Lecture - 31
Principal Component Analysis (PCA) - Model Adequacy and Interpretation

(Refer Slide Time: 00:25)

Good morning, we will continue with PCA. Today, we will discuss PCA model adequacy and interpretation, so last class we have discussed up to a extraction of PCA, extraction of principle component and we have seen that first we have used S minus lambda determinant equal to 0. Then we got the Eigen values and then using individual, suppose of the for the first one a 1, first Eigen value with the first Eigen vector we have found out and subject to a 1 transpose a 1 equal to 1. Similarly, you can find out a 2 using S minus lambda 2 i a 2 equal to 0 and a 2 transpose a 2 equal to 1. In this manner you we will be able to find out the Eigen vectors.

(Refer Slide Time: 02:06)
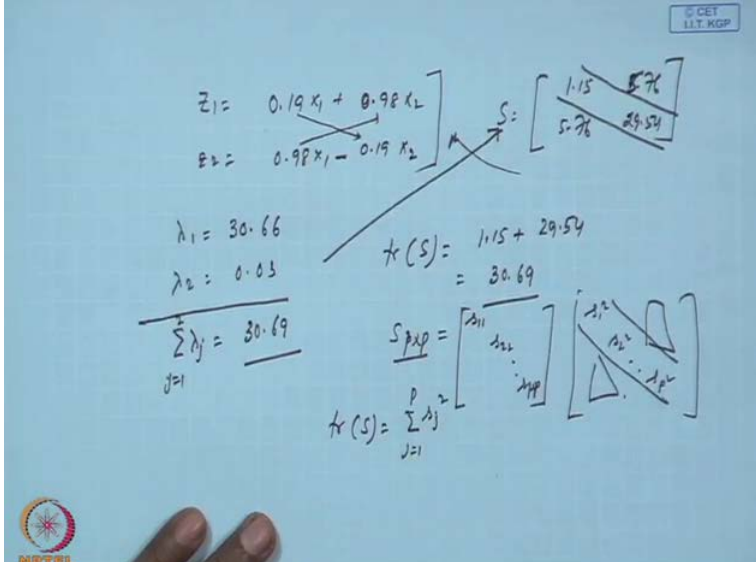

Once you estimate the Eigen values and Eigen vectors. You can write down the equations also, for example what the example we have considered here that the loadings which from profit, and sales are two variables inter related variables. We have extracted also two components PC 1, PC 2 and these are all Z 1, Z 2 and the loadings that basically these component loading the Eigen vectors 0.19 and 0.98 for your first one.

(Refer Slide Time: 02:36)

That principle component 1, 0.19 x 1 plus 0.98 x 2 and similarly, for Z 2, it is 0.98 x 1 minus 0.19 x 2, you see ultimately what is happening here just reverse, only one sign

reversal is there. So, what I ask you that you please compute in detail using these matrix S matrix like 1.15, 5.76 and 5.76, 29.54, you compute in detail and come to up to this level and you will be able to find out why what is the reason of this type of relationships. Now, under adequacy test what you want to discuss first you have identified lambda 1, I think it is 30.66 probably and lambda 2 is 0.03. Then the total is sum total of lambda j, j equal to 1 to 2, here this equal to 30.69.

Now, we have extracted this from this S matrix, now what is the stress of S stress of S is the sum total of the diagonal elements this is 1.15 into 29.54 is not this, it is 60, 30.69, this is 30.69. Now, if there are if S is p cross p matrix, then you will be getting like this suppose we are using sample S 1 1, S 2 2 like this S p p. Other way you can write S 1 square S 2 square like this S p square that you will be getting. There are definitely off diagonal values are also there, but stress is this, so I can say stress of S is nothing but sum total j equal to one to p S j square.

(Refer Slide Time: 05:30)

So, these can be proved that your sum total of lambda j j equal to 1 to p equal to sum total of S j square sum total of that means variability. If you extract the maximum number of Eigen values and what you will find out that sum total of this Eigen values will be sum total of the variance component of the original matrix. This matrix what you have considered and you have already also found out another thing that suppose I say variance of a j transpose x this one also we found out that a j transpose S a j. Now, if I

want to know what is the variance of Z 1 that will be variance a 1 transpose x which is a 1 transpose S a 1, it should be lambda 1 it will be lambda 1 that is a 1.

(Refer Slide Time: 06:57)

In this case a 1 is 0.19, 0.98 S is 1.15, 5.76, then 5.76 and your 29.54, so just you do this multiplication and find out that is a 1 transpose S a 1, you find out because this one is a 1 transpose 1 cross 2, 2 cross 2, 2 cross 1. So, resultant will be 1 cross 1 and definitely this will be 30.66 which is this is a 1, we can take a 1 is lambda a 1, we can take lambda outside and that you can do also that is also possible. Basically, what I am saying, suppose if you create a diagonal matrix of lambda 1 all 0 lambda 2 and lambda p. Then what you can write basically is sum total of this one, I think that is j a square j equal to 1 to p S j square that you can write stress of your lambda a lambda.

