Assignment 4 Group 8

[redacted]

1 Introduction

Individuals are faced with a copious amount of choices concerning their health. Examples of these are frequency of consuming alcohol or frequency of doing sports. However, do one's choices within a certain aspect of life imply that one makes certain other lifestyle choices? In this paper, we examine the relationship between lifestyle choices, using data concerning the weekly frequency of consuming several foods, beverages or cigarettes and frequency of working out. Our research question can be formulated as: "What are the relationships between lifestyle choices?"

In order to answer this question, we apply the method of nonlinear principal component analysis (NLPCA). The paper is constructed as follows. In Sect. 2, the data are introduced and the method is proposed in Sect. 3. The results are discussed and interpreted in Sect. 4 and Sect. 5 concludes.

2 Data

Before going into our method, we elaborate on the data. Our dataset consists of 390 participants in a survey who are asked to specify the number of times per week they smoke cigarettes, do sports and consume each of 7 kinds of foods and beverages: red meat, white meat, vegetables, fruits, soft drinks, crisps, and alcohol. We find that for one participant in the survey, all responses are 0, which is most likely a bogus response. Therefore, we leave out this participant, leaving 389 observations of 9 variables. The data are all integer-valued and should be treated as categorical rather than numerical. In NLPCA this is done by transforming the category values to category quantifications. However, preliminary analysis on the data using transformation plots shows outstanding quantifications for certain categories. These quantifications do not seem to have meaningful interpretations and could be due to the fact that the number of categories for some of the variables is large, and many categories only have a very small number of observations. This may cause instability in the solution, or may influence the quantification process (Linting & van der Kooij, 2012). Therefore, we merge two or more adjacent categories in the data according to the following criteria:

1. The number of categories per variable is at least 2 at most 9.
2. Each category contains at least 8 observations, with the categories in a tail forming a new category containing at least 2.5% of the total observations.

These criteria are a combination of those suggested by Linting, Meulman, Groenen, & Van der Kooij (2007) and Markus (1994). The first criteria ensures that the number of categories is not too large relative to the number of observations. The second criteria ensures that individual categories are reasonably filled, and based on the fact that observations in the tails of a probability distribution are often considered extreme. Note that we merge categories in such a way that we expect the categories to be similarly related with other variables, in order to minimize the loss of information. Therefore, a necessary exception is made on the second criterion for two particular categories that appear to contain possible outliers. This is further described in Appendix A.

3 Method

The main idea behind NLPCA is to extend PCA in a way that non-numeric measurement levels and nonlinear relations between variables can be incorporated. Category values are transformed to category quantifications while specifying variables as e.g. nominal or ordinal, with for ordinal the order of the categories retained. This quantification process is also known as optimal scaling, where optimal implies maximized variance accounted for (VAF) in the transformed variables for the first p components. We take several steps iteratively until a statistically optimal solution is obtained, using the `homals` package in `R`. We start with the least restricted feasible analysis levels, which is nominal for each variable in our data. Before going into details about the iterations of the optimal scaling process, we describe the technical details of NLPCA.

Let the data be contained in the $n \times m$ matrix $\mathbf{H}$, with n the number of observations and m the number of variables. Suppose that variable j consists of k categories after merging, with the observations contained in $\mathbf{h}_j$, the j^{th} column of $\mathbf{H}$. We introduce the $n \times k$ indicator matrix $\mathbf{G}_j$ containing dummy variables for each category of variable j. Furthermore, we consider the vector $\mathbf{y}_j$ of length k, which consists of the optimally scaled values of the categories of $\mathbf{h}_j$. Then the vector $\mathbf{q}_j$ of length n containing the optimally scaled values for variable $\mathbf{h}_j$ can be obtained by:

$$\mathbf{q}_j = \mathbf{G}_j \mathbf{y}_j. \tag{1}$$

The $n \times m$ matrix $\mathbf{Q}$ contains the optimally scaled values of all variables. The dimensionality of the data is reduced in such a way that each observation is modeled with p component scores, where $p \le m$. The $n \times p$ matrix $\mathbf{X}$ contains all of these scores. The component scores relate to the original data such that the scores multiplied by the component loadings approximate the data as well as possible, with the loadings contained in the $m \times p$ matrix $\mathbf{A}$. We use the following loss function to find the optimal $\mathbf{A}$, $\mathbf{X}$ and $\mathbf{Q}$ (Linting, Meulman, Groenen, & van der Koojj, 2007):

