Nonparametric inference in nonlinear principal components analysis: Exploration and beyond

Linting, M.

Chapter 5

The Use of Permutation Tests in Nonlinear Principal Components Analysis: Application

Nonlinear principal components analysis (NLPCA) is a nonlinear analysis method that transforms ordered and unordered categories to numeric values, and simultaneously performs linear PCA. Because in nonlinear PCA the properties of the distributions of the outcome values are unknown, classical statistical approaches cannot be used for hypothesis testing on the results. Alternatively, permutation tests can be used to establish the statistical significance of the contribution of the variables to the nonlinear PCA solution (VAF). In this study, we apply permutation tests in nonlinear PCA to an empirical data set. In our approach, the variables are independently and sequentially permuted, that is, one variable at a time is permuted, keeping the others fixed. Complementary to hypothesis testing, we propose a measure of effect size, based on the difference between the outcome value estimated from the observed data and the center of the permutation distribution. Finally, the permutation results are compared to the results from a previous bootstrap study, considering statistical significance of the contribution of the variables to the nonlinear PCA solution.

5.1 Introduction

Nonlinear principal components analysis is a generalization of linear principal components analysis (PCA). Its goal equals the objective of linear PCA, namely to explore correlational structures in a large set of variables by replacing those variables by a small number of principal components that represent the information in the observed data as closely as possible. The outcomes of nonlinear and linear PCA are much alike, both including eigenvalues, component loadings and person scores (also referred to as component scores or object scores). The main difference between linear and nonlinear PCA is that linear PCA assumes all variables to be numeric and linearly related to each other, whereas nonlinear PCA incorporates variables with ordered as well as unordered categories, and can discover and deal with nonlinear relationships between variables. Because data in the social and behavioral sciences often contain categorical variables that are nonlinearly related to each other, nonlinear PCA can be a valuable alternative to linear PCA. Nonlinear PCA is available as PRINQUAL in SAS (SAS, 1992), and as CATPCA (Categorical PCA) in SPSS Categories (Meulman, Heiser, & SPSS, 2004). In the application in this study, CATPCA is used.

Essentially, every variable can be viewed as a categorical variable, with as many categories as observed values. With numeric variables, the categories are equally-spaced, and can be used as real numbers, whereas in nonnumeric variables, the categories can only be viewed as (ordered or unordered) labels. In nonlinear PCA, the categories of nonnumeric variables are transformed to numeric values, on which linear PCA is performed simultaneously. This transformation process is referred to as *optimal quantification* or *optimal scaling.* The term "optimal" refers to the fact that each quantification is calculated so that it gives the smallest possible loss of information in the transformed variables. In other words, when performing a principal components analysis with a prechosen number of components, nonlinear PCA ensures that these components explain as much variance in the *quantified* variables as possible. This objective is achieved by maximizing the sum of the first p eigenvalues of the correlation matrix of the optimally quantified variables, where p denotes the chosen dimensionality (Gifi, 1990) (also see Chapter 2). In the approach to nonlinear PCA used in this chapter, the sum of squared errors is minimized over the object scores, the quantified variables, and the component loadings, using a least squares loss function. This sum of squared errors is indicated by the difference between the object scores on the one hand, and the component loadings multiplied by the transformed variables on the other. For more information on the mathematical procedure and the nonlinear PCA algorithm used in SPSS, see Gifi (1990), SPSS Inc. (2007) (see also Chapter 2).

The optimal quantification of the variables is carried out in accordance with the analysis (or scaling) levels of the variables. Such an analysis level is specified by the user, and need not be equal to the measurement level of a variable. If only the grouping information in the data is considered important, and nonlinear relations between variables exist, a *nominal* analysis level is called for. If, in addition, the ordering information in the observed variable should be preserved in the category quantifications, and nonlinear monotone relationships between variables are assumed, an *ordinal* level may be preferred. Finally, if the relative spacing between observed category values should be preserved and linear relationships are assumed, a *numeric* level is the most appropriate. The analysis levels and their properties are more extensively described in Chapter 2.

Despite the fact that nonlinear as well as linear PCA are often used in exploratory research, these methods need not be deprived of confirmatory statistics, such as stability measures or p-values. In Chapter 3, we established the stability of linear and nonlinear PCA solutions using the bootstrap procedure. For the particular data set used, we found that, after merging categories with relatively small frequencies, the nonlinear PCA solution was remarkably stable, also when compared to the stability of the linear PCA solution.

Another way of performing inference in PCA is to look at the statistical significance of its results. A nonparametric way to do this is by permutation tests, which involve comparing statistics for the observed data to their null distribution, which is established conditionally on the observed data set itself. A null distribution for a statistic consists of the values for that statistic, computed on a large number of Monte Carlo data sets that are generated by randomly permuting the original variables. Permutation tests have been used in, for example, homogeneity analysis (Heiser & Meulman, 1994), multiple regression and ANOVA (Anderson & Ter Braak, 2003; Ter Braak, 1992), and linear PCA (Buja & Eyuboglu, 1992), but have not been applied to nonlinear PCA before.

In linear PCA, permutation tests have proved to work quite well in determining the significance of the VAF of the solution as a whole (Buja & Eyuboglu, 1992). In Chapter 4, we proposed an effective strategy for assessing the significance of the contribution of the variables to the PCA solution, which approach will be applied to nonlinear PCA in the current study. Because nonlinear PCA does not make distributional assumptions, the nonparametric character of the permutation approach suits this method especially well.

In the remaining part of this chapter, we will first discuss the use of permutation tests in linear PCA, and then explain the strategy we used in the application to nonlinear PCA. Consecutively, we will apply permutation tests to nonlinear PCA on an empirical data set (NICHD Early Child Care Research Network, 1996). We will assess permutation distributions for these data, and establish the significance of the contribution of the variables to the solution. Complementary to p-values, we will propose a measure of effect size based on the distance from an observed outcome to the center of the permutation distribution. Finally, we will compare the permutation results to the results from the bootstrap study in Chapter 3, considering the statistical significance of the contribution of the variables to the nonlinear PCA solution.

5.2 Permutation Tests

The objective of permutation tests is to determine whether an observed statistic deviates significantly from its null distribution. This distribution is not presupposed, but is established nonparametrically from the observed data themselves by generating a large number of Monte Carlo data sets. In each of these data sets, the values of a variable are randomly rearranged, thereby destroying the correlational structure between the observed variables. Because in most data sets, variables are not interchangeable, due to differences in range, scale, or content (Good, 2000), permutation usually takes place within the columns, and not within the rows of a data set. For each permuted data set, the value of the statistic of interest is computed, and all of the computed values form the permutation distribution for the statistic. Then, the alternative hypothesis that the observed statistic deviates significantly from its permutation distribution is tested against the null hypothesis that it does not. A p-value is assessed by computing the proportion of values in the permutation distribution that is equal to or exceeds the observed statistic (Hubert & Schultz, 1976; Hubert, 1984, 1985, 1987; Noreen, 1989).

