Nonparametric inference in nonlinear principal components analysis: Exploration and beyond

Linting, M.

Chapter 4

The Use of Permutation Tests in Linear Principal Components Analysis

Principal components analysis (PCA) is often used in exploratory research and does not standardly provide inferential statistics. In this chapter, the statistical significance of the contribution of variables to the PCA solution is assessed nonparametrically by the use of permutation tests. We compare two different strategies. The first involves permuting the columns (variables) of a data matrix independently and concurrently. This strategy destroys the entire correlational structure of the data set, and is considered appropriate for assessing the significance of the PCA solution as a whole. However, for assessing the significance of the contribution of single variables, we propose an alternative strategy, which involves permuting one variable at a time, while keeping the other variables fixed. We conduct a simulation study, in which the two strategies are compared, considering proportions of Type I and Type II error. We use two corrections for multiple testing: the Bonferroni correction and controlling the False Discovery Rate (FDR). For the assessment of the significance of the contribution of the variables in PCA, permuting one variable at a time, combined with FDR correction, yields the most favorable results.

This chapter has been co-authored by Bart Jan van Os and Jacqueline J. Meulman from the Data Theory Group, Leiden University, the Netherlands.

4.1 Introduction

Principal components analysis (PCA) is a nonparametric analysis method frequently used in the social and behavioral sciences to reduce a large number of variables to a smaller number of uncorrelated underlying variables – called principal components – that contain as much information from the observed variables as possible. This objective can be achieved in different ways; in the present study, we use the eigenvalue decomposition of the Pearson correlation matrix of the observed variables. The advantage of analyzing the correlation matrix instead of the covariance matrix is that differences in variance between the variables do not have an influence on the analysis results, that is, the principal components are not sensitive to the measurement units of the variables.[1] In practice, PCA is often treated as a form of exploratory factor analysis. However, PCA has a different statistical background as well as a different objective compared to factor analysis: If the goal of the analysis is to optimally reduce a large number of variables to a smaller number of composite variables, instead of deriving a parsimonious model of the correlational structure between the variables, PCA is the more appropriate procedure (Fabrigar et al., 1999).

Despite the fact that PCA is often treated as an exploratory analysis method, it need not be deprived of inferential diagnostics. In Chapter 3, for example, we established the stability of the PCA solution using the bootstrap procedure (also see, for example, Timmerman et al., in press). In addition, asymptotic confidence intervals for the component loadings have been established for PCA based on the covariance matrix (Anderson, 1963; Girshick, 1939) as well as the correlation matrix (Ogasawara, 2004). Another way of performing inference in this context is to establish the statistical significance of the PCA results by using permutation tests (Buja & Eyuboglu, 1992).

In the current chapter, we will study the behavior of permutation tests to establish the statistical significance of the contribution of the separate variables to the PCA solution. Permutation tests involve generating new data sets by randomly and independently permuting the columns of the data matrix, which contain the variables. Subsequently, parameters of interest are computed for each of the permuted data sets. These permutation results are used to compose a permutation distribution for each of these parameters, and the observed values can be compared to the corresponding permutation distributions. Different strategies for permutation may destroy the structure of the data set to a greater or lesser extent.

[1] When PCA is performed on the covariance matrix, variables with relatively much variance dominate the first few principal components (Jolliffe, 2002).

