Data Analysis and Decision Making – I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 41
Factor analysis

Welcome back, my dear friends and dear students. This is the DADM which is Data Analysis and Decision Making – I course under the NPTEL MOOC series and as you know this is 12 week course total duration being for 30 hours which is 60 lectures because each lectures for half an hour and we are going to start the 41 – lecture which is basically starting we have already finished 40 which is 8 weeks are gone, we are gone to start the ninth week. And, each week we have as I just mentioned 5 lectures each being for half an hour and my name is Raghu Nandan Sengupta from IMA Department, IIT, Kanpur.

Now, if you remember we, but technically I am going a go a little bit in into the past starting from the 29th, 13th lecture we started studying the concepts of multivariate statistics. And for the multivariate statistics we took the concept initially for the different type of distributions like the multinomial distribution and the then wizard distribution, multivariate t distributions, the multivariate normal distributions.

Then, we slowly covered all the concepts and we also studied only related to the multivariate normal distribution, the concept of hypothesis testing and the same basic concepts continue remaining as in the univariate case that given the standard deviation you want to find out the mean. So, if the standard deviation is known we just simply take the sample mean.

Similarly, even if it is not known the best estimate is still the sample mean whether for the univariate case or the multivariate case. Then for given the mean value for the population I am all considering the normal distribution remember that. We care we want to find out something to do with the standard error of the stanard of the variance of the sample because we do not know the population variance.

So, they can be technically two formulas in one case, we use the s dash another case we use the s without the dash an actually means that if the mean value of the population is known, we take that we do not lose any degrees of freedom and we divide it by N and do the calculation the formula is no I am only is highlighting in the mean difference and obviously, inside the formula it will basically summation of in the numerator summation of the in the bracket X i minus mu whole square because mu is known.

Now, in the next case when mu is not known the best possible course of action would be try to find out the best estimate of the population mean which is the sample mean. So, as you use the set of observations X 1 to X n for the first time and try to calculate the best estimate for the population mean which is the sample mean you lose one degrees of freedom. Hence in the formula it is now it will be divided by n minus 1 and in numerator the actual equation would be summation bracket X i minus X bar n whole square because that mu has been replaced by X bar n which is the sample mean.

Now, later on we saw I am sorry, I am going to go back a little bit more in the future in the sense recapitulate whatever we had discussed even though it that may not be very relevant here you will not understand why I am doing that later on. So, given for the [po/point] point estimation problems we found out the best estimate for the normal mean, best estimate for the normal standard deviation for the population of course.

Then given exponential distribution we found out the best estimate for the parameter for the population then for the Poisson distribution we form of the best estimate and obviously, they were with proofs which we did not state, but the main contention of the of the analysis or the discussion was that you basically try to analyze the unbiasness and consistency; unbiasness mean the case where you want to find out the expected value of the sample statistic or the sample characteristics with its distribution such that in the long run the expected value is equal to the and the actual parameter value from the population.

And, the consistency means that as the sample size increases observation number increases the variance should basically tend to 0. So, what I mean by the variance actually it would be the probability between the difference between the parameter value and the statistic value as n which is the sample size increases the probability slowly of the differences being less than equal to some epsilon is very small value. So, it will always tend to 1, because it is both the values are equal; that means, the population parameter is equal to the sample mean which you are trying to utilize as the best estimate.

Now, later on we saw that we are given in this was the crux of the point distribution I will be try to use the same ideas in the multivariate case also. In the interval estimation we saw that there were basically four types of distributions coming out and the consequence why the consequence were, that given the normal distribution you want to find out something to do with the normal mean given the standard deviation of the population known we took the z distribution and obviously, there was no concept of degrees of freedom.

Then, we found out that and we gave the logical sequence of the understanding was that given the population variance is unknown and yet you want to find out something to do with the population mean we took the t distribution and the t distribution had n minus 1 degrees of freedom the question is why n minus 1 and because the t distribution coming for the case when as I have just mentioned the population variance being known, but unknown, but you want to find out something to do with the population mean.