As the total number of possible permutations is usually huge, a random sample is drawn from all possible permutations. Under the null hypothesis, the observed data set is viewed as just another random permutation. Therefore, instead of the number of permutations P, the number $P + 1$ is taken to be a round number. P should be taken large enough to obtain an acceptable amount of power. For weak effects, Buja and Eyuboglu (1992) recommend 99 or 499 permutations. In the study in Chapter 4, 999 permutations showed satisfactory results, specifically for data sets with between 100 and 200 cases or more.

5.2.1 Permutation tests in linear PCA

Buja and Eyuboglu (1992) constructed permutation distributions for the component loadings in linear PCA by independently and concurrently permuting all the variables (in the columns of the data set), entirely destroying the structure of the observed data. In Chapter 4, we proposed an alternative strategy to establish the significance of the VAF of the variables (i.e., their sum of squared component loadings across the principal components), that is, permuting the variables independently and sequentially (one variable is permuted, while keeping the other variables fixed). When this strategy is applied, only the correlational structure between the permuted variable and the other variables is destroyed, whereas the relationships between the fixed variables are preserved. This strategy helps answer the question whether the variable contributes more to the structure of the other variables than a random variable would, and thus seems theoretically the most sensible when the contribution of the variables to the PCA solution is of interest. On the other hand, the strategy of permuting the entire data set is focused on assessing whether the data set as a whole differs from a random structure.

These two permutation strategies were compared under different data conditions: data sets with a strong, moderate or random structure, varying in size from 20 variables and 100 cases to 40 variables and 500 cases. The authors used a standard significance level of 0.05 as well as two corrections for multiple testing: the Bonferroni correction and controlling the false discovery rate (FDR) (Benjamini & Hochberg, 1995). For determining the significance of the contribution of the variables to the solution, the strategy of permuting one variable while keeping the others fixed, especially when combined with the FDR correction of the significance level, proved to be favorable over permuting the entire data set. The former strategy yielded acceptable proportions of Type I error of around 0.05 or lower for all data conditions (except when the data did not show any component structure). In addition, it had much higher power than the strategy of permuting the entire data set: Permuting a single variable resulted in a power of more than 80% under all data conditions with more than 100 objects, whereas permuting the entire data set resulted in much lower power (even less than 10% in some conditions, when combined with the Bonferroni correction).

5.2.2 Permutation tests in nonlinear PCA

In the current chapter, we will apply the sequential permutation strategy (permuting one variable at a time while keeping the others fixed) to the variables in the *nonlinear* PCA solution. Different approaches might be taken to determine the significance of the nonlinear PCA solution, which are suitable in different situations. The first approach involves applying nonlinear PCA to the observed data set, and consequently performing a permutation study on the *quantified* variables. In other words, the optimal quantification process takes place only once, prior to the permutation process. This approach is sensible when the entire data set is permuted, and thus, the permuted data sets have an entirely random structure, with the variables only related due to chance. Evidently, in such a case, the optimal quantification objective of maximizing the relationships between the quantified variables should not be pursued, because relevant relationships are supposed to be nonexistent. Therefore, it seems insensible to permute the *observed* data set and perform nonlinear PCA (including optimal scaling) on each of the permuted data sets. However, this alternative approach would be preferable in the current study, in which the contribution of the variables to the PCA solution is of interest and we permute only a single variable while keeping the others fixed. In this case, the permuted data set is still expected to show a particular stucture, determined by the fixed variables. Therefore, the optimal quantification process should be performed on each permuted data set. In the current study, we perform 999 permutations per variable, which are used to compute p-values.

5.3 Effect Size

P-values are frequently reported, but they do not provide information about the size of effects. For instance, if the analysis involves a very large sample, quite small effects will obtain small (significant) p-values. Therefore, measures of so-called *practical significance* (effect size) are often reported (see, for example, Gliner et al., 2002) in addition. In the context of permutation tests, p-values can be particularly precarious, because they have a minimum bound. That is, the p-value is the proportion of values in the permutation distribution that is equal to or exceeds the observed value, which can be computed as $p = (q + 1)/(P + 1)$, with q the number of values equal to or higher than the observed value, and P the number of permutations. (In this computation, the 1 is added, because under the null hypothesis, the observed value is also considered to be a random permutation.) Thus, when 999 permutations are performed, the lower bound for the p-value is $(0 + 1)/(999 + 1) = 0.001$. All values that lie outside the permutation distribution obtain this same p-value.

However, some of these observed values will lie quite close to the permutation distribution, whereas others will lie rather far away. Obtaining a measure of effect size based on this distance between the observed value and the (center of the) permutation distribution may be of substantial interest.

Effect size can be defined as the "degree to which the null hypothesis is false" (Cohen, 1988, pp. 9-10). If the null hypothesis is true, the effect size is zero. In the literature, there are many measures of effect size, for example d for the difference of two means in a t-test context, the correlation coefficient r, q for differences between correlation coefficients, g for proportions, h for differences between proportions, f for analysis of variance and covariance, f^2 for multiple regression and correlation analysis, and so on (Cohen, 1988). One of the properties of such effect size measures is that they are pure (dimensionless) numbers, independent of the variable's measurement unit (Cohen, 1988), which can be achieved through standardization. For example, the Pearson correlation coefficient r has this property, as well as r^2. Because in the PCA context, a component loading is a Pearson correlation coefficient between a variable and a principal component, the observed squared component loading may be viewed as a measure of effect size. This measure indicates to which degree the observed VAF differs from zero, and would work well in a traditional hypothesis test setting, because all observed values are supposed to have the same population distribution. However, in the permutation test setting, the permutation distributions differ across variables. As a VAF measure gives no information on the degree to which the observed value differs from the permutation distribution, an additional effect size measure indicating the distance from the observed value to the center of the permutation distribution may be warranted.

In traditional hypothesis testing, the center of the distribution used for testing can be specified beforehand, because the distribution (for example, the normal distribution) is known. However, in the context of permutation testing, the center of the null distribution has to be assessed from the data themselves. The permutation distributions for the VAF of the variables in (nonlinear) PCA can be quite skewed: In many cases, the squared component loadings have a value close to zero, in fewer cases, their value is somewhat higher, and in very few cases, their value may be more substantial. As the mean can be highly influenced by a few outliers in the distribution, the most suitable measure for the center of the permutation distribution would be the median, instead of the mean. The difference between the observed VAF and the median of the permutation distribution may then be an insightful measure of effect size. This difference can be viewed as the difference between two r^2 measures. Cohen (1988) describes this as a measure of effect size that is similar to q, which measures the difference between two correlations, using the Fisher z-transformation. Cohen (1988) also gives guidelines for the size of the difference between two r^2 measures: values between 0.05 and 0.08 indicate small effects, values between 0.15 and 0.23 indicate medium effects, and values between 0.28 and 0.38 can be called large effects. As these indicated ranges show gaps, we slightly adjusted these criteria for the current study: We indicated effect sizes between 0.05 and 0.15 as weak, between 0.15 and 0.28 as medium, and above 0.28 as strong effects.

5.4 Relation between Statistical Significance and Stability

Statistical significance and stability of a solution are often examined concurrently. In traditional hypothesis test settings, a 95% confidence interval gives the boundary values outside of which a two-sided significance test with $\alpha = 0.05$ leads to rejection of the null hypothesis. In other words, if an observed value is more extreme than either of the boundaries of the 95% confidence interval, the null hypothesis is rejected at a two-sided $\alpha = 0.05$. In such a parametric setting, we would assume that the component loadings are approximately normally distributed, and centered at zero. The corresponding null hypothesis to be tested would be that the component loadings do not differ from zero.

