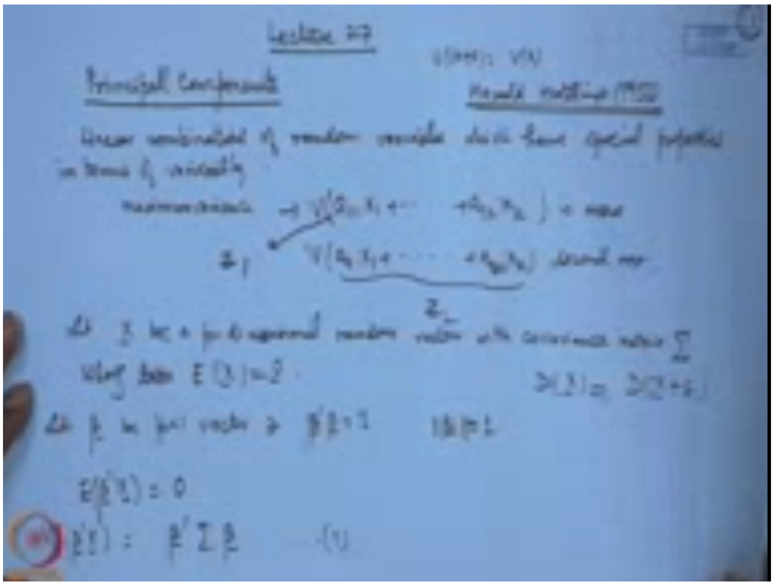
Statistical Methods for Scientists and Engineers
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology – Kharagpur

Lecture - 27
Multivariate Analysis - XII

Friends, today I will introduce a new topic in the Multivariate Analysis. This is called principal components, our principal component analysis. In principal component, so basically in any study in which we are dealing with the several variables one of the prime concern is to look at the variability of the variables. That means in those random variables how much variation is there, that means how much variability these variables are contributing to the model.

For example, you may have a regression model, you may have a time series model where variability's of importance. Now in case the random variables are the variables under considerations are many then we may have to discard some of the variables because it may not be feasible to study all the variables at a time. In that case it may be convenient to consider certain linear combinations of the variables which are more influential in terms of variability compared to other.

(Refer Slide Time: 01:31)



So the problem of principal components is the determination of linear combinations of random or statistical variables basically which have special properties in terms of variability. So for example, the first principal component is the normalize linear combination which will have maximum variants. So we can order them like we can put the one which is having the maximum variants so that linear combination let us consider, okay.

Say for example, a11 x1+and so on a1ksk, so this could be having the, that means the variance of this is maximum. Then you consider second maximum a11 x1+and so on a1ksk, second maximum and so on. So basically what happens that in many of the practical studies it is found that some linear combinations will contribute more that means they will have almost 90% or 99% of the variability and very small variability will be there in all the components.

So, basically what we can do that we can consider this as a new coordinate system we can call it Z1 this one we can call Z2 and so on. So in place of the original variables X1, X2, Xn, Xk. We can consider new variables Z1, Z2, Zk. Now out of this it may happen that only Z1, Z2, Z3 they are contributing say more than 99% of the variability then we may discard other variables and we can consider our relationship of the variables among these itself.

So, basically this is the, you can say in a nutshell the problem of principal component analysis. Now let us develop the mathematical procedure to derive this. The procedures for finding out principal components originally were discussed by Harold Hotelling, an American statistician in around the year 1933 he discussed or he derived this methodology. So we are considering let x be a p dimensional random vector with covariance matrix, say sigma.

Now this may have certain mean vector say mu but without loss of generality I can shift it to mean vector 0. Because in the discussion of the variance covariance matrix the mean does not play any role because if I shift all the observations by the same amount then the variability does not get affected. For example, if I look at the simple property like variance of x+a that is the same as variance of x.

So in general, if I consider the dispersion matrix of say x vector or the dispersion matrix of x+say some c then this is c. Therefore, we can consider without loss of generality take expectation of x to be 0. So, this will simplify in terms of calculation if we take this to be 0 otherwise you can go with mu also. But when the calculation I am presenting there will be easy if I use this one.

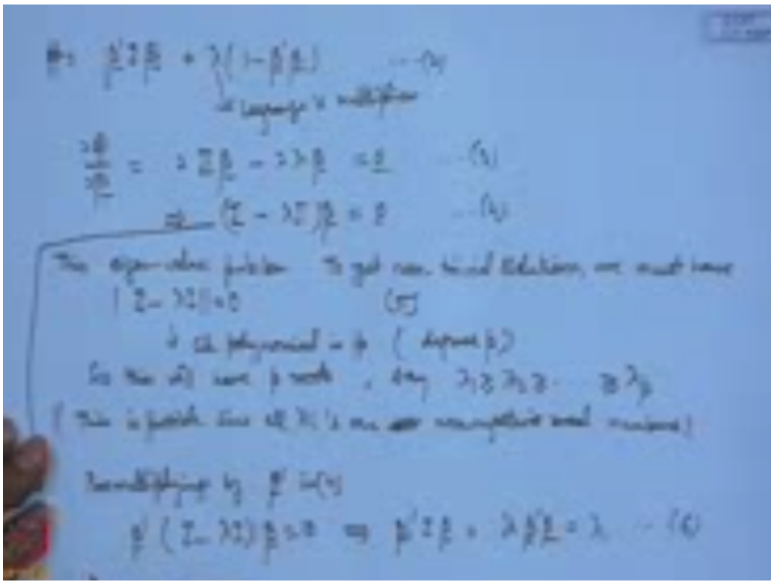
So let us consider let beta Ba Py1 vector such that beta prime beta is = 1. What does it mean? It means that the norm of beta, norm of beta is 1. So actually I will consider that linear combination which will be normalized here. So now you take expectation of beta prime x, so naturally that is going to be 0 because I have assumed expectation of x to be 0 here. So this is going to be 0. Let us look at the variance of beta prime x.