And, this degrees of freedom as you know is basically coming from the fact for s dash and hence without a dash. Now, these two formulas related to z and t would always be utilized when we are trying to just do some studies or go into the depth of analysis of some problem related to the population mean and the population variance of the standard deviation given being given in the first case standard deviation being known and in the second case standard deviation being unknown; that means, z and t.

Now, if you want to find out something to do with some studies related to the population variance. So, again there would be two cases population mean being known and population mean being not known. If the population mean is known then obviously, you will use the chi square with n degrees of freedom when there is only one population and if you want to basically compare to population standard deviations and if both the mean values of the population 1 and population 2 is known, we will use the f distribution with m comma n, m being the number of observation in the first population, n being the number of observation in the second population. So, it would be f with degrees of freedom m comma n.

Now, when we want to find out something to do with one population its standard deviation on the variance and the population mean is not known then we know that we will be utilizing the chi square with n minus one degrees of freedom and when we go into the realm of trying to understand the two different populations corresponding to their population variance given the population mean values are unknown, then we will use the f distribution again, but the degrees of freedom would be m minus 1, n minus 1. The m minus 1 being fact that we are losing one degree of freedom from the first population and n minus 1 being the case when we are using the degrees of freedom for the second population. So, this concept would come.

And, the corresponding distributions which we have for the t we had the student t distribution multivariate case, then the wizard distribution was basically replacement of the chi square which we discussed. And, if you remember I did mention very fleetingly though it is absolutely true that as we utilize the concept of the sample mean and the sample variance are independent for the case of the normal distribution this will continue to be true in the case of the multivariate case also which means that wizard distribution and the mean value wizard distribution corresponding to the standard deviation.

Then, later on we came into trying to utilize different other distribution; that means, the Poisson the exponential and hyper geometric other things considering the sample set of observation is large and we can easily replace them their mean values those distribution mean values and the variance in the case with the normal distribution and still continuing with our work. Now, this was also true in sense for the normal case for the multivariate distribution and the multivariate distribution, we did discuss of the simple examples of multinomial distribution and all those things which we have seen.

Now, later on we found out that after the proofs and few of the results we went in the concept of copula theory. Copula theories basically the linkage function which you want to find out using the marginals and the joint distributions of the of the random variables x 1, x 2, x 3 and so on and so forth and consider you have basically p or k number of random variables and you want to find out what is the copula linkages it is basically a marginal distributions concepts.

And, obviously, copula comes into the picture when we want to find out something to do with the independent structure at the extremes I had drawn that if you remember. Now, when we come to the in the concepts of trying to find out later on for the independent structures the first different type of ideas we considered was basically the principal component analysis.

Now, the principal component analysis which we went a little bit slow and I would really like you to understand that this was the first case that we are basically going through only the formulas and trying to basically utilize this formulas not derivation the concepts and try to utilize this for the problem solving. So, the main idea was basically in principal component analysis you are given a set of for data and some random variables X 1 to X p and we will consider the X 1 to X p are not independent of each other there is dependent structure. So, obviously, they would each of them would have a mean value each of them would basically the combined variance covariance matrix.

Now, our main question would be that is it possible to find out a certain set of convex combination of this random variable X 1 to X k or X 1 to X p such that we can find out the maximum variability step by step and basically formulate those variabilities in some orthogonal direction to each other such that we are able to explain the variability to the maximum possible extent using the combinations of X 1 to X p which are all orthogonal to each other. So, it looks absolutely perfect conceptually; that means we want to basically break the dependence of the variability into orthogonal direction.

And, obviously, the next question would come that how you go into that we will take basically the variability for the case which is the maximum then the next level and the third level fourth level will go on and so on and so forth. So, obviously, it becomes very intuitive that, why not basically break those principal components in the orthogonal direction can using the concept of Eigenvalues and Eigenvectors.

