SCF07 Data Set Analysis

Introduction

The SCF07 data set contains 5000 observations of 8 variables. The variables are:

1. `lsam` : the logarithm of savings
2. `linc` : the logarithm of income
3. `educ` : years of education (0-17)
4. `cacc` : the number of checking accounts (1-6; 6 meaning 6 or more)
5. `sacc` : the number of savings accounts, similar to cacc
6. `hous` : cathegorical variable for housing (9 cathegories)
7. `life` : binary, 1 if the person has life insurance, 2 if not
8. `occ` : cathegorical variable for occupation class (7 cathegories)

The first two variables are continuous. The next three are ordinal, but could be treated as continuous. The last three are cathegorical and should be treated as such. To give a better idea, these are the first few observations of the data set:

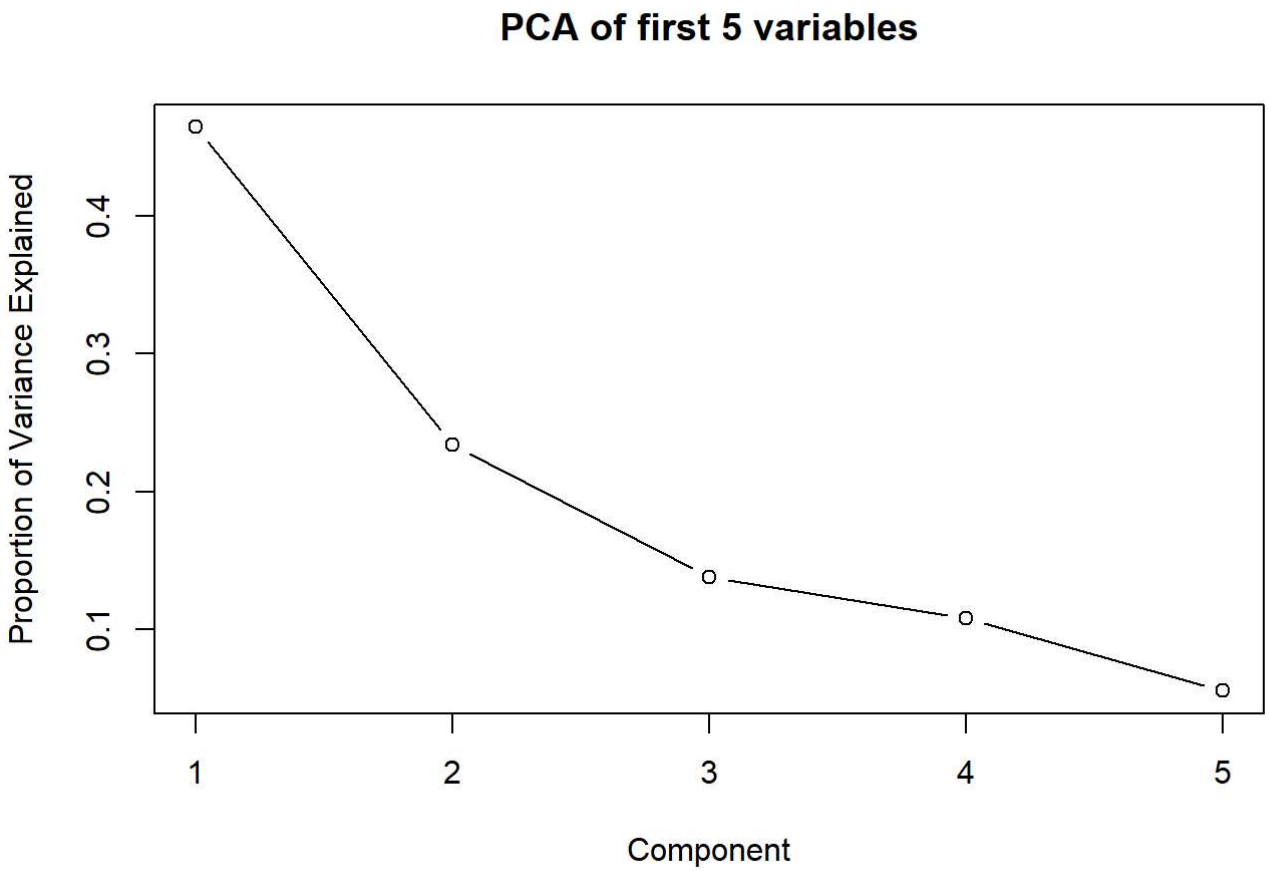
```
##        lsam     linc educ cacc sacc hous life occ
## 1 10.464531 15.34592   17    4    6    3    1   6
## 2 12.100990 13.48846   16    4    4    3    1   6
## 3  7.047517 11.30282   14    3    3    3    1   3
## 4  3.912023 12.61170   16    4    1    3    1   5
## 5 10.310618 12.32408   16    3    2    3    1   6
## 6  3.912023 10.27677   12    2    1    3    2   3
```

