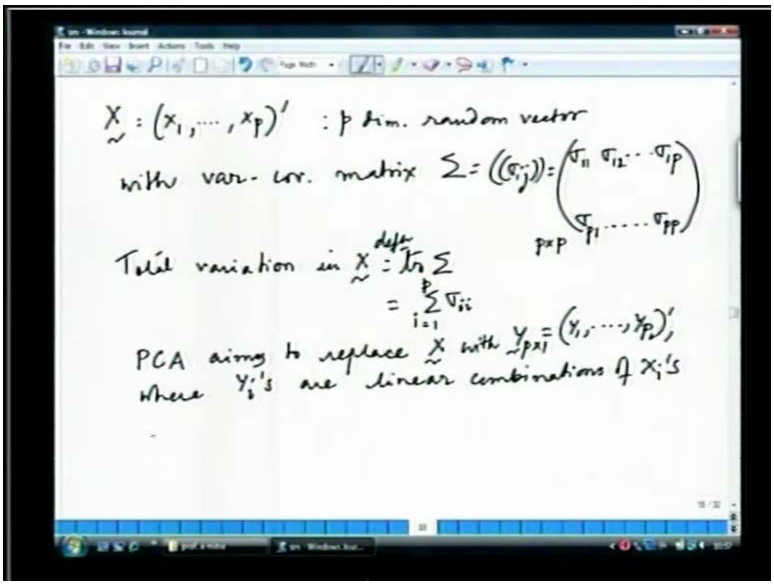


So, let us say first just very briefly write what is principle component analysis our new topic, this what it is doing in this analysis? What we are basically doing is a PCA is concerned with explaining the variance covariance structure, when we say the variability - the total variability in the data, this is through the variance covariance matrix of the random vector. How exactly that we will see once we define total variability in the data.

So, variance, covariance structure of a set of variables through a few - this few can be really very few like even k equal to 2 or one may be also good enough to explain the total variability through a few linear combinations of these variables. Why do we do a principle component analysis? So, broad objectives of PCA are the first one is data reduction rather we should say data dimension reduction, and the second one being data interpretation. Now, in between there are many thing before we can correctly interpret the data. So, we will look into all these aspects.

So, these are the 2 broad objectives, but in between there are many more other analysis that will help us. So, and besides an analysis of the PCA, it also it in time it reveal some interesting relationships among the variables - among the P variables which were not apparent otherwise. So, with the the the crack of the matter always remains the dimensionality reduction. So, once we do this, we can see that I can project the new variable - the two-dimensional variables say Y 1 and Y 2, and I can have a clear idea about the data cluster or if there is an outlier in the data. And of course of course, once we while we calculate the principle components, we get interesting relationships among the variables.

(Refer Slide Time: 05:52)

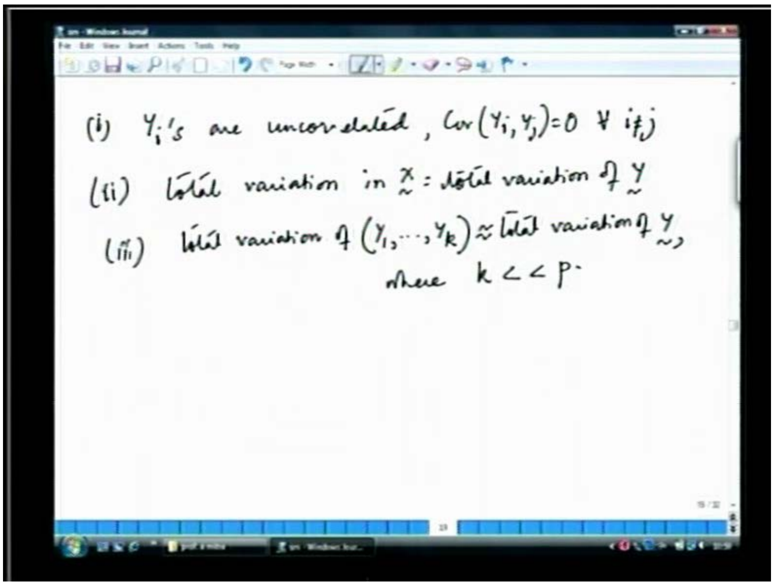


So, how is this done? So, first we have the whole thing is based on the variance covariance matrix. So, what we have is the random vector X, which is a P dimensional data vector X 1 to X P, and that is the P-dimensional random vector with variance covariance matrix sigma. A very general sigma elements of sigma i j, and we preferred to write it in the (( )) sigma 1 1, sigma 1 2 to sigma 1 P; and these these are symmetric matrix this is symmetric matrix we can write 1 P or P 1, which means basically sigma i j is equal to sigma j i. So, that is a P Y P dimensional square matrix, and we assume positive definiteness of this matrix.

Now, we have been saying that total variability in the data, total variability in X is going to be explained through the total variability of Y the new variables. So, what we exactly mean by total variability or total variation in data, and how is the variance covariance matrix coming into the picture with this concept. So, total variation or total variability also say in X, this is nothing but trace of the variance covariance matrix as simple as that so this is my definition. So, total variation in X is nothing but the trace of sigma which means that I consider some of the diagonal elements, some of the variances that is summation sigma i i form 1 to P. And then what does PCA attempt to do PCA aims to replace the X - this X the hole data vector with some Y.