Now, the Eigen vectors would basically be give would give me the. So, called values depending on the Eigenvectors which I find out the cell values will give me that in what combination I should combine X 1 to X p such that I am I am able to get the first principal component which is Y 1 then the principal second principal component it is Y 2 and so on and so forth till the p-th one. So, obviously, this Y 1 to Y p would all be in independent of each other and we can find out what is the variability.

So, the variability is found out and when we find out the variability of Y 1 or Y 2 or Y 3 then we found we calculated obviously, this those will be calculated using the covariance matrix which I have for x one to X p you find out the variable that was equal to the Eigen vector values and; obviously, they would give us the information that the orthogonality actually the breakages of the orthogonal vectors of Y 1 to Y 2 to Y t 3 Y 4 alright.

And, also we found out what is the variability. So, the variability of the actual variance for y one with the highest then the Y 2 and Y 3 and it will decrease continue decreasing we use this example for a 3 by 3 matrix and basically proved on the orthogonality the concept of how Eigenvalues and Eigenvectors would be utilized later on we come came to the factor analysis in the factor analysis what we are trying to do which we will continue discussing now on further also will continue that the factors are decided such that will break them into two parts. One of the factors which are basically have their effects and one would basically be the white noise.

Now, we will also consider that white noise concept we did not actually consider to be existing in the even though it was existing we basically assumed that they could be known. In the factor analysis we will try to utilize the concept of the factors along with an error terms and basically procedure accordingly again the same concept of the Eigenvalues and Eigenvectors would be used the Eigenvectors would be denoted by lambda 1, lambda 2, lambda 3, till lambda p.

And, obviously, the Eigenvectors direction would be given by the bold e 1, then e 2, then e 3, e 4 and so on and so forth where e 1 would basically be the Eigenvectors corresponding to lambda 1, e 2 would be the Eigenvectors corresponding to lambda 2. Now, we were basically discussing that we will come to that later on. I am sorry that had to basically recap all this thing because the concept which are being utili[zed]- discussed may be a little bit intense. So, I thought I will go a little bit slow in order to basically clarify the points as may be necessary for the students.

Another point which is are also nothing to related to the this course directly, but technically this being the a 12 week course which is about 60 lectures 30 hours. So, we will basically have a one to one session with the students and the instructor would me and obviously, there would be queries and we will try to basically split the set of queries which would have been basically after the first four lectures then next four lectures and so on and so forth which is 4, 8 and 12.

But, after the eighth lecture which were started; obviously, there would be a weekend or a weekday in the evening where based on the communication which will be sent from the NPTEL office I will definitely be available to answer queries both for the first four a set of lectures and the next set of four of lectures which will basically eight lectures and then depending on how your queries and whatever the queries in the forums we will basically try to tackle all the queries which may be coming up both from theoretical point of view as well as the practical point of view and then on the end of the twelfth lecture we will have the third session of one to one video conferencing.

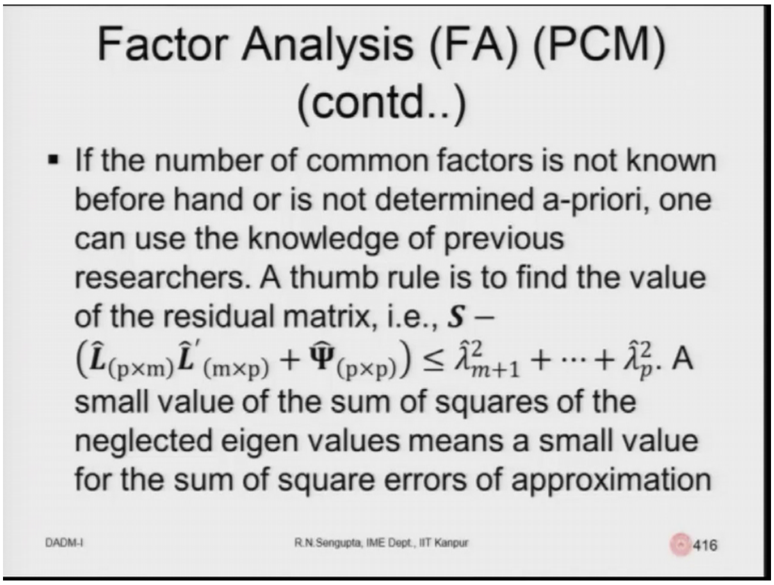
And, I am sure in both considering the assignments which will be 12 in number also the three different 1 hour sessions minimum of 1 hour sessions which we plan to have with the students for the first four then the next four and the twelfth one and 12 lectures I am sure and also obviously, they would be twelve as assignment. So, this is the eighth week starting. So, you will definitely have the solved at least seven set of assignments and by the time when you are basically coming for the one to one sessions you would definitely you would have reach received the eight week assignments and after they are solved you can definitely have queries accordingly.