Permutation tests and the bootstrap are nonparametric methods to assess the significance and confidence intervals of a specific value. The bootstrap gives an approximation of the population distribution of the parameter of interest from which a (multi-dimensional) confidence region can be established. Following the reasoning applied in traditional hypothesis testing, if a 95% bootstrap confidence interval contains the value 0 on a component, we would expect the corresponding loading to be insignificant on that component at a two-sided significance level of 0.05. If a bootstrap interval does not contain the value zero on a particular component, the loading would be significant on that component at a two-sided significance level of 0.05. Permutation tests give an approximation of the population distribution of parameters from random data, for which a confidence region may also be established. If the observed value lies outside of this confidence region, the results are considered significant. Thus, if the reasoning from traditional hypothesis testing applies, we might assume that both methods would render the same results, and the bootstrap results are sufficient to assess both the stability and the statistical signficance of the component loadings, which would leave permutation tests redundant.

However, permutation tests render specific p-values, which is an obvious advantage over the bootstrap results that would only be able to give a global indication of significance. In addition, it is not sensible to simply generalize rules that apply to traditional significance tests (in which normal distributions are assumed) to nonparametric inference in (nonlinear) PCA. For nonlinear PCA, neither the bootstrap distributions, nor the permutation distributions are normal. The latter will be shown in section 5.5.1. The fact that the distribution of bootstrap points across confidence ellipses can deviate strongly from normal has been discussed in Chapter 3 (also see section 5.5). In addition, we permute only one variable at a time, which renders different results from permuting all variables concurrently, and is less comparable to the bootstrap procedure.

Also theoretically, the results for the bootstrap and permutation tests are not equivalent, but may complement each other. This idea is supported by Buja and Eyuboglu (1992), who noted that significance is concerned with the question of whether the magnitude of a result is likely to be due to chance alone, whereas stability considers the question of whether a result would change much due to slight changes in the data. Theoretically, it is not unlikely that significant loadings are unstable, whereas stable loadings are insignificant.

5.5 Application: The ORCE Data

We used CATPCA to analyze categorical data on 594 6-month olds from the National Institute of Child Health and Human Development Study of Early Child Care (NICHD Early Child Care Research Network, 1996). These children were observed in their primary non-maternal caregiving environment (child care center, care provided in caregiver's home, care provided in child's home, grand-parent care, or father care). In this chapter, we apply our methods to the variables concerning the interactions between the caregiver and the focus child, measured by the Observational Record of the Caregiving Environment (ORCE) (NICHD Early Child Care Research Network, 1996). The ORCE provides "behavior scales" that are the averaged frequencies of specific caregiver behaviors over a particular observation period, as well as "qualitative ratings" that are averaged 4-point rating scales of overall caregiver behavior during an observation period, ranging from 1 ("not at all characteristic") to 4 ("highly characteristic").

As CATPCA is developed for analyzing integers (see Chapter 2), we rounded the scores on the behavior scales and (averaged) ratings to obtain variables with integer values.[1] In this bootstrap study, we found that categories with small marginal frequencies were quite unstable in the CATPCA solution. Therefore, we recoded the ORCE variables such that each category contained at least 15 observations. Only the variable "Negative physical actions" still contained only 8 observations in the second category, and 586 in the first, and remained quite unstable in the bootstrap study. We decided to use these recoded ORCE variables in the current study, anticipating comparisons between the bootstrap and permutation results. Bar charts of these variables are in Figure 5.1.

In accordance with Chapter 3, we performed a two-dimensional CATPCA on the recoded ORCE data, with all of the variables treated ordinally. Ordinal analysis levels were chosen, because we wished to retain the ordering information in the data, but did not assume the variables to be linearly related. The eigenvalues of the first and second dimension are respectively 7.034, indicating a VAF of approximately 33.5%, and 2.028, indicating a VAF of approximately 9.7%. The component loadings of the two-dimensional CATPCA solution for the ORCE data are presented in Table 5.1. The variables form three groups: The first and second group determine the first component, the first indicating a certain degree of positive engagement with the child (PE), and the second a degree of disengagement (DE). The third group, which determines the second dimension, includes variables indicating "overt negative behaviors" toward the child (ON). In Table 5.1, loadings of .30 or higher are in boldface, and the group abbreviations are given in parentheses behind the variable name. All variables except "Flatness" and "Negative physical actions" clearly belong to one of these groups. "Flatness" shows a substantial loading on both components, but as the loading on the first dimension is clearly higher, this variable is most associated with the second group (DE). "Negative physical actions" has a relatively small loading on both components, but distinctly higher on the second than on the first, so that it is associated most with overt negative behaviors (ON). An elaborate description of this nonlinear PCA solution is available in Chapter 2. The interpretation of these component loadings makes sense, but the question is whether or not the loadings are statistically significant. The next section addresses this issue.

[1] We could also have used one of the discretizing options available within CATPCA, but in accordance with the bootstrap study in Chapter 3, we decided to take this very simple approach to obtaining integers.