And initially, we will look into P linear combinations of the variables. So, that is Y also P dimensional Y 1 to Y P, but first we were these Y i's - these are not just any variables, these are linear combinations of the original variables X i's, linear combinations of X i's, but not just any linear combinations.

(Refer Slide Time: 08:51)



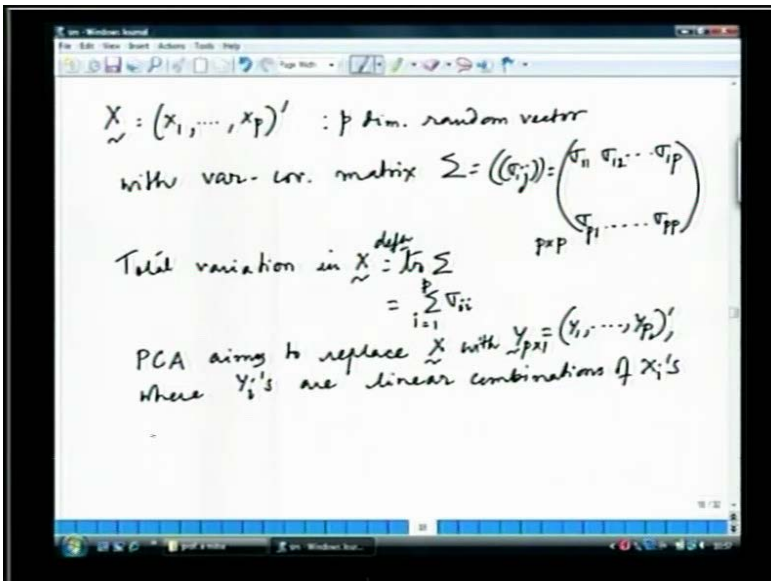
We must satisfy certain conditions, such that the first thing is I have to remember the Y i's are uncorrelated, note that we have not taken X i's to be uncorrelated, because I have not never said that sigma is a diagonal matrix, just the general variance covariance matrix which means X i's can be correlated also. But linear combinations through which now we are going to explain the total variability, these new variables Y i's they have to be uncorrelated; that is covariance between Y i and Y j, this is equal to 0 for every i not equal to j, this is the first point.

The second one is the very thing that we have started with the total variation in X, we say that the information content of Y is as good as the information content of X with respect the with with respect to the total variability present in the data. So, which means the total variation in X is equal to the total variation of Y, but then why do we choose to work with the Y, this is the situation, because here comes the most important point that the total variation of total variation of Y, which is of Y 1 to Y P, this is actually almost equal approximately equal to the total variation of Y 1 to Y k. So, this is approximately equal to total variation of Y, when I say total variation of Y i, I mean that all P members of Y are present, but the cracks of the matter is this total variation can be explained with a much less number of variables Y 1 to Y k.

So, when I say much less number of variables, I mean that k is where k is really less than P much less than the total dimension P. So, these are the three basic features of the new variables the principle components which are basically the linear combinations of the original variables that we have listed here, they form the cracks of the whole exercise, and we have to be careful in constructing our principle components in a manner, so that all these three properties are satisfied.

Now, before we formally define principle component, let us talk about some of the other uses of principle components. We had said the broad objectives are data dimension reduction, and data interpretation in between there are some other tasks that we can accomplish through the construction of principle components.

(Refer Slide Time: 11:49)

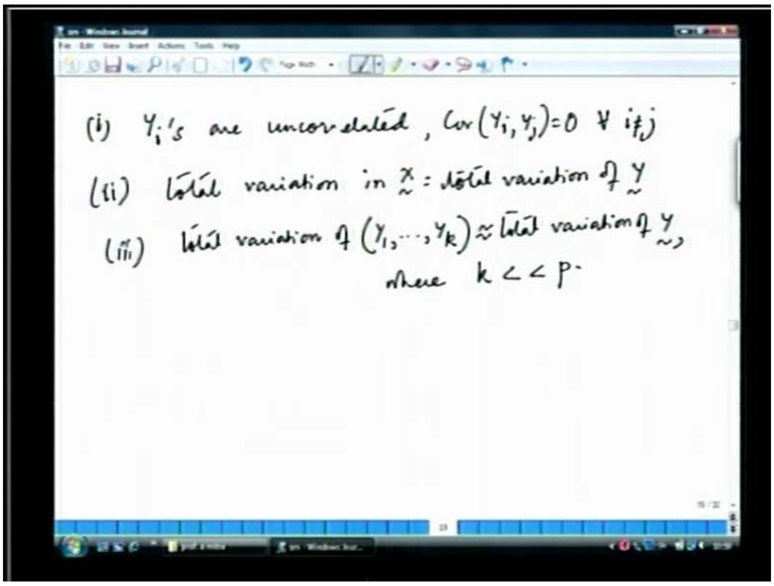


And these are write down the major uses of principle components for principle component analysis. What can we what all can be achieve through PCA is of course, the first thing is the data dimension reduction, and everything else that follows is basically is is essentially following from this fact sorry data dimension reduction. The second one of importance is… Once there is dimension a dimension reduction, we can project the data in a in our two-dimensional plane or at the most the three-dimensional plane to properly visualize the whole thing. So, data projection and visualization, this is possible if we can achieve a value of k equal to 2 at the most 3. So, that all the other all the properties that we have listed or satisfied, if that can be done then projection and visualization it can be done in really nice manner.