Principle Component Analysis

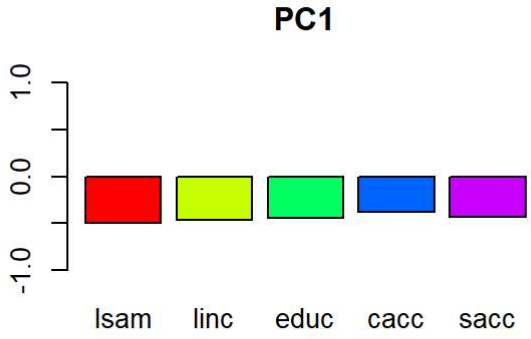
Firstly, to reduce the dimension of the data, we will perform a Principle Component Analysis (PCA). The method will help us capture the most important information in the data and reduce the number of variables. PCA can only be performed on continuous variables, so we will exclude the cathegorical variables. It is easy to see that the units of the variables are different, so we will scale the data to make it unitless.

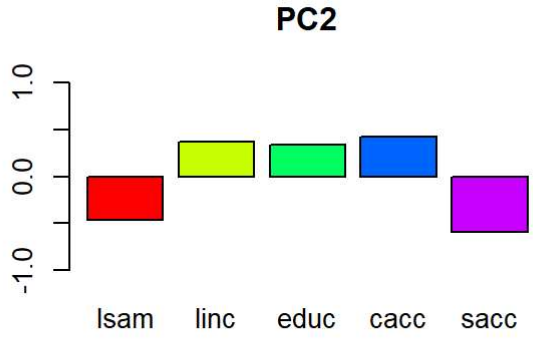
```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5
## Standard deviation     1.5238 1.0821 0.8299 0.7355 0.52664
## Proportion of Variance 0.4644 0.2342 0.1378 0.1082 0.05547
## Cumulative Proportion  0.4644 0.6986 0.8363 0.9445 1.00000
```

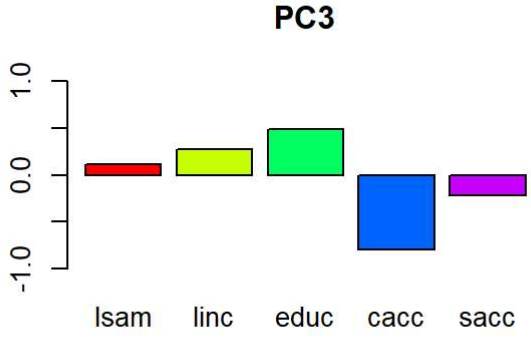
On a first glance, we can see that the first three components encapsulate over 83% of the variance in the data. The first component is the most important, capturing over 46% of the variance.

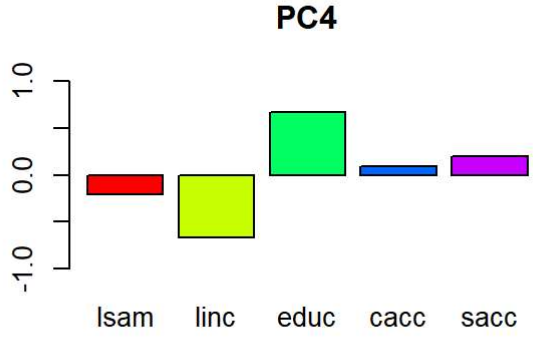


Interpreting the Components









The first component reflects the overall wealth, or “presence” of the person, as it is influenced by all of the variables in the same direction. Low value in the first component broadly speaking means higher income, savings, education, and number of accounts.

