Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 12

Lecture – 53

Principle Component and Its Graphical Analysis in R

Hello friends, welcome to the course Multivariate Procedure with R. So, you can recall that in the last lecture we had talked about principal components and its analysis and we had understood that how we are going to compute the principal components manually. And if you try to see the whole process precipitated in a very simple approach that we need to find out the eigenvalues and eigenvectors of the covariance matrix. And whatsoever are the values in the eigenvector they are going to help us in the construction of principal components. And whatsoever are the values of characteristic rules they are indicating the variance being contributed by any particular component. So, now in this lecture we will continue further and we would like to see how we can implement it in the R software or how we can do the principal component analysis in the R software.

So, the way I am going to do this lecture is that first I will try to take a very small data set where I will try to show you that how can you construct the principal components and I will try to simply take a covariance matrix or 3 by 3. And then I will try to take a bigger data set and then I will try to show you that how principal component components and its variances and other components are obtained in this data set. Also, I would try to show you some graphical procedures which helps us in taking different types of correct decision in the case of principal component analysis. For example, how to determine that which variable is contributing more etcetera.

So, let us try to begin this lecture and try to understand that how are we going to implement the principal component analysis in R software ok. So, now in this lecture we are going to talk about the principal component analysis in R as well as we will be talking about some graphics criteria which are obtained through the R software right. So, just to recall we had considered in the last lecture that let sigma be a p × p symmetric and positive definite covariance matrix of $p \times 1$ random vector say x which is x1, x2, xp. And suppose the eigenvalues or the eigen or the characteristic root of sigma be lambda1, lambda2, lambda p and we are considering that lambda1 >= lambda2 >= p. And since sigma is a positive definite matrix so all the characteristic roots or the eigenvalue, they are always going to be positive.

And based on these lambda1, lambda2, lambda p we can find out the characteristic vectors or the eigen vectors and let these eigen vectors be indicated by e1, e2, ep. So, now we can obtain the principal component as a linear combination of x1, x2, xp with this here e1, e2, ep right. So, this is how you can obtain the the i-th principal component. And variance of i-th principal component is here lambda i which is here is the the i-th characteristic root right. And this the covariance between yi and yk is actually 0 that means, all the principal components they are independent of each other.

And in case if some lambda i's are equal then the choice of ei's and hence yi are not a unique, but that will really happen in real life. So, based on that after computing the lambda1, lambda2, lambda p which are indicating the contribution of the variance of the principal components y1, y2, yp we can compute the proportion of total variance that is accounted by first principal component as the lambda1 upon the lambda1 + lambda2 + lambda p which is the total variance right. And similarly, I can compute the proportion of the total variances accounted by second principal component which is here lambda2 upon lambda1 + lambda2 + lambda p. And similarly, for the pth principal component we can compute this this proportion lambda p upon lambda1 + lambda2 + lambda p. Now what do we try to do? Based on that we will try to choose those lambdas 1, lambda2, lambda p in such a way such that they are making most of the contribution in the total variability.

So, instead of choosing all the p variables we will try to choose here only those selected principal components which can take care of most of the variation. So, in this way the principal component technique helps in the dimensionality reduction based on this proportion of variance. So, you will see that many times this principal component analysis is also called as say dimensional or say dimensionality reduction technique right. So, instead of here all the y1, y2, yp we try to choose here some selected y1, y2 etcetera such that the total variation is nearly the same as the total variation without much loss right. So, now we try to understand that how are we going to implement it in the R software.

So, in the R software we have a command here prcomp which is available in the base software and this is given by here like this either you try to give here this x which is a numeric or a complex matrix or a data frame to input the data or you can use this prcomp with here formula. So, formula is something like what we have used in the case of regression analysis also. So, this is the formula with no response available referring only to the numerical variables and then I have to give here a data which is here an optional data frame containing the variable in the formula and then you have to give here subset and there are many other things and similarly there are other option also like as has ret x center scale dot etcetera. So, they have the following meaning. So, if you try to see I will try to give you some brief idea about those things which we are going to use otherwise I will request you that you please try to look it into the help menu and try to understand the meaning and use of each of this parameter right.