$$L(\mathbf{Q}, \mathbf{A}, \mathbf{X}) = n^{-1} \sum_{j=1}^{m} \mathrm{tr}(\mathbf{q}_j \mathbf{a}_j - \mathbf{X})'(\mathbf{q}_j \mathbf{a}_j - \mathbf{X}), \tag{2}$$

with $\mathbf{a}_j$ row j of $\mathbf{A}$. This loss function is subject to a number of restrictions. First of all, $\mathbf{q}_j \in \Omega_j$, with Ω_j the set of admissible transformation of variable j. $\mathbf{X}'\mathbf{X} = n\mathbf{I}$, to ensure that the principal components are uncorrelated and to avoid trivial solutions $\mathbf{X} = \mathbf{0}$ and $\mathbf{A} = \mathbf{0}$. Furthermore, $\mathbf{1}'\mathbf{X} = \mathbf{0}$ to ensure that the object scores are centered, with $\mathbf{1}$ an $m \times 1$ vector of ones. Lastly, because $\mathbf{q}_j \mathbf{a}_j$ is not unique, $\mathbf{q}_j' \mathbf{q}_j = n$ is imposed.

Different from linear PCA, the solutions are usually not nested for different dimensionalities p. Therefore, p should be selected for analysis. In NLPCA, the sum of the eigenvalues equals the VAF, or total fit, of the transformed variables. This is used as the main indicator for the selection of the amount of components in the solution. While the maximum fit equals m, the increase in fit with each additional component can be displayed in a scree plot. The selection of p is then based on elbows in the plot along with interpretability as a criterion.

Next, we obtain the p-component NLPCA solution on our data analyzed at a nominal level. First, we check for outliers in the solution using an object plot. Outliers obtain component scores that are relatively far from other scores in the component space, specifically an absolute score of larger than 3.5 (Linting & van der Kooij, 2012). In case of outliers, we repeat the analysis without them and check for outliers again. We then reevaluate the number of components and if appropriate, we choose a new p and repeat the analysis. We then examine transformation plots to check whether the choice of a nominal analysis level is appropriate for each variable. For variables that do not show non-monotonic relations with other variables, a more restricted ordinal analysis level can be used, or even a numerical analysis level in case of no nonlinear relations. If there is only a small difference in the VAF, a more restricted analysis level is preferred, in order to obtain more stable results and simpler interpretability (Linting & van der Kooij, 2012). The steps of checking for outliers, choosing the number of components and examining the analysis levels are repeated until the final analysis options are determined. Then, we further examine the relations between the quantified variables through a loadings table and a biplot.

4 Results

Before interpreting the NLPCA solution, we must choose the appropriate number of dimensions. Therefore, we examine the adjusted scree plot in Fig. 1. Elbows occur at the first and fourth dimensions, and we see that beyond the fourth dimension the increase in fit by further dimensions is less than 1. We therefore conclude that $p = 3$ components is the appropriate choice for our NLPCA solution. Note that the increase in fit is higher for the ninth component. We may attribute this anomaly to numerical instability, as the `homals` package indicates an increase in the loss function while iterating. Taking a short look at an object plot using the current analysis options, we can see that there are no outliers in the data and we can continue using $p = 3$.