I think A transpose that you can be it can be written like this, now what you will do basically is why am I using a is because you started with this and then lambda is this transformation matrix. So, it will be lambda transpose you a sorry stress of this, now A transpose this is orthogonal I, so stress of lambda into i, this will be stress of lambda that means diagonal elements, so sum total of j equal to 1 to p j square. So, this is what the decomposition lambda j lambda j square j equal to 1 to p, it is not square lambda j lambda j. If you go for singular manner decomposition that square will be there, so this is the case, so you now understand the relationship everything is wrong.

(Refer Slide Time: 10:02)

Next is as I was telling you suppose you got lambda 1 equal to 30.66, lambda 2 equal to 0.03 this are point estimates because we have taken example that 2 by 2 matrix from n equal to certain value, I think in our case n equal 12, two variable result. Suppose, I go for next time as I told you earlier also point estimate if you want to find out, then we are going for expected value of lambda j that is equal theta j.

That means we are saying that in a population principle component analysis, if you do there you will be getting. If you know everything about the population you will be getting the exact value of Eigen value for that a particular population this is from sample this is from population. So, this one is covariance of lambda j lambda k this will be 2 theta j square j by n minus 1 for j equal k and 0 for j not equal to k.

This is the development, so what I mean to that is that is two lambda component lambda j and lambda k they R covariance when k j equal to k this is nothing but the variable component j not equal to k. It is the covariance component, so covariance is 0 and variance is 2 theta j square like this.

(Refer Slide Time: 11:57)

We can assume that it will follow a normal distribution with theta j and 2 theta j square by 2 n minus 1. When you that mean you know this solution, now this distribution is common that is definitely a question, but it is complicated question. So, there are several researchers statisticians develop this, so we are sticking to this lambda j can be approximate to normal distribution provided definitely n is large n is large.

(Refer Slide Time: 12:53)
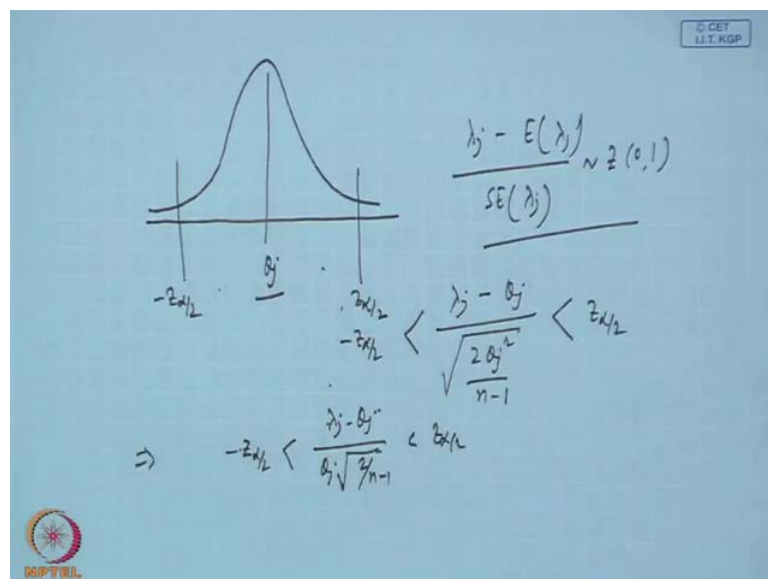
Now, when this is the situation you can find out the confidence interval for theta j, so what we will do basically you want to find out this minus Z alpha by 2 plus Z alpha by 2.

We know that lambda j minus expected value of lambda j by standard error of lambda j this follows Z 0 one the basic statistics that concept. So, that means lambda j minus theta j now what is the standard error of lambda j.

Now, what is the value here that variance part 2, theta j square n minus 1, so that means standard error if you write down here it will be 2 theta j square by n minus 1. Now, these quantity will be this side minus Z alpha by 2, this side Z alpha by 2, so ultimately this one can be written like this Z alpha by 2 less than lambda j minus theta j by theta j will come out root 2 by n minus 1 less than Z alpha by 2.

(Refer Slide Time: 14:44)

If you know mathematical manipulation, then ultimately what you will get you want something like this. So, theta j will come out here, then this theta j is there, so this theta j plus this quantity will be added and then as it is within this 1 by something will come final form will be like this one plus Z alpha by 2 by n minus 1. Then 1 by Z alpha by 2 by n minus 1, but we have this lambda j component, so it will be lambda j it will be lambda j into lambda j into lambda j you can write.

(Refer Slide Time: 15:47)

If our alpha value 0.05, then alpha by 2 is 0.025, then Z alpha by 2, this will be 1.96, now we got 30.

(Refer Slide Time: 16:03)

Our lambda value 30.66 by 1 plus 1.196 into 2 by 11 n is 12 less than equal to also you can write less than equal to 30.66 divided by 1 minus 1.962 by eleven. So, ultimately this quantity will yield 30.66 by 1.836 less than equal to theta j less than equal to 30.66 by this one minus 0.164 and this leads to 16.70 less than theta 1, because we are considering j equal to 1. So, if I write theta j also j equal to 1 here, then this value is 186.95, but do

you agree with this? There is an original see your original sample data if you see 30.69 this is the total variability, now this is because of many things your number variables is number of observation is very small this is huge difference.