So why the formula for the variability it is = beta prime sigma beta. This is the formula for the dispersion matrix. So if the dispersion of x is sigma then if I consider any linear combination of this then the variance of that will be beta prime sigma beta. So what is our problem now? I want a linear combination, so this is a linear combination because what is beta prime x you can consider it as say beta1 x1+beta2 x2+beta p xp.

So our aim is to find out those values of beta1, beta2, beta p such that the variance of beta prime x is maximum but of course we have put a condition here. We want the normalization here that is norm of beta is -= 1. Why that normalization is required? Because we can make it free from the units of the measurement here. So we can consider here we want to find beta such that variance of beta prime x is maximum subject to the condition norm of beta is = 1.

So we can consider the method of Lagrange's multiplier.

(Refer Slide Time: 07:33)



So we considered a function say phi function that is = beta prime sigma beta+lambda time 1 – beta prime beta. This is a simple Lagrange's multiplier term here. So this is the 5 function and if we consider here this is del phi over del beta that is = twice sigma beta – 2 lambda beta. So we are putting this to be = 0 here. Let us call this function 2 and this is 3. So you can take out this beta, so this is becoming sigma – lambda I beta is = 0.

Now you look at this equation you are very well familiar with that is ax = lambda x basically this is becoming the Igon value problem here, is not it? So this is Igon value problem. So to get non trivial solutions we must have determinant of sigma – lambda = 0. Now this is nothing but characteristic polynomial is not it?

Characteristic polynomial in p this is of degree p and this will have so this will have p roots, now p roots can be denoted by lambda1, lambda2, lambda P, so we can actually choose lambda1 > or = to lambda2 > or = lambda P. See there is nothing wrong in naming them in such a way that the largest one is called lambda1 then the second largest is called lambda2 and then the lowest one is called, okay there is one problem here.

How did I put this here because in general if I find out the characteristic values or the Igon values of a matrix then they can be real or complex, is not it? Then is it possible that I can do this? It is not necessary, is not it? Then how did I write this? That means there is some assumption here. Since, sigma is a covariance matrix I can make the assumption that it is positive semi definite.

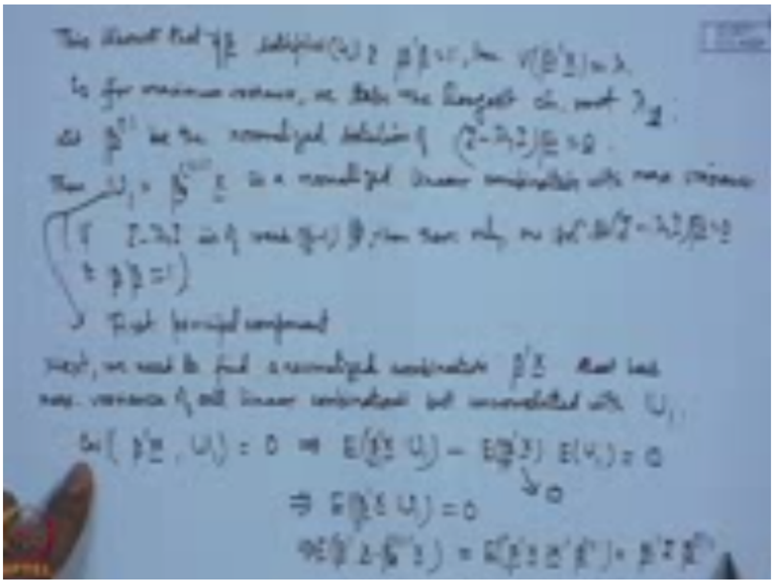
Actually, in general we can have positive definite also but all the time positive semi definite is always assured. If it is positive semi definite the Igon values will be non-negative. So this is possible since all lambda is are positive or basically non-negative real's, okay, non-negative real numbers. Now we do some manipulation here. Let us take 4 here. In the 4 1 you can consider pre multiplying by beta prime in 4.

So we will get beta prime sigma – lambda E beta = 0. This will become now scalar here. So this is giving you beta prime sigma beta = lambda beta prime beta. Now here there is an advantage here I chose beta in such a way that beta prime beta must be = 1. So this is simply = lambda. This is simply becoming = lambda. Now that mean I am having an explanation for the solution here.

That means what should be the value of lambda it should be satisfy beta prime sigma beta. So basically what I am saying is that and what is beta prime sigma beta it is actually, what was our So what I am getting here. Here I am getting beta prime sigma beta = lambda, that means original aim to maximize this quantity? So I consider the maximization problem as an optimization problem using the Lagrange's multiplier here, I am getting the solution.

So what I am saying here is that this value lambda is actually the Igon value here, okay. So what I am getting here. Here I am getting beta prime sigma beta = lambda, that means I have solved the problem. What is the solution? The solution is that I choose the Igon values and take the largest one, is not it? So the problem is solved here.

(Refer Slide Time: 12:47)



So this shows that beta satisfies 4 and beta prime beta = 1 then variance of beta prime x = lambda. So for maximum variance we take the largest characteristic root that is lambda1. Corresponding to this we can find out the Igon vector or the characteristic vector here. So let beta1 be the normalized solution of sigma – lambda1 I beta = 0. Then the corresponding beta1 prime x this is the normalized linear combination with maximum variance.