The second component represents the ballance between savings and everything else. Low value in this component means higher savings and more saving accounts.

The third component is most strongly influenced by the number of checking accounts. The number of saving accounts goes in the same direction, while everything else goes in the opposite direction; education is the most influential. Low value in this component means more accounts, but possibly lower education.

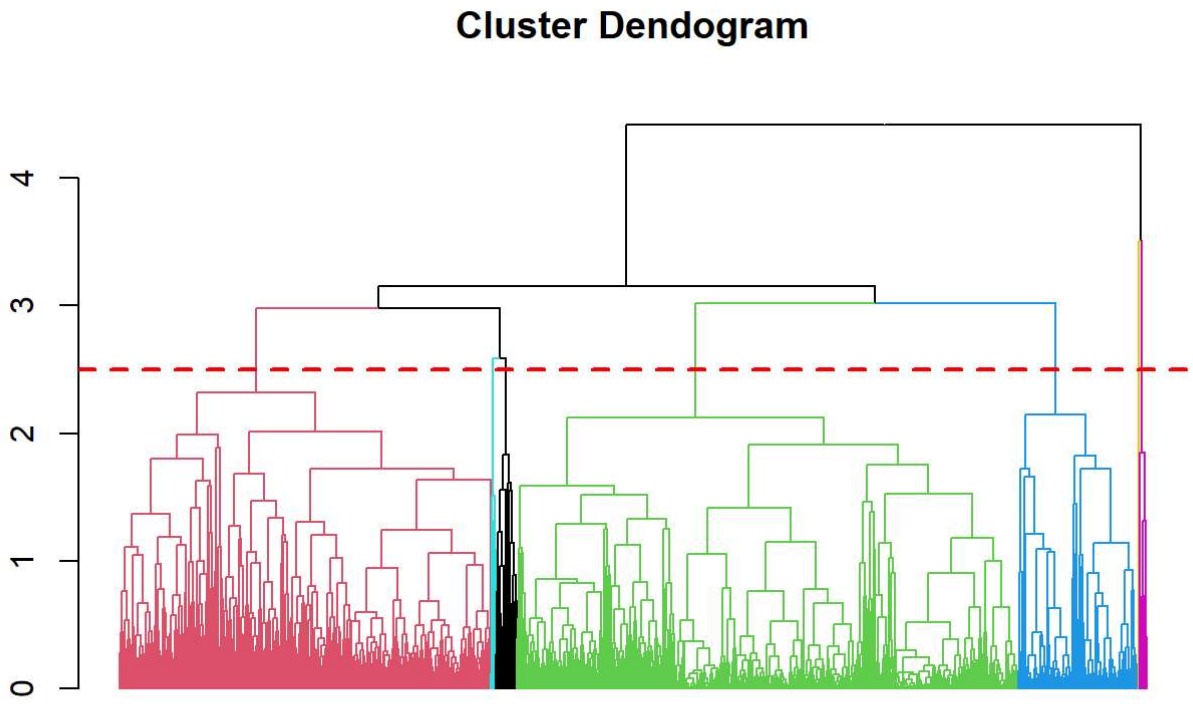
The contribution of the fourth and fifth components is diminishing, so I will settle on the first three components.