So, I am sticking to that first one point estimate one only that point estimate will it is all practical purpose is it will solve the what you want to do, but see this much variability for the population PCA is there. If I go by this simple sample piece, then this trial error method we are saying that this is possible when n is large we are assuming normal distribution our case n is very small value. So, we cannot say that this is correct, but this is the way you will be finding out the confidence interval for the population Eigen value.

(Refer Slide Time: 19:12)

Now, another one is there, similarly using S we found out that a j what is this Z j equal to a j transpose a x. Now, this a j, you have a compound root the value of a j, but this core is sample a j corresponding in population in this lambda j is there. This is the Eigen vector for from sample this is population Eigen vector Z, so similarly there is definitely the distribution possible, but that distribution again really they are all competitive distribution what is that the Eigen vector what you compute.

These follows multivariate normal multivariate normal with a j and t j where t j is theta j by n minus 1 sum total of k equal to 1 to p. Then j not equal to k theta k by theta k minus theta j square lambda k lambda k transpose not lambda that is alpha k alpha k transpose, this alpha is what this 1. So, as we are using here alpha k, k equal to 1 to p, this type of

mixed equation will come just for the sake of clarification request all of you to check this that k and j term the notation. If there is just check this one, so ultimately as we are getting as are getting that in p a j t j.

That is the multivariate normal, you will be getting 100 into 1 minus alpha percent confidence region, if it is p variable case you will be getting the confidence region. That confidence region takes this form n minus 1 alpha j transpose lambda j S inverse plus lambda j inverse S minus 2 lambda 2 i 2 into lambda j that follows chi square p minus one alpha this is the confidence region. So, I repeat that one n minus 1 alpha j transpose lambda j S inverse plus lambda j inverse S minus 2 i, this i is basically if S is p cross p. Then i is also p cross p that is the identity matrix then this quantity will follow this distribution chi square p minus one with alpha you can check this also confidence region.

Now, as it is confidence is here you may be interested to know the simultaneous confidence interval then it will be kept to simplify it further, but there are complex, but I am giving you basically what you have applied in your PCA that is from sample. Now, what is the guarantee that suppose in the texture direction what you are trying to find out that you are actually going for the clue what is your sample size it is a quite large yes very large very large. So, we are I think you can still check the particular the confidence interval for lambda j i m values. That means the dispersion must be checked and if it is very big I think it will be within the limit because variant is large. So, you mean of size 128 cross 128, then is it 128 into 128 this much data.

(Refer Slide Time: 24:10)
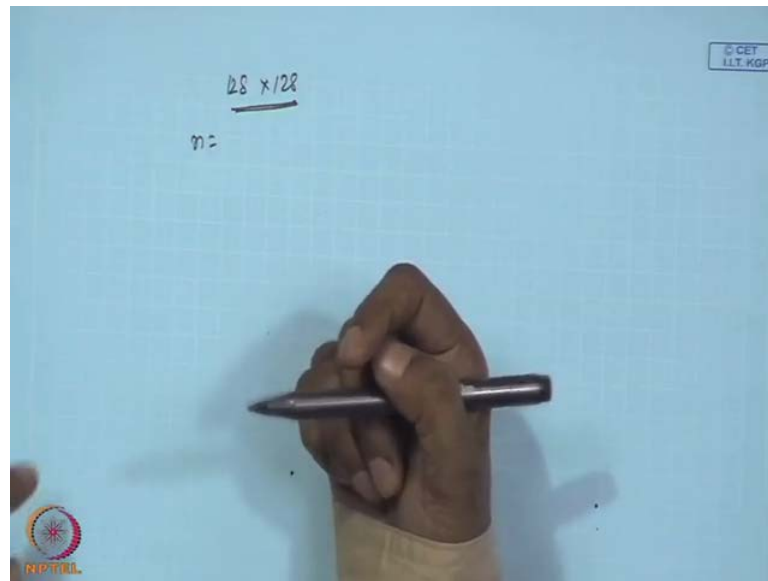
Now, what is this 128 into 128 no I understand, but 128, this itself is a matrix 128 cross 128 what about n, what is n total observation 128 into 128, so these many observations are there, it is quite larger. So, I think that as you are doing it may be we will be finding papers where this dispersion measures the Eigen values that is not considered that means uncertain part is not considered. So, you can propose the adequacy will be much better you can propose that and may be different sample size how the adequacy that your fitness will change that can be also a good addition to theoretical contribution.

(Refer Slide Time: 25:05)
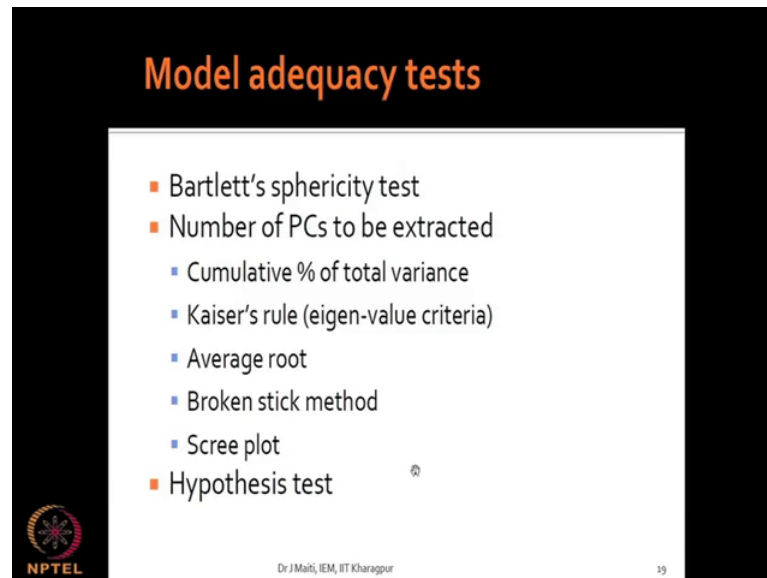
Now, let us see this what that example is, now see what happen ultimately lambda 1 is this and ultimately this is what our confidence region is. You have seen in multivariate normal distribution to the exponents are that same thing is basically when I am talking about that a j is multivariate normal. For two variable case the ellipse is formed ellipse is formed so you find out this confidentially form your Eigenvector.