The development of permutation tests for nonparametric inference has quite a long history. The procedure originated as a nonparametric alternative to the t-test for comparing the means of two different groups of objects (see Fisher, 1935; Good, 2000). The biologist Mantel (1967) developed a significance test for the congruence of two distance matrices, by constructing a null distribution from random permutations of the rows and columns of one of the matrices, while keeping the other matrix fixed. Many extensions and generalizations of this idea have been proposed (for example, see Dietz, 1983; Smouse, Long, & Sokal, 1985), and the idea has been applied and modified in several research fields, such as geography (Glick, 1979; Sokal, 1979), ecology and evolutionary research (Douglas & Endler, 1982), and psychometrics and classification (for example, see Hubert & Schultz, 1976; Hubert, 1984, 1985, 1987). Permutation tests have proved useful in multiple regression and ANOVA (Anderson & Ter Braak, 2003; Ter Braak, 1992), and in several forms of nonlinear multivariate analysis that use optimal scaling (for example, see De Leeuw & Van der Burg, 1986; Heiser & Meulman, 1994; Meulman, 1992, 1993, 1996). Permutation tests were also used to establish the significance of PCA results (Buja & Eyuboglu, 1992; Landgrebe et al., 2002).

In the present study, we consider two different permutation strategies. The first involves permuting the variables independently and concurrently, and was used by Buja and Eyuboglu (1992) to establish the significance of the eigenvalues (indicating the variance-accounted-for by the principal components in the entire data set), for which purpose we consider this particular strategy appropriate. However, these authors assessed the significance of the component loadings (indicating the variance-accounted-for per variable) by using the *same* strategy. We propose an alternative strategy for this latter purpose, that is, permuting the variables independently and sequentially, that is, permuting one variable at a time, while keeping the others fixed.

First, we will briefly discuss permutation tests in general. Then, we describe the two permutation strategies mentioned above to assess the significance of the VAF per variable. Next, we explain the design of the simulation study we conducted to compare the effectiveness of these strategies, and finally, on the basis of this simulation study, we compare the proportions of Type I and Type II error from both strategies. Since multiple testing is involved, we include correction of the significance level with the well-known Bonferroni correction, and with the somewhat less familiar correction method of controlling the False Discovery Rate (FDR) (Benjamini & Hochberg, 1995).

4.2 The Use of Permutation Tests in PCA

Permutation tests are used for nonparametric hypothesis testing. The objective of permutation tests is to determine whether an observed statistic deviates significantly from its null distribution, which is established conditionally on the data, and does not require a particular probability model. This characteristic of the permutation procedure matches the nonparametric nature of PCA, which makes the procedure suitable for inference in PCA.[2]

To establish a null distribution, first, the correlational structure of the observed data is destroyed by randomly rearranging the values within each variable (independent of the other variables). In this way, a data matrix of the same size as the original data matrix is constructed, with a random structure. If the variables are assumed to be interchangeable on the assumption of shared marginal distributions between variables, the data may be fully permuted between rows as well as columns. However, this assumption is unrealistic in most cases (Buja & Eyuboglu, 1992), because variables mostly differ in content and scaling. Therefore, usually, the data are only permuted *within the columns* of the data set, on the assumption of shared marginal distributions between the objects (Buja & Eyuboglu, 1992), or interchangeability of the persons (Good, 2000). This permutation process is repeated a large number of times. A null distribution for each parameter of interest is then composed from the parameter values estimated for the permuted data sets.

Finally, the alternative hypothesis that the observed value deviates significantly from the center of its null distribution is tested against the null hypothesis that it does not. This test is executed by calculating the proportion of the values in the permutation distribution that is equal to or exceeds the observed statistic (the p-value). The p-value is then, as usual, compared to a prechosen significance level α: If $p < \alpha$, the result is called significant.[3] A p-value is computed as $p = (q+1)/(P+1)$, with q the number of times a statistic from the permutation distribution is greater than or equal to the observed statistic, and P the number of permutations (Buja & Eyuboglu, 1992; Noreen, 1989). Because, under the null hypothesis, the observed data are assumed to be just another permutation of a random data set, the denominator in this equation is $P+1$ rather than P. The total number of possible permuted data

[2] In this study, we focus on the use of permutation tests with two-dimensional PCA solutions. For high-dimensional solutions, the results might be different, for when the dimensionality approaches the number of variables (maximum dimensionality), the eigenvalues of each of the components will all approach 1, and will thus become (almost) equal to each other.