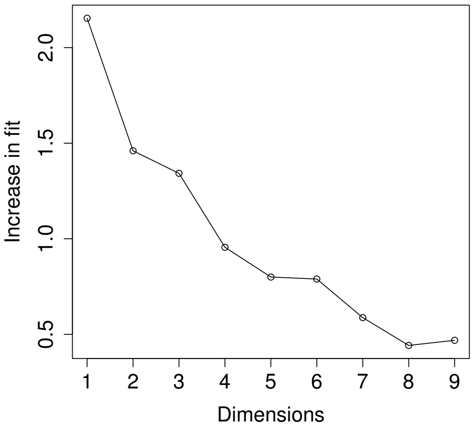

Figure 1: Adjusted scree plot

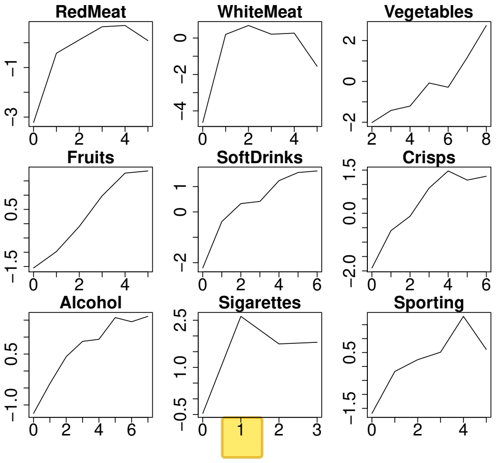

Figure 2: Nominal transformations of variables

In Fig. 2, the nominal transformations of the variables for $p = 3$ are shown. SoftDrinks and Fruits have strictly monotonically increasing quantifications and could just as well be analyzed ordinally. All other variables do not have strictly monotonic increases in the quantifications. However, for Alcohol, Vegetables and Crisps, we only see slight decreases in the quantifications that do not seem meaningful. Therefore, without substantially decreasing the total VAF, an ordinal treatment of these variables may simplify interpretation. The quantifications of RedMeat, WhiteMeat, Sigarettes and Sporting are clearly non-monotonic. However, whereas RedMeat and WhiteMeat show an approximately reversed U-shape, Sigarettes and Sporting suggest very particular patterns. Specifically, for cigarettes, participants who smoke 2 and 3 times a week are alike in terms of scoring pattern on other variables, and differ in the same way from those in the categories 0 and 1. Similarly, for Sporting, participants who do sports 3 and 5 times a week are alike in terms of scoring pattern on other variables, whereas those who do sports 4 times a week are very different. Therefore, an ordinal transformation for these two variables may also be more appropriate.

Table 1 shows the decrease in VAF as we restrict the analysis level of the indicated variables from nominal to ordinal. We do this in order of closeness to monotonic increases and interpretability of non-monotonic quantifications. Throughout the iterations, no outliers are detected and $p = 3$ remains appropriate. In line with our expectations, up to iteration 5 the VAF does not substantially decrease. While in iteration 6 the VAF decreases by 0.10%, in iteration 9 we see biggest decrease of 1.02%. As the interpretability of the analysis level is very important, we choose to draw the line at iteration 8. That is, we choose to only retain the nominal analysis level of RedMeat and WhiteMeat. With all other variables restricted to ordinal, this results in a marginal total decrease in VAF of 0.22%. Thus, with these final analysis options and $p = 3$, the three selected components explain 54.91% of the variance in the quantified lifestyle variables (22.14%, 16.78% and 15.99% for component 1, 2 and 3, respectively), indicating reasonable fit.

Table 2 shows the component loadings of each of the variables. We observe that for the first component, the variables on unhealthy consumptions have a high negative loading, whereas the variables on healthy consumptions Vegetables and Fruits have positive loadings. This indicates that component 1 distinguishes healthy and unhealthy choices. In the second component, Sporting has the highest absolute loading. Note that this variable has a very small loading on component 1. Furthermore, RedMeat and WhiteMeat are mainly contained in component 3. The biplot in Fig. 3 shows the component loadings in the space of the first two principal components as vectors along with the objects scores as dots. In line with Table 2, the loading vectors corresponding to WhiteMeat and RedMeat are relatively short in length, as they are less contained in component 1 and 2. We can also see that the first principal component indeed corresponds to healthy vs. unhealthy lifestyle choices. Clearly, the loadings of sporting are mostly on the second component, implying that this component captures the choice to exercise. This is confirmed by inspection of the label plot in Fig. 4. Here, individuals with a high object score for the second component rarely exercise, whereas people with a low object score do sports often.

Moreover, we can distinguish between four groups in the biplot: snacks (crisps, soft drinks) in the second quadrant; meats and drugs (red and white meat, cigarettes, alcohol) in the third quadrant; sporting on the border of the third and fourth quadrant and fruits and vegetables in the

Table 1: VAF for restricted analyses

Iter.	Nominal to ordinal	VAF
1	None	55.07%
2	Fruits, SoftDrinks	55.07%
3	Alcohol	55.07%
4	Vegetables	55.07%
5	Crisps	55.06%
6	Sporting	54.96%
7	Sigarettes	54.91%
8	RedMeat	54.85%
9	WhiteMeat	53.83%

Table 2: Loadings of first three components

	1	2	3
RedMeat	-0.3321	-0.3346	-0.6660
WhiteMeat	-0.3344	-0.3436	-0.6566
Vegetables	0.3753	-0.5134	0.3670
Fruits	0.3656	-0.5419	0.2686
SoftDrinks	-0.6865	0.1529	0.2829
Crisps	-0.6640	0.2492	0.3587
Alcohol	-0.6086	-0.4825	0.1555
Sigarettes	-0.4607	-0.2586	0.3133
Sporting	-0.0331	-0.5812	0.1638

