Nonparametric inference in nonlinear principal components analysis: Exploration and beyond

Linting, M.

Chapter 6

General Discussion

6.1 A Short Retrospect

The purpose of this thesis is taking the method of nonlinear PCA beyond its exploratory status. In the introduction, we provided a didactic and practical guide to this method, showing its similarities and differences to linear PCA, and giving an extensive display of an application to empirical data. Subsequently, several ways to perform inference within the essentially exploratory setting of both linear and nonlinear PCA were described. The bootstrap was applied to assess the stability of the nonlinear PCA solution, and permutation tests were used to establish the significance of the contribution of the variables to the PCA solution. The most effective permutation strategy in this context was established, and then applied to the nonlinear PCA solution. Finally, the permutation and bootstrap results were compared, considering the statistical significance of the contribution of variables to the nonlinear PCA solution.

6.1.1 Nonlinear PCA as an exploratory method

Chapter 2 of this thesis is a didactic guide to nonlinear PCA, offering an extensive description of the method including its advantages and disadvantages, followed by an application in which the analytic steps in conducting a nonlinear PCA are described in detail.

Nonlinear PCA shows many similarities to linear PCA, in data-theoretical philosophy as well as output and interpretation. In fact, if all variables are treated numerically in nonlinear PCA, the results are exactly equal to the linear PCA solution. However, nonlinear PCA is more appropriate when: (1) the data set at hand contains nonnumeric variables, and/or (2) the variables in the data set are (or may be) nonlinearly related to each other. If the variables are nonlinearly related, nonlinear PCA will be able to account for more of the variance in the data, and may enhance the interpretation of the solution compared to linear PCA. In addition, nonlinear PCA software provides useful output, even if all variables are numeric and linearly related. For example, the program CATPCA (available in the Categories module in SPSS) provides comprehensive plots of the component loadings and object scores, as well as biplots showing the variables and the objects together in the space spanned by the principal components, and even triplots displaying variables, objects, and category points simultaneously.

When applying nonlinear PCA, it is useful to try out different analysis levels for the variables, preferably starting with a nominal level (if the researcher is willing to allow the quantification that much freedom) and consecutively applying ordinal and numeric restrictions. The resulting solutions should be compared considering fit measures and transformation plots. If the

fit of the solution decreases only little when imposing linear restrictions, nonlinear quantification is not warranted, and may unnecessarily complicate the interpretation. Otherwise, nonlinear quantification may lead to new insights. It is important when comparing solutions, to not only look at the variance-accounted-for (VAF) of the solution, but also consider its interpretation. If using nonlinear analysis levels overcomplicates the interpretation of the results, linear analysis may be preferred. In addition, other fit measures than the VAF, indicated by the sum of the eigenvalues across principal components, may be considered; for example, the multivariate predictability of the data from the (nonlinear) PCA solution (Gower & Blasius, 2005).

In general, nonlinear PCA proved very useful in analyzing the data set at hand by visualizing the relationships between ordinal and nominal variables in an insightful way. Nonlinear PCA is a valuable alternative to linear PCA that can deal with a larger scope of variables without making prior assumptions about the distributions and interrelations of the variables. The method can be a helpful tool in exploring and visualizing the nature of the relations between variables in a data set.

6.1.2 Stability of nonlinear PCA

Nonlinear PCA is viewed as an exploratory method, and does not provide inferential measures. In practice, the distinction between exploratory and confirmatory methods need not be as strict as is sometimes suggested (also see De Leeuw, 1988). Confirmatory methods contain descriptive elements, and exploratory methods may provide inferential tools. One of the aspects of inference is the stability of the solution obtained by a particular data analysis method. In Chapter 3 of this thesis, we examined the stability of the nonlinear as well as the linear PCA solution by applying the nonparametric bootstrap to assess confidence intervals for the eigenvalues, the component loadings, the objects scores (component scores), and, in the case of nonlinear PCA, the quantified variables. In this context, stability is defined as the degree to which the solution stays the same after inflicting small changes to the data.

We concluded that the nonlinear PCA solution can be about as stable as the linear PCA solution, despite the fact that the number of outcome values to be estimated is larger in nonlinear PCA (i.e., for each category, a quantification is estimated). However, categories with small marginal frequencies pose a problem: The quantifications of these categories are quite unstable, in turn influencing the stability of the corresponding variables and objects. This problem can be explained by the fact that persons scoring rare categories do not appear in some bootstrap samples and appear several times in others, causing considerable differences between solutions. Merging categories with

small marginal frequencies is an effective solution to that problem.