[3] In the literature, the null hypothesis is often rejected if the p-value is smaller than or *equal to* the significance level, but in the current chapter, we chose the slightly more conservative rule of rejecting when p is smaller than the significance level.

sets is $n!^{m-1}$, with n the number of persons and m the number of permuted variables. Because this number increases rapidly with the number of persons and variables, usually a random sample of the total set of permutations is used. For weak effects, or for reporting p-values in publications, Buja and Eyuboglu (1992) suggest using 99 or 499 permutations, because permutation tests with a smaller number of permutations than 99 have too little power.

4.2.1 Two permutation strategies

Buja and Eyuboglu (1992) (from here on, for short, referred to as B & E) assessed the significance of the eigenvalues from a PCA solution by permuting all variables in a data set independently and concurrently. In the context of PCA, it is also worthwhile to look at the significance of the contribution of the separate variables, as we can use this information to interpret the solution. Thus, we distinguish two forms of significance in PCA. The first form relates to the variance-accounted-for in the entire data set by the first c principal components, with c the number of components selected to represent the data set sufficiently. This fit measure, called total VAF (TVAF), is equal to the sum of the eigenvalues of the first c components.[4] The second form relates to the contribution of each separate variable to the TVAF, which is given by the sum of the squared component loadings for each variable (a component loading being defined as the Pearson correlation between a variable and a principal component). B & E use the same strategy of permuting the entire data set for assessing both forms of significance. These authors note correctly that the two forms (significance of the summary statistic and of its constituents) do not always have to go together. In their terminology, loadings may be quite weak (insignificant), but if they are numerous, the largest eigenvalues may be relatively high (significant), whereas if strong loadings (that are significant) are few in number, the eigenvalue may be relatively low (insignificant). Thus, it is important to realize that in case the solution as a whole is not significant, the VAF for particular variables may still be significant, and vice versa.

The strategy of permuting all the variables in a data set concurrently, enables the fit of a variable in an observed data set to be compared to the fit of variables with the same univariate distributions (its permutations) in a dataset with a completely random structure. We believe that this may not be the most appropriate setting to establish the significance of the contribution of a single variable to a principal components structure, which is supported by the fact that B & E (1992) found surprisingly few significant loadings. We consider it

[4]The eigenvalues in PCA equal the eigenvalues of the Pearson correlation matrix of the variables. Principal components are ordered according to their eigenvalues, with the first component associated with the largest eigenvalue.

more appropriate to assess the significance of the VAF of a variable, *given the structure among the other variables*. To attain this, not the entire data set, but only the scores of one variable at a time should be randomly permuted, while keeping the other variables fixed. In this way, the correlational structure between the permuted variable and the other variables is destroyed, while the correlational structure among the other variables remains unchanged. A consequence of this approach is that more permutations are needed: If the first strategy is performed with 999 permutations, the alternative strategy will involve $999 \times m$ permutations (with m the number of variables in the data set).

B & E (1992) stated that establishing the significance of the component loadings was somewhat problematic, because eigenvectors are only determined up to a sign, which means that component loadings from similar solutions may have different signs. This problem is easily fixed by using the VAF as the statistic in the permutation study, which is given by the *squared* component loadings. In this study, we mainly focus on the *sum* of squared component loadings across components as a VAF measure, because this value remains constant over possible rotations or reflections of the solution. We establish the significance of the VAF for each variable, both by permuting the entire data set (B & E), and by permuting one variable while keeping the others fixed, and we compare the results of these two strategies.

4.3 Design of the Monte Carlo Study

To compare the effectiveness of the two permutation strategies described above, we performed an extensive Monte Carlo simulation study, varying several aspects of the design. We used a large number of Monte Carlo replications (R) of three different types of simulated data sets: data with a strong principal components structure, data with a moderately strong structure, and data with no distinct structure. Each Monte Carlo replication consists of the following steps: (1) generating a data set of a specific size and structure, (2) permuting the generated data set a large number of times, (3) using the permutation results to establish a null distribution for the VAF of each variable in the data set, and (4) computing p-values. The null hypothesis is that there is no distinct correlational structure, and thus no contribution of the variables to the first two principal components. Because the simulated data sets are generated with a prespecified principal components structure, we can formulate alternative hypotheses in which specific variables contribute to the solution, while others do not.