The third one is once we project the data, then there are certain features of the data that become a parent to us; that is we can see formation of data clusters. So, idea about data clusters, the the various groupings of the data; and of course, if we can project the data in a two-dimensional plane we can also see, if there is any outlier present in the data. So, that is multidimensional outlier handling, so multi-dimensional outlier detection that also can be done.

And fifth there can be some ranking of the multidimensional data also ranking of multidimensional data, and projection of the data can also tell us whether the population, whether the data comes from a multivariate normal population or not. So, that is checking for multivariate normality. So, we can handle as many things with a PCA, and all of these are very important practical application, practical uses the for the multi-dimensional data checking for last one is checking for multivariate normality.

(Refer Slide Time: 14:33)

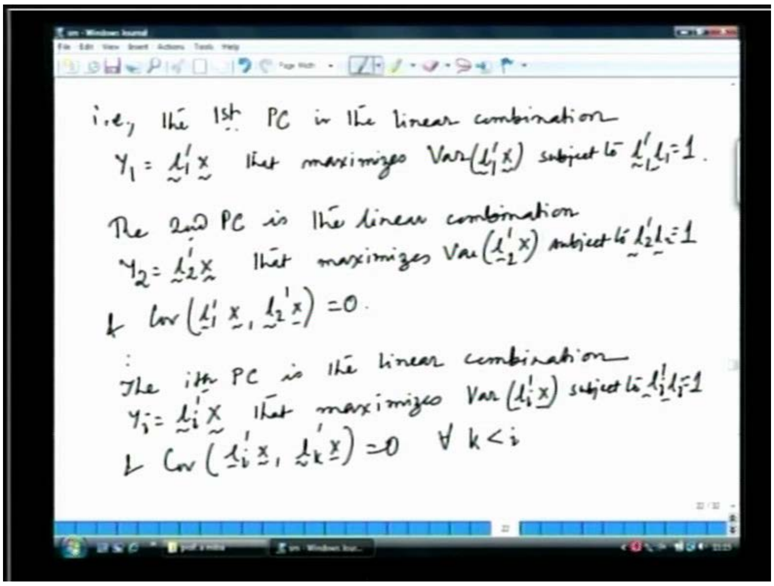


Next, we are going to define formally. What is the principle component? And we say that will we have all all we talked about if they are basically linear combinations of the original variables, such that certain properties are satisfied. So, this… So, definition of principle components. The principle components are the uncorrelated linear combinations Y 1, Y 2 to Y P, note that initially we talk about as many number of linear combinations of Y’s of of X is as there are number of X variables. So, we talk about Y 1, Y 2 to Y P, when we have X 1, X 2 to X P.

But ultimately we will work with the much fewer number of Y’s, so Y 1, Y 2 to Y k. So, initially we say that there is P such linear combinations - P is the equal to the number of data variables that we have originally. So, linear combinations Y 1, Y 2 to Y P, whose variances are in decreasing order. So now, this is something which we are saying for the first time uncorrelatedness of course, we said before. Now, something more we are saying we have Y 1, Y 2 to Y P; the variances of these Y 1, Y 2 to Y P, they are in decreasing order. So, what we have is variance of Y 1 that maximum say these are in decreasing order. So, Y 1 explaining the maximum variability, Y 2 explaining the second highest variability, and so on.

It is it is very logical that we have this criterion on the principle components, because our ultimate aim is to restrict the number of principle components to as few as possible. So, if this can be possible, if only the first one it can explain the maximum of the variability. So, it it can be so high that some sometimes we may be satisfy with the first principle component only, and then in that case we will say that the whole data dimensionality has been reduced to one, we are happy with the value of k equal to 1 as low as that. So, we we have been must have the principle components designed in this fashion, while to explaining the second highest variability, and so on.

(Refer Slide Time: 17:52)



So, what are the the things that, we are talking about that is let us now try to sum up the situation. The first principle component, the first PC is the linear combination, will we are using the notations Y's for the principle components. So, the first principle component is the linear combination Y 1 the first one, this is l 1 dash l 1 prime x. So, this is essentially a P dimensional vector known vector, it should be so that Y 1 is known (( )) what what is the linear combination of X that I am using for Y 1, that I am getting for Y 1. So, that is Y 1 is l 1 transpose X such that, that maximizes variance of l 1 X subject to l 1 transpose l 1 is equal to 1.

Now, why is this required now, I say that this Y 1 formed as l 1 transpose X it should be such that variance of l 1 transpose X is maximum. Now that can be if I consider any other as as scalar multiplication of the linear of the of this vector l 1, and I consider say some l 1 star which is c times l 1, c is the very high constant. So, then I can always have variance of c l 1 transpose X greater than variance of l 1 transpose X.

So, to put a check on that, and to achieve some uniqueness I I require this factor, I put this criterion that it is subject to l 1 transpose l 1 is equal to 1. Then the second the principle component is the linear combination Y 2, this is some other linear combination of the X is X 1, X 2 to X P, such that the variance is maximizes variance of Y 2, that is l 2 prime X subject to l 2 transpose l 2 is 1.