fourth quadrant. A high correlation between fruits, vegetables and sporting indicates that people who eat healthy foods tend to work out as well. Individuals in the fourth quadrant generally have the healthiest lifestyle, choosing to exercise while avoiding snacks, alcohol and cigarettes, as can be seen by their projections on the variables. Following the same line of reasoning, individuals in the second quadrant prefer less exercise, fruits and vegetables while consuming more snacks. The consumption of vegetables and fruits seems uncorrelated to the consumption of meats, with an orthogonal position of the corresponding loadings. In the same vein, there does not seem to be a relation between cigarette and alcohol consumption on one hand and fruit and vegetable consumption on the other hand, due to the orthogonality of these loading vectors. Consumption of snacks correlates negatively with consumption of fruits and vegetables. Lastly, all component loadings project positively on Alcohol, indicating that a positive correlation exists. Noticeably, the projection of Sporting on Alcohol is high, indicating higher alcohol consumption amongst athletes.

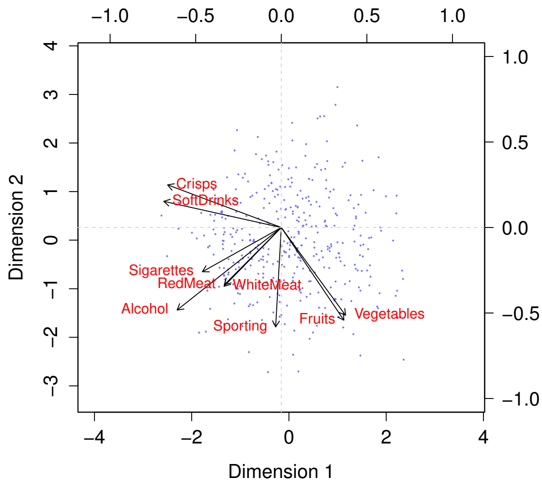

Figure 3: Biplot of the first two components

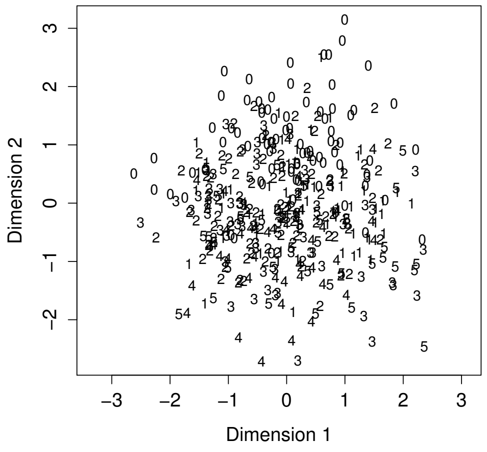

Figure 4: Label plot of the Sports variable

5 Conclusion

In this paper, we examined the relations between lifestyle choices by applying nonlinear principal component analysis on lifestyle data of 390 individuals, for which we merged categories to ensure stability in the solution. An adjusted scree plot shows that the data could best be summarized in three dimensions and we determined the appropriate analysis levels by looking at transformation plots. In our final analysis, about 55% of the variance in the quantified variables is accounted for. The first component captured the healthiness of one's consumptions, whereas the second and third mainly captured the frequency of exercise and meat consumption, respectively. A biplot in the space of the first two components showed several interesting relations between lifestyle choices. There does not appear to be correlation between the consumption of cigarettes, alcohol and meat on one hand and fruits and vegetables on the other. The healthy choices of doing sports and consuming fruits and vegetables are positively correlated, as well as consuming crisps and soft drinks and consuming alcohol and cigarettes. Lastly, the healthy options fruits and vegetables are negatively correlated with crisps and soft drinks.

References

Linting, M., Meulman, J. J., Groenen, P. J., & Van der Kooij, A. J. (2007). Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological methods*, *12*(3), 359.

Linting, M., Meulman, J. J., Groenen, P. J., & van der Koojj, A. J. (2007). Nonlinear principal components analysis: introduction and application. *Psychological methods*, *12*(3), 336.

Linting, M., & van der Kooij, A. (2012). Nonlinear principal components analysis with catpca: a tutorial. *Journal of personality assessment*, *94*(1), 12–25.

Markus, M. T. (1994). Bootstrap confidence regions for homogeneity analysis; the influence of rotation on coverage percentages. In *Compstat* (pp. 337–342).

Appendix

A Merging categories

Fig. 1 and 2 show the histograms of each variable before and after merging categories, respectively. Each bin in the histogram corresponds to a single category. We can see that after merging, only two categories that appear to contain potential outliers have less than 8 observations. That is, category 8 for Vegetables and category 0 for Fruits.