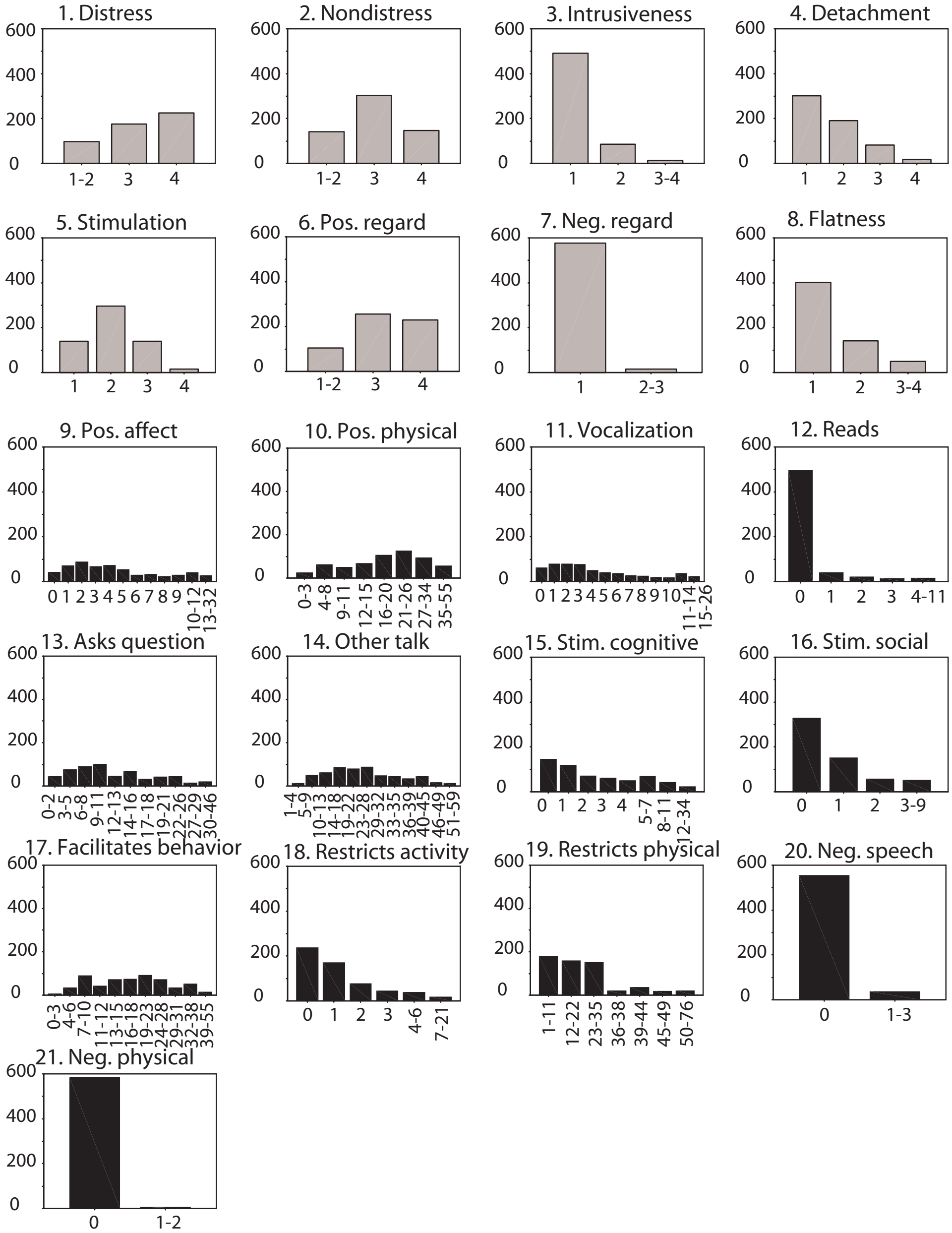

Figure 5.1: *Bar charts for the recoded ORCE variables. N=594. On the x-axis the categories of the ORCE variables after recoding are displayed.*

Table 5.1: *Component loadings for the 21 recoded ORCE variables from a two-dimensional CATPCA. Nr. Cat.= the number of categories of the variable. Loadings >.30 are in boldface. Three groups indicating positive engagement (PE), negative engagement (NE), and overt negative actions (ON) can be distinguished. Group abbreviations are in parenthesis behind the variable names.*

			Component 1		Component 2	
	Variable	Nr. Cat.	Load.	VAF	Load.	VAF
1	Distress (PE)	3	**-.561**	.315	-.253	.064
2	Nondistress (PE)	3	**-.789**	.623	-.205	.042
3	Intrusiveness (ON)	3	.047	.002	**.645**	.416
4	Detachment (DE)	4	**.763**	.582	.142	.020
5	Stimulation (PE)	4	**-.743**	.552	.026	.001
6	Positive regard (PE)	3	**-.793**	.629	-.120	.014
7	Negative regard (ON)	2	-.001	.000	**.613**	.375
8	Flatness (DE)	3	**.514**	.264	**.316**	.010
9	Positive affect (PE)	12	**-.599**	.359	.149	.022
10	Positive physical (PE)	8	**-.628**	.394	-.011	.000
11	Vocalization (PE)	13	**-.702**	.493	.087	.008
12	Reads (PE)	5	**-.335**	.112	-.075	.006
13	Asks question (PE)	11	**-.767**	.589	.114	.013
14	Other talk (PE)	13	**-.852**	.725	.129	.017
15	Stimulates cognitive (PE)	8	**-.724**	.524	.115	.013
16	Stimulates social (PE)	4	**-.352**	.124	.165	.027
17	Facilitates behavior (PE)	11	**-.742**	.550	.196	.039
18	Restricts activity (ON)	6	-.158	.025	**.618**	.382
19	Restricts physical (DE)	7	**.406**	.165	-.083	.007
20	Negative speech (ON)	2	.083	.007	**.629**	.395
21	Negative physical (ON?)	2	.006	.000	.261	.068

5.5.1 *P*-values for the contribution of the ORCE variables

The nonlinear PCA solution for the rounded and recoded NICHD data was subjected to a permutation study in which each variable was permuted 999 times, keeping the others fixed. In other words, $999 \times 21 = 20,979$ permuted data sets were constructed. For each of these data sets, nonlinear PCA was performed with ordinal transformation of the variables.

In Figure 5.2, the permutation distributions of the VAF in the variables across components are displayed. In each plot, the sample value is indicated by a star, and the corresponding *p*-value is given (for the computation of these values, see section 5.2.3). The permutation distributions are all quite skewed,

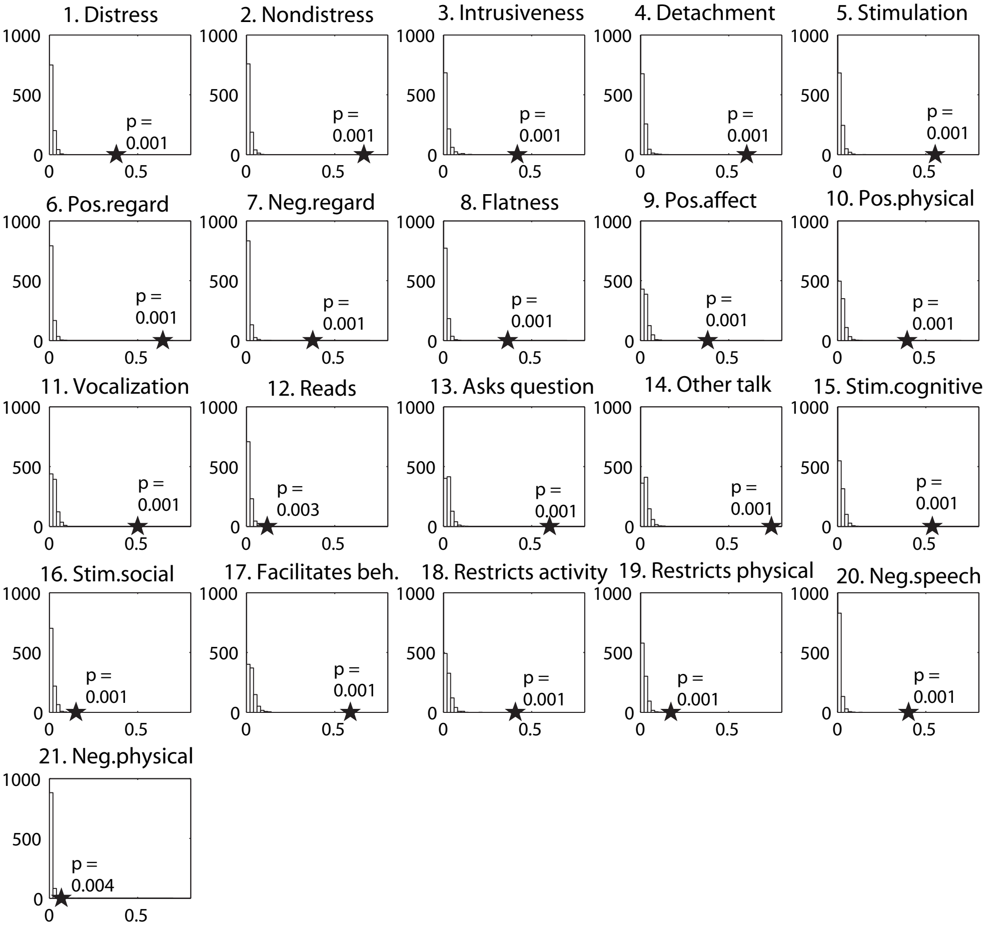

Figure 5.2: *Histograms for the permutation distributions of the VAF (on the x-axis) of the ORCE variables. N=594. The sample values are indicated by stars.*