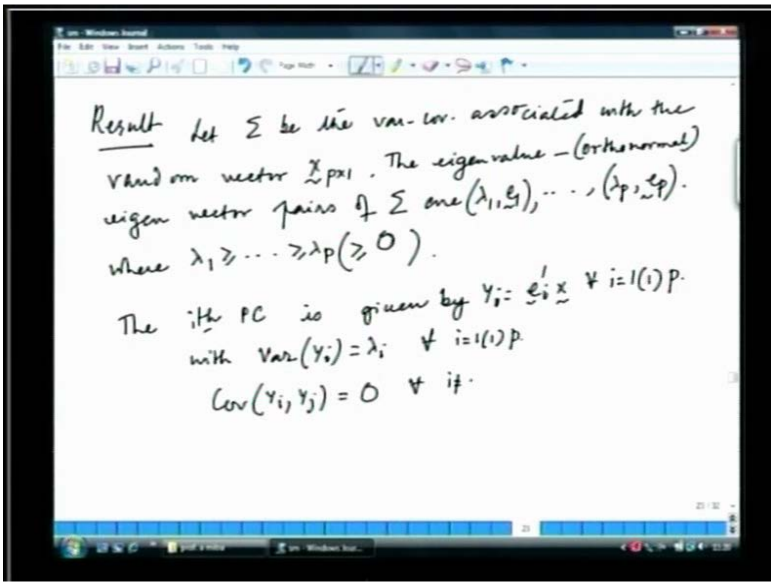
And we must have something else here, and covariance between Y 2 and Y 1 is 0, that is l 1 prime X. And l 2 prime X this is equal to 0, and in this way I go to the i th principle component, i th principle component is the linear combination Y i is l i transpose X, that maximizes the variance - variance of l i transpose X subject to as before l i transpose l i this is equal to 1. And covariance between l i X with say some l k X this is equal to 0, and now this has to be true for all k which is less than i right.

See, if I go to the third principle component I must check the covariance between the third, and the second, and the third and the first as well; and both of these have to be equal to zero. Now, what I am saying here, we will this guaranty mean the things that I have said before, that is the first thing was that the principle components have to be uncorrelated, they should maximize or the first one should have the maximum variance, the second principle components should have the second highest variance, and so on.

The total variability of these Y should be equal to the total variability of X and last, but not the least. The total variability of Y can be explained through the variability of of fewer number of a very few number of Y's.

So, all these things whether those can be satisfied with this type of a construction that I am saying for this. We go on to the next thing, let us see that the way that we are saying the principle components we are describing they can in fact, satisfy all the properties.

(Refer Slide Time: 23:05)



So, the first result is let sigma be the covariance matrix, the variance covariance or the dispersion matrix associated with the random vector X, the whole exercise will be done through the Eigen value Eigen vectors of this sigma matrix. So, the with the random vector X the Eigen value, and Eigen vector ortho normal Eigen vectors. So, corresponding or the normal Eigen vector pairs of this sigma matrix are l 1, e 1 P of them. So, up to l P e P.

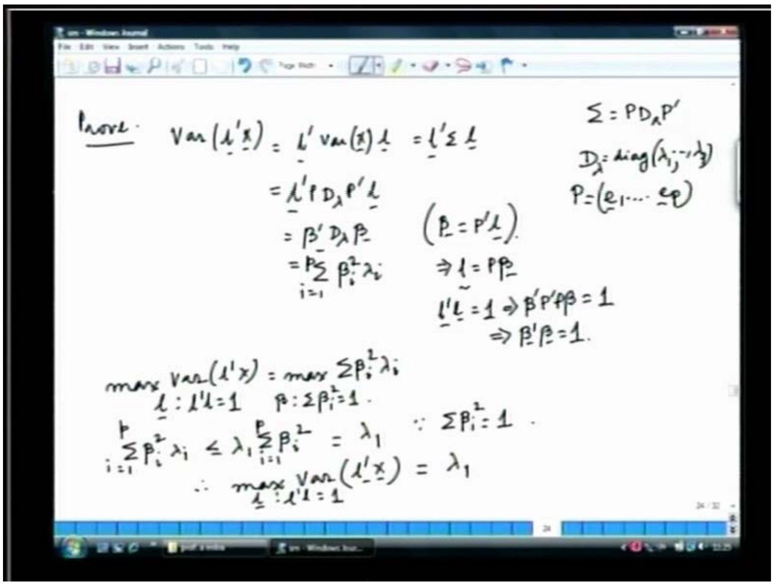
Let us say, where I have lambda 1 greater than equal to lambda 2, and greater than equal to lambda p. So, this is how I have arranged the Eigen values, and the corresponding ortho normal Eigen vectors, and I have this lambda 1 greater than equal to lambda 2 up to lambda p. So, these are in decreasing order, and each of them of course, are greater than equal to 0 means for positive semi definiteness also we can have strictly speaking, but most most of the situation will have this as positive definite matrix. So, leave it like this.

And then the i th PC is given by we say that the i th PC is given by Y i, this simply turns out to be e i transpose X for every i from 1 to p. So, after having said all these things what we do is simply consider the sigma matrix calculate the Eigen value, and the corresponding I ortho normal Eigen vector, and we see we will see that the i th principle component is nothing but a linear combination of this type, where we are taking help of the ortho normal Eigen vectors. So, the linear combination that I have is Y i is nothing but e i transpose X.