I have also taken it only from there. So, this ret x is a logical value which indicate the whether the rotated variable should be a returned or not. Center is a logical value indicating whether the variable should be shifted to be 0 centered or not. If you try to see we have used the scale command also where we have done the shifting of the data with respect to mean as well as standard deviation right. So, this is the same thing.

Similarly, the scale dot is also a logical value. Logical value means it takes the value true or false and it indicates whether the variable should be scaled to have unit variance before the analysis takes place or not. Then we have here the parameter tol which is a value indicating the magnitude below which the component should be omitted right. Then we have an option here rank dot. So, which is specify a number of the maximal rank that is the maximal number of principal components to be used.

And then we have here new data and optimal data frame or vectors in this tool look for variable for this to predict right. So, these things are there will I will not be able to show you here all, but you can experiment by taking different types of these options. And when you try to look at the outcome of this prcomp, then you will get here the value of sdev which is the standard deviation of the principal components. Rotation the matrix of variables loading that means the matrix whose column contain the eigenvectors right. And then it gives you here the value of here x and then it gives you the value of here center or a scale whether the centering and scaling is used in terms of true or false.

So, let us try to take a very simple example and try to understand how you can you can interpret it. So, look at me take here the 3 by 3 covariance matrix of our of 3 random variables x1, x2, x3. And suppose they are here and suppose and then now I need to transform it into data set up which can be given to this R command. So, I try to convert them into 3 data vector first data vector let us see here x1, second data vector of - 12, 15, 4 which is here second row and the third row will become here see here x3. And then I try to create here data frame data dot frame of x1, x2, x3 and then you have here ah this here data frame.

So, now we use here the command here prcomp and then we try to expose it over this data set datapc. So, if you try to see this prcomp is actually principal component this is the short form and I try to store the outcome in the pcdataout that we principal component data out right and this is here the outcome. So, you have to understand what it is trying to show you here. So, if you try to see here first value here is the standard deviation. So, what is this thing? This is the value of square root of lambda1 right.

These are the values of the characteristic roots. So, you know that the characteristic roots or the eigenvalues they are indicating the variance of the i-th principal component. For example, lambda i is indicating the variance of the i-th principal component. So, its square root will be standard deviation. So, this R is indicating the standard deviation.

Similarly, here the second value here this is the value of square root of lambda2 and this is here the value of here square root of lambda3 right. So, because there are 3 variables. So, we will have here 3 principal components and this are the standard deviation of the first second and third principal component. Now after this if you try to look at this first column under this PC1. So, this is eigenvector here e1 corresponding to lambda1 and similarly here this second column here PC2 this is the second eigenvector e2 corresponding to lambda2 and similarly here the third eigenvector is the value of the third eigenvector corresponding to lambda right.

So, now based on that we can create or construct here the principal component as follows. So, please try to look at this first column based on that I try to create here a this first principal component which is here e1 transpose x. So, you can see here I try to use here - 0.66 which is here the first value in the PC1 or the coefficient associated with x1 in the first principal component which is here like this. So, this will become here - 0.66 * x1 + then I try to take here 0.745 and yeah I can write it as 0.74 x2 well you can if you wish you can write 0.75 also, but anyway my objective is to explain you how these values are being used here and then I try to use here the third value - 0.07. So, this will come here this is coming here - 0.07 x3 and similarly if you try to see here for this x2 also you can create here this y2 right. So, you can see here these values are coming here like this and then similarly you can compute here the third PC also whose values are giving here right. So, now, you can see here using the eigenvectors you can create the principal components and after this if you try to look at this the value of lambda1 lambda2 lambda3 then they are obtained here like this the variance of y1 will be 17.8 whole square which is coming from here right and which is actually the value of here lambda1 then variance of y2 we will be coming from here.

So, this will be your here 5.46 whole square which is the value of here lambda2 second variance are the variance of the second principal component and third principal component has this variance lambda3. So, because the R is reporting the standard deviation. So, that is why I have to square it. So, now, this is how you can see that it is not very difficult to obtain it. So, now, we try to find out the summary of this outcome and it is essentially trying to give us an idea about the importance of the component.