So actually if you look at this can be called as the first principle component. In fact, if we can say that if sigma – lambda1 I is of rank p – 1 then there is only one solution to sigma – lambda1 I beta = 0 and beta prime beta = 1. Okay, so this is the first principle component, okay so this one is called the first principle component. Now what is the next problem?

Next we need to find a normalized combination beta prime x that has maximum variance of all linear combinations but uncorrelated with the first one, okay. So this will be the second principle component. So that means we want covariance between beta prime x and U1 = 0.

Now this is equivalent to expectation of beta prime x into U1 – expectation of beta prime x into expectation if U1 = 0.

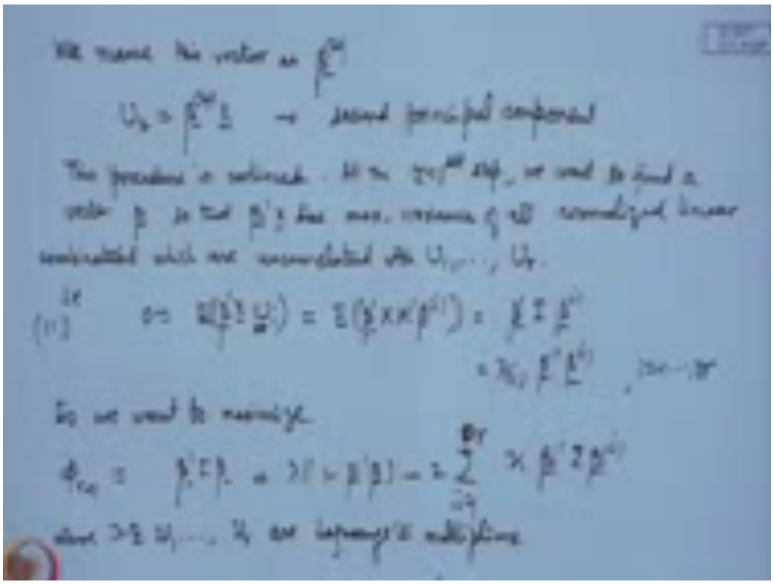
Now expectation of beta prime x is anyway 0 because if our expectation has been 0. So this one will not matter, so this means basically we are saying expectation of beta prime X U1 = 0. Now this is equivalent to beta prime. Now what is expectation of this one? U1 is beta1 prime, okay, so you write it here, so this is becoming beta1 prime x, okay. So these 2 terms I am multiplying this is a scalar this is a quantity, okay.

But what we can do if it is a scalar I can write it in the reverse way also. I can write it as beta prime X X prime beta1, okay. So this is just a simple mathematics here. But what this will give this will give the sigma. So this is becoming beta prime sigma beta1. So what I am getting the condition that this is = 0, okay. So this should be = 0. But what is sigma beta1. If you remember your original derivation here, sigma beta1 was lambda1 beta1, is not it?

Because that was the solution of the first one. So this is = beta prime lambda1 beta1. Let me call it 7. See this lambda even keep on this side so what is you are getting. So you just look at the condition here. See this is = 0 lambda1, so that is non 0 anyway. Beta prime beta1 = 0, what does it mean? That means the condition of uncorrelatedness, how did I start with the condition?

I started with that a next linear combination which should be uncorrelated with U1. That condition is reducing to that this new combination will be our vector to this, okay. So it is leading to that condition now.

(Refer Slide Time: 18:56)



So the condition of uncorrelatedness is reducing to finding an orthogonal vector, okay. So that means what is the problem now? The problem is beta prime sigma beta is to be maximized subject to the condition that you have the normalization thing so this term is the same as I wrote earlier. This term you can see it is the same but now I will write one more term that is – 2.

So I am just putting some – to even this is just for adjustment term here, okay. This is coming from here, okay. So, this should be = 0, so I am putting that here. So, these are 2 Lagrange multipliers now in place of one because 2 conditions are come in here. So, this nu one this lambda these are actually Lagrange multipliers. Let me call it function phi 2 here. So, you are having del phi 2 by del beta that is = twice sigma beta – twice lambda beta – twice nu one nu sigma beta1 that is = 0.

Now you can consider pre multiplication here, pre multiplying by say beta1 prime in the above equation, we will get twice beta1 prime sigma beta – twice lambda beta1 prime beta – 2 nu one beta1 prime sigma beta1. Now here we can use some conditions, sigma beta1 is nothing but lambda beta, so this one and this one is getting adjusted here.

So, you will get here it is simply = – twice nu one lambda1 using 7 and that is this condition that I wrote here. This condition and this condition if we write here that is sigma beta1 = lambda1 beta1. So, let me call it equation number 10 here. So, nu one is 0 and beta must satisfy the condition number 4 that is again let us sigma – lambda I beta = 0 because here this I am putting to be 0, now I am putting this to be = 0 then this is becoming same here.

This 2 will not come because I have taken out this 2 from here, so these will not come here. 2 I am writing outside here. So, beta must satisfy 4 that means lambda must satisfy this condition number 5 here that is this because it is again becoming the Igon value here, must satisfy 5. So, lambda2 will be the maximum of lambda1, lambda2, lambda p such that there is a vector beta satisfying sigma – lambda2 I beta = 0 beta prime beta is = 1 and 7.

So, these 3 conditions will be satisfied. So, we call this vector as now we name this vector as beta2, okay. The first one I call beta1, so now this is the second linear combination that is U2 = beta2 X this is the second principle component, okay. So, what is the property of the first principle component? I discovered the largest Igon value of the variance covariance matrix corresponding to that what is the Igon vector.

