Multivariate analysis - hw8

Matěj Pešek

April 14, 2022

1 Factor analysis

Our task is to use factor analysis on food data from r package SMSdata, consisting of consumption by 12 french families of 7 different food and drink types, including bread, milk, poultry and wine. We used the same dataset in the last assignment, where we used principal component analysis and found out that the first 2 principal components describe 96 % of the total dataset variance. We might therefore get an idea, that using just 2 factors may be enough.

When finding the factors by varimax procedure, we consider the case of 2 factors, and the case of 3 factors. Even though 3 factors may not be enough according to a formal test, the software won't allow us to use 4 or more factors, as our dataset doesn't contain enough variables. Using 2 factors would describe 82 % of the total variance, 3 factors would result in description of 92 % of total variance. To better describe the difference between PCA and FA, we will consider the 2 factor case, and compare it to the 2 principal components we found in the last asssignment. Before we throw the third factor out, lets take a look what would the third factor actually describe in figure 1

From figure 1, it seems that using 2 factors could still work fairly well.

2 FA vs. PCA

Firstly we should ask ourselves, how the description of the data from 2 factors differs from the description of the data from 2 principal components. In figure 2, we can see the biplot of the first 2 principal components and the variables they describe. Compare that with the estimated loadings in the case of 2 factors in table **??**. In the PCA case, the first component seemed to describe only the variables poultry, fruits and meat, with the variable vegetables being only sligtly described. In FA case, the first factor describes the first 3 variables the same, while vegetables seems to be described even better. Another difference is in the second component. Second PC seemed to describe milk and bread by basically the same margin, but in the second factor loading, we see that milk is much better described than bread. In both cases we see that wine is not well described when we use only 2 factors or principal components.

Variables	Factor1	Factor2
vegetables	0.75	0.59
fruits	0.93	>0.50
meat	0.96	>0.50
poultry	0.99	>0.50
bread	>0.50	0.85
milk	>0.50	0.99
wine	>0.50	>0.50

Table 1: Loadings in the case of 2 factors. Note that the variable wine doesn't seem to be well described when using just the 2 factors.

We can also take a look at the graph from library FactoMineR in figure 3, showing us the individuals factor map, although I'm not sure what the numbers next to the dimensions mean. The percentages doesn't equate to any of the proportional variance percentages in any of the 3 models (PCA, FA with 2 factors, FA with 3 factors), so they could correspond to the case where more factors are somoehow considered.

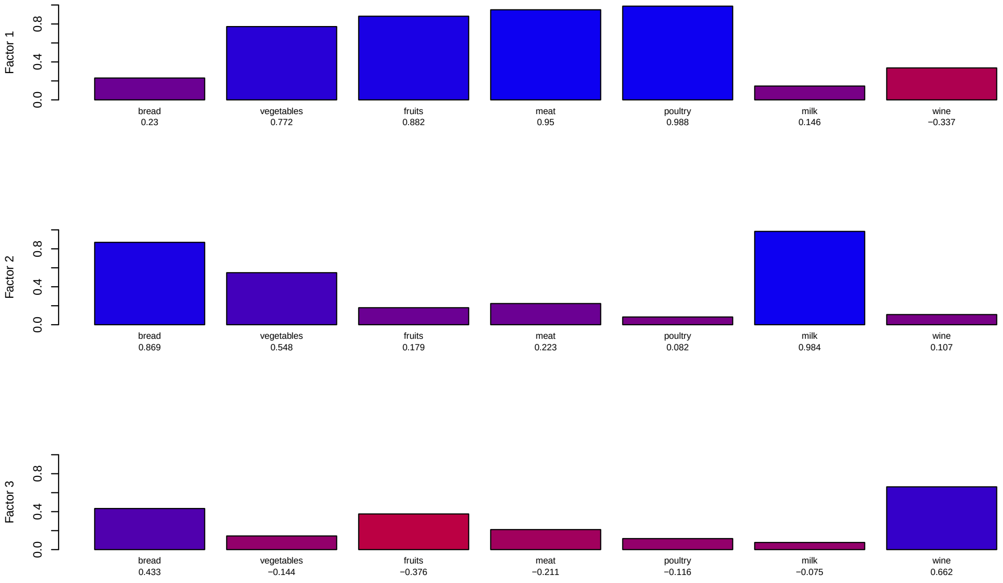

Figure 1: Visualization of the factor loadings corresponding to the 3 factors. We see that the third factor mainly describes bread and wine variables, but none of them very strongly, and for that reason we could omit it in this study.