So, if you try to see here this is there are three values which are given here is standard deviation proportion of variance and cumulative proportion. So, standard deviation you know this has been obtained from say from the eigen values. So, this is the value of here square root of lambda1 this is the value of square root of lambda2 and this is the value of square root of lambda3. And then this proportion of variance this is computed by here lambda1 upon lambda1 + lambda2 + lambda3 this value here is lambda2 upon lambda1 + lambda2 + lambda2 + lambda3 and similarly, here this is the value of here lambda3 upon lambda1 + lambda2 + lambda3.

Right So, now, then in the third row it is here the value of this the same lambda1 upon lambda1 + lambda2 + lambda3. Then, the second value here this is the value of here lambda1 + lambda2 upon lambda1 + lambda2 and so on. This is the value of here lambda1 + lambda2 + lambda3 upon lambda1 + lambda2 + lambda3. So, that is why it you can see here it is trying to give us the cumulative proportion and by looking at this thing you can decide that how many principal components would you like to keep here. Well, this is a very small data set.

So, we will show you may not get here a very clear idea, but if you have a large data set then you can very easily get it. Right. So, and this is the screenshot of the same option here. So, let me try to show you these things on the R software also. So, that you become more confident about it.

So, let me try to first create here this data frame. So, you can see here now this is here data same and then now I try to this find out the principal components. So, you can see here this values are coming out to be here like this. So, this is the same value which I have just shown you and if you try to find out the summary of these values you will get here the same thing.

Right. So, you can see there here that executing it is not difficult the main thing here is how to interpret the outcome. Right. So, now I try to give you some more ideas about some graphics which are ah used in the principal component analysis to get some more idea. So, we have here a command for here biplot.

Right. So, before we try to understand what is biplot and how it is being used firstly let us try to understand what it is doing and how it is creating. So, actually this biplot helps in visualizing the similarities and dissimilarities between the samples and shows the impact of each attribute on each of the principal component. Right. So, the command to create the biplot here is biplot B I P L O T in parenthesis you have to give the here the data object to be fitted here and then if you have more points like as then you can give it here as say x y where y is the second set of points or a two-column matrix.

Right. Usually associate it is associated with the variable then I have here different option var.x column xlab ylab etcetera. So, I will try to give you here some basic idea, but again I will request you to look into the help menu to get the more details. So, var.x is it is a logical variable if it is true then the second set of points have a rows representing them as a scale axis. I will try to show you that in the biplot there is a axis there is an arrow like in case 1. With this here col the vector of length 2 giving the color to the first and second set of points respectively can be used.

Then cex is the character expansion factor used for labeling the points arrow.len, this indicates the length of the arrows heads on the axis plotted in variance var.x is if it is true. The arrow head can be suppose can be suppressed by arrow.len = 0. Right. So, you can and then xlim, ylim as usual they are the limits of the on the x and y axis in the units of the first set of variables. So, that we have used earlier also, but now you can use see that there are different components of the graphic they can be controlled very easily from this biplot.

Right. Now, the next question is this how to understand it and how to take inference out of that. Right. Well, before I move look at me try to tell you that it depends on your practice also that how much you practice to understand what a graphic is going to going to explain you. Right. So, in the biplot if you will try to see well, I can show you here the first how it looks like.

Right. This looks like here like for example, this is the biplot of the same data set what we have obtained this pcdataout. So, you can see here you have here this arrow whose length is here longer than this arrow is here shorted it has only one color say red and then this this arrow head also can be controlled with respect to the size etcetera. And this is here on the x axis we have here PC1 on x axis we have here PC2. So, this is biplot. So, now, you can understand what I am trying to say that the biplot of the principal component shows both PC scores of sample and loadings of the variable.

So, this PC scores of sample that means, the they are indicated by here dots and the loading of the variable that is indicating by here vectors. Right. Some of the aspect I will be able to show you here in this is smaller in a small data set in a very simple plot after that I will take another data set where I will be able to show you more aspects. Right. So, now, you can see here in this PC there is here is origin you can see here this is here at the 0 0.

So, this is here the origin from where all the arrows are originating. So, the further away these vectors are from a PC origin the more influence they have on that PC. Right. So, that means, if you try to have a look here if this length is here more then it is trying to indicate this is about x2. So, x2 is making the maximum influence on the principal component.