(Refer Slide Time: 25:45)

Then, I will show you the that model adequacy test in terms of Bartlett's sphericity test, then we will go for what are the different criteria that can be used to find out that what are the number of PC's that can be retained. Finally we will show that some hypothesis test what is this what is happening suppose there are p x variables you are extracting PC's m PC's you are extracting let m is much less than p. For example, I am taking twenty variables and then finally compacting into two dimensional using PCA. So, that means you are you are excluding 18 dimensions, you are removing this if your data is very highly correlated, then only it is possible.

That means the 18 dimensions are not required, they are too much highly correlated, the Eigen values will be 0 for all the remaining 18 only for 2 Eigen values will be significant others will be almost negligible. So, that eighteen PCs what you are not considering are not the considering are not there, they give you something wrong signal that if you go by subject in criteria of eliminating some of the some of the principle components it may give you some wrong results.

So, finally that is why what we want to say that some hypothesis test that we are removing at some subset of principle components. We want to see that whether those sets are contributing to the extraction or not, but to the explaining the variability in two sense that is not extraction. I mean to say that explaining the variability of the x or whether the discarded components variability is very limited or small.

(Refer Slide Time: 00:34)

What is Bartlett's sphericity test, Bartlett's sphericity test is interesting.

(Refer Slide Time: 28:20)

For example, you see your data scattered data x 1 and x 2, this is your data, this is random data if you increase this this is a random data sheet, I mean there is more relation between x 1 and x 2. So, it will resemble a circle if we go for p greater than 2 that means p greater than equal to 3 or what will happen it will become a sphere. So, for this is for p equal to 2 it is a circle for p greater than 3, it will be a sphere greater than equal to 3 it will be sphere. Now, when you find out the circle or sphere that means the variables are basically random in nature in variables are un correlated variables scattered randomly without any systematic component.

This means variables are un correlated if each of the variables are un correlated with others what will happen how many PC component you can extract. If I go for principle component analysis, for example this is my case, suppose if I rotate this, suppose this Z one this side it is Z 2 you are getting. Now, if I take the correlation matrix between these variables with no correlation between each other what will happen you will get only diagonal elements one and off diagonal elements all will be 0.

So, if true sense the variables are not correlated what is the meaning of going for principle components analysis, no need of principle analysis. In that case R will be i, that is what happens in that Bartletts test, it is like this h 0 that R is i and correlation co efficient is that matrix is a identity matrix. Now, instead of this figure if you get a figure like this high correlation is there, so in that case your this R it cannot become this cannot become 0, this will be having substantial help.

(Refer Slide Time: 31:39)

So, Bartlett says then alternative hypothesis is R not equal to i and he developed the statistics value also. What is the statistics to be tested the statistics to be tested is here minus n minus 1 minus 2 p plus 5 by 6 log base a determinant of r, this is the statistic. Now, these statistics follows chi square distribution with p into p minus 1 by 2, sorry now the example we have taken 3 for 12 observation. Then we have two variables, then our log R the determinant part the determinant if you say that the determinant that log R that determinant of R this one will be computed like this 0, 2, 5, 8, the determinant of R.

So, this quantity this quantity will become some value 9.5 into 3 that is almost 34.77, now what is your chi square that p means your 2 into 2 minus 1 divided by 2, let alpha is 0.05. So, this value will be k i square 0.05, if you go if you see the table I think this value may be very low value not this much 34. You will not get this, so ultimately what will happen your alpha values for this this one is this value and this 34 will be coming somewhere here, so for our example reject ho R equal to i. This is obvious if you see the R matrix here this R matrix is 1, 0.987, 0.987, 1 this is our matrix, I think similar may be somewhere it is there, but very high value

(Refer Slide Time: 34:25)

So, when there is high value this Barlett's null hypothesis is rejected what we are saying you can go for principle component analysis. Now, this is one of the test in fact this test can be done much before the principle component analysis, first see the correlation matrix then this correlation go for Bartlett's test. If you find that yes there is felicity no need of going back to principle component, so much of effort why should you put, then suppose Bartlett's test is rejected and you can go for several PC's.

(Refer Slide Time: 35:17)

Now, what is there how many numbers of how many numbers how many PCs you will give keep number of PCs to be retained. There are several criteria, one of the criteria is cumulative percentage variance explained what we have seen we have seen that the variance component of the original data matrix that is equal to j equal to 1 to p lambda j that we have proved. So, that is why what you do you do like this, suppose your this is lambda j that is lambda 1 lambda 2. So, like this there will be lambda p you find out the value of lambda one how much some value like this you are getting, then this is the value.