Now, if Y i's are this other other property satisfy, they will be satisfy, because simultaneously we have something for these Y i's, therefore these are for every i from 1 to P with variance of Y I, we will see that this is nothing but lambda i for every i from 1 to p. So, that another property if you recall of the principle component is a is satisfied, that is the first principle component will satisfy the maximum variability, its its variance will be higher than the variances of all other principle component. So, this is true, if I have the variance of Y i equal to lambda i.

The first principle component will have variance lambda 1 which is greater than lambda 2 to lambda P, and so on. And another thing was whether these are uncorrelated we will see that covariance between Y i, and Y j will be 0 for all i not equal to j.

(Refer Slide Time: 26:50)



So, let us prove this result, we have made a strong statement that the linear combinations or the principle components are nothing but the linear combinations of x is in the in the way along with the ortho normal Eigen vectors simply of the dispersion matrix.

So, for the proof of the result we start with let us consider variance of l prime X, I consider one linear combination of X, and I check its variance which is nothing but l transpose variance of X storing from the scratch, and this is nothing but l prime sigma l, and this I preferred to write by after using the spectral decomposition of the sigma matrix. So, I have P D lambda P transpose l.

Now, I I have already said that so this sigma is given in terms of its spectral decomposition P D lambda P prime, I have already said that Eigen values of sigma are lambda 1 to lambda P, and e 1 to e P are the corresponding ortho normal Eigen vectors. So, I know the structure of D lambda, this is nothing but diagonal lambda 1 to lambda P, and the P matrix has e 1 to e P has its columns. So, P is an orthogonal matrix. Now, this is something like I can write for this l prime P D lambda P prime l, I can write it something as beta transpose d lambda beta, where beta is nothing, but P prime l right.

And that is well, that is nothing but I have a beta vector, I have a diagonal matrix whose diagonal elements are lambda i’s, and then beta vector again, so that is nothing but summation of the type beta i square lambda i, i from 1 to p.

Now, what I am required to do is to get. So, I have beta is P prime l. So, this also implies that l is nothing but if you consider l what you have to do is simply pre-multiply this with P transpose inverse that is possible, because P is an orthogonal matrix. And since P is orthogonal this is nothing but l is nothing but P beta. And P beta this relationship gives me a very important thing that l transpose l equal to 1 implies that you have beta transpose P transpose P, and then again beta is equal to 1, and that implies beta transpose beta is also equal to 1.

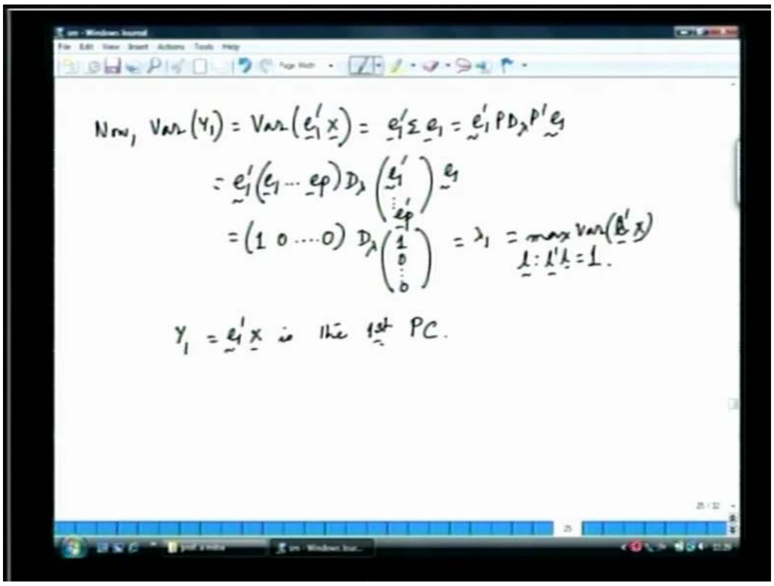
So, l transpose 1 equal to 1 is equivalent to saying beta transpose beta is also equal to 1. So, that now that I have to maximize variance of l transpose X over l, such that l transpose l is equal to 1. So, this can be said that equivalently I can maximize this expression, which I have I have shown to be equal to the variance, I have to maximize summation beta i square lambda i over beta such that well - such that beta transpose beta is equal to 1. And terms of summation I can write this as such that beta i square is equal to 1.

Now, I have summation beta i square that is that is the variance of l prime x. So, I can if I can obtain and upper bound of this expression subject to the condition that summation beta i square equal to 1 that I am true.

So, I am trying to looking into its, so I have I have summation beta I square lambda I sum from i from 1 to P, this has to be less than or equal to if I replace all the Eigen values with the maximum Eigen value. So, I am writing lambda 1 for all lambda i's, and hence I get this less than equal to sign, and then this summation beta i square remains there right. So, I have and then then when I have this is equal to lambda under the condition, since summation beta i square is equal to 1. So, I have achieved that maximum of variance l prime X maximum over l, such that l prime l is equal to 1 is nothing but lambda 1, because I have shown that this variance of l prime X is nothing but this summation beta i square lambda i, and the condition is nothing but summation beta i square equal to 1.