Then the second largest length here is this about x1. So, this x1 is going to have the smaller impact than x2 and then x3 here has the shortest length. So, that is indicating that x3 has the least influence. So, this is what I am trying to say from here. Right. So, the length of this vectors or the line of the arrow that is indicating the influence of the variable on that corresponding principal component.

Right. Then representation of angle degrees between the two variables. Right. So, for example, if you try to see here this is here the angle if I try to make it like this is here the angle this is here the angle. So, this angle is indicating here that 90 percent angle indicate no correlation between the two variable that means they are completely independent and if the angle is less than 90 degree this indicates a positive correlation between the two variables and if the angle is more than 90 degree it indicates a negative correlation between the two variables. So, you can see here now in this case if I try to say here angle number 1 and say here angle number 2.

You can see here angle number 1 is greater than 90 degrees and angle number 2 is less than 90 degrees. So, now, you can see here because the angle is greater than 90 degrees between x2 and x3. So, I can say that here x2 and x3 have a negative correlation whereas, this here angle number 2 which is between x1 and x3 right this is less than 90 degrees. So, I can say that they indicate a positive correlation between the two variables.

So, this is how you can interpret it in this biplots. The vector line length represents the variance level of the variable. So, longer the vector line the greater the variance and the shorter the line the less the variance. So, that I already explained you that this length is trying to indicate the importance right. So, and then similarly we have here the scree plot also firstly let me try to explain you about this scree plot and then I will try to show you both the plots on the R software which are very easy to compute right. So, if you try to see here what is the scree plot? The scree plot is used to visualize the importance of each principal component and a scree plot is a diagnostic tool to check whether principal component analysis works well on the data or not right.

A scree plot shows how much variation each principal component captures from that data and it is used to determine the number of principal components to be retained. Well, I can show you that how it looks like. So, this is the scree plot for the same data set that we have analyzed with x1, x2, x3 right. So, the scores the scree plot has eigenvalue on the y axis and the number of principal components on the x axis. So, you can see here this is here on this side there is information about the eigenvalues.

Here it is trying to tell that this is the percentage of the explained variance right. So, whatever variance you have obtained here this outcome right. So, these are the proportion of variances which were explained in this outcome and they are plotted here on this scree plot right. So, ideally the curve should be steep and then it should bend at an elbow and this is the cutting of point and after that it flattens out right. So, you can see here it is coming here like this, then this is here the elbow and then it is after that it is flattening out.

So, then we can take a decision that how many principal components are going to be used and this can be generated by using the command fvis underscore eig. So, I have used this command on the same data set pc data out and you can see here. So, you can see here that almost this 90 percent of the variation is being controlled by first principal component and the second principal component has only here a smaller value that you can see from this outcome also.

The first value is 0.91 and the second value here is 0.08 and third value is close to 0. So, that is what is shown here in this scree plot. So, by looking at this curve wherever there is a sharp bend after after this the curves bends out then you can take the you can decide the number of principal components to be there to be used. Surely, I have here only three variables. So, it is not looking so nice, but I but if I have large number of variables this curve will be better right.

And then another if casualization is here cost to visualization and it is used to determine how much each variable is represented in a given component and this is a similar curve which is computed from here fviz underscore cos2 function right. So, this is computed and if it gets a high value then it indicates a good representation of the variable on that component and if this fviz underscore cos2 function has a low value then it indicates that the component does not perfectly represent the variable right. You can see here if I try to create here this thing this will be here via this type of graphic. So, if you try to see here this is here the cos2 quality of representation and if you try to see here this is trying to take here x1, x2 and here x3 and it by this height here it is trying to show you that how much each variable is represented in a given component right. Now, you can see here that we have different types of here plot which are giving us different type of information.

So, now, the good option will be if they can be combined together. So, then we try to combine here the biplot at this cos2 right. So, biplot and attribute importance can be combined to create a single biplot where the attributes with similar to cos2 scores will have similar colors and this can be achieved by the function here f with underscore PCA underscore VAR. So, you will try to see here this will give you here this type of here outcome here and it will here look like this. So, you can see here these lines are coming from the biplot and their importance which is here giving by this by the pressure cos2 scores right. So, now, if you try to see here that if this is here green this is here green if this is here yellow then this belong to this band and if this is here blue then it is here somewhere here 0.