Then you find out cumulative value for example, in this case our first one is 30.66 second one is 0.03, so this is 30.66. If I see the percentage percent cumulative 36.55 by 36.69 this will be almost 99 percent 30.66 by total sum of lambda. That is the total sum of the lambda because we have two only two from the example that is 2, so that is why this into hundred it will be almost 99 percent. Now, the thing is that we just go on doing like this that lambda one divided by sum of lambda j that lambda 1 plus lambda 2 divided by sum of lambda j.

So, this into 100 into 100, in similar way then you find out what is the percentage what percentage like this then you put a cut off late cut off by 90 percent. So, those many lambda components you keep that is cumulative percentage of total variance explained this is what the first criteria you can do is. So, in this criteria what will happen you may find out that you have to go for 95, 90 percent of the total variability of x to be explained you may take some of the principle component. For example, j plus 1 to may be j plus j plus m, these many component or up to j plus 1 to lambda p, these many components these components we have taken or if I say no problem m components, we have taken these values are almost equal very negligible.

For example, if I plot like this, this is my lambda value this side the PC value PC 1 to PC p component, you may find out first lambda value is here second lambda value is here third lambda value is here fourth lambda value will be here like this. You may get a figure like this actually suppose 90 percent is coming here, this is your 90 percent because it will go for cumulative at this point it will extract here. You make a percentage column also to cumulative percentage, this will be 90 percent, so see from contribution point, if I see, I think up to here it is there is difference.

All those values are not significantly different there is many screw contribution still there because I want this. So, in that case you may not be interested in 90 percent total variability explained, you may be sacrifice little bit and then you come to some other total variability may be it is 80 percent. Then you go for 80 percent and then these many these many components you keep there are some other methods just to do this one how nicely you will be able to do this.

(Refer Slide Time: 40:34)
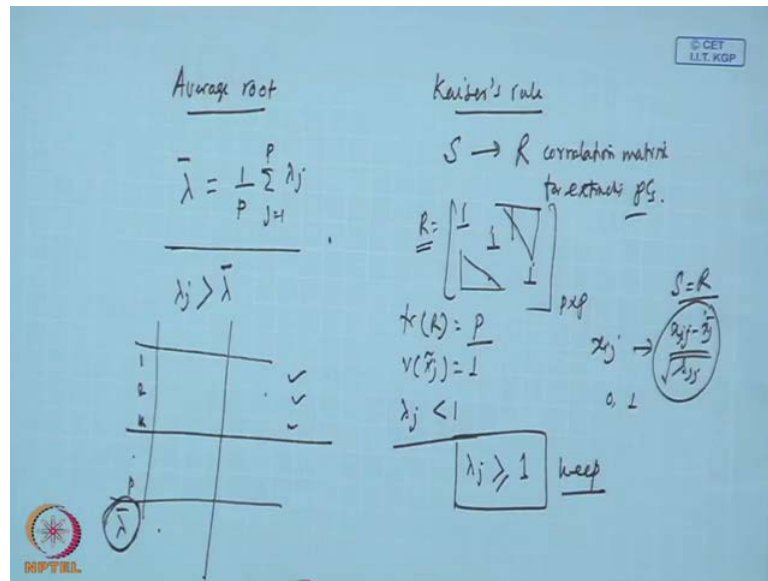
Another method is average root average root basically what will be your average root lambda average one by p sum total j equal to one to p lambda j. So, you keep those many principle components value is greater than lambda bar getting there is a, I spoke about it. So, many components where p component is extracted and the lambda bar, then component 1, 2 2 like this p component. You check suppose k components are having values which is greater than lambda bar. These many will be kept understood there is another method known as Kaisers rule Kaiser rule says that that you instead of S you use correlation matrix R matrix that is the correlation matrix for extracting PCs.

So, your correlation matrix is like this all diagonal elements will be one end of diagonal values will be there and some values will be there, then what is this stage of R. If it is p cross p matrix that will be p, so if I use correlation matrix instead of co variance my total variability is number of variables. Now, what is the each variable variability in terms of when he go for standardized variables.

So, when you are using correlation matrix what actually you are doing instead of xi j you are transforming this x i j to x i j minus x j bar by s j j square root, then if you compute x sorry S co variance matrix for this, this will be equal to R. So, standardized variable when you are taking for every standardized variable mean is 0 and standard deviation is 1 and variance is 1. Now, in this case what will happen for every variable the variability of x j that is 1, but standardization and giving that symbol.

So, if any of the lambda j that is obtained from that correlation matrix value is less than one no need of keeping this because these component is not able to explain a single variable variability. So, that is why what it says that you keep those principle component whose lambda j lambda value is greater than equal to 1, this is keep these many, so there is another criteria. I have one question regarding average root this one here we are arranging lambdas in descending order.