Although the results from Chapter 3 are based on one specific empirical data set, they are in line with previous findings about the stability of the results obtained by homogeneity or multiple correspondence analysis, and some outcomes from the SPSS program PRINCALS (the predecessor of CATPCA) (Markus, 1994). Considering these agreeing results, the following guidelines may be (cautiously) formulated:

1. In general (in nonlinear PCA as well as in other analysis methods), smaller samples will lead to less accurate results than larger samples (the analysis level and number of categories of the observed variables being equal). Confidence intervals for small samples will be large, because the variety of solutions for such samples is large. Thus, although a solution resulting from a small sample may deviate severely from the population solution, the bootstrap gives a correct impression of the stability (Markus, 1994). Based on the findings in Chapter 3, and Markus's recommendations, we conclude that the results for samples with a few hundred cases are quite stable.

2. Use a large number of bootstrap samples. According to Markus (1994), using 1000 samples is adequate, as we experienced in the current study. Using modern computers, performing a bootstrap study on such a large number of samples will not take much time: The bootstrap code using SPSS macro files around the CATPCA program takes about 30 minutes to run for a data set with approximately 600 cases (run on a Pentium 4 computer, 3.00 GHz).

3. Use an orthogonal Procrustes procedure to rotate the bootstrap component loadings towards the original component loadings. Otherwise, the stability of the solution is underestimated, particularly if the eigenvalues of the principal components are close together (and thus there is, for instance, no obvious dominance of the first over the second dimension).

4. Merge categories with small marginal frequencies. Such categories can cause instability in the nonlinear PCA solution. The number of observations needed in each category may differ across data sets. In the bootstrap study in Chapter 3, we draw the line at 2.5% of the cases, which is in accordance with the theory behind normal distributions, where the 2.5% of the observations in the tails of the distribution are usually considered extreme. Markus (1994) suggests an absolute minimum of eight observations per category, but also states that even merging categories

with a marginal frequency as large as 45 can be benificial to the stability of the results. Specifically for small samples, limiting the number of categories is recommendable.

5. Besides examining the bootstrap ellipses, also check the distributions of the bootstrap points. These distributions may be quite nonnormal, and thus an ellipse not always gives the best imaginable representation of the spread in the bootstrap cloud.

6.1.3 Statistical significance of the contribution of variables to the linear PCA solution

Where the bootstrap procedure may provide information about the stability of an exploratory analysis method, permutation tests can be used to obtain p-values to assess the statistical significance of particular elements of the solution. In Chapter 4, we used this method to establish the statistical significance of the contribution of the variables to the total VAF of the linear PCA solution. We performed simulation studies to compare two permutation strategies: (1) permuting all variables in the data set independently and concurrently, and (2) permuting the variables independently and sequentially, that is, one variable at a time is permuted while the others are kept fixed. We varied the pre-imposed structure and size of the generated data sets, and used a 5% significance level as well as significance levels corrected for multiple testing with either the Bonferroni correction or control of the false discovery rate (FDR) (Benjamini & Hochberg, 1995). P-values are calculated as $p = (q+1)/(P+1)$, with q the number of permutation values equal to or more extreme than the observed sample value, and P the number of permutations.

Based on the results from the simulation study, we can formulate the following advice for establishing the significance of the contribution of variables:

1. Always use the strategy of permuting the variables independently and consecutively. In that way, the correlations between the permuted variable and the other variables in the data set are changed, but the structure among the other variables remains intact. Theoretically, this is the most sensible approach, because it is focused on the contribution of one variable on top of that of the others, instead of establishing whether the contribution of a variable improves on the contribution of a similar variable from an entirely unstructured data set. This approach, in practice, also rendered more favorable results than permuting all variables independently and simultaneously, considering proportions of Type I and especially Type II error.

2. Do not use the Bonferroni correction of the significance level. Specifically for small data sets (with 100 objects or less), this form of correction leads to a dramatic loss of power.

3. For large data sets, with more than 100 objects, use the FDR correction of the significance level. For smaller data sets, consider the gravity of the risks of making each type of error: If the risk of making a Type I error is considered more serious than the risk of a Type II error, use the FDR correction of the significance level. Otherwise, an uncorrected significance level is to be preferred.