So, you can see here. So, this is how you can have a better information about the importance of the variable on the PCA right. So, first let me try to show you these things on the R console it is a very simple thing for you. So, if I try to first create here the biplot you can see here. So, now, if you try to see if you try to create here this biplot it comes out here very easily right. Now, similarly if you want to create here this scree plot here then in order to create this scree plot here, I can share with you that you need here the there here the package see here fact to extra right.

So, I can just write down here itself. So, let me just copy this library function here although I already have installed this package and now if you try to create here this scree plot here it will here come like this you can see here like this right. This is the same scree plot which I have shown you I video on the R console. Now, if you want to have here this one type of a cost to. So, this is here the cost to variables plot and then if you try to combine here both the things together although I have here given some nice functions which you can see from the help menu also you can see here this is here this these are the values right. So, you can see here it is not a very difficult thing actually to do and now I try to show you the same exercise I will not explain you much, but I will simply try to show you that if you try to take the same data set, but if you spinterize it because I had told you that when we are trying to use the principal component then we have two option either we try to make the units free data or we try to make a data with units.

So, if I want to make the variables unit free then I can use here the option here center = true and scale dot = true and using the scaling of the data the data will become unit free. So, definitely it will give you principal components, but the nature and behavior of those principal components may be different. So, I just want to show you here in this lecture that you will get a different outcome and how it is working. So, I am now going to repeat the same exercise that I just use the same data set with the same command pr comp over the data pc, but now I have added here center = true and scale = true and if I try to do its principal component analysis we get here the this type of data you can see here now this values of square root of lambda1 square root of lambda square root of lambda3 this e1 this e2 and here e3 are changed from earlier and now. So, now based on this we can obtain the principal components here the first principal component and third principal component using this e1, e2, e3 and these are here the variances of Y1, Y2, Y3 and you can see here they are different than earlier.

Now, if you want to have here the summary of this outcome you can see here these are the standard deviation proportion of variations and cumulative proportion, but they are now different than others. If you want to have the biplot then it is also looking like this which I will show you in more detail in the next slide, but you can see here these values are different than your earlier values and this is here the biplot with the scale data. So, you can see here now these lines are almost the same whereas earlier it was they were of different lengths and this is here the screenshot of the same outcome that you can do yourself and then if you try to see here the scree plot also. This is also changed when we have the scale data. Earlier the second bar was not so high it was something like somewhere here, but now here you can see that the contribution of the second principal component has significantly increased as compared to when we use the unscaled data.

And similarly, if you try to see here the cos 2 function is giving you all the three variables have the same important same contribution whereas it was different in the earlier case. So, if you just try to compare in the case of unscaled data, we had a scree plot like this one and in the scale data the scree plot is here like this one. So, this is changed and similarly if you try to compare the scree plot in the case of unscaled data it was like this and in the scale data it is here like this right. So, I just wanted to show you that you can conduct the principal component analysis with this scale as well as unscaled data and in both the cases you will get a different outcome right. Now, I try to take here a smaller beta set which is not so big and that we have used in the cluster analysis also and we would like to conduct here the principal component analysis.

So, this is the data is USArrests. So, this is the data which is available in the R software and it is about the violent crime rates by the U.S. state and it has 50 observations on four variable the four variables are murder, assault, urban population and rape. And it has some data set which I have taken from actually this reference. Anyway, my objective here is to consider these 50 observations on four variable and try to see what happens right.

So, now this is the data set that you quite we will see in the R software. Now, I try to load this data set and then you will see here this looks here like this there are four variables say 1, 2, 3, 4 and then you have here states, Alabama, Alaska, Arizona, etc. And we have such 50 observations 50 states right.

So, now, I try to do here the principal component analysis on this USArrests data. So, it is straightforward right. I will just repeat the same steps what I used earlier you will use here the command here prcom on the data set USArrests and you try to save it in this data vector or say this this outcome variable pc data U.