So, the reason which I am why I am spending so much time in the first part is basically to discuss what we have covered and the next part in order to basically tell how things will be handled from our and in order to answer the queries was because in the forum there are many different questions which are coming up which may be a little bit easy, but unfortunately due to either positive of time from your end or on the way the whole concept of the course are considering that a huge set of coverages are to be done.

Due justice has to be done from your end in the sense that spend some time for with the concepts, clear your doubts read daily related the books. If you remember I have given you a lot of books and the references and still if you are not able to do that after solving the assignments and trying to basically get your queries answer in the forum and also after the one to one session video conferencing still if you have some questions you are most welcome to basically collate them and send on the forum; we will definitely take our time out and answer all the related questions which may arise for this course, ok.

So, I am basically continuing with the factor analysis sorry for a little bit more exhaustive introduction and what things we should be planning in order to basically cover the course in a much better way. So, in let me read it.

(Refer Slide Time: 18:57)



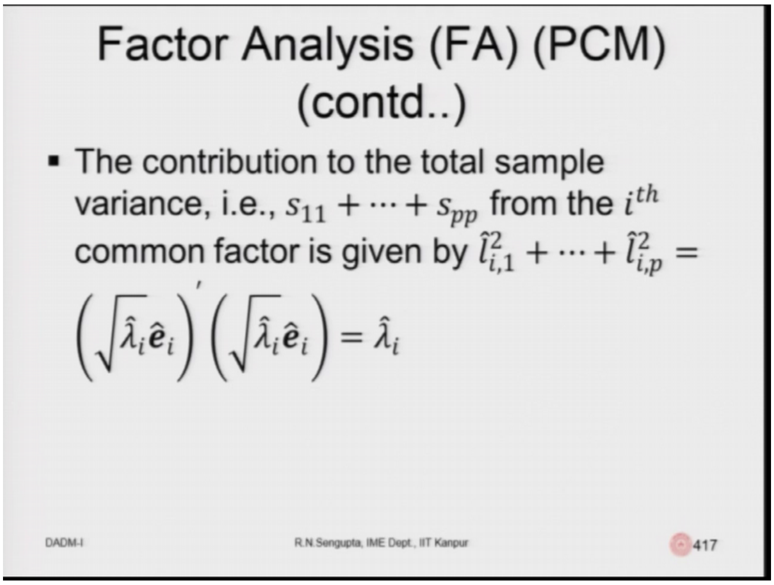
For the factor analysis I have come back to the course. So, if the number of common factors; that means, the factors based on which we will try to basically do the analysis is not known before hand or is not determined a priori one can use the knowledge of previous researchers. So, the set of information which will be there from by the previous researchers can be utilized in order to answer the question.

A thumb rule is to find the value of the residuals in the matrix S. S if you remember if the and the variance covariance matrix from or the square root of that from the point of view of the sample because we are nothing to do with base the population. So, L 1 L L 1, L 2, L 3 till the values of L which will basically be till m because we are taking m out of the p number of such factors.

So, they would basically would have two sets one is the known level of dependence one is the white noise. So, that should technically be in the long run if the divisions of the amount information so called or how should I put it information being relevant from the point of view where your able to give the reasons for those aberrations and one set of white noises for which you are not able to give the information's so, they should technically be equal to the sum of the Eigenvector Eigen a value square.