The most important factor in the design is the permutation strategy: To

obtain a permuted data set, either each column of the observed or generated data set is permuted independently and concurrently (i.e., the correlational structure of the entire data set is destroyed), or each column is permuted independently and sequentially (i.e., one variable at a time is permuted, while keeping the others fixed). We refer to the first condition as *permutation of the entire data set*, abbreviated as *PermD*. This condition requires 999 permutations in each Monte Carlo replication. The second condition is called *permutation of a single variable*, abbreviated as *PermV*, and requires $999 \times m$ permutations in each replication. We studied the behavior of each strategy with respect to its incorrect indications of the (in)significance of the contribution of each variable (VAF) to the principal components structure (TVAF). The overall performance of each strategy is assessed in terms of proportions of both Type I error (incorrectly marking a result significant) and Type II error (incorrectly marking a result insignificant).[5]

The results of the Monte Carlo study are expected to vary with the size of the data set, as a principal components structure may be more easily detected in larger data sets. Therefore, for each principal components structure, we varied the size of the data set, from quite small (with 20 variables and 100 objects) to relatively large (with 40 variables and 500 objects). In Figure 4.1, the complete design is presented schematically. We estimated the p-values for the VAF per variable by using a large number of Monte Carlo replications for each cell of the design. The details of the study are fully described in the next paragraphs.

4.3.1 Generating data matrices with different principal component structures

To examine the behavior of the two strategies under different conditions, data matrices have been constructed with three different types of prespecified principal components structures. Each type can be represented by two blocks of variables: one block with variables that may contribute significantly to the principal components of $\mathbf{C}$ (the *signal* variables), and another block with variables that do not (the *noise* variables). We assume that in the population there is no correlation between the two blocks of variables. The blocks of variables are created as follows.

We generate a block-diagonal population correlation matrix $\mathbf{C}$ (see Figure 4.2) with two blocks on the diagonal. The first block, $\mathbf{C}_1$, contains the correlations between m_1 (possibly) signal variables, and the second block, $\mathbf{C}_2$,

[5]The probability of making a Type II error is also referred to as β, where $1 - \beta$ approximates the power.

Permute entire data set (PermD)

	Strong structure		Moderate structure		No sign. structure	
	$m = 20$	$m = 40$	$m = 20$	$m = 40$	$m = 20$	$m = 40$
$n = 100$	Type I, Type II	Type I, Type II	Type I, Type II	Type I, Type II	Type I	Type I
$n = 200$	Type I, Type II	⋯	⋯	⋯	⋯	⋯
$n = 300$	Type I, Type II	⋯	⋯	⋯	⋯	⋯
$n = 500$	Type I, Type II	⋯	⋯	⋯	⋯	⋯

Permute separate variables (PermV)

	Strong structure		Moderate structure		No sign. structure	
	$m = 20$	$m = 40$	$m = 20$	$m = 40$	$m = 20$	$m = 40$
$n = 100$	Type I, Type II	Type I, Type II	Type I, Type II	Type I, Type II	Type I	Type I
$n = 200$	Type I, Type II	⋯	⋯	⋯	⋯	⋯
$n = 300$	Type I, Type II	⋯	⋯	⋯	⋯	⋯
$n = 500$	Type I, Type II	⋯	⋯	⋯	⋯	⋯

Figure 4.1: *Complete design of the Monte Carlo study; m is the total number of variables in a data set, n is the number of objects. In each cell, Type I and II errors for the uncorrected (UNC), Bonferroni corrected (BF), and FDR corrected (FDR) 5% significance levels are computed.*

the correlations between m_2 noise variables. The off-diagonal blocks of $\mathbf{C}$ consist of zeros, thus the variables in $\mathbf{C}_1$ and $\mathbf{C}_2$ do not correlate.