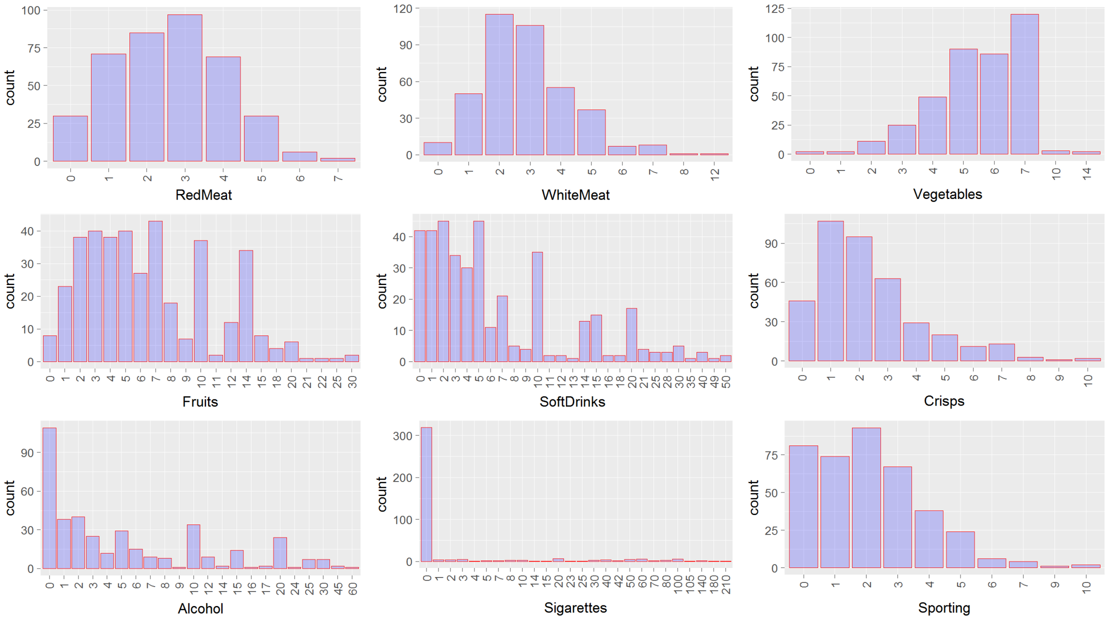

Figure 1: Distribution of the variables before merging categories

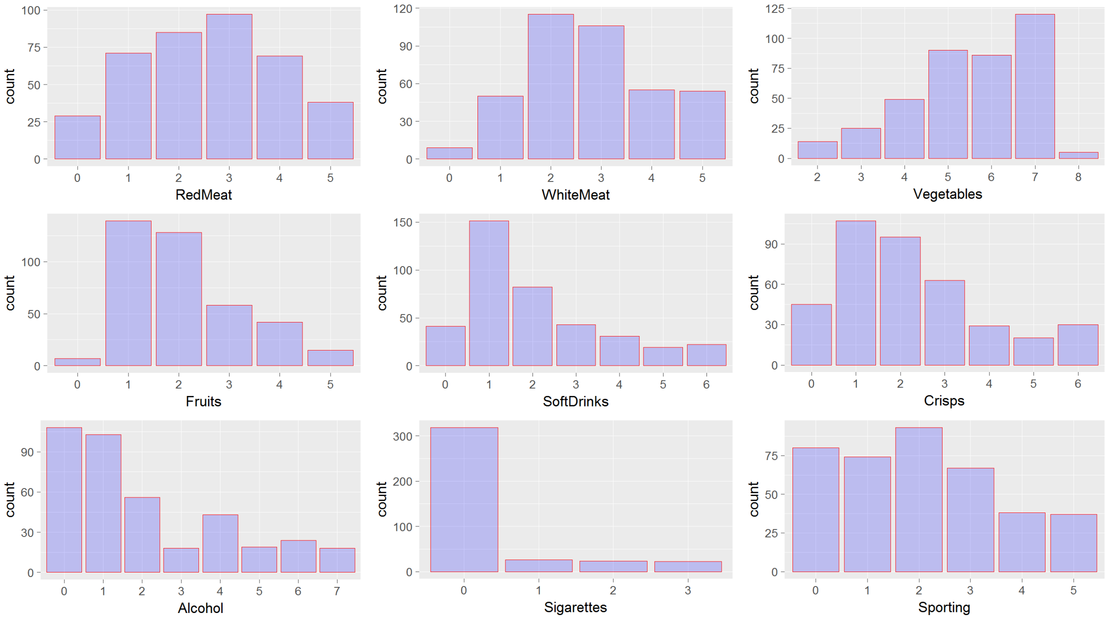

Figure 2: Distribution of the variables after merging categories