So, a small value of the sum of squares of the neglected Eigenvalues means a small value of for the sum of the squares errors of approximation. So, obviously, depending on the problem we are able to answer it or not we will come out from that value.

(Refer Slide Time: 20:43)

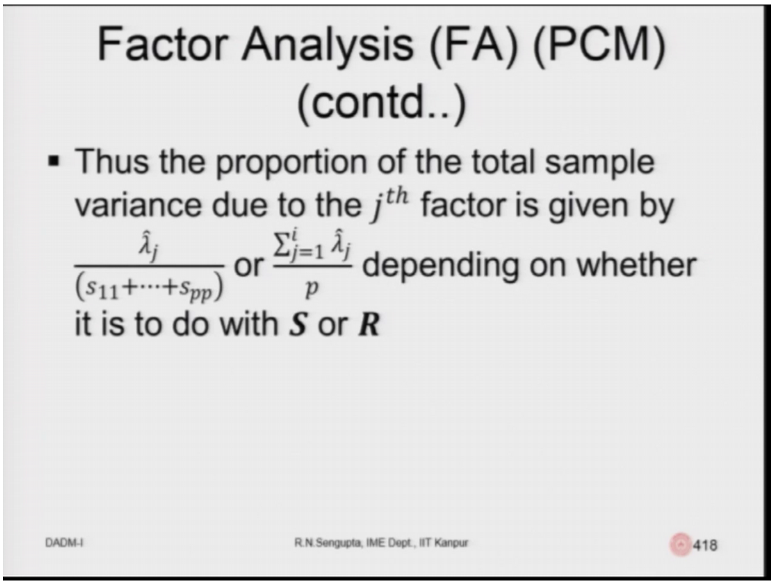


The contribution to the total sum sample variance would technically be if you consider the principal diagonal. So, the first value would be s 1 1 second would be s 2 2 which is basically the counterpart from the population if you remember it will be sigma 1 1 which is the covariance of the first with respect to itself.

Then, the 2 comma 2 element would be sigma 2 2 which is basically s 2 2. So, sigma 2 2 is basically the covariance of the second with itself and till the last element which in the principal diagonal will be s pp which is basically sigma pp which is base all these are suffix sorry my mistake. So, which are basically the covariance of the p-th vector with itself. So, they so, this the total sample variance which is s 1 1 plus s 2 2 plus dot dot till s p p from i-th common factor is given by this I.

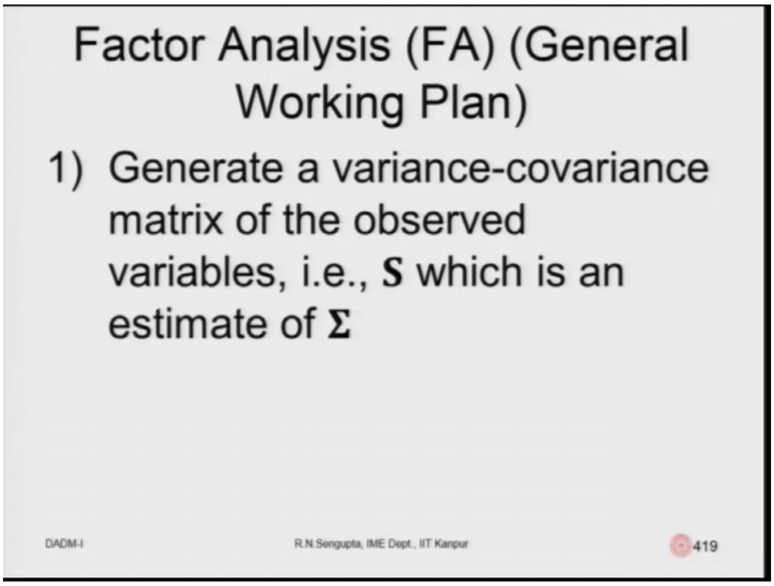
So, these are basically coming out from the values of the lambda. So, they would basically be given by i square plus i 1 square plus i 2 square plus i 3 square till like p square. So, p is in the suffix and the squares are on the on the p is are in the in the base and they would basically be power in the square. So, they technically they should be equal to the values of lambda 1, plus lambda 2, plus lambda 3 multiplied by square root of them multiplied by their Eigenvectors and there should basically equal to the lambda i; so, obviously, then the estimated value.