Table 5.2: *VAF across components, and corresponding* p*-values, and the distance between observed value and median permutation value (DMP).*

Variable	VAF	p	DMP
Distress	0.379	0.001	0.37
Nondistress	0.665	0.001	0.65
Intrusiveness	0.418	0.001	0.41
Detachment	0.602	0.001	0.59
Stimulation	0.553	0.001	0.54
Positive regard	0.643	0.001	0.63
Negative regard	0.375	0.001	0.37
Flatness	0.364	0.001	0.35
Positive affect	0.381	0.001	0.36
Pos. physical	0.394	0.001	0.37
Vocalization	0.500	0.001	0.48
Reads	0.118	0.003	0.11
Asks question	0.602	0.001	0.58
Other talk	0.742	0.001	0.71
Stim. cognitive	0.537	0.001	0.52
Stim. social	0.151	0.001	0.14
Fac. behavior	0.589	0.001	0.56
Restricts act.	0.407	0.001	0.38
Restricts phys.	0.172	0.001	0.16
Negative speech	0.402	0.001	0.40
Neg. physical	0.068	0.004	0.06

as could be expected, because they represent squared values. Variables with few categories (see Table 5.1) show relatively little spread in the permutation distributions, because the distortion of these variables due to permutation is less than for variables with many categories.

The p-values for the VAF across components are also displayed in Table 5.2. (This table will be discussed in more detail in section 5.5.2) This table, along with Figure 5.2, shows that all variables except “Reads” and “Negative physical action” obtain a p- value of $(0 + 1)/(999 + 1) = 0.001$, which is the smallest possible p-value with 999 permutations. This result is due to the fact that all these observed values lie completely outside of their permutation distribution.

The VAF in the variables across components indicates the contribution of the variables to the total VAF of the solution. However, for the interpretation of the solution, it is more interesting to look at the VAF in the variables *per component*. The VAF in the variables by both components separately is displayed in Figure 5.3. This figure shows boxplots for each variable on each principal component. All values larger than $Q_3 + 1.5\text{IQR}$ (i.e., 1.5 times the interquartile range above the third quartile) are displayed as outliers outside the boxplot, and the observed VAF is indicated by an S. The VAF's of "Intrusiveness," "Negative regard," "Restricts activity," and "Negative speech" lie outside the permutation distribution on the second component. For "Negative physical actions," the observed VAF lies within the permutation distribution on both dimensions. The contribution of that variable may therefore be questioned. All other variables show VAF's outside of the permutation distribution on the first but not on the second component. The *p*-values of the VAF of the ORCE variables on both components separately are displayed in Table 5.3. (This table will be discussed in more detail below.) Whether the values corresponding to these *p* values are statistically significant is dependent on the significance level used. This issue will be discussed in section 5.5.3. As mentioned above, *p*-values do not indicate the importance (practical significance) of a variable in the solution. For instance, "Flatness" has rather low VAF (thus little practical significance), especially on the second component, but is statistically highly significant (see the small *p*-values on both components). Therefore, measures of effect size will be the focus of the next section.

5.5.2 Effect sizes for the ORCE variables

In Figure 5.2, the *p*-values are 0.001 for almost all variables. However, the distance from the observed VAF to the permutation distribution varies considerably. In Table 5.2, the difference between the observed VAF and the median of the permutation distribution is displayed as a measure of effect size (see section 5.3). This measure is abbreviated as DMP (distance to the median permutation value). As DMP is calculated by subtracting two r^2 measures, we can apply the criteria derived from Cohen (1988) (see section 5.3) to the DMP values in Table 5.2. Using these criteria, across components, "Negative physical action," "Reads," and "Stimulates social" show weak effects, "Restricts physical" shows a medium effect, and all other variables show strong effects.

Table 5.3 shows VAF and DMP for the variables per component. On the first component, as expected, the variables belonging to the first and second group of variables – indicating degree of positive engagement with the child – show strong effects, except "Flatness" and "Restricts physical" which show medium effects, and "Reads" and "Stimulates social" which show weak effects. The variables from the third group indicating overt negative behaviors ("Intrusiveness," "Negative regard," "Restricts activity," "Negative speech," and "Negative physical") show no effect on the first component. However