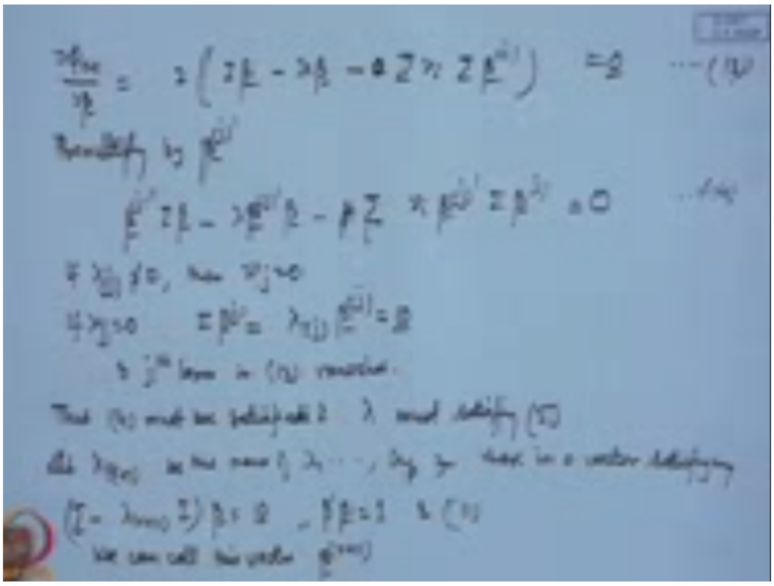
That Igon vector gives you the combination that is beta1 prime X then so that will have the largest variance because the variance that largest Igon value itself. Then you find out the second one and then second one will have then Igon vector, so basically what we are doing is we are orthonormalizing those Igon vectors. That is, you can actually use Gram–Schmidt process also. So, the second one can be continued.

So, this procedure is continued at the r plus first step we want to find a vector beta so that beta prime X has maximum variance of all normalized linear combinations which are uncorrelated with U1, U2, Ur. That means we are saying 0 that is = to expectation of beta prime X Ui as we have already written that how this term is coming this term is coming from here.

That covariance between beta prime X and U1 that was coming as and = 0 this condition gave you expectation of beta prime X U1 = to 0 because of this term being anyway 0. So, if we use the similar thing here this will actually be coming = 0 which is nothing but given you beta prime X, X prime beta I that is = beta prime sigma beta i and that is = lambda i beta prime beta I where i = 1 to r.

So, our aim is now to maximize, we want to maximize now the term let us call it phi r+1 beta prime sigma beta and the Lagrange multipliers will give you. Now there will be r+1 Lagrange multipliers. For i = 1 to r you will be getting this condition here where lambda and nu 1, nu 2, nu r are Lagrange's multipliers. Now you consider the vector of partial derivatives.

(Refer Slide Time: 27:35)



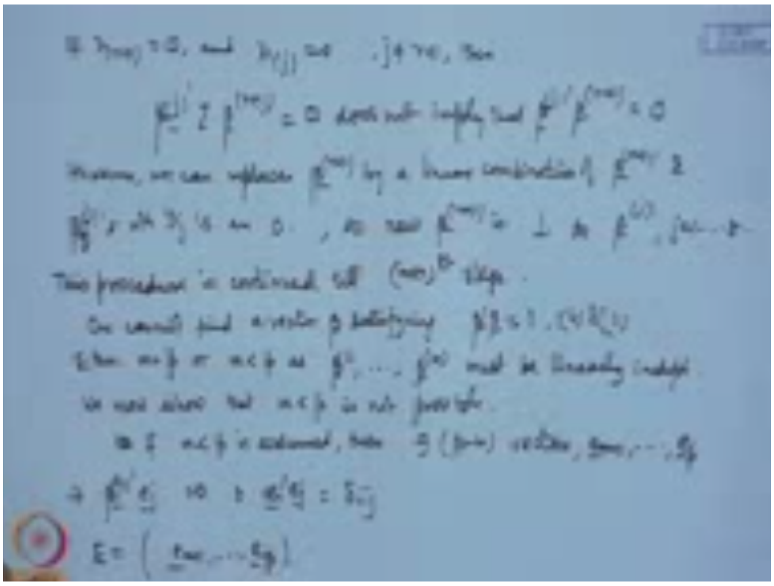
So, you will consider del phi r+1 by del beta that is = twice sigma beta minus lambda beta minus twice sigma nu I sigma beta I. So, this 2 will also go outside. We are putting it = 0 say. Let me call this function as 12 and this function as 13 here. So, pre multiply by beta J prime. If you do that then you will get beta J prime sigma beta – lambda beta j prime beta – beta this is summation here sigma.

In the previous term I had this after differentiation that, okay with respect to beta if I do then this term will give me the, this one, okay. So, that is coming out to be twice sigma nu I beta J prime sigma beta I. This summation is over I here. This is = now scalar here, okay. This term will become scalar because I have multiplied by a row vector on the left hand side. So, now if lambda J is not = to 0 then we will get nu J's to be 0, okay.

And of course you can have the condition that lambda = 0 then sigma beta I that is J that is = lambda J beta J that will be = 0 and J^{th} term in 13 will vanish. J^{th} term will not be there. So, what we are ultimately getting, finally the same condition here that is sigma beta 4 must be satisfied and lambda must satisfy that means it is again the r+1 Igon value. So, let me give you a formal proof of this.