(Refer Slide Time: 22:13)



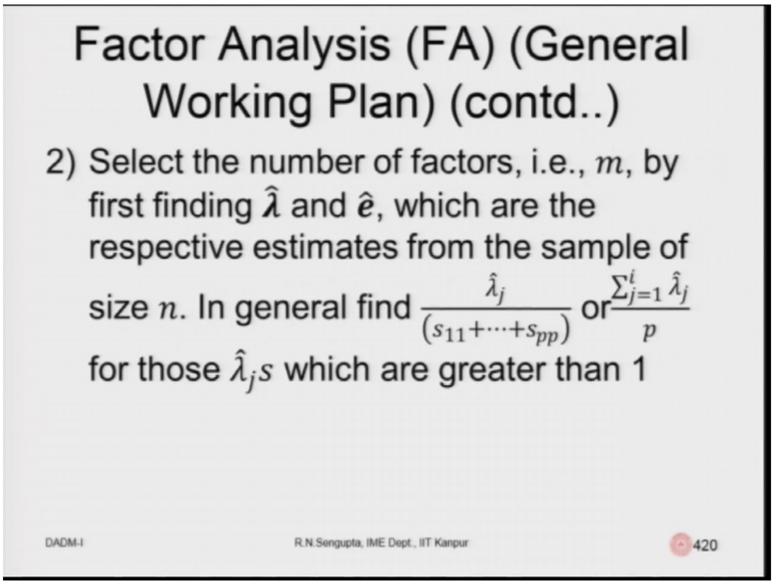
Thus the proportions of the total sample variance due to the j-th factor would be given by the ratio of the Eigenvalues divided by the sum of the principal diagonal which is s 1 1 till s p p such that depending on whether it is to be is to do with S or R. S would basically be corresponding to the case when we have the sample and R would be the case when we have corresponding to the population value. So, the R I am what I am saying is that population means that I have the counterpart from the population which is the correlation coefficient. So, based on that, we can basically do the calculations.

(Refer Slide Time: 22:57)



So, what is the general plan of working for the factor analysis? So, what first will generate the variance covariance matrix of the observed variables that is S which is basically the estimate of summation or the mu value and this capital mu value is basically the variance covariance matrix of size p cross p.

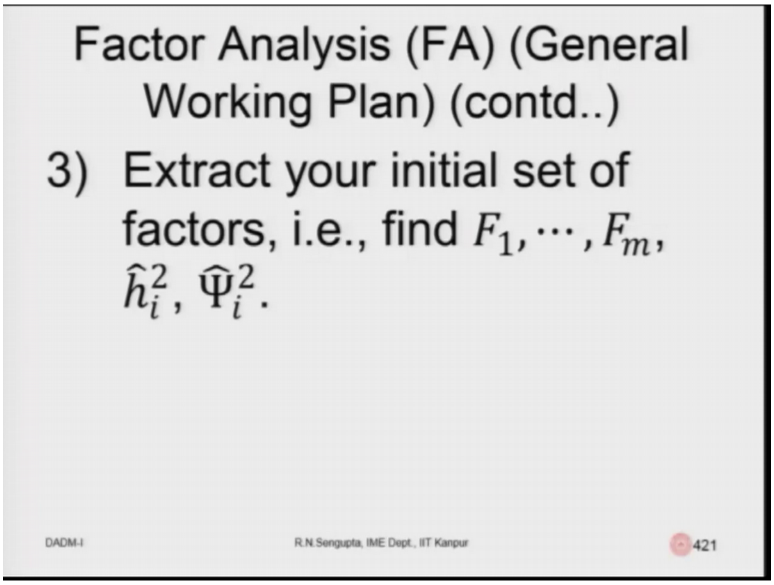
(Refer Slide Time: 23:15)