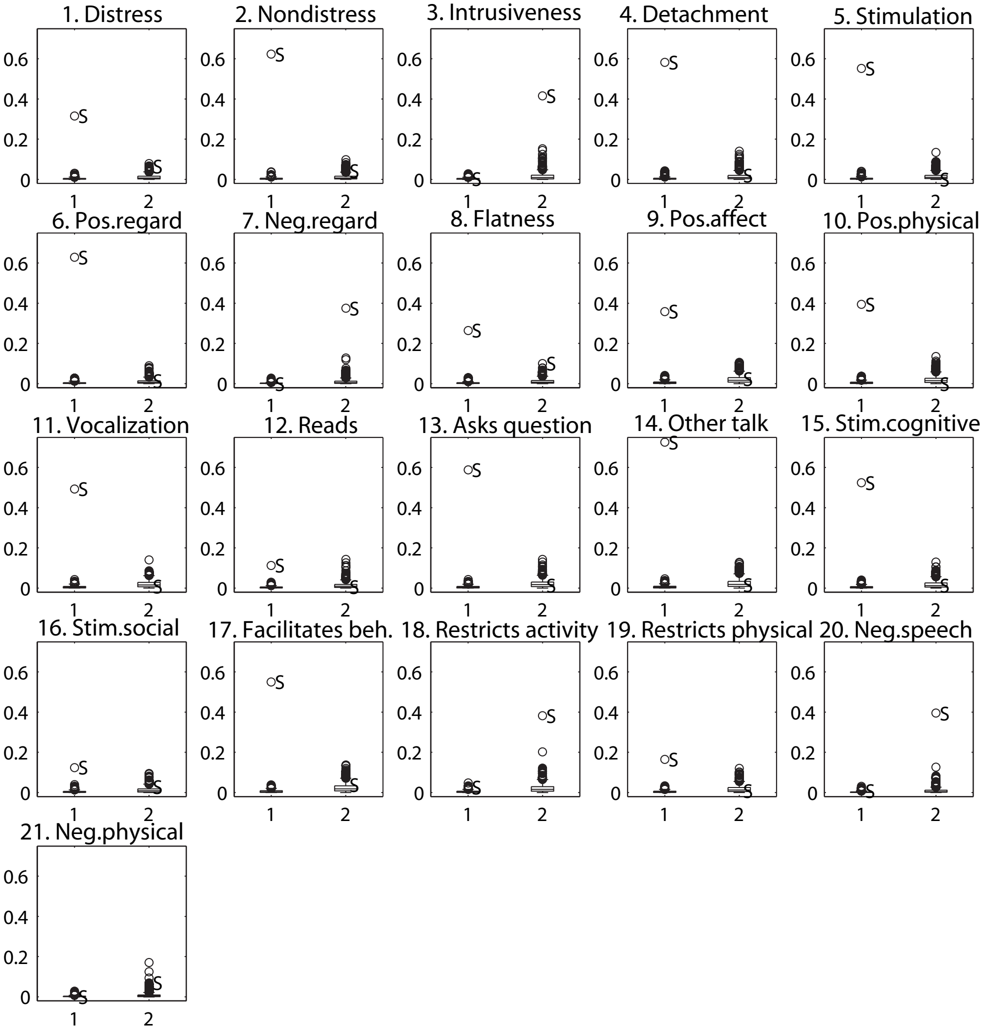

Figure 5.3: *Boxplots for the permutation distributions of the VAF for the ORCE data per dimension. N=594. The observed sample values are indicated by the letter S. On the x-axis, the dimension numbers are displayed.*

Table 5.3: *VAF per component, and corresponding p-values, distance from observed value to the median permutation value (DMP), and FDR corrected signficance levels (c_{fdr}) for the ORCE variables. Rank numbers for assessing FDR significance (r_p) are assigned over components. Significant VAF's after FDR correction are in boldface.*

	Component 1					Component 2				
Variable	VAF1	p	DMP	r_p	c_{fdr}	VAF2	p	DMP	r_p	c_{fdr}
Distress	**0.315**	0.001	0.31	1	0.0095	**0.064**	0.004	0.06	22	0.0274
Nondistress	**0.623**	0.001	0.62	2	0.0024	0.042	0.038	0.04	25	0.0298
Intrusiveness	0.002	0.445	0.00	32	0.0381	**0.416**	0.001	0.41	17	0.0226
Detachment	**0.582**	0.001	0.58	3	0.0036	0.020	0.253	0.01	30	0.0357
Stimulation	**0.552**	0.001	0.55	4	0.0048	0.001	0.916	-0.01	40	0.0476
Positive regard	**0.629**	0.001	0.63	5	0.0012	0.014	0.248	0.01	29	0.0345
Negative regard	0.000	0.965	0.00	41	0.0488	**0.375**	0.001	0.37	18	0.0179
Flatness	**0.264**	0.001	0.26	6	0.0119	**0.100**	0.001	0.09	19	0.0250
Positive affect	**0.359**	0.001	0.36	7	0.0155	0.022	0.387	0.01	31	0.0369
Pos. physical	**0.394**	0.001	0.39	8	0.0143	0.000	0.969	-0.01	42	0.0500
Vocalization	**0.493**	0.001	0.49	9	0.0131	0.008	0.712	-0.01	38	0.0452
Reads	**0.112**	0.001	0.11	10	0.0214	0.006	0.608	0.00	36	0.0429
Asks question	**0.589**	0.001	0.59	11	0.0071	0.013	0.599	0.00	35	0.0417
Other talk	**0.725**	0.001	0.72	12	0.0060	0.017	0.541	0.00	34	0.0405
Stim. cognitive	**0.524**	0.001	0.52	13	0.0083	0.013	0.476	0.00	33	0.0393
Stim. social	**0.124**	0.001	0.12	14	0.0202	0.027	0.140	0.02	27	0.0321
Fac. behavior	**0.550**	0.001	0.55	15	0.0107	0.039	0.196	0.02	28	0.0333
Restricts act.	**0.025**	0.010	0.02	24	0.0286	**0.382**	0.001	0.37	20	0.0238
Restricts phys.	**0.165**	0.001	0.16	16	0.0190	0.007	0.644	-0.01	37	0.0440
Negative speech	0.007	0.086	0.01	26	0.0310	**0.395**	0.001	0.39	21	0.0167
Neg. physical	0.000	0.906	0.00	39	0.0464	**0.068**	0.004	0.06	23	0.0262

all of these variables, except "Negative physical") show strong effects on the second component. All but two ("Distress" and "Flatness") of the variables from the first and second group show no effect on the second component. In general, the DMP shows quite well which variables are important on the first and second component. Most variables only have an effect on one of the two components. The importance of the variables "Reads," "Stimulates social development," and "Negative physical action" might be questioned.

For all permutation distributions in this example, the median of the permutation distribution is a small value, resulting in similar orderings and magnitude of DMP and VAF. However, these measures are not exactly equal: Variables with higher VAF do not necessarily show a larger difference between observed VAF and the center of the permutation distribution. Thus, VAF and DMP are not simply interchangeable (which will be explained in the Discussion).

5.5.3 Significance of the contribution of the ORCE variables to the nonlinear PCA solution

Whether the ORCE variables have significant VAF should be assessed using a particular significance level, for example, the conventional 0.05 or 0.01 level, or some corrected level. In Chapter 4, we showed that permutation of separate variables combined with controlling the false discovery rate (FDR) (Benjamini & Hochberg, 1995) leads to acceptable proportions of both Type I and Type II errors with structured data sets. Also, the procedure is theoretically sound in the context of exploratory data analysis (Keselman et al., 1999; Verhoeven et al., 2005). The FDR procedure involves sorting the p-values for all of the variables in an ascending order, and, starting with the highest p-value, testing each p-value by a significance level of $(r/t)\alpha$, with t the number of tests (in this case equal to the number of variables m), and r the rank number of the p-value. Thus, smaller p-values are subjected to stricter significance levels.

With permutation tests, all observed values that lie outside of the permutation distribution obtain the same p-value, because p-values have a minimum bound of $1/(P+1)$. In other words, ties in the ranking of p-values occur. With FDR correction, the comparison of p-values to the FDR corrected significance level (c_{fdr}) starts with the highest p-value (with the largest rank number). Going down the list of p-values, the results corresponding to the first p-value smaller than c_{fdr}, as well as the results corresponding to all p-values with smaller rank numbers are marked significant. Thus, if p-values are tied, they will either all indicate signficance or all insignificance, and the ordering within groups of equal p-values is irrelevant. We simply assigned rank numbers to equal p-values in the order of the variables in the data set.

The VAF for the variables across components, as displayed in Figure 5.2, is significant for all variables, which is immediately evident, as the smallest FDR controlled significance level is equal to $(r/t)\alpha = (1/21)0.05 = 0.0024$, and all p-values are below that value.

In Table 5.3, the significance of the VAF of the variables is determined for the components separately, using the FDR corrected significance level. Significant VAF values are in boldface. Because the p-values for the two principal components are established on the same data, they are sorted across components, in an ascending order (the smallest p-value obtaining the lowest rank number). The FDR significance level is calculated as $(r/t)\alpha = (r/2m)\alpha = (r/42)0.05$. From rank number 24 down, all p-values indicate significance. Table 5.3 shows that the variables from the first and second group (degree of positive engagement with the child) load significantly on the first and not on the second component, whereas the variables from the third group (overt negative behaviors) load significantly on the second and not on the first component. The only exceptions to this rule are "Distress" and "Flatness" from the first group and second group, which also have significant (although small) loadings on the second dimension, and "Restricts activity" with a (small) significant loading on the first dimension. In other words, most variables in this data set load significantly on only one dimension.