A $\mathbf{C}_1$ block is generated according to one of three types of structure:

1. A *strong* principal components structure;

2. A *moderately strong* principal components structure;

3. *No* principal components structure.

In the $\mathbf{C}_1$ blocks with a strong structure, the m_1 variables in the block correlate highly with each other, and the first principal component accounts for approximately 50% of the TVAF, while the second principal component accounts for approximately 30%. The remaining variance is approximately equally divided among the remaining components, thus the associated eigenvalues are approximately equal. In $\mathbf{C}_1$ blocks with a moderately strong structure, the first two components account for 30% and 10% of the TVAF respectively. $\mathbf{C}_1$ blocks without a particular principal components structure are generated such that the proportion of variance-accounted-for by each component approximately equals $1/m$.

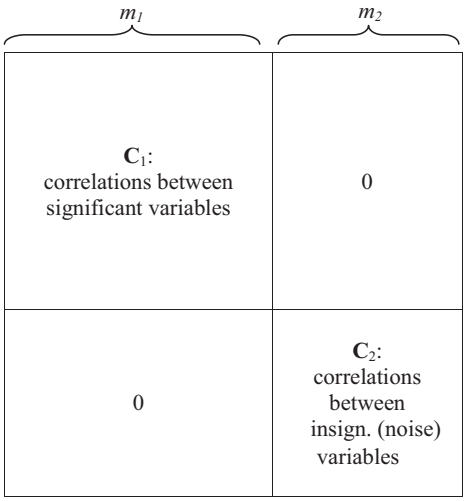


Figure 4.2: *Block structure of the simulated correlation matrices.*

The noise variables in $\mathbf{C}_2$ do not contribute to the principal components structure of $\mathbf{C}$. In reality, the population data will never exactly concur with the null hypothesis, and thus the noise variables will not be perfectly uncorrelated. Therefore, to obtain a more realistic situation, we impose a very weak one-dimensional structure among them (for example, as in a simple form of method effect). Consequently, the proportion of variance-accounted-for by each component in $\mathbf{C}_2$ is not exactly $1/m_2$ (which is the value we would expect when the variables are uncorrelated), but is slightly higher for the first component. The details of the procedure developed for generating the data are fully described in Appendix D.

To keep the results as comparable as possible, and to not further complicate the design, the same ratio of variables from $\mathbf{C}_1$ to $\mathbf{C}_2$ has been used to create small and large data sets. The ratio of m_1 to m_2 is $3:1$, resulting in $m_1 = 15$ and $m_2 = 5$ for the data sets with 20 variables, and $m_1 = 30$ and $m_2 = 10$ for the data sets containing 40 variables.

When assessing the significance of the contribution of the variables to the PCA solution, we are testing the general null hypothesis that none of the variables contributes, which can be specified in the following null hypothesis for each of the variables in the data set: The observed VAF value does not differ from the center of its permutation distribution. When we perform a two-dimensional PCA on data with a strong or moderately strong two-dimensional principal components structure, the null hypothesis should be rejected for the

variables in $\mathbf{C}_1$, and not rejected for the variables in $\mathbf{C}_2$. For the data without a distinct principal components structure, the null hypothesis should not be rejected. To find out whether the two permutation strategies are able to detect the imposed components structure in terms of the VAF, a large number of p-values are computed. These p-values are used to establish the number of times each variable is found to be significant. We use different significance levels, either corrected or uncorrected for multiple testing, and record the number of errors in marking variables significant.

4.3.2 Correction for multiple testing

It is common practice to consider a test result significant if its p-value is smaller than 0.05. We will refer to this criterion as the *uncorrected* significance level (UNC). When we consider the VAF for each variable, however, several tests are performed on the same data set, and when a significance level of 0.05 is used for each separate result, the chance of incorrectly marking one of the results significant will be inflated. To overcome this problem in multiple testing, we use two different methods to correct the significance level.