Let lambda r+1 be the maximum of lambda1, lambda2, Lambda p such that there is a vector satisfying sigma - lambda r+1 I beta = 0 beta prime beta = 1 and the condition 11. That is this condition that lambda I beta prime beta = 0 here. So, we can call this vector beta r+1. Let us define the linear combination U r+1 is beta r+1 prime X, okay.

(Refer Slide Time: 31:20)



Now if lambda r+1 = 0 and lambda J = 0 for J0 = r+1 then the corresponding beta J prime sigma beta r+1 that is also 0. This does not imply that beta J prime beta r+1 = 0. So, we can replace however, we can replace beta r+1 by a linear combination of beta r+1 and those beta J is whose lambda J's are 0. So, again this nu beta r+1 this is orthogonal to beta J for J = 1 to r.

So, we can continue this procedure this procedure is continued till m+l-th stage such that we cannot find a vector beta which will satisfy the beta prime beta = 1. So, now either m = p or m is < p where beta1, beta2, beta m must be linearly independent. We can show actually that the m < p is not possible, it will lead to a contradiction.

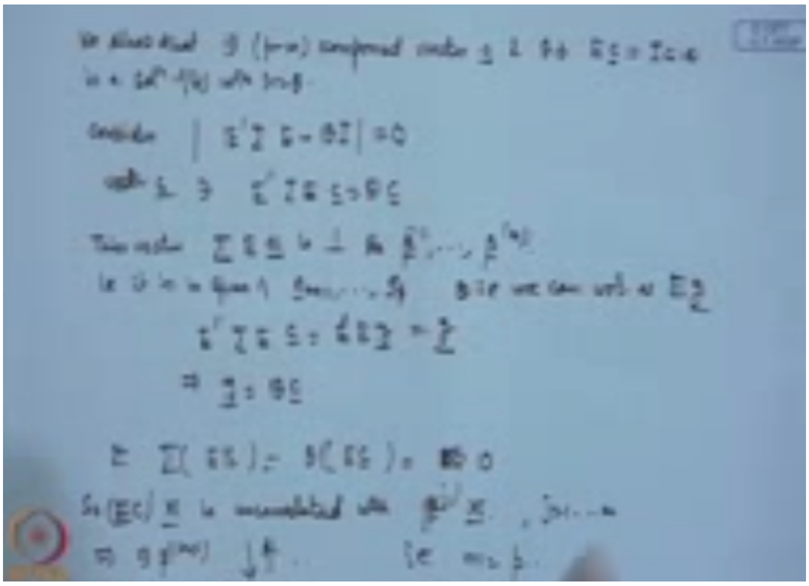
Because if m is < p that means there will be p – m vectors let us call them em+1, em+2 ep such that beta I prime ej will be = 0 and ei prime ej will be equal to something like delta Ij which is the characteristic function that means it is = 1 if I = J and it is equal to 0, If I != J. So, ultimately this can lead to a contradiction here. So, let me not discuss this full thing and let me just give you a brief hint here.

So, this procedure is continued till m+first stage and then one cannot find a vector beta satisfying beta prime beta = 1 4 and 11. So, either m = p or m is < p as beta1 and so on beta m must be linearly independent. We now show that m < p is not possible.

So, if m < p is assumed then there exists p – m vectors, let me call them e1m+1 and so on up to ep such that beta I prime ej = 0 and ei prime ej is = delta i j.

So, let us consider the set E as e m+1 and so on e p.

(Refer Slide Time: 35:48)



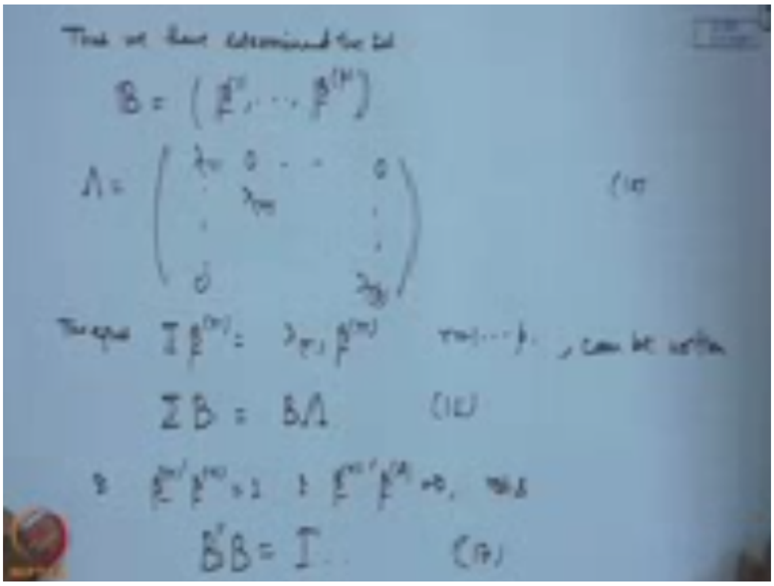
We show that there exist a p – m component vector c and theta such that e c is = sigma c i e i is a solution of 4 with lambda = theta. So, let us consider a root of say e prime sigma e – theta i = 0. And let us also consider a vector c which satisfies e prime sigma E c is = theta c. That means it is the Igon value of this. Theta is the characteristic root and c is the characteristic vector here.

So, this vector sigma E c this is orthogonal to beta1, beta2, beta m. That means it is in a span of e m+1 and so on up to e p. That means we can write as e g if it is in the span of this because e is the e m+1 up to e p that means it is in the column space of that. So, you multiply by e prime, so you get e prime sigma e c that = e prime e g that is = g because what is e prime e, e prime e is the identity matrix.

