Regularization of category quantifications in NLPCA

Flora Katsiki (s1298658)

Thesis advisors: Dr. Anita van der Kooij and
Prof. Dr. Jacqueline J. Meulman

MASTER THESIS

Defended on September 28th, 2017

Specialization: Statistical Science

Universiteit Leiden

Statistical Science

Leiden University Medical Center • Mathematical Institute • Institute of Psychology
Wageningen University • VU Medical Center

STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Chapters and Sections

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER

1

INTRODUCTION

In many fields data sets are usually considered as a collection of variables with numerical measurement level. In the Social and Behavioural Sciences, however, variables are often categorical, thus of nominal or ordinal measurement level. Categorical data are not suitable to be analyzed with methods that require data to be numerical measurements, such as standard Principal Components Analysis (PCA) or Regression. A categorical variable does not consist of numeric values, but of numbers that in the best case are ordered (ordinal measurement level), and in the worst case merely labels (nominal measurement level). An example of the first is education level or Likert scale variables and of the latter blood type or gender. So, the "values" of categorical variables are not real values but just category indicators and thus statistics like mean and variance are meaningless and an alternative to standard methods is called for. The analysis method that will be discussed in this thesis is the technique of Optimal Scaling in standard analysis methods. Optimal Scaling converts category indicators to numeric values, which is achieved by nonlinearly transforming the observed variables. A consequence of nonlinear transformation of variables is that nonlinear relations among the observed variables are taking into account in the analysis. Note that this makes the Optimal Scaling methods suitable for data of numeric measurement level as well, since linearity of relations is an assumption of standard methods.

The particular nonlinear analysis method that is the focus of this thesis is nonlinear PCA (NLPCA) as implemented in the CATPCA program in the SPSS Categories module (Meulman, Heiser, & SPSS Inc., 2015).

1.1 Linear and Nonlinear Principal Components Analysis

1.1.1 PCA

Linear principal components analysis (PCA) is a statistical procedure where the original variables (possibly correlated) are converted to a new set of lin-

early independent variables, which are called Principal Components. Principal components result from the linear combination of the original variables using weights that are called loadings. Graphically stated, the objective of PCA is to convert the high dimensional data space, (the number of dimensions being the number of variables), to a lower dimensional space, with new axes representing the principal components while accounting for as much variance as possible. Therefore, PCA performs dimensionality reduction, summarizing the original data set in a few principal components.

Cases (also called objects or subjects) are represented as points in the component space and the coordinates are called component scores. Variables are represented as vectors in the components space with endpoints given by the loadings. Loadings are the correlations of the original variables with the principal components. Therefore, the squared loading is the variance accounted for (VAF) of a variable by a component, and summing over the components gives the squared length of the vector, called communality, thus the loadings vector represents the total variance of a variable accounted for by all components. As mentioned above, PCA assumes linear relations among variables. So, even in case all data are of numerical measurement level, when nonlinear relations are considered a possibility or when one does not know but wants to check the linearity assumption, NLPCA is called for. Practically PCA is an exploratory technique that tries to depict in an informative and concise way the relations among variables and between variables and cases.

1.1.2 Nonlinear PCA

The objective of NLPCA is the same as for PCA but with respect to transformed variables. Thus summarize the transformed variables in a lower dimensional space while accounting for as much of the variance of the transformed variables as possible. Thus, NLPCA generalizes PCA by incorporating variable transformations. In NLPCA an optimal scaling transformation is performed where each category for each observed variable obtains an optimally scaled numeric value (called category quantification).

1.1.2.1 Optimal scaling levels

The properties of a transformation depends on the optimal scaling level (also called analysis level). This level is the choice of the researcher and may, however does not need to, correspond to the measurement level. CATPCA offers a variety of analysis levels, differing in the type of information contained in the original variable that is retained in the transformed variable. A variable of nominal measurement level only contains grouping information; of ordinal measurement level grouping and ordering information; of numeric measurement level grouping, ordering, and interval information. By choosing the numerical analysis level, all information will be retained in the transformed variable, with the ordinal analysis level interval information will be lost in the transformation, and with nominal analysis level only the grouping information is retained. Note that with numerical analysis level for all variables, NLPCA is equal to standard PCA.

1.1.2.2 NLPCA algorithm

In NLPCA, as stated above, the objective is to reduce the transformed variables to a lower dimensional space where as much variance in the transformed data is accounted for by the components. For this reason, estimation of the transformations (also called optimal quantifications) and the model parameters (loadings and component scores) is performed simultaneously. This procedure is performed via an ALS (Alternating Least Squares) algorithm, which is an iterative process that alternates in each iteration between estimation of the model parameters and estimation of the transformation (category quantifications). The iterative scheme is stopped once the algorithm has converged to a stationary point where the estimates of quantifications and the model estimates are not changing anymore.

1.1.2.3 Centroid and Vector quantifications

Categories can be quantified according to either the vector model or the centroid model, depending on the desired way of representation of variables.
The scaling levels described above result in quantifications according to the

PCA model, which is a vector model: variables are represented in the component space as lines going through the origin. The variable categories, or in standard PCA the variable values, are on this line (ordered and equally spaced for standard PCA). These scaling levels are called single, because they result in one-dimensional quantification (and a loading for each dimemsion). In contrast, variables quantified according to the centroid model are not represented as lines but as points that are allowed to scatter throughout the component space. This is the Correspondence Analysis model. The optimal scaling level for this kind of transformation is called multiple nominal level, since the resulting quantification is multi-dimensional (and no loadings). In this thesis, the multiple nominal scaling level will not be used, so further details are not given here, and in the remainder "nominal" will mean "single nominal".

1.2 Visualizing results

1.2.1 Representation of variables

1.2.1.1 Transformation plots

Optimal scaling transformations are optimal in the sense that the relationships of variables to other variables is optimally expressed. The transformation of a variable (quantification) can be visualized in transformation plots. Such a plot shows the transformation curve, that is, the form/shape of the relation between the original and the transformed variable, and this form/shape of the relationship of the original variable with other variables. The relationship of a variable with other variables can be fully expressed when the least restrictive optimal scaling level (nominal) is chosen. In this case the transformation curve is allowed to be non-monotonically nonlinear, that is, to go up and down. For ordinal scaling level the curve is not allowed to go down, thus is restricted to be monotonically increasing. For the numerical scaling level the transformation is restricted to be linear, so the curve is restricted

to a straight line. The shape of the transformation curve for nominal and ordinal scaling level depends on the form of the relationships of a variable with the other variables. For variables analyzed at the ordinal scaling level the transformation plots will show a curve somewhere between a straight line and a monotonically increasing curve (the category quantifications (y-axis) are required not to decrease with increasing category numbers (x-axis), but interval information is dropped, thus equal spacing is not required). The exact shape of the curve depends upon the relation of the variable with the others. For variables analyzed at the nominal scaling level, the transformation curve is allowed to be anything between a straight line and a non monotonic curve, because in addition to interval information also ordering information is dropped.

1.2.1.2 Category points

Another way of displaying quantifications is in a category points plot. Category points are the quantified variables in the principal components space. In this type of representation variables are depicted as vectors where category points are lying on this vector and take as coordinates the category quantifications multiplied by the coresponding component loadings. Category points give useful information of the association of categories within and between variables.

1.2.2 Object Scores - Representation of cases

As in PCA components scores are the new variables that result from weighing and combining the analysis variables. Individuals/cases/objects are displayed as points in the new dimensional space. In NLPCA, object scores and variables (as vectors) can be represented in the same plot, referred to as biplots. From these type of plots relationships between variables and objects can be detected.

1.3 Research question

A problem that can occur in NLPCA is that categories with low frequency can have too much influence and cause solutions to be unstable. A category quantification for such a category can be many standard deviations from the mean (transformations are z-scores). Furthermore, these categories when dominant can result in altering radically the solution compared to the case when objects scoring in such categories are left out, or treating such categories as missing. This problem has been investigated by others in the past, such as Linting, (Linting et al., 2007), who suggested the remedy of merging categories with low frequencies with neighboring categories. Alternatively, in our research we will investigate the solution that when a quantification exceeds a preset value, it is regularized by restricting it to a maximum or minimum value. In order to asses if this regularization made results more stable, balanced nonparametric bootstrap will be used. The research question of this thesis is: **"will a suggested regularization of the quantifications makes the results of NLPCA more stable?"**. More detailed explanations about the problem and the suggested solution will be presented in chapter 2.

1.4 Outline

In this research it is investigated if NLPCA's stability is improved by regularization of the quantifications. In the next chapters the problem motivating this research will be presented and a possible solution is discussed. Finally, results and some discussions with remarks and suggestions for future research will be presented. Below an outline of topics covered in each chapter is given.

- **Chapter 2** to establish the stability of NLPCA results nonparametric bootstrap used. The chapter begins with definitions about stability and continues with the validity of nonparametric bootstrap to assess stability. Also, a method for visualization of bootstrap results (confi-

dence ellipses) is described.

- **Chapter 3** Data sets are simulated with and without low-frequency categories and analyzed with different scaling levels, illustrating the influence of such categories and how they affect the stability.
- **Chapter 4** focuses on the comparison of the results of NLPCA when category quantifications are regularized and when not.
- **Chapter 5** Technical considerations and recommendations for further research.

CHAPTER

2

STABILITY IN NONLINEAR PRINCIPAL COMPONENTS ANALYSIS

Principal Components Analysis (PCA) is used to investigate the structure of data sets consisting of linearly related numerical variables. Alternatively, a generalized version of PCA, nonlinear Principal Components Analysis (NLPCA) is used to investigate the structure of data sets that consist of non numerical or numerical variables which can be linearly or nonlinearly related.

Both in PCA and NLPCA questions about stability of their outcomes can arise. Under the assumption of multivariate normality, stability can readily be assessed analytically. However, both for PCA and NLPCA this assumption is often not realistic for data in the social sciences where techniques such as nonparametric bootstrap are needed to evaluate the stability of the results. In this chapter the focus lies on the generalized version of PCA, NLPCA and cases where NLPCA reproduces unstable results are presented. Specifically, cases where data sets contain categories with low marginal frequencies which result in instability. The concepts of stability (internal and external) will be explained below. The nonparametric balanced bootstrap will be explained and applied as well as stability measures such as confidence intervals and confidence ellipses.

2.1 Stability

In this thesis we are concerned with stability defined as the sensitivity of the analysis method to small unimportant data variations. The solution of NLPCA can be regarded as stable when small changes in the data result in small changes in the results.

Greenacre (1984) and Markus (1994) made a distinction between internal and external stability. Internal stability is when the analysis result can capture the sample's structure well enough. Outliers can be a source of internal instability. External stability is when the analysis result can be generalized to the population. If samples are large enough and representative for the population the results from equally sized samples are expected not to vary much.

Under the framework of external stability, sample values can be viewed as an estimate of the population value. This estimate is expected to differ between samples. In order to measure how much the population estimates variates across samples, confidence intervals with prespecified accuracy are calculated. These confidence can give valuable information both for internal and external stability. Confidence regions are computed by a method called nonparametric balanced bootstrap. In the following section details about nonparametric balanced bootstrap are presented.

2.2 Assessing Stability

2.2.1 Nonparametric balanced bootstrap

Nonparametric bootstrap

A common approach to assess stability is through relying on distributional assumptions, where confidence intervals can be derived analytically called as parametric methods. When distributional assumptions do not seem realistic a nonparametric approach is used where confidence intervals are not derived analytically and a resampling method is applied. For categorical data multivariate normality often does not hold and parametric approaches can seem unrealistic, for this reason approaches as nonparametric are often preferred.

In nonparametric bootstrap, B bootstrap samples are randomly drawn with replacement from the original $n \times m$ data with n the number of objects and m the number of variables. Each bootstrap sample has the same size as the original data set (parent sample), where some objects may be drawn multiple times and others not at all. For each bootstrap sample the analysis is performed which results in B values in total for each parameter. These B values are used to estimate the confidence regions.

Balanced bootstrap

In the bootstrap samples some individuals may occur many times and some

just a few times or not at all. Thus over all B samples individuals may not have occurred with the same frequency, thus the individuals may not be represented equally in the bootstrap procedure. Efron and Tibschirani (1993) found not much difference between the balanced and simple bootstrap. However, Markus (1994) found better performance of the balanced bootstrap for variables, although this improvement decreases as the sample size (n) increases. As our goal is to evaluate the stability of individuals as well (object scores) as variable parameters, equal representation of objects in the bootstraps is of importance, so we used the balanced bootstrap.

In order to perform balanced bootstrap a vector $\mathbf{k}$ is created of length $n * B$ with values from 1 to n, repeated B times. Thus, the values in $\mathbf{k}$ are row indices of the parent sample copied B times. Next, the $n * B$ values in $\mathbf{k}$ are permuted. Then the first bootstrap sample is created by picking from the $\mathbf{k}$ vector the rows indexed by the first B values of $\mathbf{k}$, the second sample uses the $B{+}1$ through $2 * B$ values of $\mathbf{k}$, etc. For example, to create 1000 bootstraps samples from a data set with 500 individuals, $\mathbf{k}$ will have a length of $500 * 1000 = 5000$, with the numbers from 1 to 500 repeated 1000 times and the 5000 values are randomly permuted. The values in $\mathbf{k}$ index specific rows in the parent sample and each row occurs 1000 times in $\mathbf{k}$. Using the first 1000 values of $\mathbf{k}$ mimics drawing a sample of size 1000 with replacement, and using values 1001 through 2000 mimics drawing a second sample, etc. Thus, using the row indices as constructed in vector $\mathbf{k}$ is a way of drawing with replacement row indices from the parent sample 1000 times, under the restriction that over the 1000 draws each row index occurs with the same frequency. So, In this way it is ensured that each object is represented equally in the bootstrap procedure.

2.2.1.1 Validity of bootstrap in NLPCA

Even though in the NLPCA literature bootstrap results have not been thoroughly evaluated for all possible combinations of analysis levels, the conclusions of previous research convinced us that the bootstrap can sufficiently

establish NLPCA stability. For example, Markus (1994) assessed the validity of bootstrap results to establish stability in Multiple Correspondence Analysis (MCA) by conducting a Monte Carlo study. MCA is identical to NLPCA when all variables are analyzed as multiple nominal. She also investigated if the bootstrap provides valid results about stability for NLPCA outcomes where variables were treated as ordinal. Markus' concluded: firstly, coverage percentages[1] of bootstrap confidence regions are more satisfactory when 90% or 95% confidence regions are computed and not pushed in more extreme such as 99%. Secondly, bootstrap provides sufficient results about external and internal stability when the number of bootstrap samples are at least 1000. Thirdly, n should be no less than 200 to provide acceptable range of coverage percentages. The fact that Markus came to the conclusion that bootstrapping is valid in assessing stability of NLPCA for multiple nominal or ordinal variables and that many researchers have used bootstrap in assessing stability for nonlinear multivariate data methods (Gifi 1990; Greenacre, 1984; Heiser and Meulman, 1983; Vand der Burg and De Leeuw, 1988; Linting, Meulman, Groenen, and van der Kooij, 2007; Linting 2007), we regard as compelling evidence that bootstrap gives valid results when used under some conditions.

2.2.1.2 Rotation

PCA solutions can be freely rotated, without their overall fit being changed (Cliff, 1966; Jolliffe, 2002). There are several methods of rotations differing in their goal. Procrustes rotation is used when one wants to rotate a solution orthogonally to resemble as much as possible another solution. Varimax rotates orthogonally the solution making each variable to load as highly as possible only in one component, which simplifies interpretation. A restricted form of rotation is reflection. Reflection is caused from sign arbitrariness of eigenvectors, i.e. positive/negative signs of loadings in a component can be changed to negative/positive signs (reflected) without affecting the solution. When using the bootstrap with PCA, rotational indeterminacy has to be

[1] "the proportion of times that the population value lay within the estimated confidence region"(Linting, Meulman, Groenen and van der Kooij (2007))

taken into account, otherwise artificial differences between solutions might occur. For instance, when in the extreme case of a bootstrap sample solution that is exactly equal to the parent sample solution, the signs are opposite in each component, a big difference seems to appear, while there is actually none. So, solutions for bootstrap samples not to be rotated to be as close as possible to the parent solution. With Varimax and Procrustes rotation the total fit is redistributed over the components, so, strictly speaking, they are not principal components anymore. Therefore we chose to use reflection as rotation method, leaving the bootstrap components in PC-orientation, only taking care of the signs. The next chapters when referring to bootstrap solutions, we always referring to solutions that are reflected when necessary to resemble as much as possible the parent solution.

2.2.1.3 Confidence intervals and ellipses

From the bootstrap distributions confidence intervals, meazns, and standard deviations are calculated for each model parameter (eigenvalues, loadings, quantifications, component scores). The bootstrap distribution indicates how stable a model parameter is: the higher the standard deviation and the wider the confidence interval the more unstable. In NLPCA, parameters for the variables are loadings and quantifications. We expect the stability of these two kinds of parameters to be related, thus when loadings are stable then also the quantifications are stable, and vice versa.

Except for quantifications, the model parameters are multidimensional. Confidence intervals can be evaluated separately per component, but this has the disadvantage that information about the components simultaneously is disregarded. Such joint information is given by confidence ellipses (in the two dimensional case, or ellipsoids for higher dimensions. In this thesis we use two dimensions throughout, so we will only discuss ellipses). The construction of confidence ellipses was developed by Meulman and Heiser (1983) and used and described in Linting (2007). Confidence ellipses have some attractive features. They give a graphical representation of joint information, which is

easily perceived and interpreted. Further the method of constructing confidence ellipses is nonparametric, can be applied fast and implemented easily. A confidence ellipses can be understood as follows: The bootstrap results for a parameter, for example the eigenvalues, can be depicted in a plot, with the coordinates of the B points being the B two-dimensional vectors of eigenvalues. Those plots show a cloud of points, called the bootstrap cloud. A 95% confidence ellipse is the ellipse that best covers the 95% of the points that are closest to the centroid of the cloud (for example, see Figures 3.6 and 4.1). In addition to visual inspection of ellipses, or for comparison purposes, two joint-dimensional diagnostic measures can be used. The first one is the area of the ellipse, a large area indicating instability. However, the reserve is not necessarily true; a small area not always indicates stability because a small area can be the result of a (very) long major ellipse axis and a (very) small minor axis. Therefore, additional diagnostics like CI lengths or standard deviations are used as well.

CHAPTER

3

INFLUENCE OF LOW FREQUENCY CATEGORIES ON NLPCA RESULTS AND STABILITY

In this chapter we illustrate the influence of low frequency categories on NLPCA results and stability by using simulated data. We generated a data set with a predefined structure and altered a copy of these data slightly to obtain a second data set with unique (frequency of 1) categories. Moreover, we compared the output of NLPCA for different analysis options, specifically for nominal and numerical, to see the difference between the least and most restrictive analysis.

3.1 Data simulation

3.1.1 Procedure

A simulation algorithm was used in order to create data with prespecified principal components structure, details about the algorithm steps can be found in Appendix B. The simulation was set up to create a data set consisting of 200 cases and 15 variables with moderately strong two principal component structure, scree plot can be found in Appendix C, with a total VAF 60%. The variables consisted of 7 uniformly distributed categories. This data will be referred as data A. Data B was created from data A, with the only difference being that for one case on all variables the category numbers were set to 8 to introduce unique categories (frequencies of one). Data A and B are used in order to be compared and present a case where an outlying category, in our case one, can cause instability to quantifications, eigenvalues and person scores.

3.2 NLPCA results for data sets A and B

3.2.1 Variance Accounted For and eigenvalues

We expect that the influence of unique categories is stronger when variables are analyzed at the least restrictive level, which is nominal, and weaker with the most restrictive, numerical level. So, we performed analyses with both levels. To check whether the unique categories in data B affect the solution

in the intended way, we compare the eigenvalues and total VAF for data A and B. Tables 3.1 and 3.2 show the results.

In the last two rows of table 3.1 can be seen that the eigenvalues for data A with nominal and numerical scaling levels respectively are 5.57 and 5.38 for the first component, indicating that 37.13% and 35.88% of the variance in the transformed variables is accounted by the first principal component. In the second principal component 25.41% and 25.09% of variance is accounted for. For data B, refer to the last row of table 3.2, the eigenvalues in the first principal component for nominal and numerical scaling levels are 97.77% and 36.52%, respectively. In the second principal component 2.23% and 25.06% for nominal and numerical, respectively. Thus for the nominal scaling level the effect of the unique categories is that the solution becomes 1-dimensional while having little effect with numerical scaling level. This extreme change shows that NLPCA results are severely affected by the unique categories introduced in data set B when the nominal analysis level is used.

Moreover, from those two aforementioned tables we can see that VAF for data B with nominal scaling level is affected for each variable. Comparing the results we see that for data A numeric and nominal, and data B numeric the variable VAF's are very much the same, while for data B nominal all VAF's are perfect in the first dimension.

3.2.2 Transformations

To see the effect of the unique categories on the quantifications we can compare transformation plots for A and B. Figure 3.1 shows the optimal quantifications for variable 5, for data A and B in transformation plots (because the transformation plots for all variables are much alike, we display the plot for only one variable). In such plots the relationship between the optimally scaled variable (y-axis) with the original categories (x-axis) is depicted. It can be seen that in the numerical case the transformations for data A and B do not differ a lot. In contrast, when variables were treated nominally the

Table 3.1: VAF and eigenvalues for data A

	Nominal			Numerical		
	dim1	dim2	Total	dim1	dim2	Total
variable 1	0.77	0.00	0.77	0.76	0.00	0.76
variable 2	0.68	0.00	0.68	0.66	0.00	0.66
variable 3	0.73	0.00	0.73	0.71	0.00	0.71
variable 4	0.80	0.00	0.80	0.79	0.00	0.79
variable 5	0.39	0.00	0.39	0.37	0.00	0.37
variable 6	0.52	0.00	0.52	0.45	0.00	0.45
variable 7	0.52	0.00	0.52	0.52	0.00	0.52
variable 8	0.39	0.00	0.39	0.37	0.00	0.37
variable 9	0.42	0.00	0.42	0.38	0.00	0.39
variable 10	0.36	0.00	0.36	0.36	0.00	0.36
variable 11	0.00	0.89	0.89	0.00	0.89	0.89
variable 12	0.00	0.77	0.77	0.00	0.76	0.76
variable 13	0.00	0.68	0.68	0.00	0.66	0.66
variable 14	0.00	0.64	0.64	0.00	0.64	0.64
variable 15	0.00	0.82	0.83	0.00	0.82	0.82
Total VAF (=eigenvalue)	5.57	3.81	9.38	5.38	3.76	9.15
% VAF	37.13	25.41	62.54	35.88	25.09	60.97

Table 3.2: VAF and eigenvalues for data B

	Nominal			Numerical		
	dim1	dim2	Total	dim1	dim2	Total
variable 1	1.00	0.00	1.00	0.76	0.01	0.77
variable 2	1.00	0.00	1.00	0.66	0.00	0.67
variable 3	1.00	0.00	1.00	0.71	0.01	0.72
variable 4	1.00	0.00	1.00	0.79	0.01	0.79
variable 5	1.00	0.00	1.00	0.39	0.00	0.39
variable 6	1.00	0.00	1.00	0.46	0.00	0.46
variable 7	1.00	0.00	1.00	0.52	0.00	0.53
variable 8	1.00	0.00	1.00	0.38	0.00	0.38
variable 9	1.00	0.00	1.00	0.39	0.01	0.40
variable 10	1.00	0.00	1.00	0.37	0.00	0.37
variable 11	0.68	0.32	1.00	0.01	0.87	0.89
variable 12	1.00	0.00	1.00	0.01	0.75	0.76
variable 13	1.00	0.00	1.00	0.01	0.66	0.67
variable 14	1.00	0.00	1.00	0.01	0.63	0.64
variable 15	1.00	0.00	1.00	0.01	0.81	0.82
Total VAF (=eigenvalue)	14.67	0.33	15.00	5.48	3.76	9.24
% VAF	97.77	2.23	100.00	36.52	25.06	61.58

difference in the quantifications between data A and B is much greater. Not only the shape of the transformations is greatly different, but also the quantified value of category 8 takes a value of 14, which is extreme considering that category quantifications are standardized values. This value not only is a very serious outlier, but also influences the rest of the values, causing them to be all about the same, close to zero, while for the other cases the quantifications for category 1-7 range between -2 and 2. Thus we see that the effect of a unique category on the transformation is dichotomization; heavily emphasizing the difference between category 8 and all other categories.

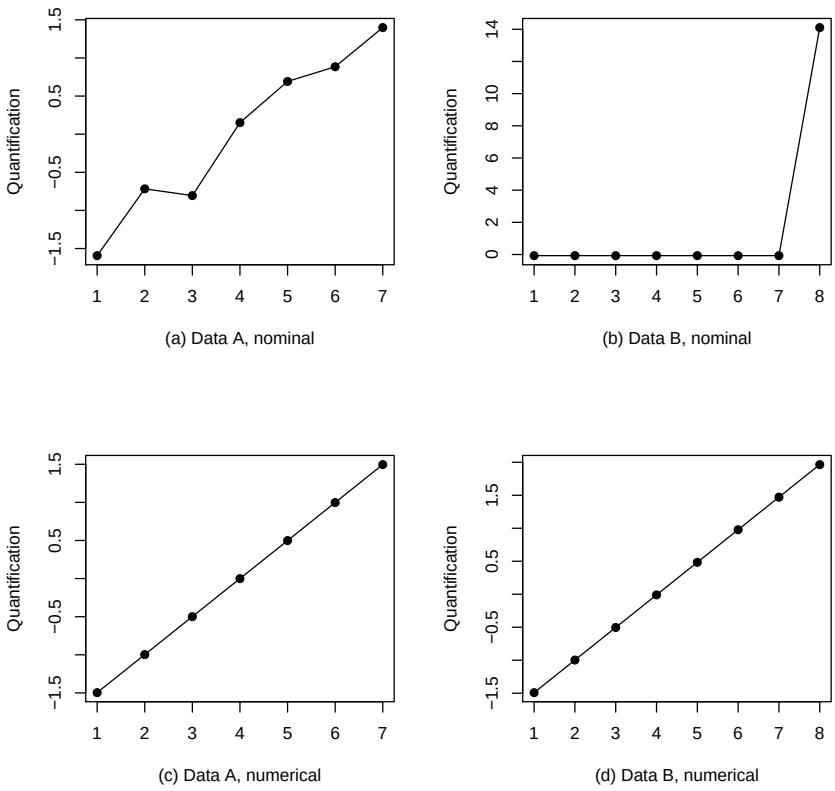


Figure 3.1: Transformations variable 5 for data A and Data B. On the x-axis, the category numbers are displayed. On the y-axis are the optimal quantifications (category quantifications)

3.2.3 Category quantifications

Another way of displaying transformations is category points plot. Our results show two groups of variables: variables 1 to 10, loading mainly in the first dimension, and variables 11 to 15, loading mainly in the second. We give the category plot for one variable from each group; variable 5 in figure 3.2, and variable 12 in figure 3.3. From these plots we see for both variables, 5 and 12, that sub-figures (a), (d) and (e) are pretty much alike while for the data B nominal case in sub-figure (b) the plots looks very different, note the different axis ranges in (b). For better comparison to the plots for the other cases, the category plot for data B nominal with the same axis ranges as for the other cases is displayed in sub-figure (c). For variable 5 (figure 3.2) we see that in (a), (d), (e) the categories spread out along the first dimension, while in (b) and (c) we see essentially 2 groups of points: the vector points of categories 1-7 all almost in the origin and category 8 far away. For variable 12 in figure 3.3 we see the same difference, there in (a), (d), (e) the categories spread out along the second dimension, and the vector points in (b) and (c) look the same as for variable 5. There is a difference to be seen however for the centroid points, in (c): for variable 12 categories 1-7 spread as much in the second dimension as in the other plots, while they do hardly spread along any dimension for variable 5. This is caused by the dominance of the first dimension in the data B nominal case. To synopsize, also for category points results are substantially different when variables are treated nominally in the case of outlying categories.

3.2.4 Component loadings

To check the effect of unique categories on the loadings we can compare the component loadings plots for data A and B, see figure 3.4. A loading is the Pearson correlation between transformed variables and a principal component. Thus, the squared loading is the VAF of a variable in a component. Summing the squared loadings over the components gives the total VAF of a variable. Thus if we plot the loadings of a variable as a vector in space from the origin to the point with coordinates the loadings, the length of this vec-

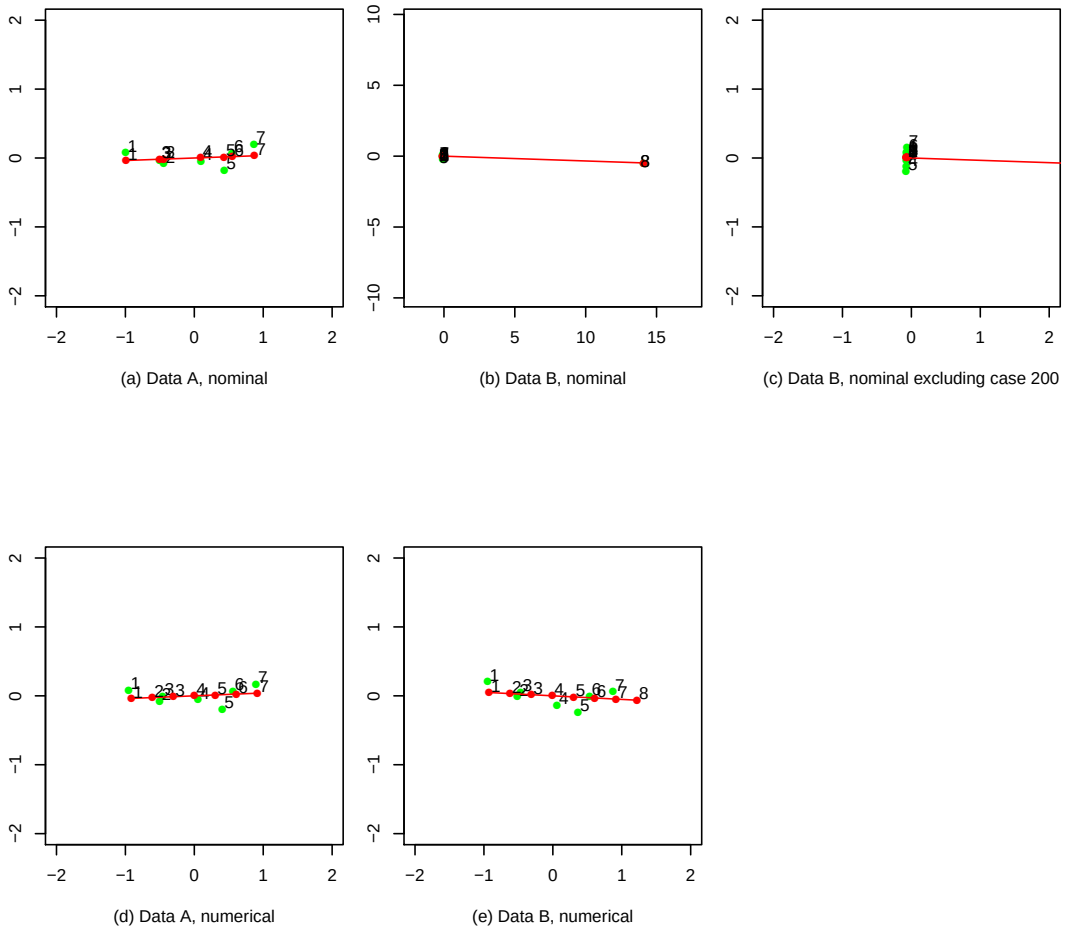


Figure 3.2: Category points of variable 5 for data A and B where variables are treated either as numerical or as nominal. Centroid Coordinates are presented with green and Vector Coordinates with red

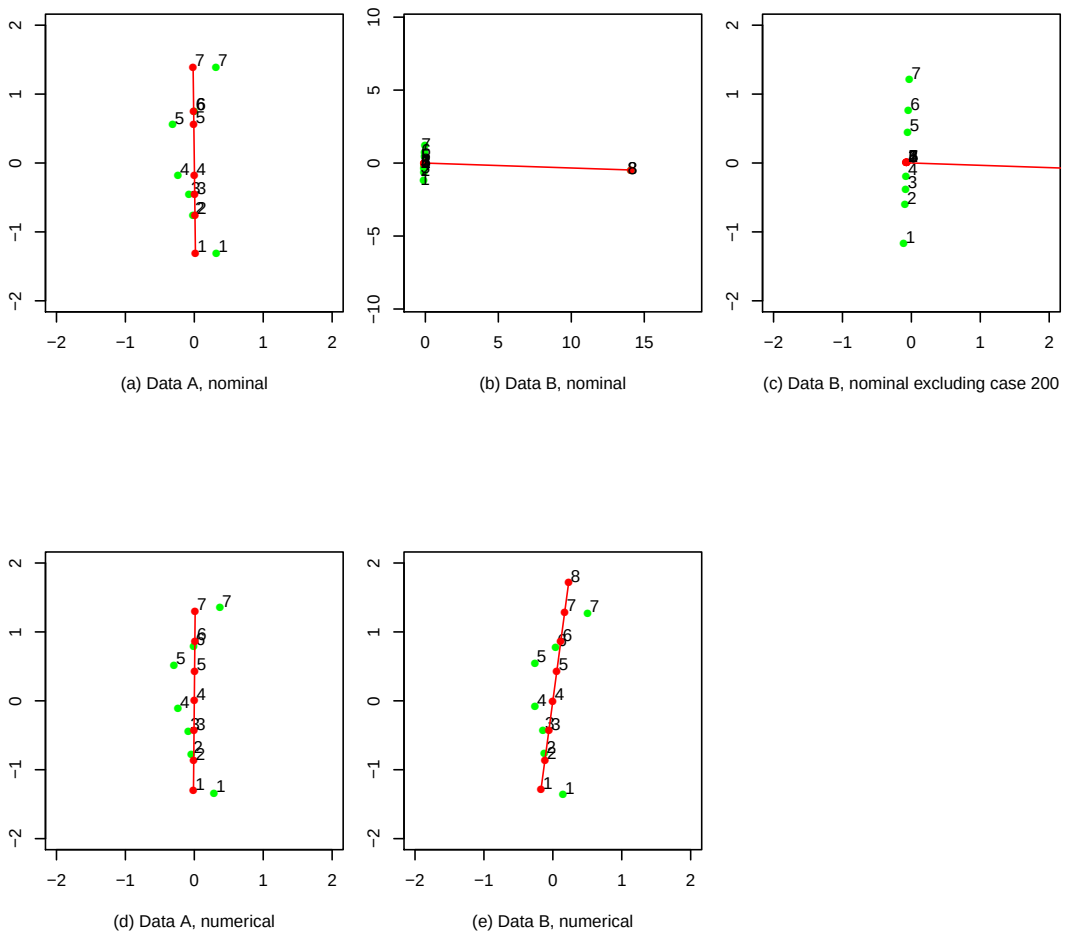


Figure 3.3: Category points of variable 12 for data A and B where variables are treated either as numerical or as nominal. Centroid Coordinates are presented with green and Vector Coordinates with red

tor indicates the total VAF. So, the longer the loading vector, the higher the VAF. Also, for variables with fairly long vectors, the angles between vectors indicate the correlations among the (transformed) variables, with 90° indicating zero correlation, and 0° maximum correlation. We see similar structures in sub-figures (a), (c) and (d): two distinct groups of variables with strong association within the groups and no association between the groups (the angle between the two groups is about 90°). Sub-figure (b) shows a different structure with all variables except var11 loading in the first component only. From the component loadings as well it is apparent that the results differ when variables are treated as nominal in the case of unique categories in the data.

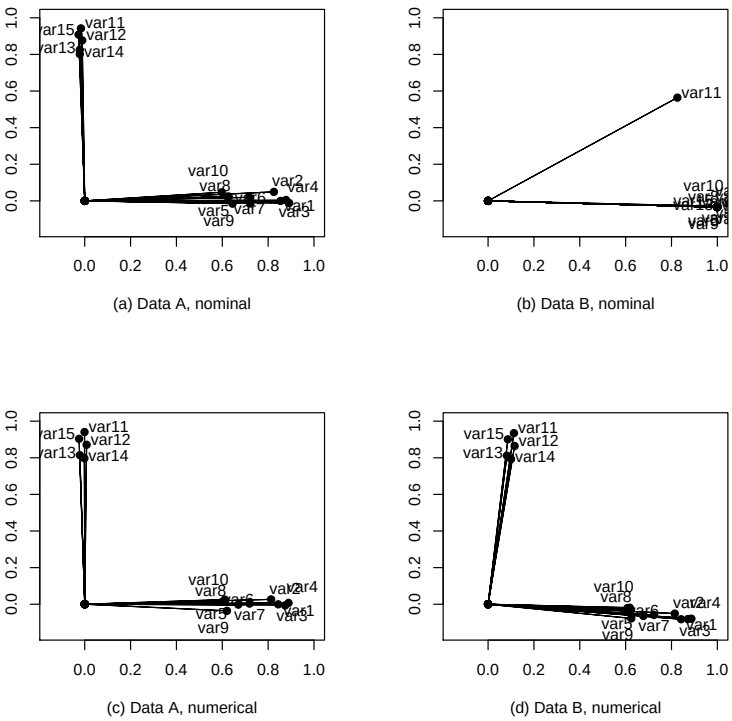


Figure 3.4: Loadings for data A and Data B. On the x-axis, the component loadings for all variables on the first component are displayed. On the y-axis, the component loadings for all variables on the second principal componet are displayed

3.2.5 Component scores

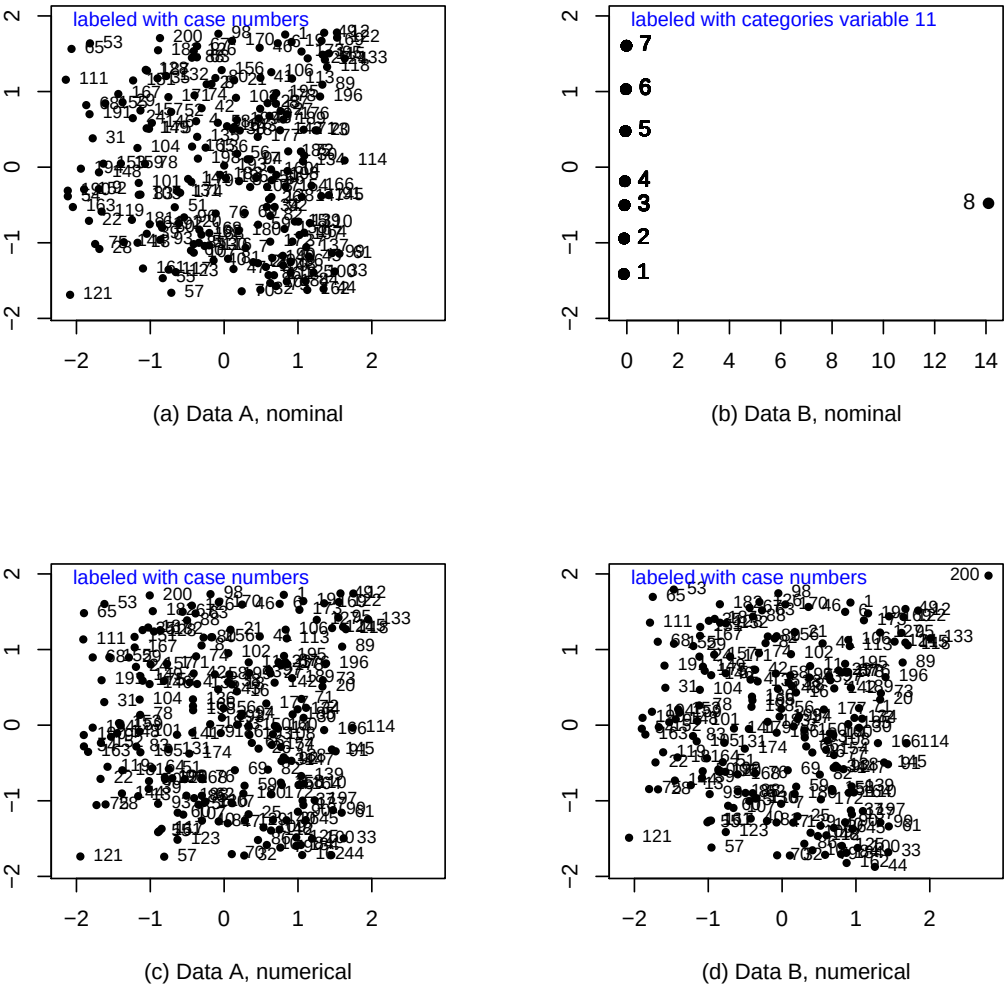
The influence of the unique categories can also be seen by looking at the component scores. Figure 3.5 (b) shows that the component scores for data set B with nominal analysis levels are dominated by the difference between the one object with unique categories and all the others, while the discrimination among all the other objects is in the second component. When variables are treated as numerical there is not much difference between data A and B, although in (d) case 200 (the case with category 8) lies somewhat away from the cloud, indicating that is somewhat different from the other cases, but not as far as in (b). Our results are expected, because numerical is too restrictive to allow the solution to focus solely on the difference between one case and all others. On the other hand when variables are analyzed at the nominal scaling level, or ordinal (also checked), the results are highly affected, since these levels are liberal enough to reveal that object 200 is extremely different from all the other objects.

3.3 Stability of NLPCA results for data sets A and B

To demonstrate the effect of low-frequencey category on the stability for each data set (A and B) and for each type of analysis levels (nominal or numerical) 1000 nonparametric balanced bootstrap were performed.

3.3.1 Stability of Variance Accounted For and eigenvalues

Tables 3.3 and 3.4 display univariate stability measures (confidence interval and standard deviation) as well as the joint-dimensional measure of ellipses area for the eigenvalues and VAF from the two-dimensional NLPCA for the example data set A and B in the nominal and numerical cases respectively. The confidence ellipses can be seen at figures 3.6. The confidence ellipses gives an insightful multidimensional representation of the stability of the



(a) Data A, nominal

(b) Data B, nominal

(c) Data A, numerical

(d) Data B, numerical

Figure 3.5: Object scores for data A and B where variables are treated either as numerical or as nominal

eigenvalues.

From the last line in tables 3.3 and 3.4 and from figure 3.6 can be seen that the ellipse areas are fairly small for data A and B, indicating a quite stable solution. Exception to this is data B when nominally treated. In the figures, exception is sub-figure (b), the major and minor ellipse axes are almost aligned with the x and y axes, indicating that the total eigenvalue varies, as well as the distribution of the eigenvalue over the dimensions. Also, the major ellipse axis is about twice the length of the minor axis, indicating that the second dimension is more stable than the first. The same conclusions can be drawn form the diagnostics in table 3.3 and 3.4. In the nominally treated variables case for data B it is clear from looking at the ellipse (note the large axes ranges) that in the first component the bootstrap eigenvalues vary very much more than the second. The confidence interval length in the first is 9.22 and in the second 3.66 and standard deviation in the first is 4.25 and in the second 1.67. Moreover, those values are around ten times higher compared to the three other cases where the confidence intervals for the first components are ranging from 0.95 to 1.08 and the second from 0.49 to 0.52 and standard deviations 0.24 to 0.25 and 0.13 for the first and second component respectivelly. In terms of VAF we notice that VAFs in nominally treated variables for data B are fluctuating 61.5% and 24.39% in the first and second component respectively, when in the same time VAFs are varying at maximum 7.21% in the rest cases. Same discrepancies in magnitude can be seen by looking the ellipse areas as well, where in data B when nominally treated variables ellipse area is 8.78 and for all the rest is around 0.60.

Furthermore, when plotting the bootstrap points (not shown), we see that in (a), (c), and (d) the points lie scattered everywhere in the ellipse, while in (b) they lie in two clusters, one at about the point (5,5) and the other at about (15,0); the first cluster representing bootstrap samples not containing the object 200, and the other cluster samples including object 200. So, the much greater instability for data B nominal is due to the outlying object 200; when the object 200 is not in a bootstrap sample, the solution is

much more like the other cases, while when included, the solution becomes 1-dimensional. Finally, the distance between the original eigenvalues point (red and the bootstrap mean (green point) indicates the bias. For the numerical cases, there is almost no bias, for data A nominal we see a little bias, while for data B nominal the bias is considerable.

Table 3.3: Stability diagnostics for eigenvalue and, in parenthesis translated to %VAF for two dimensional linear PCA and NLPCA for data A.

		Nominal	Numerical
1st Component	CI lengths (VAF)	0.95 (6.35)	0.96 (6.41)
	Standard Deviation	0.24	0.25
2nd Component	CI lengths (VAF)	0.49 (3.25)	0.49 (3.24)
	Standard Deviation	0.13	0.13
	Ellipse Area	0.61	0.60

Table 3.4: Stability diagnostics for eigenvalue and, in parenthesis translated to %VAF for two dimensional linear PCA and NLPCA for data B.

		Nominal	Numerical
1st Component	CI lengths (VAF)	9.22 (61.50)	1.08 (7.21)
	Standard Deviation	4.25	0.28
2nd Component	CI lengths (VAF)	3.66 (24.39)	0.52 (3.46)
	Standard Deviation	1.67	0.13
	Ellipse Area	8.78	0.68

3.3.2 Stability of transformations

In figure 3.7 are shown the transformation plots with the 95% confidence intervals for variable 5. The intervals can be interpreted as follows. The closer

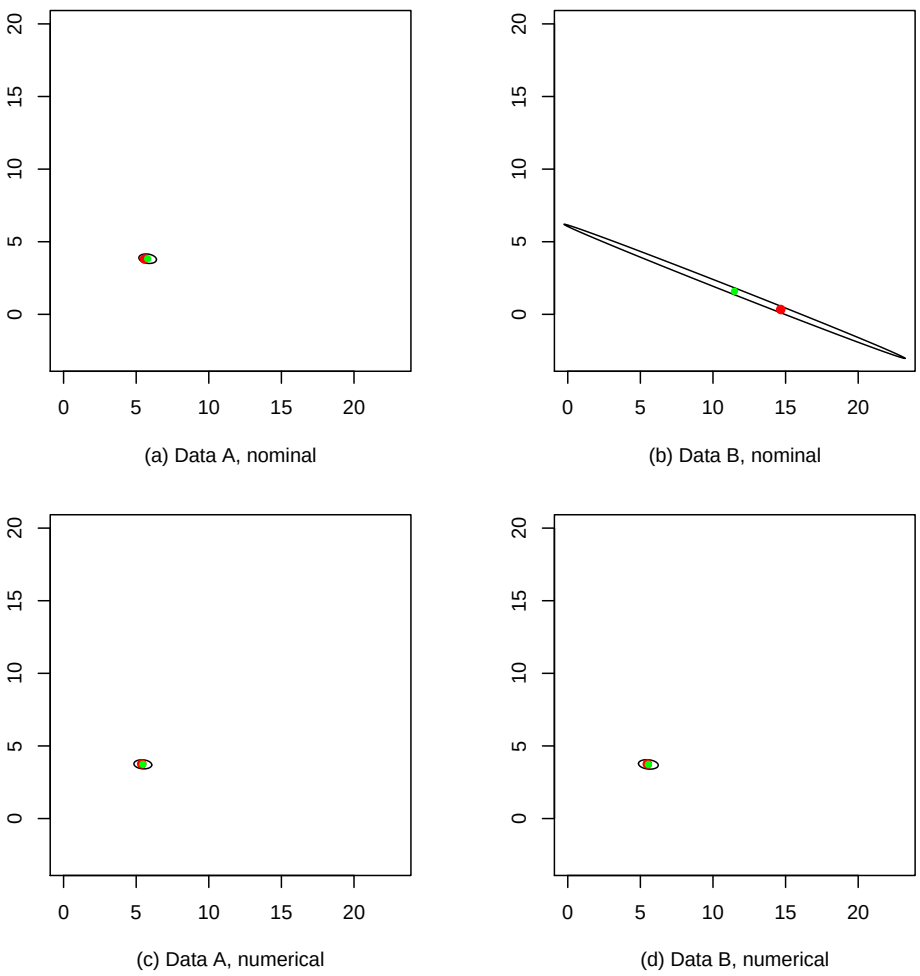


Figure 3.6: Eigenvalue confidence ellipses plot for data A and B where variables are treated either as numerical or as nominal. With red are displayed the eigenvalues in the original data set and with green the bootstrap means of the eigenvalues

the lines of the confidence intervals limits to the parent sample (red line) the more stable the quantifications. These figures show that the transformation when data are treated as numerical are considerably more stable compared to the nominal case (a and b compared to c and d), since in the nominal cases the confidence intervals are wider. In sub-figure (b) it is notable that the categories are more unstable than in (c), and quantification of category 8 is much more unstable than the other categories. Similar results were produced for the rest of the variables.

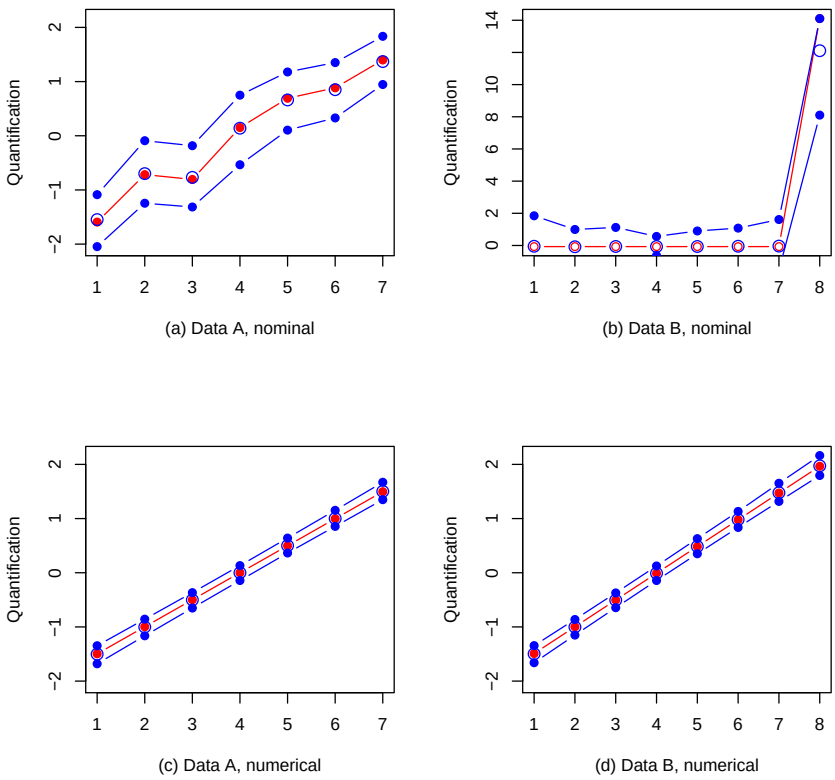


Figure 3.7: Confidence intervals for variable 5 transformations (nominally or ordinally treated) for data A and B. The red line indicates the original variable transformation from the observed data. The outer lines indicate the 95% confidence intervals. The x-axis depicts the original category numbers and the y-axis the category quantifications

3.3.3 Stability of category quantifications

The influence of unique categories on stability of category points can be seen by observing the confidence ellipses in the categories plot. As in section 3.2.3, we display plots for only variables 5 (figure 3.8) and 12 (figure 3.9), since the results for the other variables are much the same. For the numerical analysis (sub-figures (d) and (e)), variable 5, loading highly in the first dimension, the ellipses are longer in the second dimension than in the first, while the reverse is true for variable 12, loading highly in the second dimension. Thus, categories quantifications are more stable in the dimension the variable belongs to, as they are expected to be when they firmly belong to the dimension. The ellipses for data A nominal are somewhat bigger due to less flatness. This reflects the fact that due to the higher freedom of nominal quantification, the dimensions are less firmly established. For the data B nominal case we see a huge, extremely flat ellipse spreading along the first dimension for category 8 in sub-figure (b), and in sub-figure (c) (plot on the same scale as for the other cases) we see that the ellipses for categories 1-7 are also much bigger. Thus, the quantification of the outlying category is extremely unstable in the dimension it dominates, causing the other category to be very unstable as well, in the dimension the variable belongs to, thus indicating much uncertainty about which dimension the variable belongs to. These results show again that category quantifications are more unstable when there are categories with low frequencies that have a high impact on the analysis due to treating the variables nominally.

3.3.4 Stability of component loadings

In figure 3.10 the confidence ellipses for the component loadings for data A and B are shown. We can notice that all loadings ellipses for data B with nominally treated variables are bigger and thus are more unstable compared to the other cases. It can be seen from the figures that all cases showed similar ellipses, except (b), where in sub-figure (b) ellipses were wider and showed different shapes and orientations. More specifically, for (a), (c) and (d) all ellipses are small, so all loadings are quite stable, and as explained

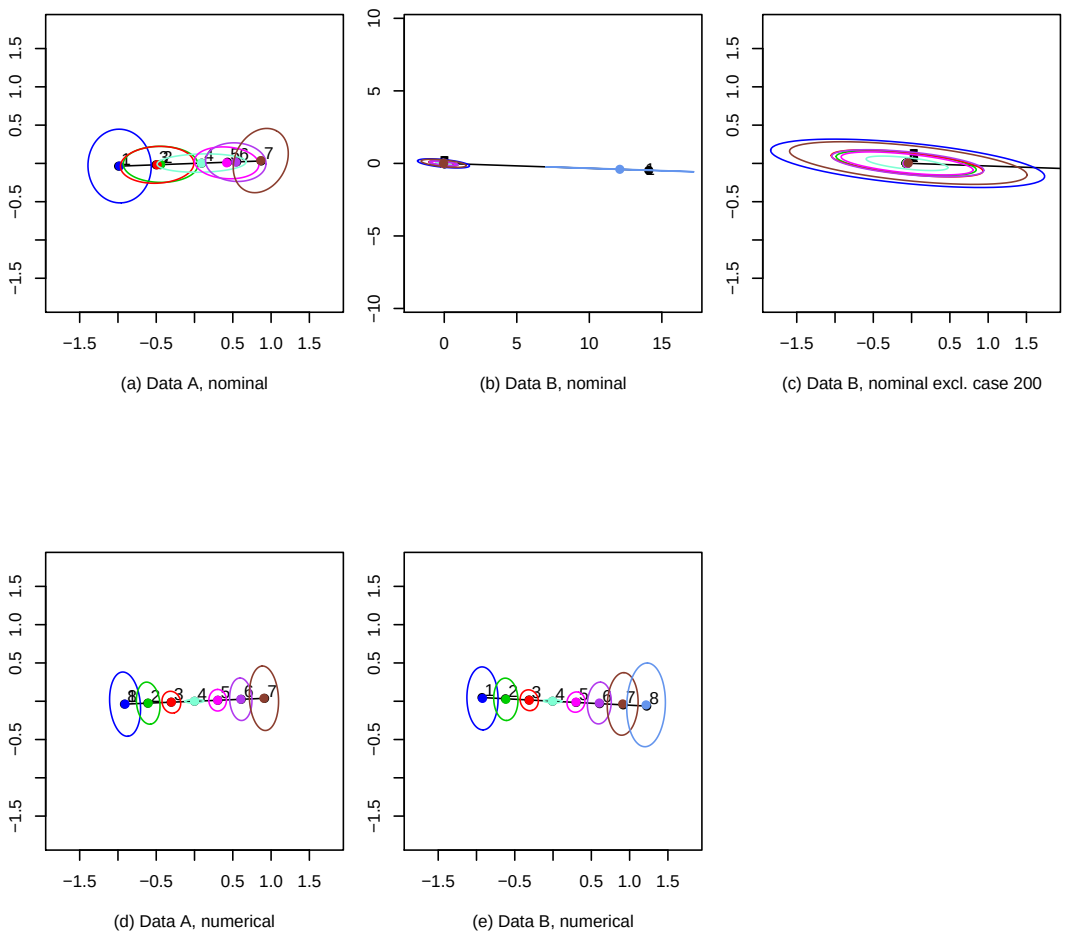


Figure 3.8: Confidence ellipses for category points of variable 5 for data A and B where variables are treated either as numerical or as nominal

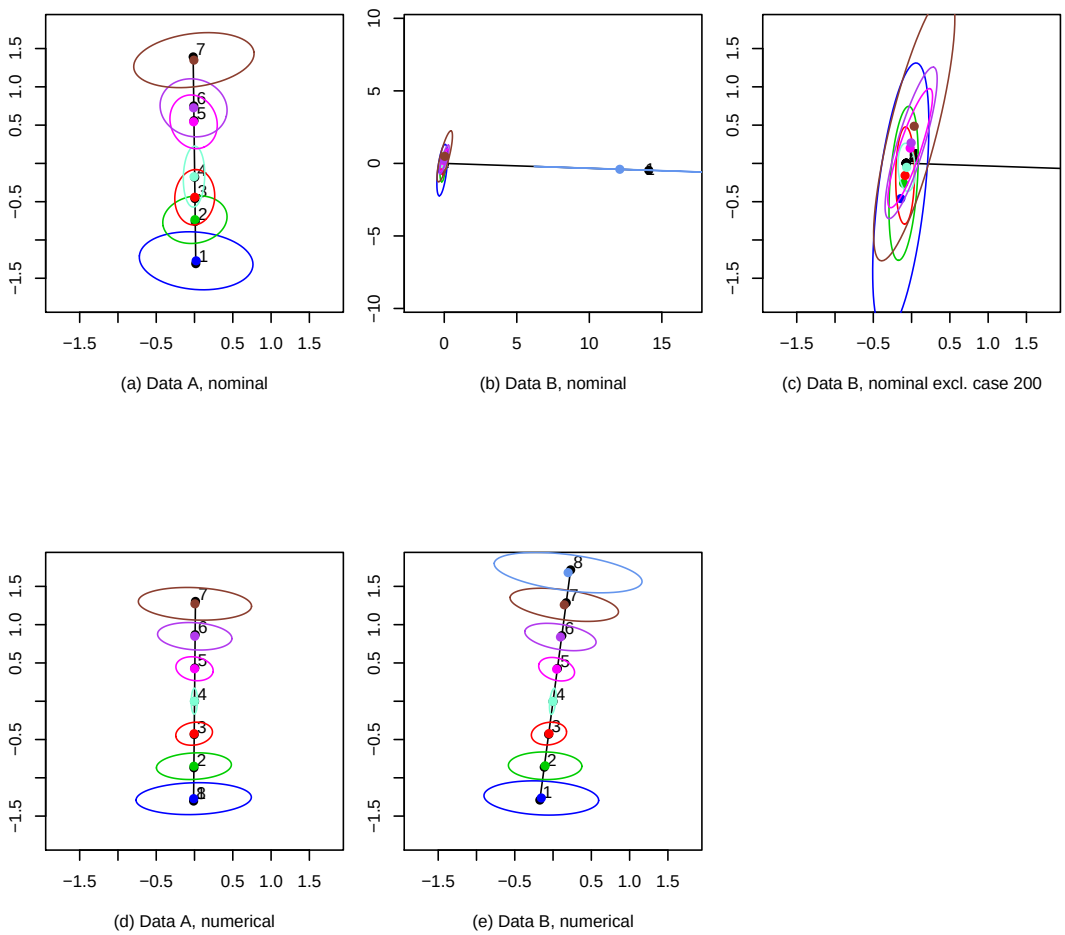


Figure 3.9: Confidence ellipses for category points of variable 12 for data A and B where variables are treated either as numerical or as nominal

before, the slight variation mostly appears in the dimension a variable does not belong to. Although, in sub-figure (b) we can notice that the loadings of the variables belonging to the first dimension are quite stable, while they are very unstable for the variables in the second dimension. This is due to the two 'kinds' of bootstrap samples: in/or excluding case 200, when included variables 11-15 end up in the first dimension, when excluded in the second dimension. The results showed that component loadings are more sensitive to low marginal frequencies when variables are treated nominally.

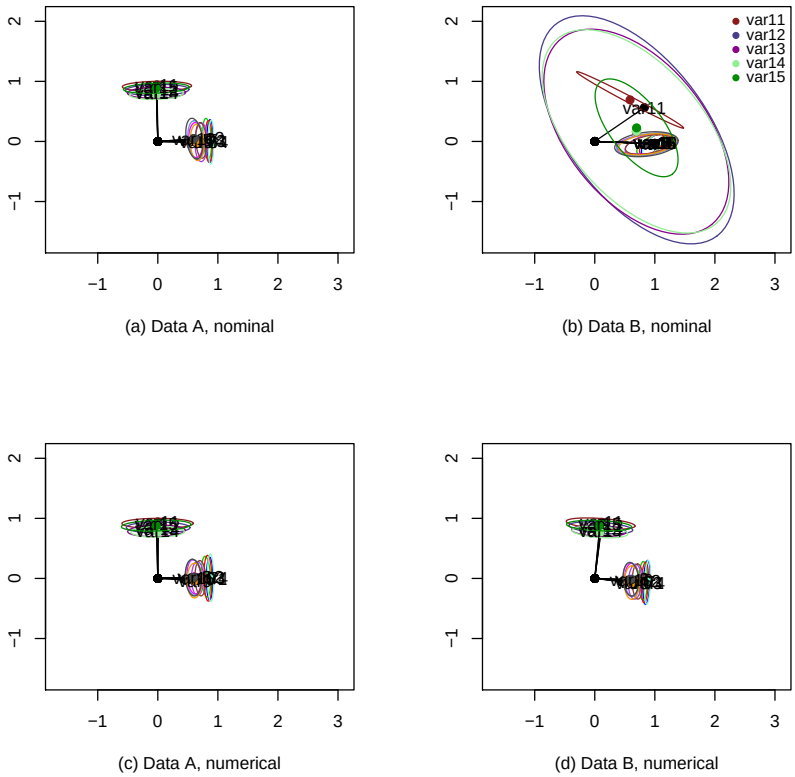


Figure 3.10: 95% confidence ellipses for the component loadings for data A and B when variables are treated either as numerical (c and d) or as nominal (a and b) respectively. Black circles indicate the loadings for the original component loadings and the colored indicate the centroids of the bootstrap clouds.

3.3.5 Confidence intervals and ellipses for component scores

In figure 3.11 we can see the ellipses for data A and B (nominally and numerically treated). Sub-figures (a), (d) and (e) look much the same. Note that although case 200 in (e) lies somewhat at the fringe, the size of its ellipse is not different from the others, indicating that despite the fact that case 200 is somewhat different than the other cases, this difference does not cause its component score to be unstable. Figure 3.11 (b) illustrates that ellipse areas is not always useful indicator of instability and that visual inspection is always necessary or taking into consideration other indices, like confidence interval lengths or standard deviations. For the purpose of comparing the overall stability of two or more results for component scores, when N is too large to inspect plots, it is helpful to sum stability indices over the objects.

In order to compare the stability of component scores for data A and B, we calculated the total sum of the confidence ellipse areas over the cases and compared those values. In table 3.5 the sums of ellipse areas and the sums of the lengths of confidence intervals are given. From this table it can be seen, that the sums of ellipse areas were ranging from 92.05 to 145.99 for data A and B (except when the variables were treated as nominal) and for data B, nominally treated, it is taking a value of 455.14. Noticeable, from these results, is that when variables were treated nominally for data B the sum of ellipse areas was 3 times more compared to all other three cases, indicating much higher instability for this case. Since the ellipse area for the extremely outlying object 200 is not large, we know that the much larger sum of areas in case B nominal is not due to this object. So, we can conclude that the stability of many or all object scores is affected by the outlying object.

To compare stability of component scores within a solution we looked at indices for object 200 and the maximum of these values of the other cases, in table 3.6. From this table it is shown that when variables were treated as numerical and nominal (exception is data B nominal) the maximum CI

lengths were about equal. In contrast, data B nominally treated had much higher CI maximum length but only in the first component. Finally, from table 3.6 is seen that object 200 has an outlying confidence region only when the low category frequency is introduced (data B) and specifically only in the nominal case. So, not only the object score of the object with extreme categories and its stability is affected but also the component scores and their stability of the other objects.

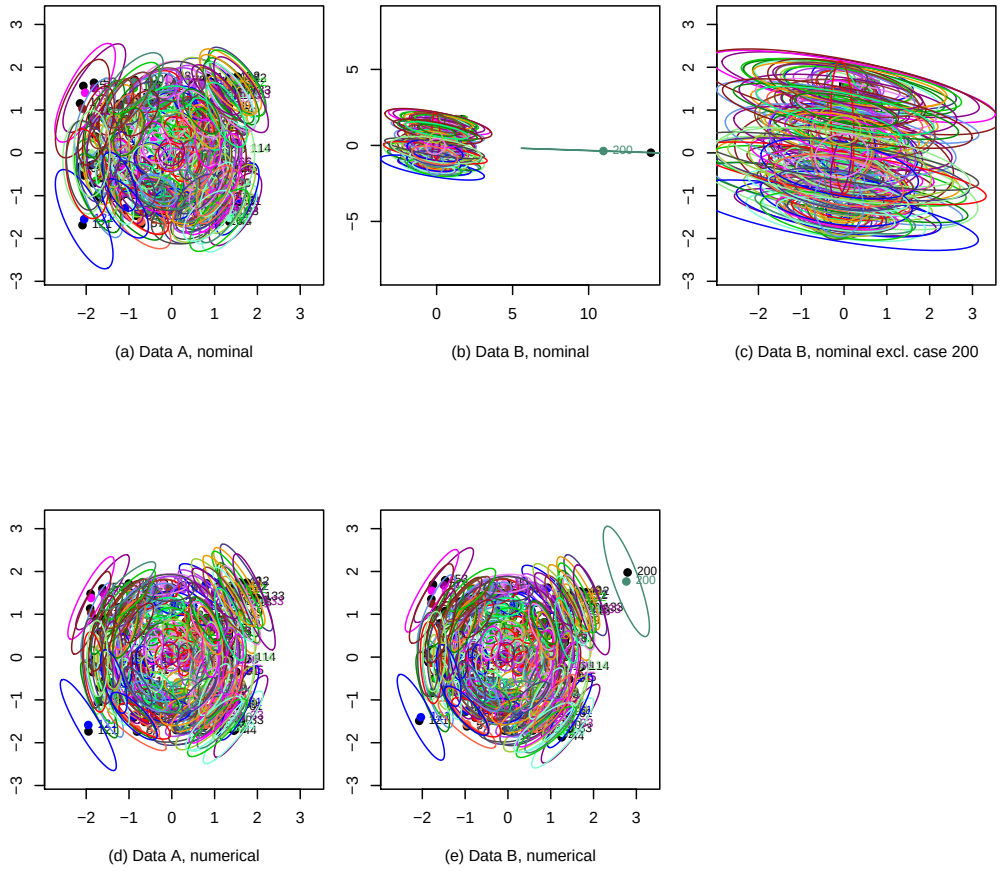


Figure 3.11: 95% confidence ellipses for component scores for data A and B when variables are treated either as numerical (d and e) or as nominal (a, b, c). With black are displayed the original object scores

Table 3.5: Sum of confidence interval lengths and sum of ellipses areas for component scores in each dimension for data A and B

	data A		data B	
	nominal	numerical	nominal	numerical
1st Component	147.77	122.42	403.8	122.27
2nd Component	186.75	173.41	165.7	171.08
Ellipse Areas	145.99	92.86	455.14	92.05

Table 3.6: Maximum confidence interval length of the cases, excluding 200 object and confidence interval length for object 200 in each dimension for data A and B

	data A		data B	
	nominal	numerical	nominal	numerical
Maximum CI length over 199 cases				
1st Component	1.19	1.09	4.61	1.14
2nd Component	1.88	1.73	1.59	1.68
CI length of 200th case				
1st Component	1.06	0.98	7.1	0.89
2nd Component	1.1	0.89	0.25	2.09

3.4 Conclusion

To summarize, all CATPCA results are sensitive to categories with small frequencies when specific less restrictive analysis options are chosen. Specifically, when the analysis level is nominal CATPCA produced more unstable results compared to numerical. Thus in the next chapter our suggested regularization will be examined for the case of nominally treated variables.

CHAPTER

4

REGULARIZATION OF CATEGORY QUANTIFICATIONS

In this chapter we will investigate whether regularization of category quantifications will diminish the influence of (extreme) outliers due to low frequency categories, as well as whether regularization will help improve stability. In order to do so, we compared the results of NLPCA with and without the suggested regularization described below. Finally, variables were treated nominally, since from previous research and as shown in chapter 3, with this analysis option results are the most affected by low frequency categories and are the most unstable.

4.1 Regularization of category quantifications

In the CATPCA algorithm category quantifications are normalized in each iteration step (Appendix A, section 3.2.1). So, they are standard scores, and therefore it is sensible to expect that category quantifications will lie within values of approximately [-3, 3], or at least not very much outside that range. However, when a data set contains category quantifications with low marginal frequencies, as in our simulated data, then category quantifications can take rather extreme values. Thus, based on this finding, we think that regularization of category quantifications by limiting quantification values to be in a specified range sensible for Z-scores, will eliminate the dominating effect of low-frequency categories. We have implemented this regularization as follows: in each iteration was checked whether category quantifications were within [-3, 3] range, and if not then they were regularized and set to -3 or 3 (the details can be found at the end of Appendix A).

4.2 Stability of NLPCA results for data sets with low marginal frequencies

For the simulated data set NLPCA with nominal scaling levels and 1000 nonparametric balanced bootstrap were performed. We compared the results when the quantifications were regularized and when they were not.

4.2.1 Stability of Variance Accounted For and eigenvalues

In figure 4.1 the confidence ellipse for the eigenvalues is plotted and in table 4.1 stability indices are displayed for the regularized and unregularized cases. Without regularization the ellipse area is 8.78 while with regularization it is 1.51. Moreover, confidence interval length in first dimension is 9.22 and in the second 3.66 when quantifications are not regularized, while when category quantifications are regularized they are 1.96 and 0.71 in first and second dimension respectivelly. It is clear from the indices above, that CI lengths and ellipse areas are taking much higher values when the category quantifications are not regularized, which shows that when quantifications are regularized, it results to much more stable results regarding eigenvalues (VAF). Furthermore, we can notice that sub-figure (b) confidence ellipse, is similar in shape to sub-figure (a) in chapter 3 figure 3.6. Also, when plotting the bootstrap points (not shown), they do no lie in two clusters anymore but scattered everywhere in the ellipse, like for data A nominal. It appears that our method of regularizing category quantifications diminishes the influence of the outlying category, and makes NLPCAs results more stable.

Table 4.1: Stability diagnostics for eigenvalue and, in parenthesis translated to %VAF for two dimensional NLPCA.

		Not regularized	Regularized
1st Component	CI lengths (VAF)	9.22 (61.50)	1.96 (13.07)
	Standard Deviation	4.25	0.48
2nd Component	CI lengths (VAF)	3.66 (24.39)	0.71 (4.73)
	Standard Deviation	1.67	0.18
	Ellipse Areas	8.78	1.51

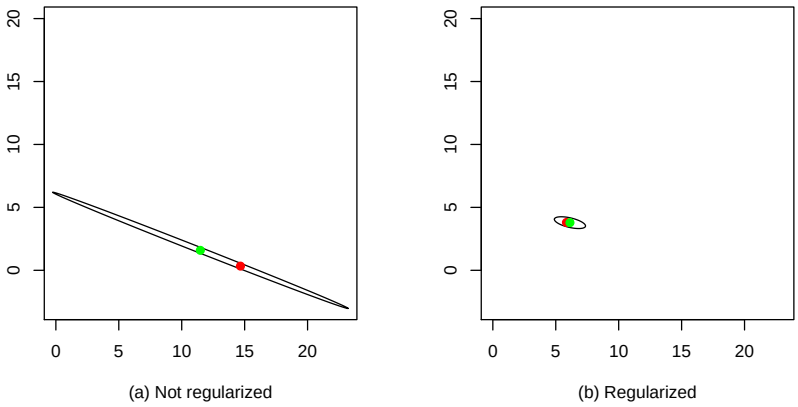


Figure 4.1: Eigenvalue confidence ellipses plot for data B, when category quantifications are regularized and unregularized. With red are displayed the eigenvalues in the original data set and with green the bootstrap means of the eigenvalues

4.2.2 Stability of transformations

To evaluate the effect of regularization on stability of category quantifications, we can look at transformation plots. In figure 4.2 are shown the transformation plots with the 95% confidence intervals for variable 5. With regularized category quantifications confidence intervals are clearly smaller (note the difference in axis scales). For the regularized case category 8 has CI length of 1.09, with lower and upper limit of 3 and 4.09, and for the unregularized case CI length has value of 6, with lower and upper limit of 8.10 and 14.10 respectively. Also, as a consequence of the quantification of category 8 only slightly outlying in the regularized case, the quantifications for categories 1 to 7 instead of being all the same are differentiated now. In conclusion, category quantifications are less influenced by the outlying category and show more stability with regularization compared without.

4.2.3 Stability of category quantifications

To further display the effect of regularization on stability of category points, we can look at category plots. As in chapter 3 we only display plots for

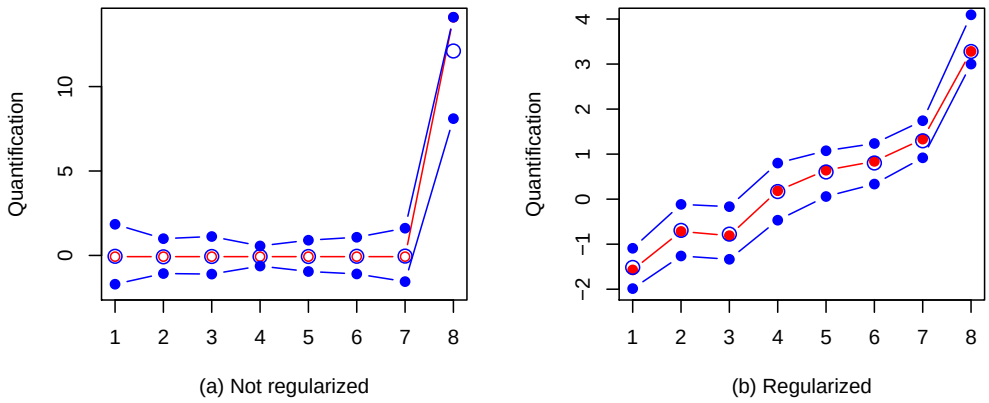


Figure 4.2: Confidence intervals for variable 5 transformations (nominally treated) for data B. The red line indicates the original variable transformation from the observed data. The outer lines indicate the 95% confidence intervals. The x-axis depicts the original category numbers and the y-axis the category quantifications

variables 5 (figure 4.3) and 12 (figure 4.4), considering that for the rest of the variables, results are about the same. In variable 5 category points and ellipses for categories 1 to 7 are similar to data A nominal case, figure 3.8 (a), and category 8 is much less extreme for the regularized case. The same applies to variable 12, where ellipses and category points are similar to data A nominal case, figure 3.9 (a), and category 8 is much less extreme for regularized case. The results are indicating that when regularized the categories are less influenced by the outlying category quantification compared to the non regularized case.

4.2.4 Stability of component loadings

In figure 4.5 are shown the confidence ellipses for data B loadings when category quantifications are either regularized (b) or not (a). It can be seen that, again, the result are very much alike the data A nominal case (figure 3.10), both in loadings and ellipses, when category quantifications are regularized

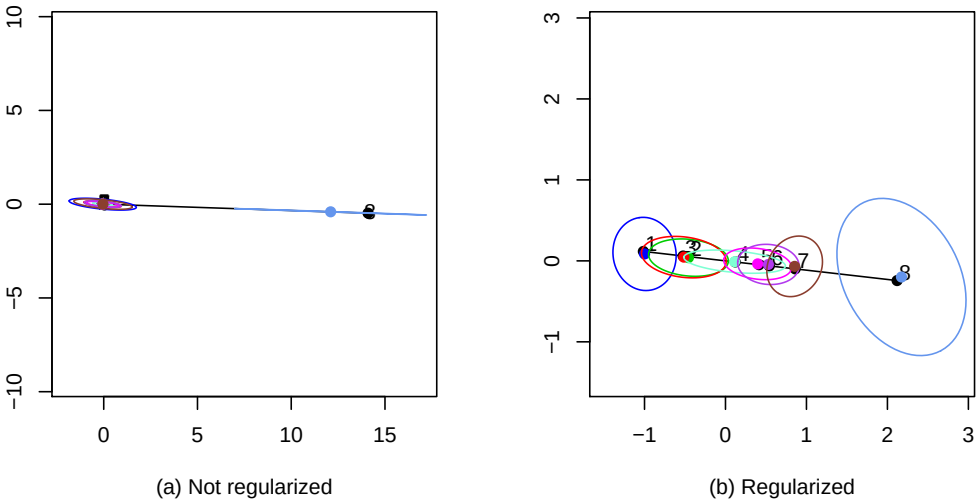


Figure 4.3: Confidence ellipses for category points of variable 5 for data B where variables are treated nominally and category quantifications are either regularized or not

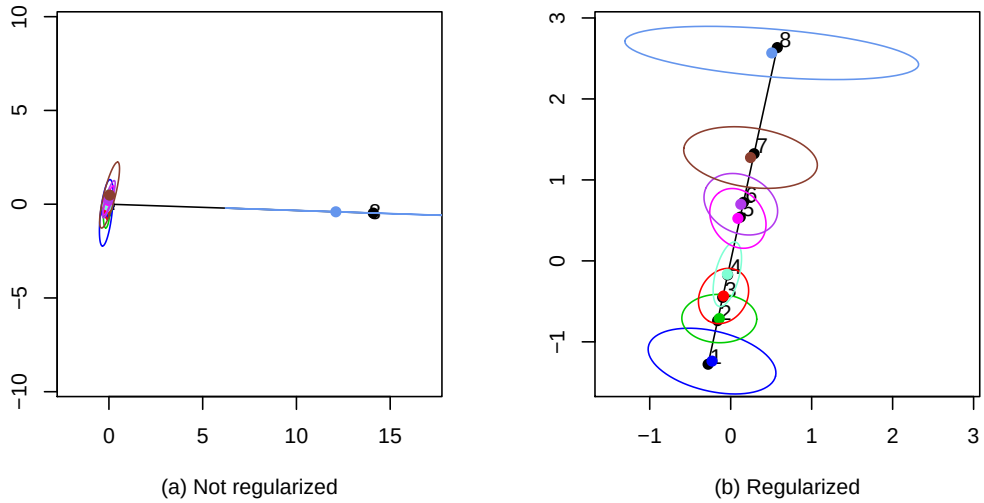


Figure 4.4: Confidence ellipses for category points of variable 12 for data B where variables are treated nominally and category quantifications are either regularized or not

compared to the unregularized. The results showed that component loadings are less sensitive to low marginal frequencies when category quantifications are regularized.

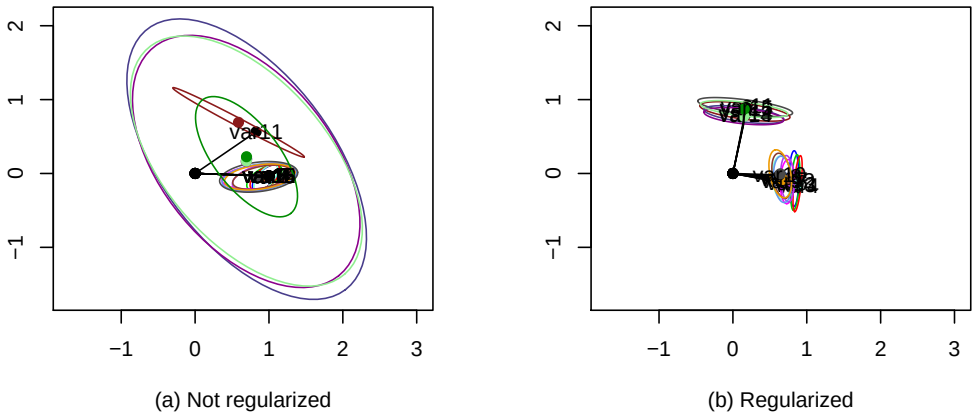


Figure 4.5: 95% confidence ellipses for the component loadings for data B when variables are regularized (b) or not (a). Black circles indicate the loadings for the original component loadings and the colored indicate the centroids of the bootstrap clouds.

4.2.5 Confidence intervals and ellipses for component scores

In figure 4.6 we can see the ellipses for data B (nominally treated with and without regularization). For sub-figure (b), the ellipses for object 1-199 are much the same as for data A nominal (figure 4.6 (a)), and for object 200 the location and ellipse is like for data B numerical (figure 4.6 (e)). Furthermore, from table 4.2 can be seen that the sum of ellipse areas for the regularized case is taking much lower value (164.90), about a third, compared to the unregularized (455.14) and close to the sum for data A nominal (147.77) suggesting that regularization increases considerably the overall stability. We can notice that the increase in stability is mainly due to the first dimension

where the sum of CI length is 155.11 compared to 403.8 and in the second is 195.63 compared to 165.7 in the unregularized. Noteworthy is that the results when regularized are closer to the ones we got in table 3.5 for data A (when nominally treated). In addition, to compare stability of component scores within a solution we looked at indices for object 200 and the maximum of these values of the other cases, in table 4.3. From this table it is seen that when category quantifications are regularized the maximum range for the first dimension is much lower, and (somewhat) higher for the second dimension, more resembling the data A case (table 3.6). Thus, all object scores are more stable when the outlying object has less influence due to the regularization.

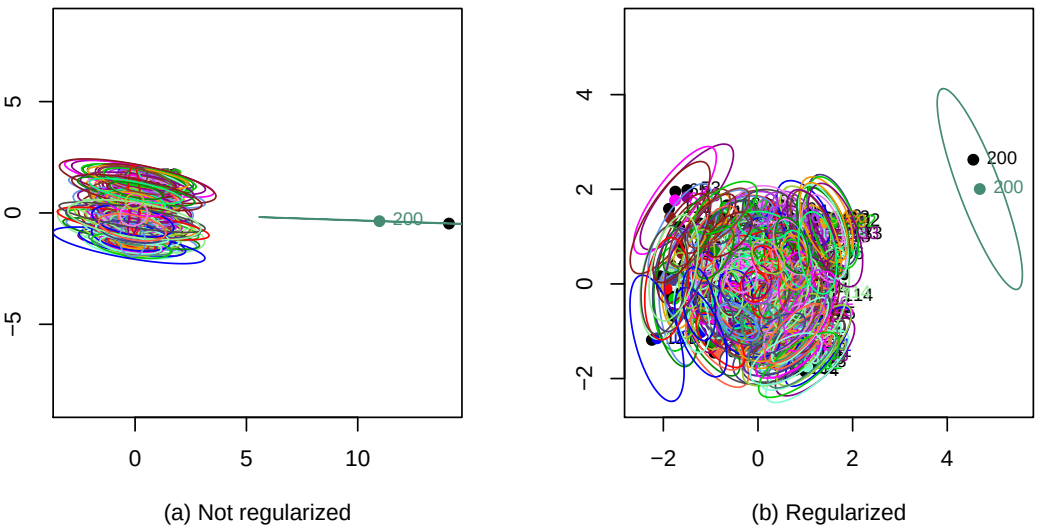


Figure 4.6: 95% confidence ellipses for component scores for data B when variables are treated either as nominal and category quantifications are regularized (a) or not (b). With black are displayed the original object scores

Table 4.2: Sum of confidence interval lengths and sum of ellipses areas for component scores in each dimension for data B

	data B	
	Not regularized	Regularized
1st Component	403.8	155.11
2nd Component	165.7	195.63
Ellipse Areas	455.14	164.90

Table 4.3: Maximum confidence interval length of the cases, excluding 200 object and confidence interval length for object 200 in each dimension for data B, when quantifications are regularized and not

	data B	
	Not regularized	Regularized
Maximum CI length over 199 cases		
1st Component	4.61	1.41
2nd Component	1.59	2.00
CI length of 200th case		
1st Component	7.1	1.51
2nd Component	0.25	3.21

4.3 Conclusion

The aim of this thesis has been to establish if NLPCA produces more stable results when category quantifications are regularized in the case of low marginal frequencies that have much influence on the solution and cause instability. In our investigation we compared bootstrap results when category regularization is applied and when it is not. In all our results (eigenvavalues (VAF), transformations, category quantifications, component loadings and object scores) NLPCA showed considerable less influence of outlying categories and more stable results when regularization is applied. Specifically, the solution becoming one-dimensional due to the outlying categories treated nominally, is remedied by the regularization. When regularized, the solution and its stability for data B nominal resemble data A nominal, the stability for B being only somewhat less.

CHAPTER

5

CONSIDERATIONS AND RECOMMENDATIONS

Technical considerations and recommendations for further research

This last chapter includes technical considerations and proposals for further research.

Technical considerations

In Figures 3.5 (b) the question might be raised why the object points lie exactly clustered according to the category points for variable 11 and looking at Figure 3.4 (b) why only variable 11 is loading on both dimensions. There is no substantive interpretation for this. It is just an artifact, resulting from the fact that when case 200 is included a perfect solution in one dimension exist when the nominal scaling level is used, allowing a transformation that is solely focused on the contrast between case 200 and all the other cases. For a two-dimensional analysis, the VAF is also perfect, but the solution is not unique because any solution with VAF per dimension summing to 100% is feasible: 100% VAF in the first dimensions and 0 in the second is as good (in terms of overall fit) as a VAF in the first dimension of 100-c and VAF of c in the second dimension. How exactly the fit is distributed over the two dimensions, and thus which variable(s) loads in the second dimension, depends on the starting values. With the starting values we used variable 11 is picked to serve as the second dimension object scores, purely by coincidence; with other starting values it can be any of the other variables, or a combination of several variables, or, with clever starting values the second dimension can come out with zero VAF. The randomness of the second dimension for data B nominal also explains the ellipse sizes for variables 11-15, see Figure 3.10 (b). The ellipse for variable 15 being smaller than for variables 12-14 has no substantive meaning, but is coincidence. Repeating the 1000-bootstrap analysis with different bootstrap samples results in one of variables 12-15 having a smaller ellipse, or more of variables 11-15 smaller, or all of 11-15 about the same size. The only constant result is the ellipses for variables 1-10 being much smaller than for 11-15.

Additionally, due to this regularization the algorithm is not always monotonically convergent anymore, we decided to use it anyway, because the deviance from the monotonicity seems to be not very serious. With regularization the algorithm first converges total maximum fit, and then in subsequent iterations the fit can start to decrease, but only very slightly after which it can start to increase again. When converged, the difference between the fit before decreasing and the fit when converged is tiny. So, we think that the results we obtain with the algorithm adjusted to our regularization method are reliable, at least for the purpose of this thesis.

Recommendations for further research

In our analysis only one simulated data set was used, which can result to data specific conclusions, even though bootstrap was used to avoid that it seems sensible in further reseach to increase the number of data sets.

Also, in further research different type of outliers, such as less extreme that are not outlying on all variables or outlying in only 1 dimension, could be introduced and assess how our suggested regularization performs.

Appendices

APPENDIX

A

TECHNICAL BACKGROUND OF NONLINEAR PCA

In this appendix a more technical explanation of Nonlinear Principal Components Analysis (NLPCA) will be presented. This description is based on how NLPCA is performed in SPSS referred to as CATPCA. CATPCA stands for the acronym of Categorical Principal Components Analysis.

In CATPCA numerical values (named category quantifications) are assigned to each variable through a procedure called optimal scaling. Optimal scaling is an iterative process that minimizes a least squares loss function while simultaneously quantifying categories. This procedure converges to a stationary point. The main outputs in CATPCA are eigenvalues, loadings, object scores, vector and centroid coordinates, which will be explained later. The outline of this appendix is as follows. Firstly, in section 1, the notation will be presented. Secondly, in section 2, the objective function of CATPCA will be presented, that follows the loss function. Thirdly, in section 3, the specific steps of the algorithm will be presented.

A.1 Notation

We write matrices in bold capital, vectors in bold lowercase, and values in italic lowercase.

$\mathbf{H}$ denotes the data, n by m, m is the number of variables and n is the number of cases, which are also referred as individuals or objects. $\mathbf{X}$, denotes object scores, n by p, with $p < m$. $\underline{\mathbf{Y}}_j$ is the $\kappa_j \times p$ matrix of category coordinates for variable j, κ_j is the number of distinct categories for variable j. When the scaling level is multiple nominal these are centroid coordinates, which are also the category quantifications. For the other scaling levels $\underline{\mathbf{Y}}_j$ are vector coordinates, which are the category quantifications $\mathbf{y}_j\mathbf{a}_j$, where $\mathbf{y}_j$ is κ_j by 1 and $\mathbf{a}_j$. Furthermore, $\mathbf{A}$ denotes the loadings of dimensions $m_2 \times p$ with m_2 standing for the number of single variables. Single variables as variables with scaling level other than multiple nominal. m_2 denotes the number of variables with scaling level multiple nominal. m_1 stands for the multiple

nominal variables. The total number of variables will be noted as m and so $m = m_1 + m_2$. As well, $\boldsymbol{y}_j$, which stands for the vector called single category quantifications and has length equal to κ_j and with $j \in m_2$.

$\mathbf{G}_j$ denotes the indicator matrix for variable j with dimensions $n \times \kappa_{\mathrm{j}}$. The indicator matrix for a specific variable is determined as row i of $\mathbf{G}_j$ has a *1* in the column that corresponds to the category that is scored by object i on variable j, and zero's in the other columns. (When the missing option for variable j is passive, row i contains only zero's if object i has a missing on variable j). Thus, the elements of $\mathbf{G}_j$ are:

$$\mathbf{G}_{(j)ik} = \begin{cases} 1 & \text{when object } i \text{ is in the } k\text{-th category of variable } j \\ 0 & \text{otherwise} \end{cases}$$

$\mathbf{m}_{(j)ii}$ incorporate passive treatment of missings (i.e. excluding missing cells of the data matrix; thus not excluding a complete row when a missing occurs for a case, and also not implementing something for a missing) a diagonal matrix $\mathbf{M}^*$ is included which is the sum of $\mathbf{M}_j$ matrices with elements:

$$\mathbf{m}_{(j)ii} = \begin{cases} 0 & i \text{ object has a missing on variable } j \text{ and the} \\ & \text{missing option for variable } j \text{ is passive} \\ 1 & \text{otherwise} \end{cases}$$

$\mathbf{M}^*$: is the $\sum_j \mathbf{M}_j$

In addition the feature of weighting cases is implemented by incorporating a diagonal $n \times n$ $\mathbf{W}$ of case weights. For an unweighted analysis all values of $\mathbf{W}$ are 1. Moreover, also variables can be assigned a weight. The subscript w in m_{w1}, m_{w2}, m_{w} and n_{w} stands for "weighted", thus for instance m_{w1}

denotes the weighted number of variables and nw the weighted number of cases. Finally, $\boldsymbol{u}_n$ denotes a column vector of length n containing ones and $\mathbf{I}_p$ denotes an identity matrix of order p.

A.2 Loss Function

CATPCA is performed via an iterative procedure where the objective loss function is minimized. This loss function and its minimization will be presented in this section.
The loss function is:

$$L(\mathbf{X}; \underline{\mathbf{Y}}_1, ..., \underline{\mathbf{Y}}_j; \mathbf{A}) = m_w^{-1} \sum_{j=1}^{m_w} tr(\mathbf{G}_j \underline{\mathbf{Y}}_j - \mathbf{X})' \mathbf{W} \mathbf{M}_j^* (\mathbf{G}_j \underline{\mathbf{Y}}_j - \mathbf{X}) \quad (1)$$

The loss function (1) is minimized under the condition of centered and orthonormalized object scores $\mathbf{X}$:

$$\mathbf{X}\mathbf{M}^*\mathbf{W}\mathbf{X}' = n_w m_w \mathbf{I}_p \quad (2)$$

$$u_n' \mathbf{M}^* \mathbf{W} \mathbf{X}' = \mathbf{0} \quad (3)$$

These conditions avoid the trivial solution of $\mathbf{X} = \mathbf{0}$ and $\mathbf{Y}_\mathrm{j} = \mathbf{0}$ (3) requires the object scores to be centered. From (2) and (3) is implied that the columns of $\mathbf{X}$ are centered and orthonormal, which means orthogonal with standard deviation 1 and mean 0. Also, transformed variables are required to be standardized. The loss function (1) is minimized in an iterative process, in which alternatingly $\mathbf{X}$ is updated keeping $\mathbf{Y}$ fixed, and $\mathbf{Y}$ is updated keeping $\mathbf{X}$ fixed.

A.3 Minimization of the loss function

The CATPCA algorithm consists of the following steps:

- initialization (section 3.1)
- update parameters (section 3.2)
  - quantifications and loadings (section 3.2.1)
  - object scores (section 3.2.2)
- check converge (section 3.3)

A.3.1 Initialization

Initial object scores are obtained either (1) as random values and the initial loadings as the cross products of the the standardized variables with the orthonormalized random object scores, or (2) the results of an analysis with numerical scaling level for all variables are used as the starting values for the object scores and the loadings.

A.3.2 Parameters

The loss function (1) is minimized in an alternating least square (ALS) way. The ALS algorithm minimizes this function with the three unknowns by updating each one of the parameter matrices (via partial derivatives) in turn and keeping the others fixed. Firstly, category quantifications are updated keeping the object scores fixed. Then object scores are updated, keeping the quantifications (and loadings if applicable) fixed. Finally, it checks if the algorithm has converged and if not, the steps are repeated.

A.3.2.1 Update quantifications and loadings

Variables with multiple nominal scaling level are fitted according to the centroid model (as, for instance, in Correspondence analysis). Then multiple sets of quantifications are obtained; a set of quantifications for each dimension. The multiple quantifications are the coordinates of the categories in the p-dimensional component space. Thus the categories are represented as points that lie scattered in the component space. When the scaling level for variable j is not multiple nominal, the variable is fitted according to the vector model (as, for instance, in PCA). Then $\underline{\boldsymbol{Y}}_j$ is required to be of rank one and is obtained as:

$$\underline{\mathbf{Y}}_j = \mathbf{y}_j \mathbf{a}_j \tag{4}$$

Thus the categories for a non-multiple variable are represented in the component space as a line (vector).

For updating quantifications, first the centroids $\mathbf{Y}_j$ of variable j are computed as:

$$\mathbf{Y}_j^{(t+1)} = \mathbf{D}_j^{-1} \mathbf{G}_j' \mathbf{X}_w^{+(t)} \tag{5}$$

where t denote the iteration number. $\mathbf{D}_j = \mathbf{G}_j^{\mathrm{T}} \mathbf{G}_j$ so, $\mathbf{D}_j$ is a diagonal matrix of dimensions $\kappa_j \times \kappa_j$ with entries the marginal category frequencies on the diagonal.

If variable j has multiple nominal scaling level, then the quantifications are the centroids (5), thus

$$\underline{\mathbf{Y}}_j^{(t+1)} = \mathbf{Y}_j^{(t+1)}$$

For nominal scaling level only rank-one restriction (4) of the centroids is required to obtain $\mathbf{y}_j$:

$$\mathbf{y}_j^{(t+1)} = \frac{\mathbf{Y}_j^{(t+1)} (\mathbf{A}_j^{(t+1)})'}{\mathbf{A}_j^{(t+1)} (\mathbf{A}_j^{(t+1)})'} \tag{6}$$

If the scaling level is ordinal, the rank-one quantifications (6) are further restricted to meet the order requirement:

$$\mathbf{y}_j^{(t+1)} < -\text{WMON}(\mathbf{y}_j^{(t+1)}) \tag{7}$$

in ordinal quantifications WNOM() is used to denote the weighted monotonic regression process.

If the scaling level is numerical, the additional interval restriction is applied to (7)

$$\mathbf{y}_j^{(t+1)} < -\text{WLIN}(\mathbf{y}_j^{(t+1)}) \tag{8}$$

In numerical single quantifications WLIM() denotes the weighted linear regression process. The weights which are used are the diagonal elements of $\mathbf{D}_j$ (marginal frequencies).

Variants of nominal and ordinal scaling level are available in Catpca via splines. Ordinal and nominal scaling level results in quantifications that are step functions, and can be quite irregular when the number of categories is large. The spline levels result in continuous quantifications and thus in a smoother transformation curve.

If the scaling level is splines nominal or ordinal then $\mathbf{y}_j$ is:

$$\mathbf{y}_j^{(t+1)} = d_j^{(t+1)} + \mathbf{S}_j^{(t+1)}\mathbf{b}_j^{(t+1)} \tag{9}$$

splines nominal quantifications are non monotonic piecewise polynomials, that are smoother than in the nominal case. The smoothness is determined from the degree of the spline polynomial (s_j). The number of the piecewise polynomials is defined from the number of interior knots $\in [0, \kappa_{j\text{-}2}]$. t_j is the number of interior knots of the variable. Also, d_j is the intercept of splines.

$\mathbf{b}_j$ is splines coefficient vector and $\mathbf{S}_j$ is defined as I-spline basis for a variable with order $\kappa_j \times (s_j + t_j)$. The only difference between splines nominal and splines ordinal is a restriction on $\mathbf{b}_j$ that results monotonic piecewise polynomials.

Finally, $\mathbf{y}_j$ is normalized:

$$n_{\mathrm{w}}\mathbf{y}_j^{(t+1)}[(\mathbf{y}_j^{(t+1)})'\mathbf{D}_j\mathbf{y}_j^{(t+1)}]^{-0.5}$$

Then loadings are updated as:

$$\mathbf{A}_j^{(t+1)} = \frac{(\mathbf{Y}_j^{(t+1)})'\mathbf{D}_j\mathbf{y}_j^{(t)}}{(\mathbf{y}_j^{(t)})'\mathbf{D}_j\mathbf{y}_j^{(t)}}$$

A.3.2.2 Update object scores

The object scores are updated as:

$$\mathbf{X}_{\mathrm{w}}^{(t+1)} = m^{-1}[m_1 \sum_{j \in m_1} \mathbf{M}_j\mathbf{Y}_j^{(t+1)} + m_2 \sum_{j \in m_2} \mathbf{M}_j\mathbf{y}_j^{(t+1)}\mathbf{A}_j^{(t+1)}]$$

Then $\mathbf{X}_{\mathrm{w}}^{(t+1)}$ is centered with respect to $\mathbf{W}$, $\mathbf{M}$ and orthonormalized.

A.3.3 check converge

After updating the quantifications, the loadings, and the object scores, the loss is computed according to equation (1). Then the loss is compared to the loss of the previous iteration and if the difference is lower than the convergence criterion or when the number of iterations has reached the maximum, the algorithm stops (both the convergence criterion and the maximum number of iterations can be specified by the user or the default can be accepted)

Suggested regularization

The regularization is inserted between the normalization of $\mathbf{y}_j$ and updating the loadings as follows: If values of the normalized $\mathbf{y}_j$ are outside the range

[-3,3], they are restricted to -3 or 3. If this restriction is applied, the result is centered, and normalized again. Due to the centering and re-normalization the resulting restricted values are not exactly -3 of 3.

APPENDIX

B

SIMULATING MULTIVARIATE DATA FROM A CORRELATION MATRIX

To generate data set with pre-specified principal component structure from correlation matrix we followed the steps bellow. The correlation matrix will be noted as $\mathbf{C}$.

1. Generated random normally distributed variables $\mathbf{A}^0 \sim \mathbf{N}(0,1)$, with order $n \times m$. n is the number of cases in the in the simulated and m stands for the number of variables. This matrix will be noted as $\mathbf{A}^0$.

2. In order to compute a set of variables, referred as Principal Components (PCs), which will be multivariate normal and orthogonal to each other (independent). PCA is performed with extracted PCs equal to m (not reduced). These variables will be Multivariate Normally distributed. Orthogonal and with mean 0 and standard deviation 1. This step will give the final $\mathbf{A}$.

3. Calculation of the Cholesky decomposition of $\mathbf{C}$. The Cholesky decomposition gives an upper triangular matrix known also as a "square root" of the correlation matrix. $\mathbf{L}$ will be noted as the output from Cholesky decomposition. $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, where the notation T denotes the transpose of a matrix.

4. The simulated data set, referred as $\mathbf{X}$ is calculated as the post multiplication of $\mathbf{A}$ from step 2 with $\mathbf{L}$ from step 3. $\mathbf{X} = \mathbf{A}\mathbf{L}$

From the aforementioned steps, multivariate normally distributed data has been calculated from pre-defined correlation matrix. Last step is to descritize the data with a procedure known as binning. Binning is used to group the interval values to corresponding categoricals.

APPENDIX

C

CHECKING STRUCTURE OF SIMULATED DATA

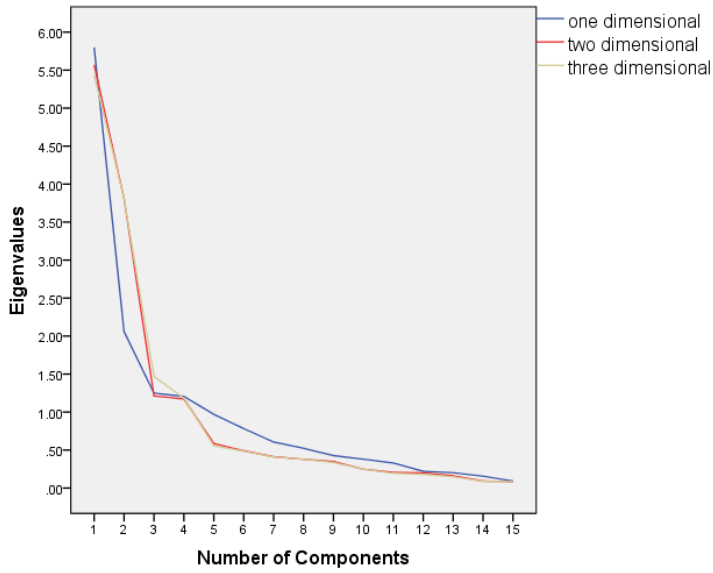


Figure C.1: Scree plot for one, two and three dimensional principal component analysis on the simulated data. On the x-axis is the number of components. On the y-axis are the eigenvalues when variables are treated nominally.

To check if data set A has the desired structure as described above, we performed NLPCA with numerical scaling level for all variables (thus, standard PCA). In NLPCA solutions are not nested meaning that, the first two components of a higher to 2 component solution are not equal to the two components of the 2-component solution. The exception is when all variables have numerical scaling level. So, we performed 1-, 2-, and 3-component analyses using both numerical and nominal scaling level. The eigenvalues are depicted in a scree plot in Figure C.1 for nominal scaling level, the results for numerical scaling level are much the same. Figure C.1, showing that indeed data set A has a 2-component structure. The scree plot shows how the eigenvalue of the components declines. What is checked in a scree plot is an 'elbow'. An 'elbow' is the position where the decline in size of the eigenvalues begins to level out. The number of components to choose is the number before the elbow. In Figure C.1, we see that the scree plot suggests two principal components. Thus is shown that the intended two principal

components structure exists.

References

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika*, *31*(1), 33–42.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap.* New York: Chapman and Hall.

Gifi, A. (Eds: Heiser, W.J., Meulman, J.J., Van den Berg, & G.). (1990). *Nonlinear multivariate analysis.* Chichester UK: John Wiley. (Original work published 1981).

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* New York: Academic Press.

Heiser, W. J., & Meulman, J. (1983). Constrained multidimensional scaling, including confirmation. *Applied Psychological Measurement*, *7*(4), 381–404.

IBM Corp. (n.d.). IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.

Jolliffe, I. (2002). *Principal component analysis. New York Springer-Verlag*, 338–372.

Linting, M., Meulman, J. J., Groenen, P. J., & Van der Kooij, A. J. (2007a). Nonlinear principal components analysis: Introduction and application. *Psychological methods*, *12*(3), 336–358.

Linting, M., Meulman, J. J., Groenen, P. J., & Van der Kooij, A. J. (2007b). Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological methods*, *12*(3), 359–379.

Linting, M., et al. (2007). *Nonparametric inference in nonlinear principal components analysis: Exploration and beyond.* Thesis, Leiden University. (https://openaccess.leidenuniv.nl/handle/1887/12386).

Markus, M. T. (1994). *Bootstrap confidence regions for homogeneity analysis; the influence of rotation on coverage percentages.* In R. Dutter and W. Grosmann (Eds.), COMPSTAT 1994 (pp. 337-342). Heidelberg: Physica Verlag.

Markus, M. T. (1994). *Bootstrap confidence regions in nonlinear multivariate analysis* Leiden, The Netherlands: DSWO Press.

Meulman, J. J., & Heiser, W. J. (1983). *The display of bootstrap solutions in multidimensional scaling* Murray Hill, NJ: Bell Laboratories. (Technical memorandum).

Meulman, J. J., Heiser, W. J., & SPSS Inc. (2015). *IBM SPSS Categories 23.* Armonk, NY: IBM Corp. (ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/23.0/en/client/Manuals/IBM_-SPSS_Categories.pdf).

Van der Burg, E., & de Leeuw, J. (1988). Use of the multinomial jack-knife and bootstrap in generalized nonlinear canonical correlation analysis. *Applied Stochastic Models in Business and Industry*, *4*(3), 159–172.