Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 12

Lecture – 54

Canonical Variables and Concepts

Hello friend, welcome to the course Multivariate Procedure with R. From this lecture we are going to begin a new topic. This is about Canonical Correlation Analysis. So, the first question come what is this and what are we going to learn in this topic. So, you can recall that in the case of principal component analysis we had ah, transform the original variables and we had obtained the principal component in such a way such that the total variability of the older and new coordinates system remains the same. That means, the total variance due to x1, x2, xp will remain the same as the total variance due to the principal components y1, y2, yp.

And this way we had used it for the dimensionality direction that we ah, tried to find that what are the important variable which are contributing more towards the variability in the system. Now, the principal component analysis works when you have only one variable. What will happen if you have two sets of variables which are interrelated correlated with each other? Although you can understand that when we considered the principal component analysis and when we use the scaled data on the covariance matrix then the covariance matrix was automatically converted into the correlation matrix. But that was the correlation matrix of the single random vector.

But now suppose you have two random vectors then how are you going to understand that which are the important variable which is going to contribute more so that you can reduce the dimension of the data. This can be achieved by the canonical correlations. So, what we try to do that we have two random vectors on which we have two sets of data and we try to explore the degree of interrelationship between the variables of these two random vectors based on the given set of observation. And then we try to transform them in such a way such that the total correlation in the original coordinate system will remain as the ah, as the same in the total correlation in the new system. Just like from x1, x2, xp we had obtained y1, y2, yp in the case of principal component analysis.

Now, we have here two sets of data and we would try to find out two different sets of linear combination in such a way such that the total correlation structure of both the variables remain the same. And this can be achieved by the canonical correlation analysis. And you can think that it resembles to the multiple linear regression model also. So, in this lecture I am going to explain only the basic concept behind this canonical correlation analysis that what are we going to do and what we want to achieve. And then in the next lecture I will try to take up the mathematical part, mathematical analysis, but I thought that ok it is first it is important for you to settle down this concept in your mind then the mathematical analysis will become very simple.

So, this lecture is going to be a non-mathematical lecture where I will try to give you the different aspects of this canonical correlation analysis. So, let us begin our lecture. So, now we are going to talk about these canonical variables and the related concept in this lecture. So, the first question comes here what is canonical correlation analysis? So, suppose there are two large sets of variables which affect the outcome of an experiment and the experiment want to study the interrelation by considering only a few linear combinations of the variables from each of the set. Similar to what we did in the case of principal component a component analysis also.

And the experimenter wishes to study that linear combination which are highly correlated as in the case of principal component analysis we took those variables which have the higher variance. So, similarly we are going to think about the say this higher correlation in the case of canonical correlation analysis. So, this is the way it will help us in the dimension correction that instead of considering all the variables we will be considering only a few variables. So, let us try to consider two sets of variables. So, definitely they are observed on some random vectors.

So, they are expected to have a joint probability distribution function. So, we consider that those two sets of random vectors or the sets of observation they are obtained from the jointed distribution. And then we try to analyze the correlation between the variables of one set with those of other sets. And then we find a new coordinate system in the space of each set of variables. So, that the new coordinate displays the system of correlation unambiguously.

So, we find the linear combinations of variables in each set that have the maximum correlation. In the case of principal component analysis, we did a similar exercise I am not saying the same exercise similar exercise keeping in mind the concept of variance. And these combinations are the first coordinate in the new system right. Then after this we consider the second linear combination in each set and such that the correlation between those in is the maximum correlation between the such linear combination right. So, you have to find out the second linear combination such that which has got the second largest correlation coefficient among all the correlation coefficient of different linear functions.

And this procedure is actually continued until the two new coordinate system are the are completely specified. So, if you have p variables then you will have here say here p such linear combinations also. Definitely the question will come if one random vector has less number of variable and another random vector have most number of variables then what to do. So, I will try to address them gradually right. So, the question before us is when the observations are taken on the large number of correlated variables then how do we reduce the number of variables without sacrificing too much information.

And then you also have to define that how are you going to define this information concept. In the case of principal component analysis, it was variance in the case of canonical correlation analysis it will be correlation. So, when variables are regarded as belonging to a single set of variables then we use the principal component analysis. And when the variables fall naturally into two sets then we use the canonical correlation analysis that is the basic difference between the principal component analysis and this canonical correlation analysis. Now if you try to see this is quite similar to the multiple linear regression also.

So, let me try to address the case issue that how it is different from the multiple linear regression analysis. So, the canonical canonical correlation analysis facilitates the study of interrelationship among the sets of multiple dependent variables and multiple independent variables. Whereas in the case of multiple linear regression and analysis we have only a single dependent variable from a set of multiple independent variables. So, the main difference is that in the multiple linear regression analysis we had a model like y equal to x beta plus epsilon where y was the single variable it can be yield of the crop or this can be the effect of the medicine right. But in the case of canonical correlation analysis there can be more than one dependent variable right.

So, this canonical correlation analysis simultaneously predicts multiple dependent variables from multiple independent variables and that is the basic difference between the two right. And when you are trying to conduct the canonical correlation analysis then the appropriate data for canonical analysis are the two sets of variable right. And if we assume that each set of variables can be given some theoretical meaning and at least to the extent that one set could be defined as the independent variables and the other set can be defined as the dependent variable. And once this distinction is made then the canonical correlation can address a wide range of objectives right. For example, we can determine whether two sets of variables are independent of each other or determining the magnitude of the relationship that may exist between the two sets.