5.6 Permutation and Bootstrap Results Compared

In Chapter 3, a bootstrap study was performed on the ORCE data to establish the stability of the nonlinear PCA results, including the component loadings. We compare the results from this bootstrap study to results from the current chapter to learn more about the relationship between the bootstrap and permutation tests.

To find out how the bootstrap and permutation tests relate in the nonparametric practice, we compare the p-values for the squared component loadings as reported in the current chapter to the 95% bootstrap confidence intervals for the component loadings obtained from Chapter 3. The bootstrap results are computed for the component loadings, whereas the p-values are calculated from the permutation results for the squared component loadings. This difference in calculation does not pose a problem, because the relative position of the loadings is equal to the relative position of the squared loadings: If, for example, four loadings in the permutation distribution are more extreme (farther from zero) than the observed loading, also four squared loadings will be more extreme than the observed squared loading, and thus, in both situations, the resulting p-value would be $5/1000 = 0.005$ (with 999 permutations).

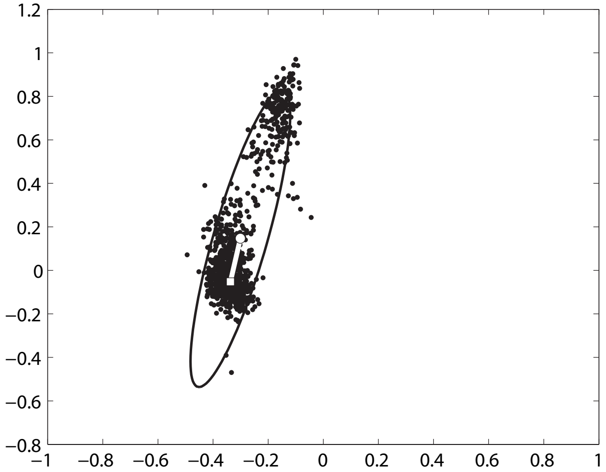

Figure 5.4: *Distribution of the bootstrap points of the variable "Reads" from the* ***original*** *ORCE data within the corresponding confidence ellipse. The white square indicates the center of the bootstrap distribution, and the white circle indicates the observed component loading point. The white bar indicates bias.*

The 95% bootstrap confidence intervals and corresponding p-values for the component loadings obtained for the recoded ORCE variables are in Table 5.4. We focus on the bootstrap intervals per component instead of the graphical representation by two-dimensional ellipses, because the permutation results are also per component, and ellipses can be somewhat conservative in displaying the distribution of the bootstrap points when the distribution of the bootstrap points across the ellipse is not normal. An example of an extremely nonnormal bootstrap distribution is given in Figure 5.4. This figure displays the confidence ellipse of the variable "Reads" from the *original* ORCE data set (before recoding). The bootstrap points clearly form two subgroups: the group of points with high loadings on the second component is obtained from bootstrap samples containing relatively many children that experienced overt negative interaction, and the group with low loadings on the second component corresponds to bootstrap samples obtaining relatively few of such children. Because such "dual" solutions are not desirable, and should not be interpreted, we recoded the ORCE variables in Chapter 3. However, the possible occurrence of such nonnormal bootstrap distributions indicates that assuming (multivariate) normally distributed results is not always sensible with nonparametric confidence intervals.

In the comparison of the bootstrap confidence intervals and the permu-

Table 5.4: *Lower (Low) and upper (Up) boundaries of the 95% bootstrap confidence intervals (BCI) for the component loadings for the* **recoded** *ORCE variables in a two-component nonlinear PCA, and corresponding p-values from a permutation test. BCI's that do not contain zero and p-values smaller than 0.05 are in boldface.*

	Component 1					Component 2				
		BCI					BCI			
Variable	Load	Low	Up	VAF	*p*	Load	Low	Up	VAF	*p*
Distress	-.561	**-.618**	**-.493**	.315	**.001**	-.253	**-.355**	**-.121**	.064	**.004**
Nondistress	-.789	**-.818**	**-.758**	.623	**.001**	-.205	**-.290**	**-.116**	.042	**.038**
Intrusiveness	.047	-.048	.138	.002	.445	.645	**.503**	**.739**	.416	**.001**
Detachment	.763	**.725**	**.780**	.582	**.001**	.141	**.040**	**.240**	.020	.253
Stimulation	-.743	**-.784**	**-.702**	.552	**.001**	.026	-.104	.142	.001	.961
Positive regard	-.793	**-.821**	**-.762**	.629	**.001**	-.120	**-.200**	**-.046**	.014	.248
Negative regard	-.001	-.072	.065	.000	.965	.613	**.437**	**.732**	.375	**.001**
Flatness	.514	**.440**	**.580**	.264	**.001**	.316	**.190**	**.425**	.100	**.001**
Positive affect	-.599	**-.657**	**-.550**	.359	**.001**	.149	**.017**	**.296**	.022	.378
Positive physical	-.628	**-.690**	**-.569**	.394	**.001**	-.011	-.162	.128	.000	.969
Vocalization	-.702	**-.748**	**-.664**	.493	**.001**	.087	-.020	.189	.008	.712
Reads	-.335	**-.412**	**-.252**	.112	**.001**	-.075	-.200	.185	.006	.608
Asks question	-.767	**-.806**	**-.734**	.589	**.001**	.114	.000	.217	.013	.599
Other talk	-.852	**-.875**	**-.828**	.725	**.001**	.130	**.044**	**.200**	.017	.541
Stimulates cognitive	-.724	**-.762**	**-.687**	.524	**.001**	.115	**.024**	**.214**	.013	.476
Stimulates social	-.352	**-.427**	**-.277**	.124	**.001**	.165	-.023	.357	.027	.140
Facilitates behavior	-.742	**-.780**	**-.708**	.550	**.001**	.196	**.096**	**.279**	.039	.196
Restricts activity	-.158	**-.230**	**-.072**	.025	**.010**	.618	**.501**	**.715**	.382	**.001**
Restricts physical	.406	**.334**	**.493**	.165	**.001**	-.083	-.247	.119	.007	.644
Negative speech	.083	**.013**	**.154**	.007	.086	.629	**.471**	**.733**	.395	**.001**
Negative physical	.006	-.062	.070	.000	.906	.261	**.060**	**.464**	.068	**.004**

tation *p*-values, we assume the simple null hypothesis that the component loadings do not differ from zero. In accordance with that assumption, bootstrap confidence intervals *not* containing the value zero as well as *p*-values smaller than 0.05 are displayed in boldface in Table 5.4. Using a one-sided 5% significance level for the squared component loadings is equivalent to a two-sided 5% significance level for the component loadings, as the upper 5% tail of the permutation distribution for the squared component loadings will contain the upper and lower 2.5% of the distribution for the component loadings together.