Once we have that we will basically select the number of factors that is out of this p; obviously, m would be less than equal to p by finding out the lambda hats. So, lambda is basically vector and which is the Eigenvalues and e is bold e's are basically the Eigenvectors. So, technically the e 1 value would be corresponding to the first element of the first vector then corresponding to the second value would be in the first Eigenvector would be basically e 1 2, then e 1 3, e 1 4, till e 1 p. Then, similarly we will basically have e 2, e 2 1 e 2 2, e 2 3 e 2 p where the first value here would basically be corresponding to the Eigenvectors which is the second level and for lambda 2.

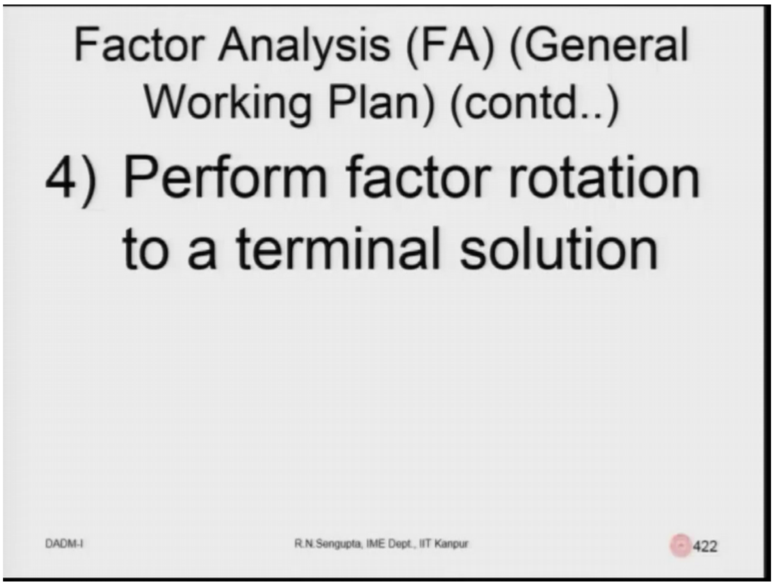
Now, with which respect to the estimates from the sample size n we get it in general trying to find out as I mentioned lambda hats mean basically the estimates divided by the sum of s 1 1 till s p p would be utilized when we have basically the standard error of the sample and in case we want to basically divide by the value of p; that means, summation of lambda j is divided by p would come be coming out when we have basically the correlation matrix corresponding to the problem of the sample.

(Refer Slide Time: 24:49)



So, we will extract our initial set of factors and based on that we will try to find out F 1, F 2, F 3 until fm remember that what we are trying to do is that trying to utilize a set of factors m which is less in size with respect to p and utilize them in such a way that we get the maximum amount of information which is plausible and the other type of white noise would basically be putting into the error term.

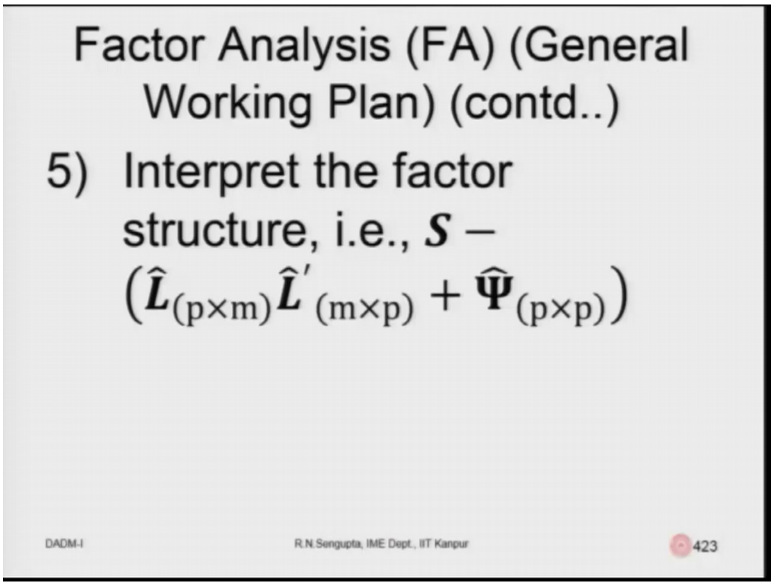
(Refer Slide Time: 25:13)