Hierarchical Clustering

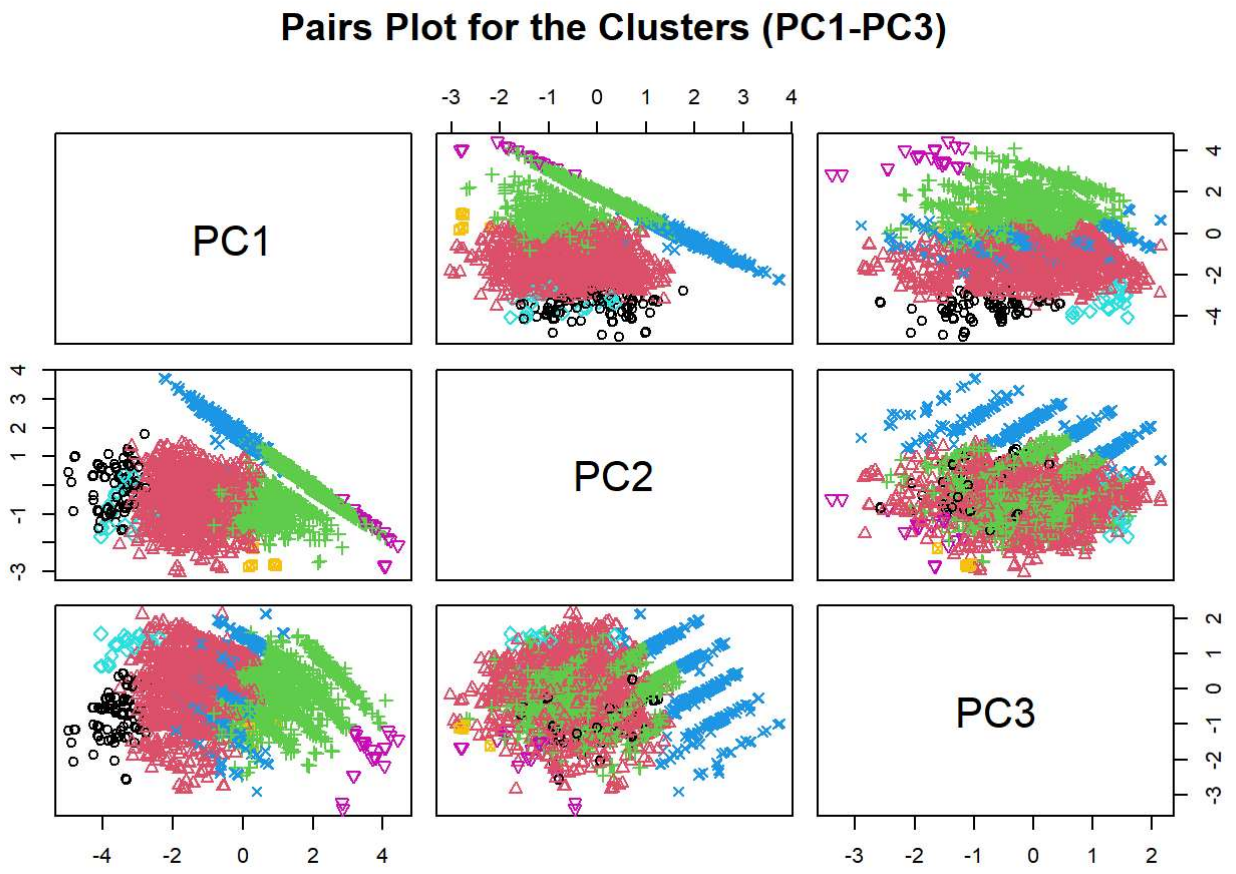
Next, I will perform a hierarchical clustering on the data. I will use the first three components from the PCA to cluster the data.

Performing the Clustering

Looking at the dendogram, it seems reasonable to cut the tree using the height of 2.5. This will give us 7 clusters, which is really nice, because the number of cathegories in the ‘occupation class’ variable is also 7. This might help us interpret the clusters.