So, this means that g = theta c and sigma e c that will be = theta e c = theta, okay that is this one. So, what we are getting that this is actually = 0 here. So, e c prime x is uncorrelated with beta j prime x for j = 1 to m. That means there exists beta m+1. Because we have started with m < p now I am saying there is an m+1, so this is a contradiction. So, that means m must be equal to p.

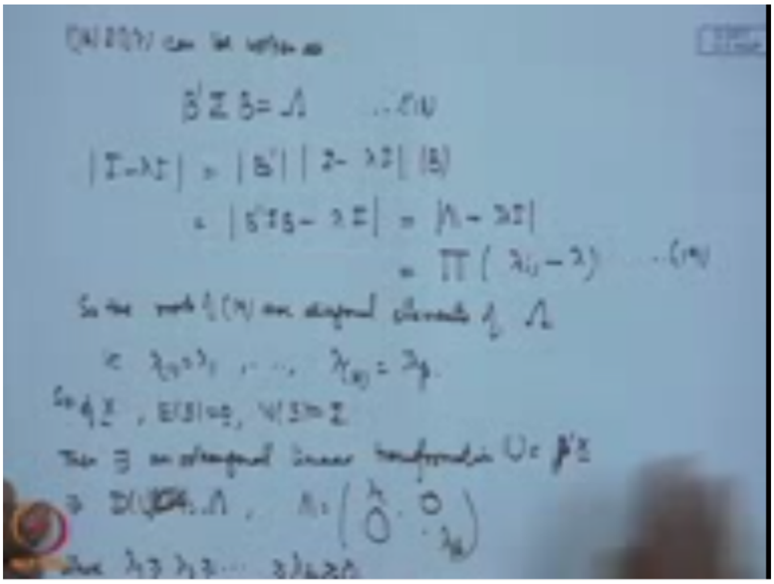
That means our process of finding out the principle components is systematic. That means we are actually going up to that route. So, now let us consider this set now.

(Refer Slide Time: 38:55)



Thus we have determined the set let us call it say a script B that is = beta1, beta2, beta p. These vectors we have obtained and the corresponding lambda I will call lambda1, lambda2 lambda p. These equations sigma beta r = lambda r beta r for r = 1 to p can be written as sigma this B = B lambda. And this conditions that we obtained for normalization and orthogonality that can be expressed as this can be written as B prime B = identity matrix.

(Refer Slide Time: 40:25)



So, this 16 and 17 will then give me B prime sigma B = lambda. Because I can multiply here B prime sigma B is = B prime B that is identity, so that is = lambda. Now this determinant of lambda sigma – lambda I we can write as B prime sigma – lambda I B. Because B prime B =

I and I can split it on both the sides because determinant of say 2 matrix p q = determinant p into determinant of q and that can be written as determinant of q into determinant of p.

So, this can be a spilt. So, this can be then written as B prime sigma B – lambda I because B prime B is again I. That is lambda – lambda I. This is product of lambda i – lambda. So, the roots of 19 are diagonals elements of lambda. That is we are getting lambda1 = lambda1 and so on. Lambda p = lambda p.

So, we have determined the principle components that means if I have a random vector x with expectation x = 0 and the dispersion matrix as = sigma then there exist an orthogonal linear transformation U = say beta prime x not beta let me call B prime x such that dispersion matrix of U, U prime = dispersion matrix of U = lambda and lambda = lambda1, lambda2… lambda p. This is all 0's here.

Where these lambdas are ordered these are roots of phi and the rth column of B satisfies sigma – lambda r I beta is = 0. This U are that = beta r transpose x this has maximum variance among all linear combinations and uncorrelated with U1, U2 Ur – 1. So, this vector U that I have determined from x this is called the vector of principle components.

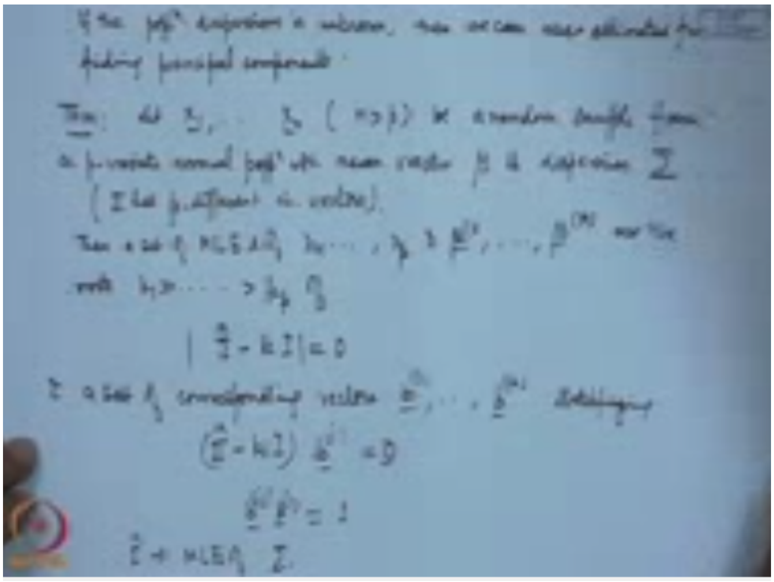
So, in a practical problem what happens that you replace the original coordinates system by the new coordinate system which is U1, U2, Up and then you choose the one's which are relevant that means it may turn out that the variance of say U1 is say 90% of the total variance, variance of U2 may be another 8%.