(Refer Slide Time: 44:30)
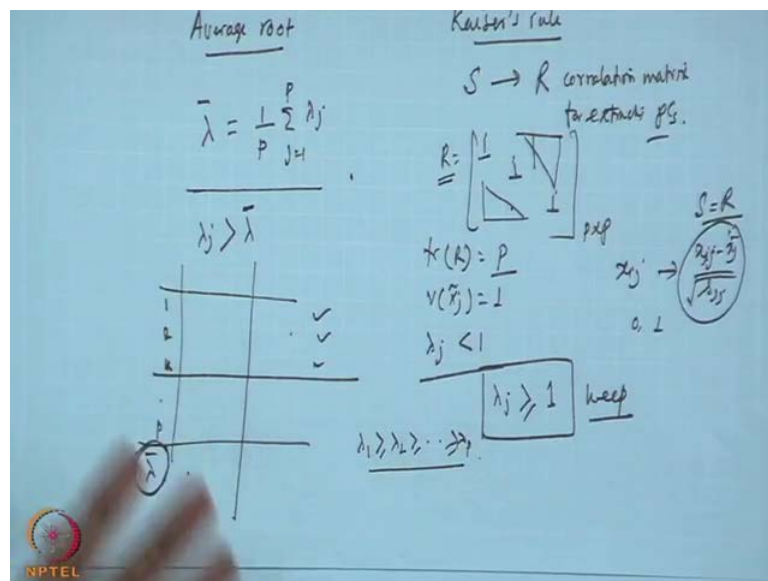
Now, when you that like when you extract your lambda one is always greater.

Student: Then equal to lambda 2.

Yes, in any ways.

Student: So, we arranging Eigen values in descending order which is what it is obvious it is obvious because this lambda one must be greater than the average lambda.
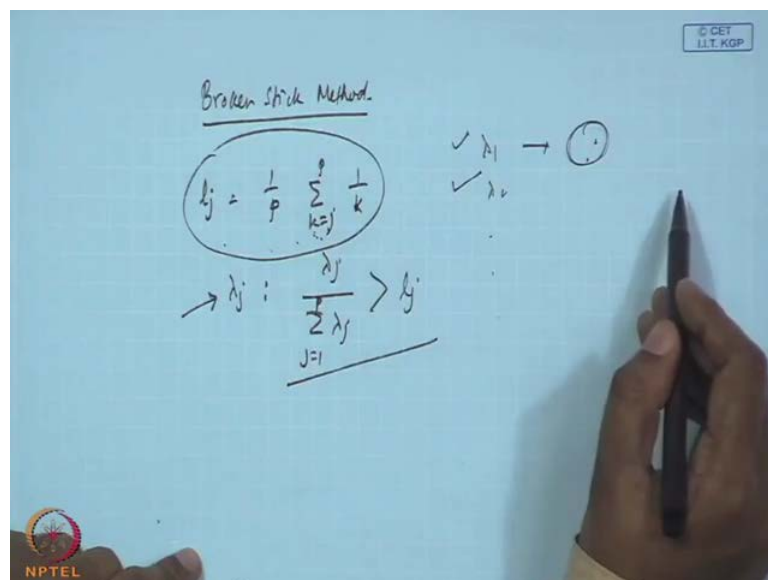
It has to be.

Student: It has to be, so one component is sufficient according to that criteria.

No, that is not the criteria you take all those components whose lambda values is greater than lambda bar not 1, 1 not one. It is not all the components have been Eigen value greater than the average Eigen value.

Student: Ok.

(Refer Slide Time: 45:25)

Another criteria is broken stick method, here this combination is first to find out a quantity called l j which is 1 by p sum total k equal to j to p 1 by k this quantity. Here, it is just like a stick broken into several components that it is random randomly that all the components in any, any size you can get. So, that is why this type of 1 by k this things are coming, now your lambda j the cumulative value for each lambda j you find out so that cumulative value is lambda j by sum total of lambda j j equal to 1 to p. You keep those lambda j whose value is greater than l j, so again the descending order is there in lambda 1 lambda 2 like this.

So, you find out for the first one k equal to what found out this quantity is this quantity this quantity is less than the cumulative percentage explained here, then keep this. Similarly, keep this when you are going forward at the end k is changing if it is first one is 1 second one is 2. So, from 2 to p 1 by k that way you are taking a broken stick

method, this is another criteria there is the popular criteria one more popular criteria is there that is

(Refer Slide Time: 47:09)
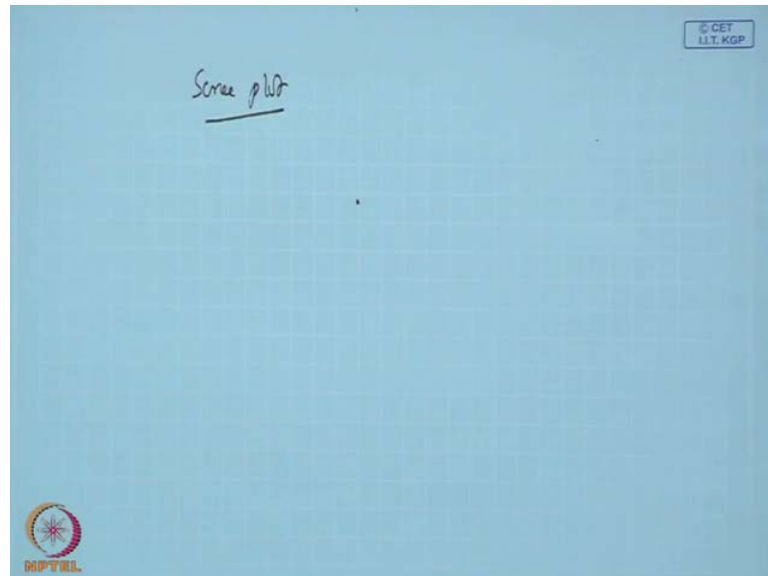
Now, see the scree plot here.