4. In general, it is best to perform permutation tests on data sets with more than 100 objects. For smaller data sets, the power is somewhat low.

5. Do not randomly perform permutation tests on a data set, when no particular structure is assumed. For unstructured data sets, the false discovery rate (FDR) and the family-wise error rate (FWE) are highly inflated.

6. Use at least 999 permutations. With a smaller number of permutations, the lower bound of the p-values that may result from the permutation study becomes too small. For example, with 999 permutations, this lower bound equals $1/1000 = 0.001$, whereas with 99 permutations this bound becomes $1/100 = 0.01$. Thus, in the latter case, if for instance a 1% significance level is used, p can never become smaller than α. When a corrected significance level is applied, one should calculate the smallest corrected significance level that will be used in a particular study, and make sure that the minimum possible p-value is smaller than that level.

6.1.4 Statistical significance of the contribution of variables to the nonlinear PCA solution

In Chapter 5 of this thesis, the strategy of permuting the variables independently and sequentially has been applied to an empirical data set analyzed by nonlinear PCA, ending up with p-values for the contribution of the variables to the VAF. Specifically in the context of permutation tests, the examination of p-values alone is not sufficient, and measures of effect size can provide an important contribution to the interpretation of effects. This specific importance of effect size is due to the fact that p-values in permutation tests have a lower bound of $1/(P + 1)$, that is, all observed values lying outside of the permutation distribution obtain the same p-value, regardless of their distance

to the permutation distribution. Therefore, in Chapter 5, effect size measures have been proposed that complement the information conveyed by the p-values.

One of these effect size measures is simply the contribution of a variable to the total VAF of the solution, indicated by the sum of its squared component loadings across components. This measure is actually an r^2 and thus can be used as a measure for effect size (also see Cohen, 1988). However, this measure of effect size does not take into account that some observed values lie further away from their permutation distribution than others. As an alternative to the VAF, the distance between the observed VAF and the median of the permutation distribution (DMP) can be computed.

For the data set used in Chapter 5 (and in Chapter 2 and 3 as well), the DMP effect size shows only slight differences with the observed VAF in the variables, due to the fact that the median is small for all permutation distributions. Considering that the permutation distributions are established by the permutation of one variable while the structure among the other variables is preserved, the VAF of the permuted variable is expected to only slightly vary around a small value. Thus, the DMP and variable VAF will lead to similar conclusions for most data sets, but DMP is somewhat more nuanced.

In a data set with n observations, the total number of possible permutations of a variable equals $n!$, but some of these permutations are equal, and thus render equal analysis results. The number of possible *different* permutations for a permuted variable j is $n!/\prod_{k=1}^{k_j}(f_k!)$, with k_j the number of different values in variable j, and f_k the frequency of persons scoring a specific value k. When the number of possible different permutations is small, permutation results will show very little spread. In nonlinear PCA, besides numeric variables, ordinal and nominal variables are incorporated. Specifically these nominal and ordinal variables tend to have a smaller number of different categories than the number of observations (n). Thus, the number of different permutations of a variable will usually be smaller in nonlinear than in linear PCA. However, according to the formula above, the number of possible permutations will only become small when the number of objects is small *and* the distribution of an observed variable is very unevenly spread over a small number of categories. In general, such variables are not particularly informative, and pose a problem for any analysis method.

Finally, from the comparison of the results from the bootstrap and permutation studies on the same empirical data we may conclude that the methods may provide essentially different information. Unlike in traditional hypothesis testing assuming normal distributions of the variables, the examination of confidence intervals and the examination of p-values do not lead to the same

conclusions concerning statistical signficance. We advise to use both methods concurrently: the bootstrap to assess confidence intervals, and permutation tests to establish exact p-values for the outcomes of interest.

6.2 Ideas for Further Research

The studies performed in this thesis have answered some questions about inference in nonlinear PCA, but they just as well raised some interesting new questions that may provide a fruitful basis for further research. In this section, we give some suggestions for future research projects.