Table 5.4 shows that, mostly, the results from the permutation tests and bootstrap study agree with each other: Bootstrap confidence intervals *not* containing the value zero correspond with p-values smaller than 0.05, and confidence intervals containing the value zero correspond with p-values larger than 0.05. However, there are some exceptions: The variable "Negative speech" has a confidence interval for the loading on the first dimension that does not contain the value zero, but the p-value for its VAF on that component exceeds 0.05. On the second component, the same apparent contradiction applies to the variables "Detachment," "Positive affect," "Other talk," "Stimulates cognitive," and "Facilitates behavior." Thus, bootstrap confidence intervals and permutation p-values do not always lead to the same conclusion concerning significance. However, the lower boundaries of the confidence intervals for all the variables mentioned above are very close to zero, and the corresponding VAF values are relatively small (ranging from 0.013 to 0.039).

From Table 5.4, we can also conclude that stability and small p-values, and instability and large p-values do not necessarily go together. Some variables, like "Intrusiveness," "Negative regard," and "Negative speech" have quite small confidence intervals on the first component, but show p-values larger than 0.05. In addition, some variables – for instance, "Negative regard" and "Negative physical" on the second component – show relatively large confidence intervals, but still have p-values smaller than 0.05. In Chapter 3, we started performing the bootstrap study on the original ORCE data (before recoding), and found that some variables – specifically, "Intrusiveness," "Negative regard," "Flatness," "Positive physical," "Reads," "Stimulates social," "Restricts activity," "Restricts physical," and "Negative speech" – had very unstable loadings on the second component, due to categories with relatively small marginal frequencies (see Figure 3.3). We did a permutation study on the original ORCE data to find out whether these variables also showed problems with establishing the statistical significance. We found that the VAF values on the second component of these unstable variables do not per se show small p-values: "Intrusiveness" ($p = 0.003$), "Flatness" ($p = 0.019$), "Restricts activity" ($p = 0.011$), and "Negative speech" ($p = 0.011$) all have p-values smaller than 0.05.

In conclusion, obviously there is a stronger relation between the significance of a variable's VAF and the magnitude of the VAF value than between the significance and the stability of a VAF value. In general, relatively high VAF values are significant, although it is not always true that higher VAF values obtain smaller p-values, due to differences in permutation distributions.

5.7 Conclusions and Discussion

Permutation tests seem to be an effective method for assessing the statistical significance of the contribution of variables to the nonlinear PCA solution. The results of this study show that, as expected, high VAF's turned out significant, and low VAF's did not. The distance from the observed value to the median permutation value (DMP) as a measure of effect size, judged by Cohen's (1988) criteria, offers information additional to *p*-values, that is useful for the interpretation of the results.

The information derived from the DMP effect size measure did not differ much from the information that can be derived from looking at the VAF in the variables: Especially for the variables that contributed highly to the solution, VAF values and DMP values were almost equal. This small difference is due to the fact that the median of the permutation distribution was a small value for all variables (which will probably be the case in most data sets, because permutation distributions are expected to have small spread). Thus, only a small value was subtracted from the VAF in all cases. However, the ordering of the variables according to VAF did not perfectly correspond to the ordering of the variables according to effect size, and thus the two measures provide somewhat different information.

As we would expect, for the ORCE data, significant VAF's always show some degree of effect (weak, medium or strong), and insignificant VAF's show no effect at all (see Table 5.3). Specifically in the context of permutation testing, effect size provides information complementary to *p*-values (or significance), because all variables with VAF's lying outside of the permutation distribution obtain the same *p*-value.

The distance from the observed value to the median of the permutation distribution is not the only plausible measure of effect size for permutation results. Depending on the question to be answered, alternatives, like the distance from the observed value to the maximum or mean permutation value, or a distance measure divided by a spread measure, such as the interquartile range or the standard deviation, might also work well. Another option might be to use the mean result from a bootstrap study (see, for example, Chapter 3) and compare that to the mean of the permutation distribution.

In the data set used in this study, almost all variables had a significant VAF on only one of the two components. In other data sets, however, this might not be the case, and rotation might be warranted to obtain a simple interpretation of the components. If the solution is rotated, the VAF of the variables across components stays the same, but the VAF of a variable per dimension changes. In case the researcher wishes to rotate the original solution (for example toward a simple structure), it seems the most accurate to rotate the permutation solutions toward the rotated original solution. However, rotation will probably not make much difference for the permutation distribution of a variable, because in unrotated as well as rotated distributions, the permuted variable is expected to obtain small loadings, independent of the observed data structure.

The number of possible different permutations with nominal or ordinal variables is usually smaller then with numeric variables, because numeric variables mostly have a larger number of different values. In a data set with n persons, the number of possible permutations of a variable is $n!$. However, in case a variable has less than n different values, some of these permutations are equal. The number of *different* permutations when one variable (j) is permuted 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. For instance, if a variable has three categories with respective frequencies 10, 6, 4, the number of different permutations is $20!/(10! \times 6! \times 4!) = 38,798,760$. In such a case, although the number of different values with nonnumeric variables is mostly smaller than with numeric values, the total number of permutations is still very large, and the risk of obtaining the same permutation several times is quite small. Only in cases when n is small, and almost all persons scored the same category, the number of possible permutations gets small, and caution is warranted. Note that in general, such variables with small variation, do not give much information, and are difficult to analyze properly with any analysis method.

The study in this chapter is an application of the results from a simulation study on permutation tests in linear PCA. To find out whether the strategy that performed best for linear PCA also performs best for nonlinear PCA, such a simulation should also be performed with permutation in nonlinear PCA.

In this study, we used permutation tests to assess the significance of the contribution of the variables in nonlinear PCA, but there are other useful applications of permutation studies in this context. For example, Buja and Eyuboglu (1992) used permutation tests to choose the appropriate number of components to retain in the linear PCA solution. This procedure may also be applicable to nonlinear PCA, although some adaptations may be needed.

The permutation procedure used in this study is programmed in Matlab code. Shortly, we will implement this code along with the bootstrap code into the CATPCA module in SPSS, which will make these procedures accessible to a large group of researchers. This effort will hopefully promote a wider use of nonlinear PCA by enabling researchers to report common and easily interpretable inferential statistics, which will render more easily publishable research results for categorical data.

We compared the 95% confidence intervals from the bootstrap study to the p-values from the permutation study. Considering traditional null hypothesis testing theory, this is a valid comparison. However, as we did several hypothesis tests on the same data, a corrected significance level should be used to assess the significance of the VAF values instead of the uncorrected 5% significance level. FDR correction has proved to be the most effective in linear PCA, and was used in Table 5.3. Multistage comparison procedures for assessing significance, such as controlling the FDR, have no straightforward confidence interval interpretations within traditional testing situations, but more complicated intervals can sometimes be constructed (see Shaffer, 1995).

The stability and statistical significance of the nonlinear PCA results should be taken into account concurrently. The bootstrap results cannot be used as a substitute for the permutation results regarding the significance of a variable's contribution to the total VAF of the nonlinear PCA solution. This conclusion especially applies to the current study, because we used the strategy of permuting one variable at a time, instead of the entire data set. Additionally, the specific p-values resulting from the permutation test are more informative than the bootstrap intervals considering the significance. On the other hand, permutation tests do not give information about the stability of analysis results. The significance of a variable's VAF is more related to the magnitude of the VAF value than to the stability. So, if researchers are interested in the stability and the significance of (nonlinear) PCA results, they should consider both the bootstrap and permutation tests valuable inferential equipment.