We will perform the factor rotation to a terminal solution. So, basically we will we will standardize them rotate the factors as that they would be orthogonality each other, again the same concept being taking up where we try to utilize the set of information which is coming up from the Eigenvalues and the Eigenvectors. And, rotations are being done such that the maximum variability plausibility is there and once it is for the first set is fixed we will basically keep that and then try to find out the rotation of the second set with respect to the first that they are orthogonal to each other.

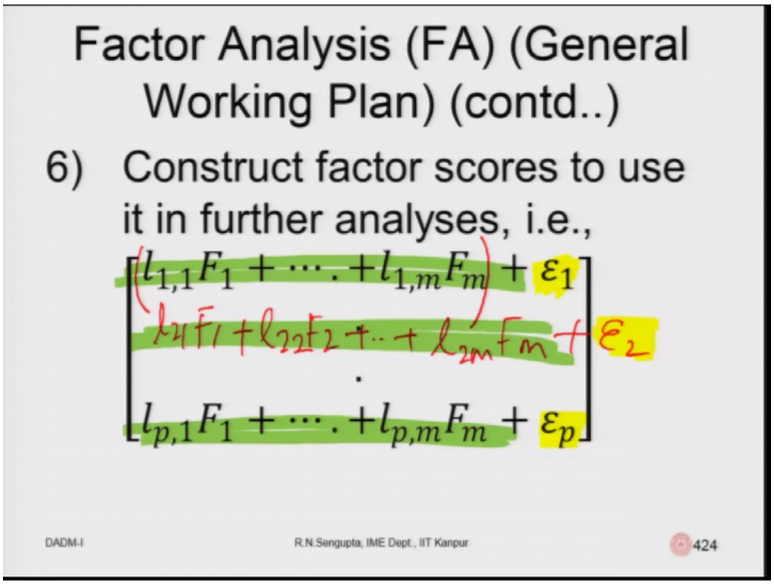
So, we will keep rotating it in this way such that each of them are orthogonal and the variabilities are basically being subsumed or coming out from the discussion the factors which are m in number m if you remember is less than equal to p.

(Refer Slide Time: 26:03)



So, once we find out the factors. So, we have the variance covariance matrix from the sample which is s and that would basically from that will try to basically subtract to two sets of information; one set of information with the factors information which is plausible and for which we do have some formulas to say that yes, the factors are coming out from the m number of. So, so called vectors corresponding to the lambda those Eigen values lambda 1, lambda 2, lambda 3 and the obviously, the Eigenvectors would be there given as e 1; e means capital all these are capital e 1, e 2, e 3 and they would basically be a set of white noise which will try to basically captured using the p cross p matrix of capital phi.

(Refer Slide Time: 26:53)



So, once we construct this you will basically have now in this case of the principal component we have the principal components such that there orthogonality each other it had nothing to do with the error terms. But, now and obviously, they were the convex combination we were being utilized because we consider the relationship was linear based on this we proceeded. So, we will consider the factor scores in such a way that the first set of factors which will we have would basically be the convex combination corresponding to the first factor which will be L 1 into F 1 plus L 1 2 into F 2 so on and so forth till the m-th one which is L 1 m into m F m.

So, this whole set which will, we have this will give me the plausible reasons which are for which we I can assigned reasons and this would be the error part. Similarly, when we go to the second one it will be x s so, if I have this I will highlight this is the error term this is all the error term of the p-th one and the factors for which we will have the information is for the second, for the first and for the last one. So, we are trying to divide the overall variance covariance into two parts the variability one on the factors which are possible another is the white noise.

So, with this, I will close the forty first lecture and continue discussing more about factor analysis and multiple linear regressions later on. Have a nice day and thank you very much.