(Refer Slide Time: 47:16)
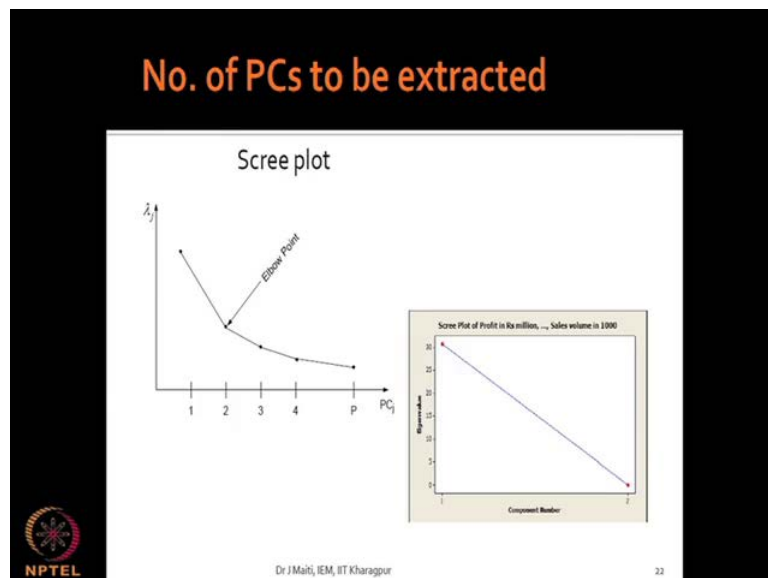
Scree plot is nothing but you just decide you keep the lambda value this side the PC, then put for every lambda values you put here and you will find out that. Ultimately, it will create the carbon create in elbow type of shape elbow will be there elbow is the point. If you take the normal posture the standard normal posture sitting posture in that case what

will happen that whole arm will be and upper arm should make 90 degree angle perpendicular that is the thing. So, that means this elbow point you find out you find out the point where this type of 90 degree will not get here, but if you get that this the best one so you are finding out where this elbow lies.

You take principle component number of component up to that elbow level, this is the usual one the reason is if I say my elbow and then this is parallel all this is perpendicular parallel to horizontal my fore arm. That means other point here, they are equally contributing there is no improvement in terms of addition of some other principle component.

Then, you have to add everyone all the things because they are parallel now in this in our case this is two variable case only, so two principle component we have found out that only one component is enough we do not require more, any question here up to this.

Student: What we should study for all those terms?

Which one you want to follow?

Student: Now, how many so many things are there see ultimately what happens I can say that each are basically almost the same thing they are talking about one way or other thing it will difference, now cumulative percentage.

Student: we have to use that.

That is probably you cannot ignore that one because ultimately the original data sheet these are certain variability that co variance structure is there. What you are doing using PCA that co variance structure, you want to explain in some transformed dimensions in transformed variables, so if I am not able to explain the majority of the variability present in the original data sheet. Then I think it is not a good model, so cumulative percentage where that is the first one we have to look into and then you have to think what the total variability that you must explain the percentage of the total variability.

You will find out that it is very difficult to explain in fifty percent of the variability, now question is then if I require twenty variables to explain 90 percent variability. Now, if I reduce it to five dimensions, and able to explain 80 percent of variability, then definitely you will go for the five dimensions with the deduction of 20 percent variability. You are

not able to explain the 25 percent variability at 90 to 80 that is 10 percent deduction is there correct that is possible second thing is then why all those things are there. Basically, it goes for everyone, then you see which one is favoring you, but whatever you will do either broken stick or Kaiser Rule or your scree plot. Finally, you have to see that how much the total variability you are able to explain these are the scree plot is basically mutual representation.

(Refer Slide Time: 51:50)
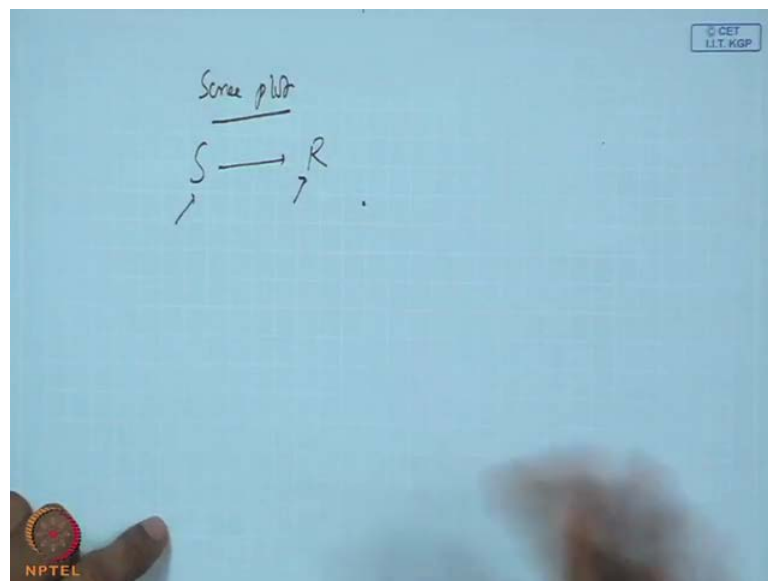
Another important question is that whether we will go for S or R which co variance which matrix You will explain see if you use S and if you use R that will get different type of different regions it is in the transformed this. Now, it all depends on that what you want you want that suppose you are measuring the variables in different units the variability is wide. Under this case, suppose your interest is there not on scree the pattern of relationship co variance relationship, then I can ask you to go for correlation matrix that mean whatever axis you give dimensions you give it is in terms of correlation matrix.