And then I have seen I have I have shown that I can obtain an upper bound of this term, and it is it is nothing but the maximum Eigen value lambda 1. So, this has been shown that variance of l 1 prime X under the conditional prime l equal to 1 maximum of that is equal to lambda 1.

(Refer Slide Time: 32:39)



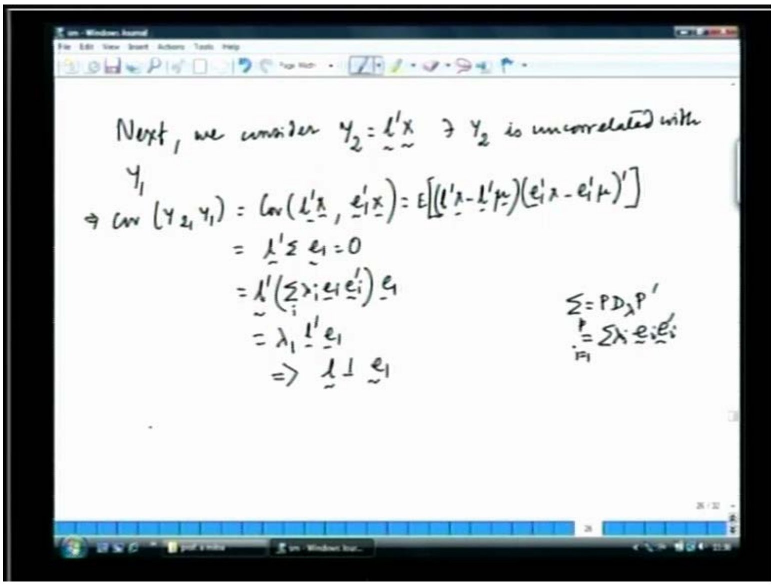
Now, I consider variance of Y 1, and Y 1 the one that is given to me that is variance of e 1 prime X; Y 1 is said to be equal to this linear combination, and this is equal to e 1 prime sigma e 1, and this is again by using the spectral decomposition of sigma this is P d lambda P prime e 1 right.

Again I can handle this, I have write this e 1 prime for the P matrix I am writing e 1, e 2 to e P, and then I have the diagonal matrix d lambda, I am writing P transpose matrix e 1 transpose to e P transpose, and then e 1 again. If this is so by the fact that this e i's are ortho normal, I will have this the this operation here combining this vector, and this matrix is going to give me the vector 1, and then followed by 0. And here, I have the diagonal matrix, and similarly I have the vector this one here.

So, this is nothing but lambda 1, because the first diagonal element of D lambda is lambda 1, and only this is coming into the picture with one has the members here. So, that is essentially 1 times re-lambda, and that is lambda 1 which is equal to maximum of variance lambda sorry l prime X maximum over l, such that l prime l equal to 1. This is actually equal to this maximum variance which we have seen just now.

So, I have Y 1 equal to e 1 prime X is the first PC, because as far as the first PC is concerned concerned, I will have to check only one thing that its its variance is having the maximum variance, and if its variance is lambda 1, and it is greater than all other lambda Eigen values, and it is actually equal to maximum of variance l prime X the maximum over this this of this choice of l with only this condition in place l prime l equal to 1. Well I have achieved to the the whatever criterion - the single criterion that was required for my first principle component, and I have Y 1 is e 1 prime X is the first principle component. Then I go to the next one that is construction of the second principle component.

(Refer Slide Time: 35:45)

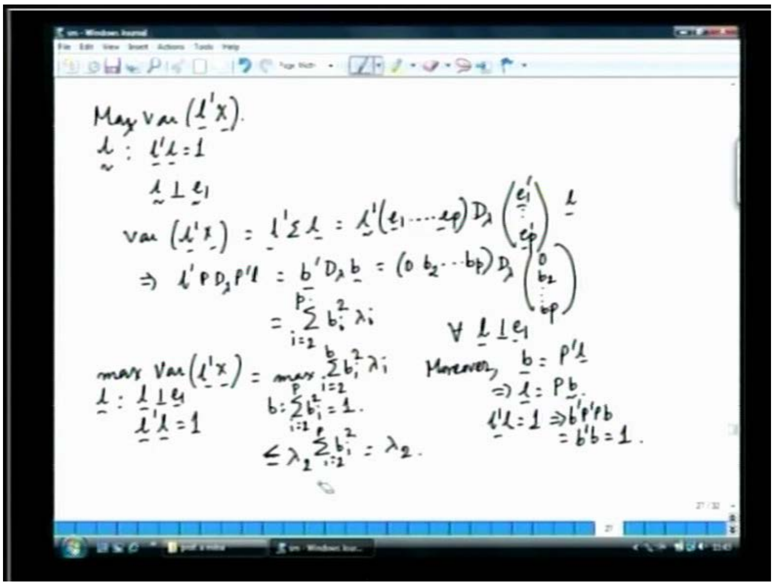


And next we consider the second principle component. So, next we consider Y 2, linear combination - another linear combination of the original variables X 1 to X P, such that Y2. Now, here we have to remember 2 things. Firstly, that Y 2 is uncorrelated with Y 1, this factor did not come when you are considering the first principle component. And secondly, the variance of Y 2 has to be less than variance of Y 1, these two a properties have to be satisfied in the construction, such that Y 2 is uncorrelated with Y 1, this is number 1, and so that is what we are getting is that implies that covariance of Y 1, and Y 2 this has to be equal to 0.