So, you just look at the one's which cover the major portion and you can discard the remaining for your problems which are like finding out the relationships between the variables that means you are setting up a regulation model, etc. or any other thing that means you are actually considering several variables but then you can keep the relevant one's only you did not consider all of them.

Now the question comes that when I do not have the variance–covariance matrix to me. In that case in the practical life we will have a data set. So, we consider the maximum likelihood estimates of that. For example, if I am dealing with the multivariate normal population. So, you have variance–covariance matrix sigma, so I can use s there that is the sample variance–

covariance matrix and I can use that to find out the principle components of that and from there I can determine.

(Refer Slide Time: 45:41)



So, let me just mention this estimation problem also. If the population dispersion is unknown, then we can use estimation for finding principle components. Actually let me consider say normal population let us consider say X1, X2, Xn a random sample from a p variate normal population with mean vector say mu and dispersion sigma. I make an assumption that sigma has p different characteristic vectors.

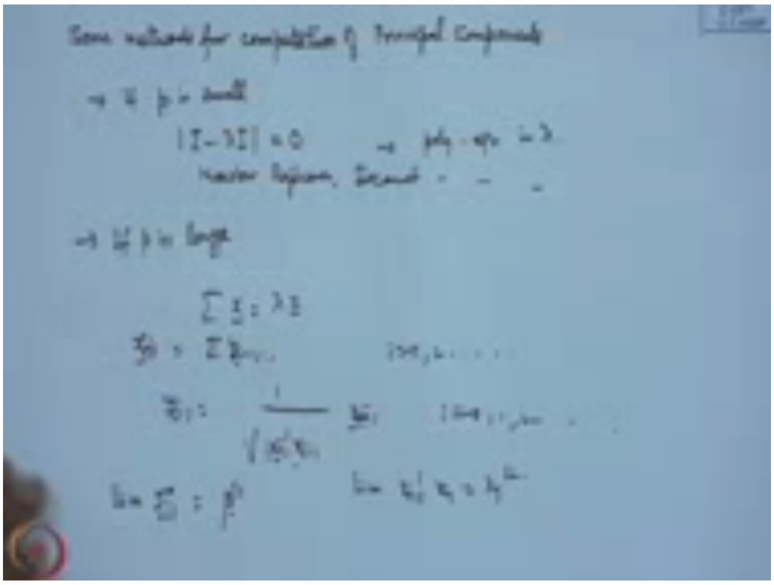
Well, you know that why this condition is required. This condition is for the diagonalization, okay. So, then a set of lambda1, lambda2, lambda p and beta1, beta2, beta p these are the roots K1, K2, Kp of sigma - ki = 0 and a set of corresponding vectors b1, b2, bp satisfying sigma - Ki I bi = 0. Bi prime bi = 1. This sigma is some set.

If the roots of sigma are not of multiplicity 1 then this method will fail actually. Now you may mention that if this method fails and what else we can do. But actually what happens that when you have a data depended problem that means when you have the observations then it is highly unlikely that any root will be repeated because the variance-covariance matrix that you will be getting will be highly diversified actually.

It will be positive but definite but the root repeated probability will be almost negligible here. Therefore, this condition is almost always satisfied here. Now let me given a methodology how to determine these principle components because I have given a theoretical method but in general if you have say dimension 5, dimension 10 because normally any problem will be of that nature only.

You cannot have a problem of 3 by 3 or 2 by 2. So, therefore we need some efficient computational procedures by which we can determine this.

(Refer Slide Time: 49:29)



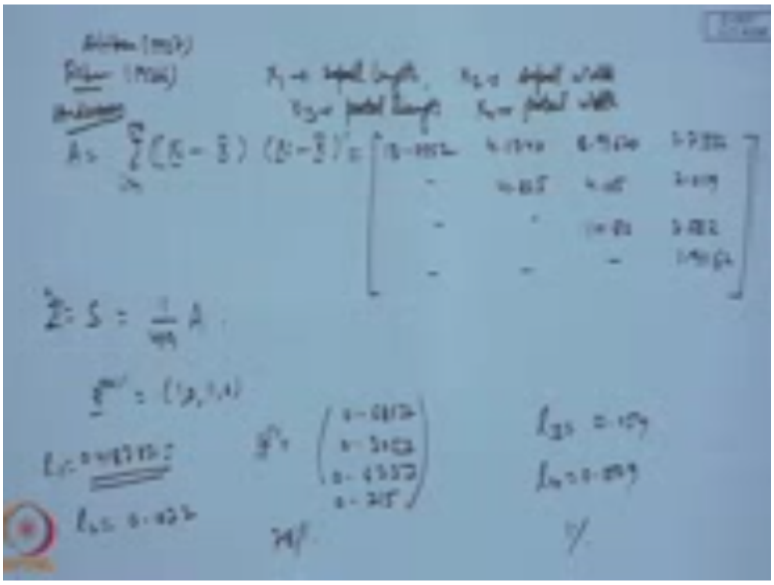
So, we consider some methods for computation of principle components. So, if p is small then we can consider this determinant of sigma – lambda = 0. So, it will be an equation well polynomial equation in lambda, okay. So, you can apply some efficient method for example Newton Raphson method or say sequent method and so on. So, they are some efficient methods which can be used here, okay.

Now if p is large that will be the usual thing that in practice you will have p large then you will look at sigma x = lambda x, okay. So, you consider some iterative methodologies for example you may consider say x i = sigma y i – 1 then y i = one by square root x i prime x i, x i, okay. So, this is for i = 1, 2 and so on then i = 0, 1, 2 and so on.