The bootstrap study in Chapter 3 as well as the permutation study in Chapter 5, both considering inference in nonlinear PCA, were performed on a single empirical data set. Despite the fact that many results from these studies are in line with results from other research – the bootstrap results confirm the findings of Markus (1994) – it is desirable to perform simulation studies to find out how well the proposed inference measures perform in general with nonlinear PCA. For instance, coverage percentages for the bootstrap ellipses and power for permutation tests could be more generally established. Also, the stability of the permutation results might be investigated by performing several permutation studies on the same data set with different starts of the random generator, and comparing the results.

6.2.1 The bootstrap

From the bootstrap study, we concluded that categories with small marginal frequencies pose a problem for the stability of the nonlinear PCA results. One solution for that problem is to merge categories with small marginal frequencies. In particular cases, another option might be to use methods to regularize the optimal transformations of the variables. Spline transformations are one way of doing so (Ramsay, 1988; Winsberg & Ramsay, 1983) (also see Chapter 2), which are already incorporated into the program CATPCA. Other regularization methods might also be implemented.

In the bootstrap study, we found that ellipses are somewhat conservative in representing the distribution of the bootstrap points: Because the points are not evenly divided across the ellipses, the confidence regions tend to become somewhat too large compared to the actual confidence intervals per component. Alternative representations of confidence regions should be examined. For example, convex hulls, which are irregular shapes that more precisely follow the distribution of the points, could be considered. Also, Rousseeuw (1984) suggested to select confidence ellipses with the smallest possible volume

("minimum volume ellipses"), or to use two-dimensional boxplots ("bagplots") (Gardner & le Roux, 2003; Rousseeuw et al., 1999).

The bootstrap procedure may be applied to assess the internal as well as the external stability of analysis results, with internal stability referring to whether the results give a good impression of the sample at hand, whereas external stability refers to the generalizability of the results (also see Greenacre, 1984; Markus, 1994). Besides the bootstrap, other methods of assessing the external stability of the nonlinear PCA solution could be considered. For instance, the multivariate predictability of the observed data from the nonlinear PCA solution (Gower & Blasius, 2005) might be of interest. In the same vein, the .632 bootstrap might be used (Efron, 1983). In this type of bootstrap, the prediction error of a model is estimated in a specific bootstrap sample (training set) for the objects that are not included in that sample (test set).[1] So, this method is a form of cross-validation with varying training and test sets per bootstrap sample. Also, the jack-knife procedure or leave-one-out method may be considered as a specific form of cross-validation, in which one object is left out of the analysis and is predicted using the analysis solution for the others.

6.2.2 Permutation tests

In Chapters 4 and 5, we used permutation tests to assess the significance of the contribution of the variables in respectively linear and nonlinear PCA. There are also other useful applications of permutation studies in this context. For example, Buja and Eyuboglu (1992) used permutation tests to choose the most appropriate number of components to retain in the linear PCA solution. However, the approach these authors used might not be suitable for nonlinear PCA, because in this method, the components are not nested (as opposed to the components in linear PCA). Two alternative approaches may be considered for nonlinear PCA. The first approach starts by performing a p-dimensional nonlinear PCA on a data set, with p a prespecified number of components. Then, do a linear PCA on the optimally quantified variables, and consecutively ignore the first p components. Perform a permutation study (PermD) on the last $m - p$ components (i.e., the residuals), and see whether there are any significant eigenvalues. If so, perform a $p+1$-dimensional nonlinear PCA, and repeat the procedure. If the residuals still contain a significant structure, repeat the analysis in $p + 2$ dimensions, and so on. The second approach is to perform a permutation study on, for example, a p-, $p - 1$-, $p+1$-, and $p+2$-dimensional solution, compare the results, and simply choose

[1] The .632 is a multiplication factor used in estimating the prediction error, which approximates the probability that an object appears in a bootstrap sample of size n.

the dimensionality which gives the best fit and interpretation. This latter approach is somewhat more intuitive, and does not account for the interdependent relationship between the eigenvalues.

Permutation tests could also be applied to establish the significance of the optimal quantification of the variables in nonlinear PCA. One way to do this is to first analyze the original data numerically and with some nonnumeric, for example ordinal, quantification. Then, calculate the differences between the (for example) ordinal and the numeric category quantifications. These differences can be randomly permuted and added to the observed variables, such that each score is changed according to a random "residual" of ordinal over numeric treatment. The permutation distribution is established by analyzing the changed data set numerically. Finally, the results from the nonlinear PCA with the variable treated ordinally can be compared to this permutation distribution.