S. arrest and which is like here. So, you can see here these are the values of square root of lambda1, square root of lambda2, square root of lambda3, square root of lambda4. So, now you can see here this is here pc1 this is the value of Eigen vector, corresponding to lambda1. This pc2 is the Eigen vector e2 corresponding to lambda2. This pc3 is the Eigen vector e3 corresponding to lambda3 and pc4 is the Eigen vector corresponding to lambda4. And based on that we can exactly in the same way we can create here the this first second third fourth principal component from the first value e1 we can create here this 0.04 * murder + 0.99 * assault then 0.046 which is 0.05 urban pop and then fourth value here 0.75 to be here 0.08 * rape. And similarly, you can consider other variable also without any problem. And if you try to find out it as summary then you can see here these are the standard deviations of the principal component, these are the proportions then these are the cumulative proportions.

Exactly in the same way what we did earlier and this is here the screenshot and if you want to now create here a biplot of this data outcome, this principal component analysis you can see here now they are here because there are here 50 observations so all those observations are given here name you can also give them in the form of here plots and then you can see here these are here the rows which are trying to give you here the biplot. But as I told you that if you try to increase the number of observation and the number of variables this picture will become more clumsy but it is not difficult to understand. But I wanted to show you that how it will look like. Now if you try to make it here the scree plot you can see here that the first principal component is trying to take care of the maximum value then the second principal component is lowest and third principal component you can see here is somewhere here or here you can take the elbow point and you can consider say two principal components in this case. Similarly, if you try to create here the cos 2 values here so you can see here this is a similar graphic here and then if you try to combine the biplot and cos 2 values here together for this USRS data set it will here look like this you can see here these are here the biplots and then now the color of these arrows is indicating the values of the cos 2 functions here.

So, you can see here it is not a very difficult thing for us to do it and if you try to see here I can show you these options in the R software also so you can first have a look here like this So, you can see here if you use a USArrests you can see here this is here the data like this one but anyway we are not looking into the data because now our job is to employ the statistical tool and then we try to see how the outcomes are going to look like So, you can see here if I try to conduct the principal component analysis the outcome on this USArrests data it looks like here like this and if you try to find out here the summary then it will give you the same outcome which I shown you on the slides and similarly if you want to create here the biplot of this one you can very easily create here is a biplot you can see here this will here like look like this right and similarly if you want to create here this scree plot here it can be done here also very easily you can see here this is the scree plot and if you want to combine here find out here this cost two values you can easily use this command and we get here these are the cost two values well I can repeat it again so that you can see clearly right and similarly if you want to have this both biplot and scree plot together then it will be here like this here you can see here it is like this right So, we now we come to an end to this lecture and yeah it was a little bit long lecture, but I had no options because I wanted to show you this analysis for a scale data as well as for a scale data in a single lecture otherwise there will be confusion in the understandings So, in this lecture what I have done I have shown you that how to do the principal component analysis in R number one then I have shown you that how to interpret the outcome how you are going to obtain the values of variances of principal component and how are you going to construct the principal components.

Then I have taken different types of plots like a biplot scree plot cost two plots etc which are going to give you some more information about the variable collection or the dimensionality reduction and I have taken it first on a small data set and then on the this data set I have considered in in two ways scale data set and unscaled data set So, remember I have converted here the covariance matrix into the scale data set that will give you simply the correlation matrix right So, then after this I have used it I have tried to show you that there will be difference in the outcome when you try to do the principal component analysis using the same function with a scale and unscaled data. Now it is up to you that which of the analysis you prefer you prefer to use right and you have to just match what is happening in the real life whether it is matching with the outcomes or not and then you try to look into different types of plot and then try to take a wise decision. And then I have taken a small data set where I have repeated exactly the same exercise but there you can see the complexity when the number of variables and number of data values increase then what really happens.

So, now it is your turn that you try to have some data set try and then try to practice it. The more you practice the more you will understand and I will stop here in the topic of principal component analysis and in the next and I will try to try to take up another topic on canonical correlation analysis. But it is very important for you to understand this principal component because you will see that the canonical correlation analysis is almost similar to what we have done in the case of principal component analysis but definitely it is different but the concepts will be extended in a different setup So, you please try to understand it and then try to believe that what outcome is going to tell you how it is related to the information hidden inside the data So, you practice and I will see you in the next lecture till then goodbye.