Interpreting the Clusters

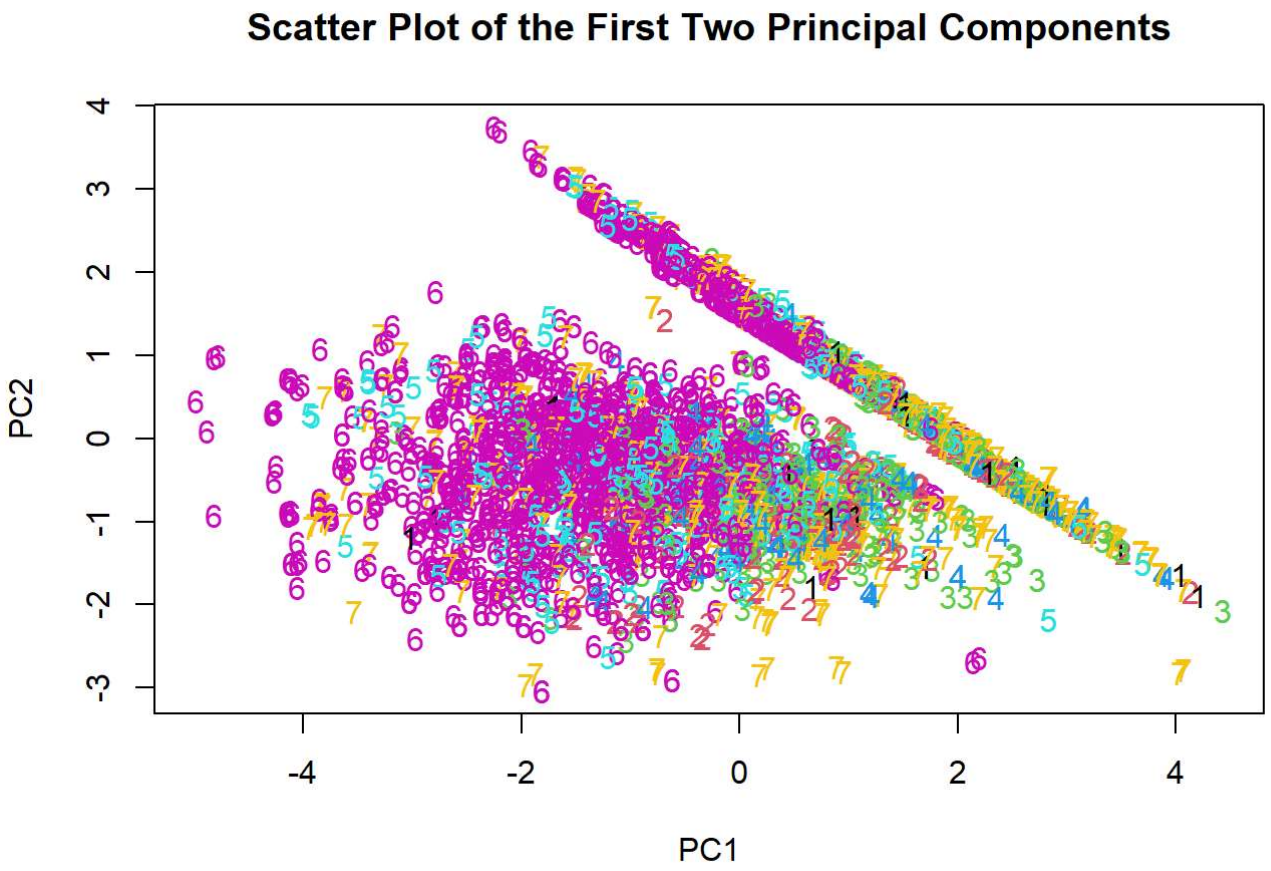


Using the interpretation of the PCA components, we can try to draw some conclusions about the clusters. The big cluster (red) probably represents the average person. The green cluster represents lower class people, while the black and cyan clusters represent the upper class. I am mostly basing this on the first component, which is the most important. The blue cluster also seems to be in the middle class, but with less of a focus on savings (using PC2). The yellow and purple clusters are harder to interpret. Purple seems to represent the most poor people. The yellows seems the be middle class with high focus on savings. The difference between the rich clusters (black and cyan) is in the number of accounts, as can be seen from PC3.

Comparing the clusters with the ‘occupation class’

I will use the Rand Index to compare the clusters with the ‘occupation class’ variable, to see if there is any connection between the two.

The adjusted Rand Index is 0.03, which is quite low, but it is higher than zero. This means that there might be some connection between the clusters and the ‘occupation class’ variable, but it is not strong. We can try to compare the first two components to see if we can find a connection there.



The most prominent class (purple) is the 'manual workers and operators' class [1]. It seems to contain the middle class, but is too spread out to draw any conclusions. The other occupation classes are a mess and there is no clear connection between them and the clusters we studied before.

Conclusion

In this analysis, we performed a Principle Component Analysis on the SCF07 data set, reducing the number of variables we looked at from 5 to 3. We then performed a hierarchical clustering on the data, using the first three components from the PCA. We found 7 clusters, which we interpreted using the PCA components. We then compared the clusters with the 'occupation class' variable and found a weak connection between the two.

---

1. I assume this is the case from the class descriptions which were given to us in the assignment.↩