8. Assignment NMST 539

Frantisek Helebrand

April 2022

1 Introduction

In this assignment we will be analyzing the data from the US, specifically, we will be focusing on the crime data from different US states in 1985. We have observations from 50 different US states with 11 variables. We will use only 7 of them (number of murders, rapes, robberies, assaults, burglaries, larcenies and auto thefts) for factor and principal analysis. Principal analysis was made in the last assignment, in this assignment, we will focus on factor analysis and compare outcomes from both.

2 Principal component analysis

This section just summarizes the results from the previous assignment. The tables and figures are the same.

By using principal components analysis we obtain the following table which shows standard deviation, proportion of variability and cumulative proportion of variability.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
St. dev.	2.0191	1.1965	0.7945	0.5832	0.4984	0.3737	0.3636
Prop. of Var.	0.5824	0.2045	0.0902	0.0486	0.0355	0.0199	0.0189
Cumul. Prop.	0.5824	0.7869	0.8771	0.9257	0.9612	0.9811	1.0000

We can see that for an explanation of 90,2 % of variability it is enough to use only the first 4 components. If we would like to explain 95 % of variability we will have to use the first 5 components.

From the figure 1 we can see clusters for data. We have used the first two components. It can be said that in the South prevail assaults and murders, in the west, there are more lancers and burglaries and in the Northeast prevail robberies and autothefts.

Overall we can say that the South region is dangerous in comparison with others.

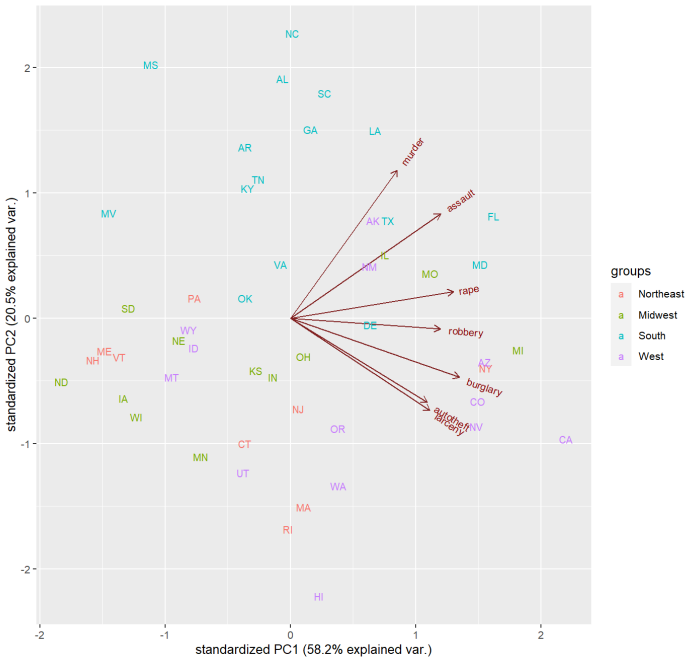

Figure 1: PCA - first and second component.

3 Factor analysis

Our vector is 7-dimensional, so we can use only factor analysis with 1,2 and 3 factors. If we would use more factors the factor analysis (factanal in R) would have too many factors and could not be executable. The table 1 shows cumulative variance for analysis with 1,2 and 3 factors. The table also shows the p-values from the test that the number of factors is sufficient.

Number of factors	Cumulative variance	p-value
1	0.51	3.92e-12
2	0.72	0.146
3	0.79	0.798

Table 1: cumulative variance and p-values for the different numbers of factors.

We can see that the p-value for the two factors is higher than 0.05. Thus two factors could be enough.

The first factor is mostly correlated with burglary, lancery and autotheft, which are "less serious" crimes. On the other hand, the second factor is mostly correlated with murder and assault which are "more serious" crimes. This can be also seen in the figure 2.

The results are quite similar to PCA. Murders and Assaults have similar

directions as can be seen in the figure 1. Similarly burglary. lancery and autheft have similar direction.

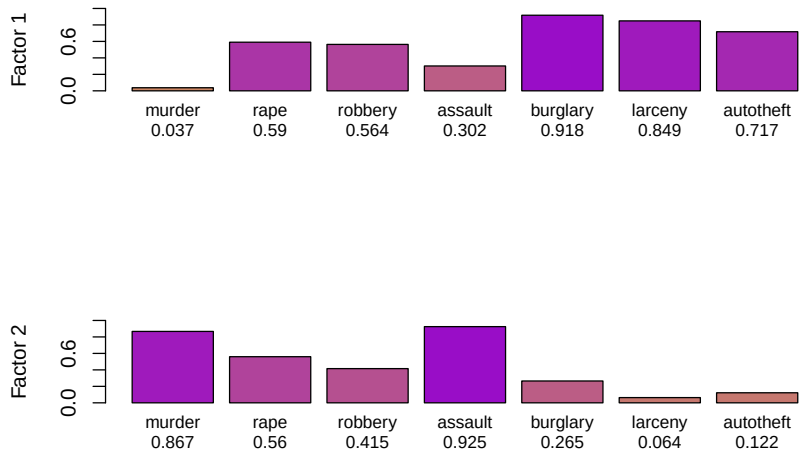

Figure 2: Two factors.