In practice, the PCA solution is often rotated towards a simple structure, for example by the use of VARIMAX rotation, which does not change the fit of the solution, but only simplifies the interpretation of the solution. In nonlinear PCA, the quantified variables may be rotated in exactly the same way. If the solution for the observed data is rotated, it would seem the most accurate to also optimally rotate the permutation results towards the rotated solution for the observed data, for example using Procrustes rotation. However, for a permuted variable, rotating the results is not expected to make much difference, because this variable in the permutation results has a random relation with the other variables and is expected to obtain small loadings that point in a random direction.

In the permutation study for linear PCA, we used simulated data sets with a particular pre-imposed structure. In this structure, the noise and signal variables were uncorrelated in the population. In further research, other types of data structures, with for example, moderate correlations between subsets of noise and signal variables, might be considered. Investigating such structures might give a more complete picture of the effectiveness of permutation tests in PCA in practice.

In the permutation studies, we focused mainly on the calculation of p-values that are theoretically comparable to the p-values calculated in traditional null hypothesis signficance testing (NHST). However, there has been a trend in the literature recently, expressing some criticism against such traditional NHST (for example, see Cohen, 1994; Killeen, 2005, 2006). The most important argument of these opponents is that the question of how large the probability of the null-hypothesis is, given the observed data, is (wrongly) answered by calculating the probability of the data, given the null-hypothesis.

Killeen (2005) has proposed the measure p_{rep} as an alternative to traditional p-values. This measure gives the probability that the direction of a specific result can be replicated in another study under the same circumstances. In our nonparametric setting, we could easily join this trend by incorporating p_{rep} for any given signed outcome (for example, a component loading). P_{rep} is then calculated by counting the proportion of bootstrap results that point in the same direction (have the same sign) as the observed result. However, the p_{rep} measure itself has already come across some opposition that suggests that this measure shows the same type of problems as the traditional p-value, specifically considering the need to assume a proper prior distribution of the measured effect size (Macdonald, 2005; Wagenmakers & Grünwald, 2006).

In Chapter 5 of this thesis, we proposed an effect size for the contribution of a variable to the nonlinear PCA solution that accounted for the distance from an observed VAF value to the center (median) of the permutation distribution (DMP). This effect size measure appeared to add some – but not much – information to simply taking the VAF (sum of squared component loadings across principal components). Other measures of effect size may be considered. For example, the center of the bootstrap distribution might be taken as an estimate of the population value, and the distance between this bootstrap estimate and the center of the permutation distribution might be calculated as a measure of effect size. This measure may be a somewhat more accurate estimate of the effect size in the population than the DMP, but will probably not differ much from the DMP as well as the VAF value if bias is small, as the center of the permutation distribution is a very small value, particularly if only one variable is permuted while keeping the others fixed.

6.3 Implementation

The elaborate practical discussion of the method of nonlinear PCA given in Chapter 2 of this thesis is meant to guide researchers wanting to discover an underlying structure in a data set containing (possibly) nonnumeric and nonlinearly related variables through the analysis process. Its objectives are to show in which situations nonlinear PCA should be considered as an alternative to linear PCA and to provide guidelines for the decisions to be made when applying this analysis method. Nonlinear PCA is a method that may help researchers in the social and behavioral sciences produce results that are coherent with the data structure and with the nonnumeric measurement levels of the variables. The method of nonlinear PCA is available in a number of software packages, the two most commonly used being SPSS, containing the program CATPCA in the Categories module (Meulman, Heiser, & SPSS,

2004), and SAS which provides the program PRINQUAL (SAS, 1992).

The methods of performing inference for nonlinear as well as linear PCA, described in Chapters 3 to 5 of this thesis, are aimed at taking PCA methods beyond their exploratory image and backing up their results by familiar and easily computable inferential statistics. In the near future, the programs used to perform the bootstrap and permutation studies described in this thesis will be made available to the general public. For now, the bootstrap method is available through SPSS macro files, whereas the permutation procedure is programmed in Matlab code but will be converted to SPSS macro files in the near future. These programs are at the moment suitable for all analysis levels, except the multiple nominal level. On the somewhat longer term, we plan to implement both the bootstrap and the permutation procedure into the SPSS Categories module, and expand them such that they also incorporate the multiple nominal level. These implementations will ensure that a large group of researchers will gain easy access to both stability and statistical signficance measures for nonlinear PCA.