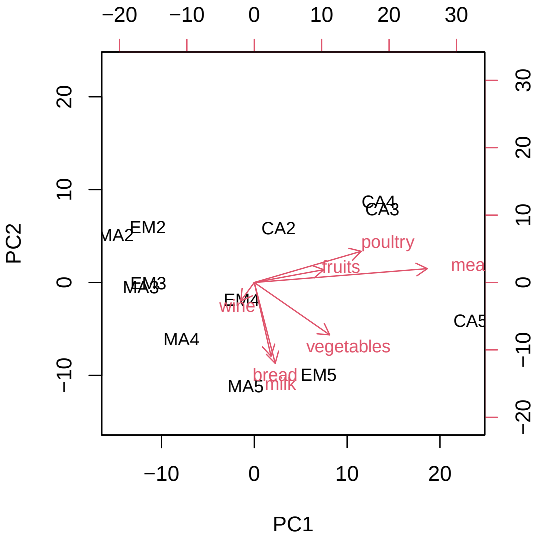

Figure 2: Projection of data on the first 2 principal components, which seem to describe 96 % of the total data variance.

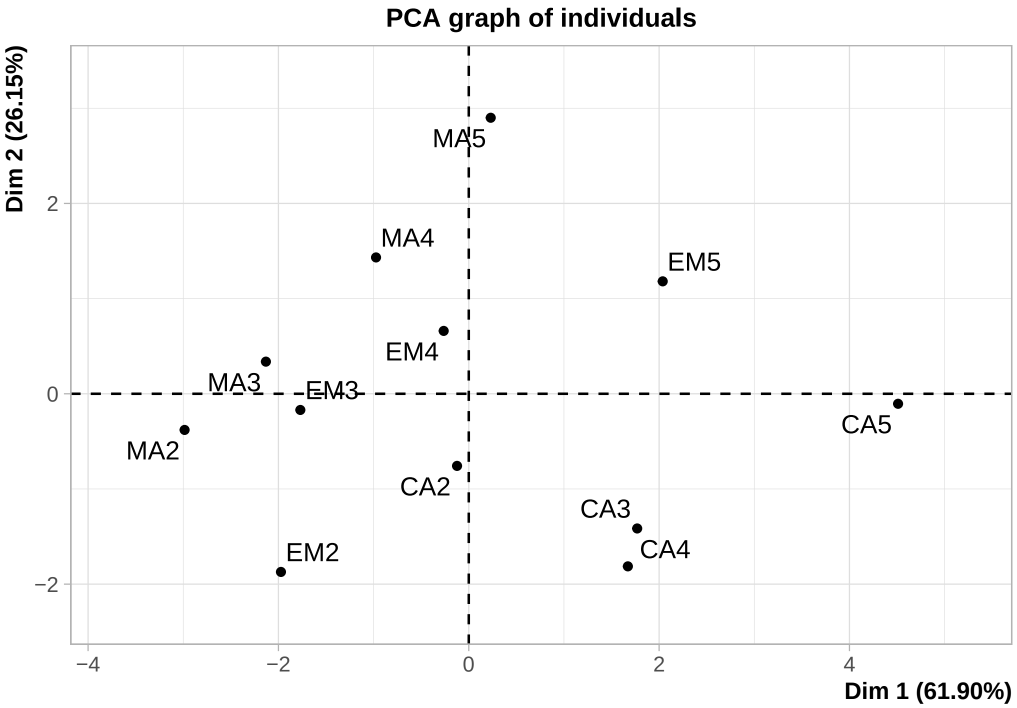

Figure 3: Individuals factor map in 2 dimensions.

3 Interpretation

Even though both methods are fairly similar, we obtained slightly different results. In both cases we discovered, that for future regression analysis, we might use the variable meat and omit the variables fruit and poultry, as they seem to be well described by the meat variable. Both methods also found a close relationship between variables bread and milk, which was well described by the second PC and second factor. The variables wine and vegetable were described differently in both cases, and in both cases we found that more components may be needed. While the PC case resulted in description of 96 % of total variance, the FA case with two factors yielded only 82 % of total variance described, also with harder interpretation. If I were to use only one of the analyses on the data, I would stick with principal component analysis.