So, we are considering covariance between l prime X, and now we know what is Y 1. So, I take that form of Y 1, e 1 prime X, and if you see that this is nothing but its its nothing but you have l prime X minus its expectations. So, this is something, we we are introducing here, we are assuming that the mean vector is of X is mu. So, that is there. And then I have e 1 prime X minus e 1 prime mu, this expectation is nothing but l prime sigma e 1, this is equal to 0. So, this is giving me l prime sigma e 1, this is nothing but l prime, and I consider another alternative form of this spectral decomposition of sigma. We know that this sigma which is P D lambda P prime can also be written in the summation form that with lambda i the scalars, and then the vectors coming into the picture it is not P i's, but e i's we are denoting them by e i's. So, this is lambda i e i e i prime sum from 1 to P.

So, we use this form here for sigma this form of the spectral decomposition, this is the summation lambda i e i e i prime I from 1 to P, and then you have 1. So, this factors leading leading me to this covariance being equal to 0, because this is nothing but you have lambda one and only coming out, and l prime is combining with e 1, and this is equal to 0 implies that that l is orthogonal to e 1. So, covariance of this equal to 0 is leading me to the fact that this l has to be constructed in such a way such that this is orthogonal to the vector which is coming in the first principle component that is e 1.

(Refer Slide Time: 39:09)



And then we have to consider the maximum of variance l prime X maximum over l, such that now we have as we have two properties of Y 2 to satisfy. So, similarly we have 2 conditions - two types of conditions on l. One is there already which we know that l transpose l has to be equal to 1, and the other one is something which we have seen just now that l has to be orthogonal to e1. So, these two conditions have to be simultaneously satisfied, and then we have to get the maximum variance of l prime x. So, how is this done? We have variance of l prime X, this is nothing but l prime sigma l, and let us use the usual form of spectral decomposition we have e 1 to e P, this is how I am writing the matrix P. Then I have D lambda, and then P transpose e 1 transpose to e P transpose with l in the end; this is l transpose.

So, this is giving me this is giving me l transpose, we have earlier seen that l transpose P d lambda e prime l is something like b transpose d lambda b, which is sum 0 to b 2 to b P, we have in place because all I have the condition that I have is l prime is orthogonal to e 1, and not so with other e e vectors. So, I have 0 to b b 2 to b P and then D lambda, and then I have the null vector the the vector, this b 2 transpose to b P transpose right.

So, this is sorry, these are these are essentially scalars. So, we are talking about the elements of the v vector. So, these of we have here, these are this is fine, and this is the elements of the b matrix. So, I have 0 to b 2 from b 2 to b P right. So, this is like a sum summation b i square with lambda i from 2 to P right.

So, for all now this is; obviously, for all l which is orthogonal to e 1, we have use this factor and how is this coming moreover we have we have certain order things to be followed we have b is nothing but if you see that b has been replace for P prime l. So, that l is l is nothing but P b, and l prime l equals 1 implies that you have b transpose P transpose P b which is equal to b transpose b, this is equal to 1. So, (( )) the whole conditions structure can be reduce to this fact that I have to maximize variance of l prime X subject to that l is orthogonal to e 1, and l transpose l is equal to 1.

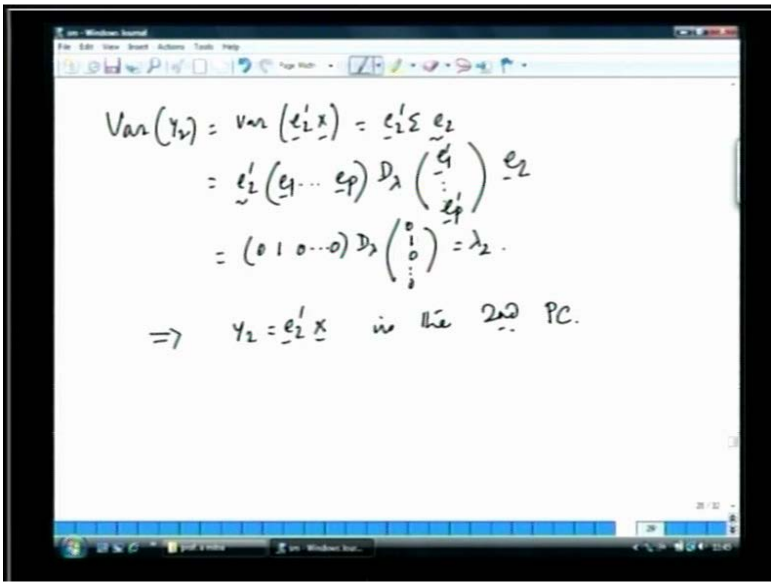
Now, with l orthogonal to e 1, I have seen and I just saw that this variance l prime, X is nothing but equal to summation b i square lambda i. Now, again I have to consider its maximum with the fact that beta prime b b prime b is equal to 1, because one condition I have already incorporated while I got the form this summation b i square is lambda i. I have already incorporated the condition that l is orthogonal to e 1, I have got yet one more condition to be satisfied that is b transpose b is equal to 1. So, I have to consider the maximum of this expression summation b i square lambda i, such that summation b i square is equal to 1.