We can also derive a set of weights for each set of dependent and independent variables so that the linear combination of the each set are maximally correlated. That means they have the maximum correlation. And additional linear function that maximizes the remaining correlation are independent of the preceding sets of the linear combinations. It is similar to principle component analysis. For example, when you found the second principle component then you had put a condition that it should be independent or say uncorrelated of the first principle component.

Similarly, when you find the third principle component then you had found it in such a way so that it has a maximum variance and it should be uncorrelated with the first two principle component and so on. So, the similar concept is implemented here also in the case of canonical correlation. Explaining the we can also explain the nature of whatever relationship exists between the set of dependent and independent variables and which is generally measured generally by measuring each variables relative contribution to the extracted canonical function. It is similarly like in the case of principle component analysis; you found the contribution of the say this first principle component by computing its contribution towards the total variability in the system. And similarly for you exercised the same thing for the principle component.

So, the same thing can be done here also in the case of principle in the canonical correlation also just like principle component analysis. And we can define that how many linear functions are really contributing towards the in towards this correlation coefficients and then they are helping us in reducing the number of variables. For example, suppose an experimenter collect data on three psychological variables, four academic variables and gender that is male versus female for say 1000 students. Now, the aim is to know how the psychological variables relate to the academic variables and gender. So, now this psychological variable will become your here dependent variable and academic variables and gender variable they will become your independent variable.

So, now here our aim is to know how many dimension or the canonical variables are necessary to understand the association between the two sets of variables. So, now how are you going to do it going to implement it? There are two aspects one you have to understand what are you going to do and the same thing we have to implement it through the mathematical procedures. So, this canonical correlation analysis focuses on the correlation between a linear combination of the variables in one set and the linear combination of the variables in say this another set. You can recall that in the case of principle you had found this type of linear functions. So, the same thing is here for the two sets of variable.

The idea is first to determine the pair of linear combination which have got the largest correlation. And next after this we try to determine the pair of linear combination having the largest correlation among all pairs which are uncorrelated with the initially selected pairs and this process continues. So, it is so what we will do that first we will try to find out the correlation among all sets of linear combination and then we will try to choose that which correlation is the maximum and that will give rise to the first set of linear combination which are having those correlation structure. And then we will try to find out the second set of linear combination which is uncorrelated with the first linear combination. So, these pairs of linear combination are called as canonical variables and their correlation what we are trying to find they are called as canonical correlations.

So, these canonical correlations actually measure the strength of association between the two sets of variables and the maximization aspect of that technique represent an attempt to concentrate a high dimensional relationship between two sets of variables into a few pairs of canonical variables. So, the same process that we had followed in the case of principal component analysis that we wrote the y1, y2, yp in such a way such that they have got a different contribution of variance and then we try to choose some of the principal components which are sufficiently contributing towards the total variability in the system. So, now the same thing we will be doing here we will have here two data sets and then if we will try to create the linear combination for both the data sets and then we will try to find out their correlation coefficient and then we will try to find out the first set of canonical variables which has got the maximum correlation. Then the second set of canonical variables will be the pair of these linear combination one from each of the data set such that they have the second largest correlation and they are uncorrelated with the first pair of canonical variables. Similarly, then we try to find out the third pair of canonical variables as the pair of linear combinations of the independent variables in the two data sets such that they have got the next highest correlation and then corresponding to which whatsoever is our, whatsoever are the linear combination they will construct the third pair of canonical variables and this third pair of canonical variables will be found in such a way such that it is uncorrelated with the first and second canonical variance.

So, this process will continue and then finally we will find a set of canonical variables. So, now we come to an end to this lecture here and then you can see here that it was a very short and brief lecture and my idea was simply to give you here some concept about the canonical correlation and I wanted to establish that the way you have understood the principal component analysis on the similar grounds you are going to understand the this canonical correlation analysis also. Now, the question comes here if you have got here two data vectors and one has suppose a smaller number of variables and another has larger number of variable. So, obviously when you are trying to find out the correlation coefficient the correlation coefficient is always obtained between the pair of random variable correlation coefficient between X and Y. So, then in that case the total number of canonical variables that can be constructed will be the minimum of the numbers in the two data sets.

So, if one data sets has five variable and another data sets has suppose eight variables then five pairs of canonical variables can be considered. Now, we have understood the concept, but definitely until we try to convert them into a mathematical setup and we try to understand it will not help and the next objective will be that how are you going to implement in the R software. So, these two aspects we will try to take up in the forthcoming lecture, but before that my request is that you try to look into the book and try to understand that how this canonical correlation is going to help in different ways. Well, I have told you about the dimension reduction only, but there are different types of areas in which this can help. For example, the principal component analysis or this type of canonical correlation analysis is very popular in the image analysis.

So, although I am not going to cover it here, but those who are working in the real-life data they have to understand and the same tools can be applied over there also. So, you try to increase your knowledge try to read the books and try to understand these concepts and I will see you in the next lecture where I will try to explain you the mathematical setup of this canonical correlation, till then goodbye. Thank you.