Then one can actually show that limit of this y is will be = beta1 and similarly limit of this x i prime x i that will be lambda1 x square, okay. And basically the rate of convergence will be depended upon the ratio lambda2 by lambda1 and closer this ratio is to 1, the slower the rate of the convergence. And for the second one then again we consider sigma 2 – lambda1 beta1 beta1 prime and from there we determine.

So, in general my comment is that one can use various numerical methods for determining the principle components here. I will close this by one example here.

(Refer Slide Time: 52:01)



This example is discussed by it in 1937, so and of course one can consider like new algorithm, like algorithm there are many algorithms for solving the system of equations. So, one can actually use this. So, let us consider the famous say example of Fischer, which is having the data on the petal lengths petal widths staple length and staple widths here. So, x1 was staple length x2 is the staple width x3 is petal length and x4 is the petal width.

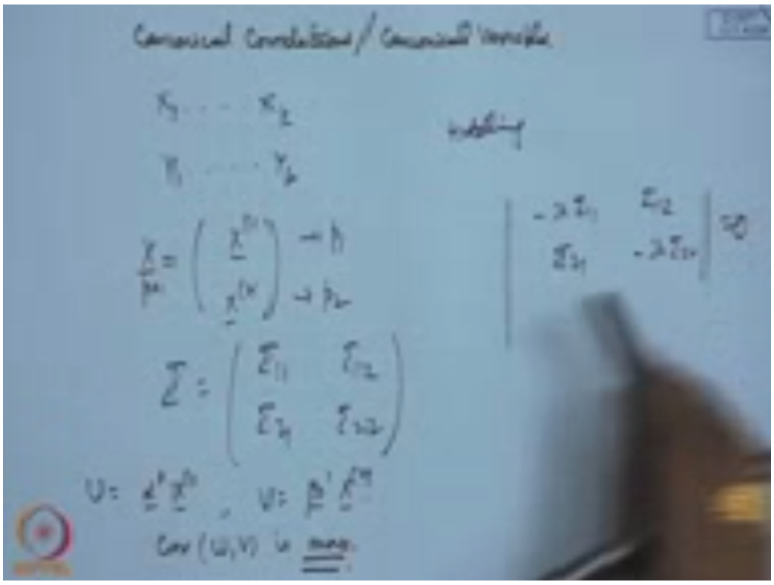
This data Fischer had considered for the certain trees Iris or something like that and for this the xi – x bar xi – x bar prime i = 1 to 50. There were 50 data sets here and this was given by 13.0552, 4.1740, 8.9620, 2.7332, 4.825, 4.05, 2.019, 10.82, 3.582 and 1.9162 and these values are repeated here. This is a symmetric matrix here. This example is worked out that way. So, sigma had is actual equal to as that is 1 by 49 A, so that can be calculated.

If we consider say initial approximation as say 1, 0 then we get the value of L1 as 0.487875 and the corresponding principle component is coming out to be 0.6867, 0.3053, 0.6237 and 0.215. This L1 turns out to be this is L1 let us call it. This is actually the largest root here coming out to be. It is more than 3 times the sum of other 3 roots.

Similarly, if I find the L2 that is turning out to be 0.072 and the corresponding b2 vector can be obtained and L3 is 0.054, L4 is 0.009. So, you can easily see the value of this is 4 exceeding these values here. So, this linear combination will be the principle component actually. You can say the first principle component which can be used.

This is corresponding to basically more than 78% of the total variation the last component is corresponding to < 1% of the variation here. So, the ultimate aim of this principle component analysis is to determine that linear combination which is leading to the maximum variability which explains the maximum variability in the data. So, therefore one should use that thing.

(Refer Slide Time: 55:58)



Another related concept is that of canonical correlations or canonical variables. So, now in general we have some correlation between X1, X2, Xp but what we do we find out those linear combinations which will have more correlations among themselves, okay. So, this type of thing will be leading to that means we transform by X1, X2, Xk to a new sets say Y1, Y2, Yk and these are arranged in the order of the correlations.

So, this problem is a similar problem that is of the problem of principle components and the principle component we are considering the variability. Here we are considering the correlations. So, for example I consider say x as a vector say these are having p components then we spilt into 2 parts. So, this is say P1 this is P2 and the corresponding sigma matrix is sigma 1 1, sigma 1 2, sigma 2 1, sigma 2 2.

So, we develop a transformation of the first P1 and last P2 coordinates in such a way that the inter correlation between X1 and X2 will be correlated. That means if find U = alpha prime X1 and V = say beta prime X2. Such that this correlation between U and V is maximum, okay. So, this is the problem of canonical correlation. So, we can if we normalize the thing then it is same as correlation or covariance both will be the same.

And like the problem of this, this can also be simplified and it turns out to be problem of Igon values of that this type – lambda sigma 1 1, sigma 1 2, sigma 2 1 – lambda sigma 2 2. So, there is a little variation from the previous problem but it is similar in nature here. So, we find then this one then next we find those linear combinations which are uncorrelated with the first set but again having the maximum correlation so that way we can determine the canonical correlation.

So, these are called canonical correlations. This problem has also been fully solved by Hotelling and the solution is given. The problem has been applied to various real life data sets. These topics are also discussed in detail when we are not having the known variance-covariance matrix that means we can have maximum estimates for that if we had any with the multivariate normal populations or if we are dealing with some other type of estimates when the form of the distribution may be something different.

I wind up the multivariate analysis portion of this course at this stage. There are many more topics in multivariate analysis like factor analysis is there. So, but in this particular course we will not be touching upon that and the next class I will take up a new topic.