That vectors Eigen vectors will be created and there the pattern of relationship is more important not the strength if it is the strength is also equally important. Then you will go for co variance matrix because strength co variance is in the original domain not transforming when you are going for R. You are going for the standardized variables so that you have to look into where you want to do.

(Refer Slide Time: 53:05)
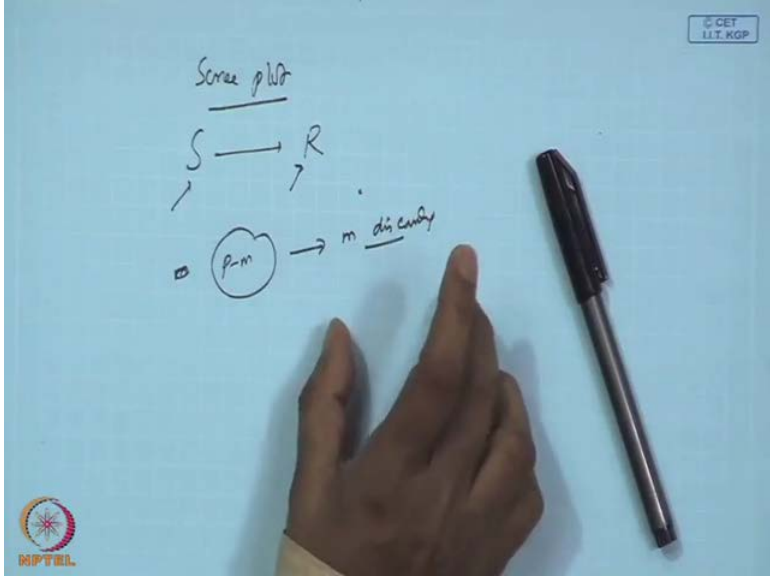
Then, final one is that as I told you the hypothesis test that Bartlett 1950 developed this one, so suppose by these traditional method like scree plot broken sticks and other things.

(Refer Slide Time: 53:29)

You are keeping out of the P components, you are keeping m components or let p minus m components you are keeping that means m components you are discarding. Now, my question is I want to test the m component collectively what this contributing in explaining the variability of the original data matrix. Significantly, in that case our

hypothesis is that lambda m plus 1 equal to lambda m plus 2 equal to lambda m p means the scree plot you make the scree plot is horizontal that is what you are trying to test.

Then, you are saying they are not equal for at least one pair in the hypothesis the statistics is d equal to n p minus m log lambda m bar and where lambda m bar is this one because the component you are discarding the average of that. Then this quantity will follow high square distribution with this divisional freedom and then you know that chi square the table you have to follow find out this values. Then you go for a change accordingly you take your decision we should accept all lambdas which ever greater than lambda m bar greater than lambda m bar, but not that is straight forward here.

It is saying that if actual thing is that if they are equal then lambda m bar will be individual lambdas, so then using this but in sample you will not get exact symbol there will be little bit of difference. So, if I assume that lambda suppose lambda m is lambda j plus one or m plus 1 is greater than lambda m bar and you will take this. Then you are going by point estimate what it is saying you just see that there is a difference or that difference is significant or not if that lambda m plus 1 minus lambda m bar is significantly different.

Then, you keep otherwise you do not keep if say that can be discarded that is why the sample this distribution is very important and using this distribution. You are first finding out the statistics then you are finding out the value of statistics, then you know that what the chi square digit of freedom is. Then you go to table find out this value if you find that d value is greater than this, then you reject null hypothesis h 0 what is all the lambda values here the discarded lambda values are equal.

So, these many tests are there and I think and we should nothing problematic these are all simple to calculate although very difficult to derive for example, Bartlett's test and all this are difficult, but very simple to calculate from the application point of view. All those things when we come to this multivariate domain what will happen, ultimately that you will find out that a large number of statistics are used to test the same thing.

Now, what is the reason is all those statistics are based on certain assumptions now and there will there will be deviation from reality in raw data when you collect the data. You will not find out that they are actually following the assumptions of the statistics of the models you are using. So, you will use several such statistics and see that all are favoring

or some are majority favoring or not if majority favors to a 0, go for a 0 otherwise do not go for a 0, yes Rahul any question?

Student: No, sir.

So, I think we will stop here and afternoon I will show you in that given data how to go for your multiple regression multivariate regression. Then principle component analysis and also we will see part very simple add data things will click business, things will be done even then once I show it will be easy for you to use the Softwares.

Thank you very much.