CSC 424

Yiyang Yang

11/14/2015

"The PCAmixdata R package" review

The lecture I choose to review is "Multivariate analysis of mixed data". From the abstract, I know the data is characterizing conditions of life of cities of Gironde, the datasets are the mixture of numerical and categorical variables.

From the instruction, I find that the restriction of the dataset is the data type, which means each dataset the authors talk about have numerical and categorical data combining together, so plenty of standard methods, such as Ade4, FactoMineR, ExPosition, can not be applied to their dataset to multivariate analysis of mixed data. Then focus this restriction, they prepare to use the three techniques, PCAmix, PCArot and MFAmix. According to these methods, the first two is going to compute the principle components of the mixed data, and MFAmix is going to compute the multiple factor of the dataset. From the description of the authors, PCAmix and PCArot are based on generalized singular value decomposition (GSVD), also involve standard PCA and the multiple correspondence analysis (MCA).

GSVD of the real dataset is used to introduce weights to rows and columns of the dataset in PCA, standard PCA in PCAmix package is going to apply on the numerical part eigenvalues in here is the sum of correlation ratios between the numerical variables and principle component, and MCA is going to apply on the categorical part eigenvalues here is the sum of correlation ratios between the categorical variables and principle component, if the dataset is only type either of them, PCAmix will apply the relative method to get the result.

PCArot is similar with PCAmix, just a rotation of it, to the numerical part, PCArot is the standard varimax roatation procedure in PCA; to the categorical part, PCArot is the corresponding varimax rotation procedure for MCA.

MFAmix is very similar with standard MFA, the main difference is MFA could only be used on the same data type group, however, MFAmix could directly apply on the mixed data. In the algorithm of MFAmix, GSVD is used alike PCAmix, but there is a difference in the factor scores processing step, in PCAmix, the GSVD of matrix Z with two metrics N and M, in MFAmix, the GSVD of matrix Z with two metrics N and M = MP, which P is the diagonal matrix of weights of the columns of Z.

The outputs of PCAmix have two parts, numerical and graphical, from the numerical part, the function could be used to predict the scores of new observations. In the graphical output, the first graph divides the dataset into two parts base on the percentage of houses infimum and supremum 90%; from the second graph, I know that if an area has more houses then it will have less council housing, vise versa; the third graph tells me the relationship between the numerical variables, the obvious relationship is the density and owners have negative relationship, and the fourth graph is a combination result of both numerical and categorical data. To the outputs of PCArot, the numerical output is similar with the one in PCAmix, according to the description of author in the PCArot graphical output, although the result is not same as PCAmix, the benefits of using rotation on this dataset are very limited. The outputs of MFAmix, the numerical output still could predict the scores of new observations. In the graphical part, these graphs are descripting the relationships between numerical and categorical variables, and then get some results about how the cities will influence the citizens'

life

Besides the three methods introduce above, in the R codes of these method, I find that “predict.PCAmix” and “predict.MFAmix” are the new R functions to predict the coordinates of cities, another R function used in PCAmix is “splitmix”, which could split the mixed dataset into different parts based on whether it is numerical or categorical.

After reading this article, I learn three very useful methods on mixed dataset, and how to apply them and interpret the result of them.

Since our dataset doesn’t have enough categorical variables, I only try the PCAmix in our final project, and get the following result.

Attempt:

Source Code:

```
library(PCAmixdata)
weather <- read.table("/Users/Yiyang/Documents/CSC 424/Final
Dataset/Dataset.csv",sep=",",header = T)
w1 <- weather[c(-470, -710, -461, -462, -184, -527, -33), -1]
head(w1)
wsplit <- splitmix(w1)
wsplit
x1 <- wsplit$X.quanti
x2 <- wsplit$X.quali
w.pcamix <- PCAmix(x1, x2, rename.level = TRUE, graph = FALSE)
w.pcamix$eig
```

w.pcamixindcoord

w.pcamix$coef

event <- x2$event

plot(w.pcamix, choice = "ind", axes = c(1, 2), coloring.ind = event, lable = FALSE, posleg = "bottomright", main = "Observation")

plot(w.pcamix, choice = "levels",axes = c(1,2), xlim = c(-1.5, 2.5),cex = 0.9, main = "Levels")

plot(w.pcamix, choice = "cor", axes = c(1, 2), main = "Numerical Variables")

Ouput: