Nonparametric inference in nonlinear principal components analysis: Exploration and beyond

Linting, M.

Chapter 1

General Introduction

1.1 Categorical Data Analysis in the Social and Behavioral Sciences

During the period of time I have been analyzing data from the social and behavioral sciences, I have noticed that these data sets contain a mixture of different types of variables, quite some of which are measured in ordered or unordered categories. For example, many background variables, such as the subject's gender, religion, profession, and hobbies, but also analysis variables like attachment of a child to its caregivers (Ainsworth & Bell, 1970; Bowlby, 1969) or the clinical diagnosis of psychological disorders, are measured in unordered categories. Variables akin to level of education, ratings on observation scales (for instance, quality of child care, ranging from "very poor" to "very high"), or a person's opinion on a specific subject (for example, assessed using Likert-type scales) are measured in ordered categories. And finally, data sets contain variables with numeric values, like age, length, income, heart rate, IQ, and so on. These different types of variables call for different treatment when analyzing the data, which is not always as self-evident as it may seem. In addition, these data sets often contain variables that are or may be nonlinearly related to each other, which should also be accounted for in the analysis results. Thus, data sets in the social and behavioral sciences cannot always be analyzed as straightforwardly as many researchers would wish. This thesis offers a solution to some of the questions that may be raised in this context, making use of an analysis technique that is part of a tradition referred to as *Nonlinear* or *Categorical Multivariate Data Analysis*. The thesis aims to aid applied researchers in performing a particular categorical data analysis method that in some cases will do more justice to their data than the traditional approaches they standardly use. It also offers some simple and easily implementable tools to strengthen the conclusions drawn from this analysis method.

Usually, a distinction is made between *categorical* variables on the one hand, and *numeric* variables on the other. The term *categorical* either refers to nominal and ordinal variables, or to nominal variables alone. *Nominal* variables are measured in unordered categories and use category labels to define specific groups of subjects. Variables that use labels with a specific ascending or descending order are referred to as *ordinal* variables. *Numeric* variables have values with actual numeric properties and can be measured on interval or ratio scales, with *interval* scales containing values with a specific order, and having equal distances between subsequent values; *ratio* scales additionally include a clear definition of the value 0.0: if this value occurs, the measured quantity is completely absent.

This distinction between categorical and numeric variables should not lead to the misconception that Categorical Multivariate Data Analysis is focused only on the analysis of categorical variables in this strict sense. Alternatively, all types of variables mentioned above can be seen as *categorical variables*, with each observed value or label indicating a category. For instance, the variable age, with observed values 4, 5, 8, 10, and 11 years old, can be perceived as a categorical variable with five categories. Consistent with that line of thought, all types of variables (and not just nominal and ordinal) can be analyzed by the type of categorical (or nonlinear) analysis technique described in this thesis.

There is an important difference between variables with and without real numeric values: Variables with real numeric properties (interval and ratio variables) can be used in standard calculations, and thus, for example, their mean and standard deviation can be computed without difficulty. For nominal and ordinal variables, however, the category labels cannot be interpreted as numbers. Therefore, such common calculations do not lead to sensible conclusions, inducing the following problem: Nominal and ordinal variables are measured frequently, but researchers often have problems analyzing them properly. The possibility of the existence of nonlinear relationships between variables further complicates that situation.

Many different methods have been described to deal with the analysis of variables without real numeric properties, among which are methods using dummy variables. In such methods, a categorical variable is split up into marker variables for each category, where for each such dummy variable, a value '1' indicates that a person scored that particular category, and a '0' indicates that the person did not. These dummy variables are then used as numeric variables in the analysis of interest. Such methods, however, are often quite intensive, especially when variables have many categories. Alternatively, specific methods – for example, nonlinear equivalents of regression and principal components analysis – have been developed for the analysis of mixed categorical (nominal, ordinal, and numeric) data, using optimal scaling (or optimal quantification) (see Gifi, 1990). Such techniques quantify category labels in an optimal way, maximizing the relationship between the quantified variables.

In the remaining part of this introduction, a brief explanation of optimal scaling methods will be given, focusing on nonlinear principal components analysis (nonlinear PCA). In addition, the emphasis will be on how such methods can be taken beyond their exploratory basis by assessing inferential measures in a nonparametric way.

1.1.1 Optimal quantification

Optimal quantification, also referred to as *optimal scaling*, transforms category labels into numeric values (which is called quantification), such that the strength of the relationships between the quantified variables is optimized. In other words, the variance-accounted-for (VAF) among the quantified variables is maximized. At the same time, a standard (numeric and linear) analysis technique is performed on the quantified data. This idea is founded in the work of several different authors. For example, Horst (1935), Fisher (1938, 1940), and Guttman (1941) provided important work on the quantification of nominal variables, and Shepard (1962b, 1962a) and Kruskal (1964) were pioneers in the area of quantification of ordinal data. Many followers of these authors have used optimal scaling in a number of different contexts (for example, see Kruskal, 1965; Kruskal & Shepard, 1974; Nishisato, 1980, 1994; Ramsay, 1988; Roskam, 1968; Shepard, 1966; Van der Burg & De Leeuw, 1983; Winsberg & Ramsay, 1983; Young, 1981; Young, de Leeuw, & Takane, 1976; Young, Takane, & de Leeuw, 1978). Gifi (1990) gives a historical overview, and a discussion of the techniques using optimal quantification. This thesis will be focused on nonlinear principal components analysis using optimal quantification. An extensive description of the objectives and procedure of optimal quantification will be given in Chapter 2.

1.1.2 Nonlinear PCA as an exploratory technique

When researchers have collected a data set, and wish to explore the correlational structure among the variables (i.e., without starting from a specific theory), principal components analysis (PCA) or exploratory factor analysis (EFA) may be the method of choice. When the goal of a study is to model the structure in the observed data set, incorporating the relationships between the variables (common variance) as well as the unique contribution of separate variables (unique variance), EFA is the most appropriate method. When the goal is data reduction, that is, reducing a large number of observed variables to a smaller number of composite variables, without considering each variable's unique contribution, PCA is more suitable (see Fabrigar, Wegener, MacCallum, & Strahan, 1999). This thesis will be focused on PCA, the most popular of the two approaches (Fabrigar et al., 1999).

In many cases, PCA matches the goal of a study, but the assumptions of PCA are not met by the observed data. Specifically, the assumptions that variables have at least an interval measurement scale and are linearly related to each other are often violated. Sometimes, especially when the variables selected for the analysis are measured on an ordinal scale, researchers

simply ignore these violations, and perform PCA regardlessly. This decision may or may not lead to substantial problems, depending on whether these variables are approximately linearly related to each other. If PCA is performed without checking its assumptions, one can never be sure whether the results are trustworthy.

In such situations, nonlinear (or categorical) PCA with optimal quantification is a helpful alternative; it pursues the same objectives as linear PCA, but incorporates nominal and ordinal as well as numeric variables, and can help discover and deal with possible nonlinear relationships among these variables. The method can analyze variables at their measurement level, but when nonlinear relations between variables exist, other analysis levels can also be specified so these relations can be handled most effectively.

Nonlinear PCA generalizes some more specific analysis methods by its ability to handle data of different analysis levels simultaneously. When all variables are at a numeric analysis level, nonlinear PCA equals linear PCA. With all variables at a multiple nominal level (this level is explained in Chapter 2), nonlinear PCA is equivalent to multiple correspondence analysis (for example, see Greenacre, 1984), dual scaling (Nishisato, 1980, 1994), or homogeneity analysis (Gifi, 1990). Nonlinear PCA is described more extensively in Chapter 2.

1.2 Inference in Principal Components Analysis

The primary goal of this thesis is to move beyond the exploratory status of nonlinear PCA by introducing nonparametric inferential measures, based on the observed data instead of preassumed population distributions. In general, the gap between exploratory and confirmatory analysis methods often seems wider than it should. Exploratory techniques are not as unregulated as advocates of confirmatory techniques may suggest, and confirmatory methods contain descriptive properties just as well as exploratory methods. In addition, confirmatory analysis of complicated multivariate models should not be idealized, because the framework in which hypotheses are tested – including, for example, the assumption of multivariate normal distributions – is generally implausible in practice (De Leeuw, 1988). In this thesis, we use methods of establishing inferential measures for exploratory techniques that do not rely on such assumptions.

1.2.1 Methods for nonparametric inference

In traditional inferential methods, certain known distributions (for instance, the normal distribution) are assumed to represent the population distribution. Without making additional assumptions, however, the properties of the population distributions of the outcomes of exploratory methods, such as PCA, are unknown. Therefore, for such analysis methods, a nonparametric approach where the observed data itself instead of the (unknown) population distributions are used as a basis for inferential measures, seems more appropriate.

A first aspect of inference that we will focus on in this thesis is stability. In line with Gifi (1990), we define stability as the degree of sensitivity of an analysis technique to changes in the data, where small changes should lead to only small changes in the output of the analysis. Stability should not be mistaken for estimation accuracy: A sample estimate can be very close to the population value (i.e., it can be accurate, or unbiased), but still have large confidence intervals, and the reverse (inaccurate or biased estimates that are very stable) is also possible.

The nonparametric bootstrap (Efron & Tibshirani, 1993) procedure can be used to establish the stability (robustness) of analysis results, and can provide an estimate of the shape, spread and bias of the sampling distribution of a specific statistic. In this thesis, the focus will be mainly on the former use. The bootstrap procedure involves treating the observed sample as if it were the population, and drawing, with replacement, a large number of new samples from it, keeping the rows of the observed data set intact. Each new sample contains as many cases (rows) as the observed data set, but some cases may appear several times, whereas others may not appear at all. For each bootstrap sample, the statistic of interest is computed, and the results form a bootstrap distribution, which can be viewed as an approximation of the sampling distribution of that statistic. Confidence intervals for the statistic can be computed from the bootstrap distribution. If the solution is stable, these confidence intervals will be small. In addition, if there is little bias (i.e., only a small difference between the center of the bootstrap distribution and the observed statistic) and if the bootstrap distribution is approximately normal, these confidence intervals will cover the population parameter with a pre-specified probability (for instance 95%).[1]

A second aspect of inference is the statistical significance of analysis results. Statistical significance is essentially different from sampling stability: Whereas stability refers to whether an observed statistic would vary much

[1]Bootstrap procedures are available in statistical packages, such as JACKBOOT in SAS and macros in SPSS; SYSTAT specifically provides bootstrap results for linear PCA, as well as the asymptotics when appropriate.

over different samples, statistical significance refers to whether the magnitude of that statistic could be attributed to chance (also see Buja & Eyuboglu, 1992). A significant statistic may be unstable, and a stable statistic is not necessarily significant.

Statistical significance can be established nonparametrically by use of permutation tests. With permutation tests, as with the bootstrap, a large number of new samples are drawn from the observed data, but independently for each variable, and without replacement. In other words, the correlational structure of the observed data set is destroyed by permuting the order of the persons independently for each variable in the data set. The resulting permuted samples have a random correlational structure, with the univariate distributions equal to those of the observed variables. The analysis results from the permuted data sets form permutation distributions for each outcome value of interest. Using these permutation distributions, p-values are computed for each observed statistic. If the original sample result takes an extreme value in its permutation distribution, it obtains a small p-value, and is marked significant.

In summary, the bootstrap and permutation tests are used to achieve clearly different objectives, in a similar nonparametric framework based on the data at hand. With the bootstrap procedure, the bootstrap results are favorable if they differ only slightly from the results for the observed sample, whereas with permutation tests, the results for the observed sample should differ substantially from the (center of the) permutation distribution.

1.2.2 Linear PCA

Over time, attention has been paid to parametric inference in the context of linear PCA under the assumption of multivariate normal distributions. For instance, Girshick (1939) and Anderson (1963) have established asymptotic distributions of component loadings derived from the covariance matrix (also see Anderson, 1984). Ogasawara (2004) has derived asymptotic standard errors for component loadings derived from the correlation matrix. However, the multivariate normality assumption made by such approaches may not apply in practice.

Alternatively, different versions of the bootstrap (parametric and nonparametric) have been used to establish the stability of the linear PCA or exploratory factor analysis results (for example, see Efron & Tibshirani, 1993; Lambert, Wildt, & Durand, 1991; Milan & Whittaker, 1995). Timmerman, Kiers, and Smilde (in press) compared the asymptotic approach to the bootstrap approach, and found that the bootstrap is more flexible and under most conditions more accurate than the asymptotic approach.

To establish statistical significance of linear PCA results, permutation tests have been applied (Buja & Eyuboglu, 1992; Landgrebe, Wurst, & Welzl, 2002). Buja and Eyuboglu used the same permutation approach, that is, permuting the variables independently and concurrently, to establish the significance of the eigenvalues as well as the component loadings. We believe that an alternative approach to establish the significance of the contribution of the variables to the PCA solution by permuting the variables independently and sequentially might be more appropriate. Chapter 4 of this thesis is focused on comparing these two permutation strategies.

1.2.3 Nonlinear PCA

For linear PCA, inferential measures may be established under the assumption of multivariate normality. For nonlinear PCA, considering the fact that it does not make distributional assumptions, it seems natural to use a nonparametric approach, such as the bootstrap procedure and permutation tests.

In a not very widely known study by Markus (1994), the effectiveness of the nonparametric bootstrap in establishing the stability of homogeneity analysis (which can be viewed as a special case of nonlinear PCA with only nominal variables) has been examined. She also examined the stability of some nonlinear PCA results with ordinal variables. In this study, 100 bootstrap studies were performed on 100 different samples from the same known population. Coverage percentages – the proportion of times the population value lay within the bootstrap estimated confidence interval – were satisfactory. Also, the standard deviations for category quantifications from homogeneity analysis from some bootstrap samples were compared to corresponding asymptotic standard deviations, assessed by the so-called "delta method" (Gifi, 1990; Meulman, 1984; Van der Burg & De Leeuw, 1988). The bootstrap and asymptotic results were quite similar, but the bootstrap was more accurate in estimating the standard deviations in the population, and more conservative in estimating the variability of the category quantifications. These results are in line with the conclusions from Timmerman et al. (in press) for linear PCA. Markus's study is described more extensively in Chapter 3.

Permutation tests can be applied to assess the statistical significance of the nonlinear PCA results just as they are to linear PCA. In fact, permutation tests have been applied to several other forms of nonlinear multivariate data analysis that use optimal scaling (for example, see De Leeuw & Van der Burg, 1986; Dijksterhuis & Heiser, 1995; Heiser & Meulman, 1994; Meulman, 1992, 1993, 1996). When applying permutation tests to nonlinear PCA, it is important to think about at what point during the permutation process the optimal quantification should take place. If one is interested in the signifi-

cance of the VAF by the nonlinear PCA solution as a whole, as expressed in the eigenvalues, it may be appropriate to permute the variables independently and concurrently (see Buja & Eyuboglu, 1992). In that case, the permuted data sets have no particular correlational structure, and therefore it does not seem sensible to perform optimal quantification on each of these data sets. Instead, optimal quantification may be performed on the observed data, and the quantified data may be subjected to a permutation study. If however, the variables are permuted independently and sequentially, only one variable is permuted at a time while keeping the others fixed. Then, a permuted data set still has a structure, and it makes sense to perform optimal quantification on each permuted data set. These and related issues are addressed in Chapter 5.

1.3 Outline

This thesis is a collection of four individual papers. As these papers will be published separately, some overlap between the chapters is inevitable, especially considering the description of the method. However, the primary content for each chapter can clearly be distinguished:

- Chapter 2 offers an elaborate didactic description of the method of nonlinear PCA. Optimal quantification and analysis levels are explained, and the decisions to be made when applying the method are clarified. Also, the strengths and limitations of the method are discussed. An extensive example, applying the nonlinear PCA program CATPCA (Meulman, Heiser, & SPSS, 2004) to an empirical data set is provided.

- In Chapter 3, the stability of the nonlinear PCA solution is established using the nonparametric bootstrap procedure. First, the definition of stability used in this thesis is clarified, and the validity of the bootstrap procedure to assess this type of stability for nonlinear PCA is discussed. Then, confidence intervals for the quantified variables, and confidence ellipses for the eigenvalues, component loadings, and person scores are established. The balanced bootstrap, bias estimation, rotation, and category merging are discussed. To provide a benchmark to judge the stability of the nonlinear PCA solution, the same procedure is applied to linear PCA.

- Chapter 4 focuses on statistical significance of standard *linear* PCA results. The statistical significance of the contribution of the variables to the PCA solution is assessed nonparametrically by permutation tests, using two different strategies: (1) independently and concurrently permuting all columns (variables) of the data set, and (2) independently

and sequentially permuting the columns, that is, permuting one variable at a time, while keeping the other variables fixed. A simulation study is conducted, varying the component structure and the size of the generated data sets. The statistical significance of the results is judged by an uncorrected 5% significance level as well as two types of significance levels corrected for multiple testing: the Bonferroni correction, and the less well-known control of the false discovery rate (FDR) (Benjamini & Hochberg, 1995). A large number of replications of permutation studies is used to assess proportions of Type I and Type II error for the two permutation strategies.

- In Chapter 5, the permutation strategy that performed the most satisfactorily in Chapter 4 is applied to nonlinear PCA. Optimal scaling is performed for each permuted data set. The data from Chapters 2 and 3 are used again to illustrate the performance of permutation tests in assessing the significance of the variables in nonlinear PCA. In addition, a measure for effect size is proposed. Finally, the results from the permutation test are compared to the bootstrap results from Chapter 3, considering statistical significance of the contribution of the variables to the nonlinear PCA solution.

- Chapter 6 starts with an overall summary and discussion of the results described in Chapters 2 to 5, and then focuses on some suggestions for further research. At the end of the chapter, we give a discussion of the implementation of the methods described in this thesis.