So, this implies that I have maximum of variance l prime X maximum over l, such that l is orthogonal to e 1 let us write this first, because this condition has been taken care of in the in the first place, and then I have l prime l is equal to 1 is nothing but maximum of summation b i square lambda i; i from 2 to b summation over maximum over b such that sum of b i square, i from 2 to P is equal to 1 right.

So, this can be achieved, if again I replace the lambda i's by they are maximum value, now here the lambda i's are form 2 to P. So, the maximum value of this lambda 2 to lambda P is nothing but lambda 2, and then I have this as. So, this here I can replace this equality by less than or equal to, and this by summation b i square 2 to P, and this is equal to lambda 2.

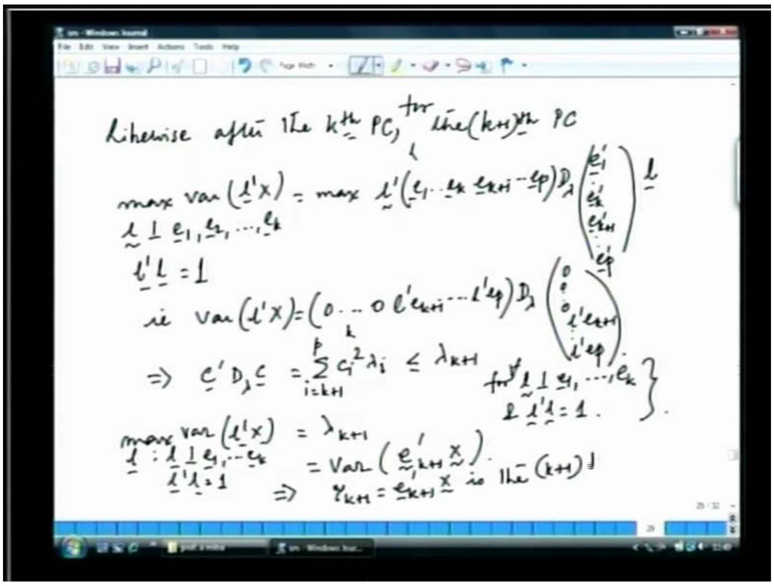
So, I have seen that, if I consider this second principle component its variance is lambda 2 which is less than lambda 1. So, its variance is actually less than variance of Y 1, not only that this covariance of Y 1 and Y 2 consider in this way is also equal to 0. So, I have successfully constructed the second principle component; therefore, I can write it here just one line, this implies that Y 2 equals to e 2 X is the second before that let us let let us just check the variance of e 2 X, just the way we have done in the case of the first principle component, and then only we can comment on that.

(Refer Slide Time: 46:12)



Now, next variance of Y 2 which is variance of e 2 X right, this is nothing but e 2 transpose sigma e 2, and this is e 2 transpose you have the P matrix, we are using the same form P D lambda P prime just spectral decomposition. So, we have e 1 to e P, then D lambda, then we have e 1 transpose to e P transpose l 2 sorry it is not l, but e 2 right, and this is nothing but because e 2 is ortho normal to all the in the e 1 to e P are all ortho normal Eigen vectors. You you will have e 2 combining with e 2 only, and since these are othho normal you get one here, so this is basically 0, 1, 0; then you have D lambda, and then again 0, 1, 0 to 0 which gives you lambda 2. So, all these in sum up to the conclusion that this implies Y 2 is e 2 prime X is the second principle component.

(Refer Slide Time: 47:49)



So, in this way we can go up to the k plus 1 (( )) one say likewise after the k th principle component, the k plus 1 th principle component. So, after going to the first, the second, and then we go to the third principle component is for the k plus 1 th principle component. We must have maximum of variance of l prime, X l is now orthogonal to e 1 to e 2, all these e k's right, k plus 1 is less than all 1 2 to e k plus 1 less than 1 2 to k.

So, l has to be orthogonal to each of these, and of course the original condition that l transpose l is equal to 1. And this will be nothing but maximum of l with e 1 to e k, and then you have e k plus 1 to e P, then D lambda transpose of these e 1 to e k e k plus 1 to e P, and then you have l right.

So, this is going to give you. So, I have you have variance of l prime X is now going to be 0 for k times, and then you have l prime e k plus 1, and then up to l prime e P, l is not orthogonal with these, then you have D lambda. And similarly, you have this 0, and then again you have l prime e k plus 1 up to l prime e P right.

So, this variance is nothing but this implies that you have a situation, where this you can define some vector to let us call this as some C vector. So, we have C transpose D lambda C, this is equal to C i square lambda i, now i is going from k plus 1 to P now. And this obviously, has to be less than equal to lambda k plus 1.

Now, note that while we are writing this we are considering the fact that for for l for every l orthogonal to e 1 to e k, and as well as l transpose l is equal to 1. After considering these 2 set of criteria we obtain this, and this gives us the maximum of variance of l prime X is maximum over l, such that l is orthogonal to e 1 to e k. And l prime l is equal to 1 this is nothing but lambda k plus 1; and lambda k plus 1 can again now shown to be equal to variance of e k plus 1 prime X. Giving us that Y k plus 1 is e k plus 1 prime X is the k plus 1 th principle component.

We have talked about certain other properties of the principle components, if you recall the an important such property was the total variation of X is equal to total variation of Y. So, we will begin our next session by proving that result.