The first correction method we use is the simple *Bonferroni correction* (BF), which is aimed at controlling the so-called *familywise error rate* (FWE). The FWE is defined as the probability of making one or more false rejections in a "family" of hypothesis tests (see Shaffer, 1995). Thus, the probability that *any* of the results is incorrectly marked significant is controlled. The BF divides the significance level α by the number of tests performed on one data set. In this study, applying the BF reduces to dividing the significance level by the total number of variables in the data set (α/m), because we perform a significance test on each of the variables. As this type of correction is easy to understand and implement, it is quite often used by applied researchers.

The second correction method that is used is a less conservative alternative to the Bonferroni correction, developed by Benjamini and Hochberg (1995). Instead of controlling the FWE, this method is aimed at controlling the *false discovery rate* (FDR), which is the proportion of falsely rejected null-hypotheses within the total set of rejections. Suppose we obtain a list of results from several hypothesis tests for one data set, with a 5% significance level α. Methods controlling the FWE ensure that the probability of the list containing at least one false rejection is at most 5%, that is, that the probability of the entire list being correct (does not contain false rejections) is at least 95%. Alternatively, controlling the false discovery rate (from here on referred to as the FDR correction) ensures that the proportion of rejections that is false is kept below 5% (i.e., that at most 5% of the *significant* results on the list is incorrect). The FWE is always larger than or equal to the FDR, thus

FWE control automatically implies FDR control, and FDR control will lead to a gain in power compared to FWE control. In the context of exploratory research, FDR control seems more sensible, as accepting a certain amount of error is common practice (Keselman, Cribbie, & Holland, 1999; Verhoeven, Simonsen, & McIntyre, 2005).

The FDR procedure starts with sorting the p-values for all the variables of one data set in an ascending order, and assigning each of them a rank number. Next, starting with the largest p-value (the one with the highest rank number, i.e. the last one in the sorted list), each p-value is tested by a significance level of $r/t \times \alpha$, with t the number of tests (in this study, t equals the number of variables m), and r the rank number of the p-value. When a p-value is smaller than the FDR corrected significance level, the VAF value corresponding to this p-value as well as the VAF values corresponding to all p-values with lower rank numbers are marked significant. Benjamini and Hochberg (1995) have shown that the performance of this procedure with respect to the proportion of Type I error is quite satisfactory. The major advantage of the FDR procedure over the Bonferroni correction is that it attains greater power (i.e., a smaller proportion of Type II error). In addition, FDR control proved to be more powerful than several other correction procedures, specifically when the number of hypotheses tested increased (Keselman et al., 1999).

4.3.3 Computing proportions of Type I and Type II error

To establish which permutation strategy has an overall better performance in establishing significance of the VAF per variable, Type I and Type II error rates are calculated for each cell of the design, both without correction as well as with BF and FDR correction. For each type of correction, the specific form of Type I error rate that is supposed to be controlled is calculated.

For the uncorrected results, the Type I error rate is defined in the following way. Variables involved in a random correlational structure (the variables in $\mathbf{C}_2$) are supposed to show no significant VAF. The proportion of times these variables do show significant VAF gives the Type I error rate. This proportion is computed as $TypeI = sig_u\mathbf{C}_2/m_2$, where $sig_u\mathbf{C}_2$ is the number of times that the VAF of a variable in $\mathbf{C}_2$ is incorrectly found significant with an uncorrected significance level of 0.05, and m_2 is the number of significance tests with a possible false positive outcome that are performed on one data set. (In data sets with no structure, m_2 equals the total number of variables in the data set, m.) The Type I error rate is averaged over all R data replications. For the Bonferroni corrected results, the Type I error rate is computed as follows: $FWE = dat_{sig_b\mathbf{C}_2>0}/R$, with $dat_{sig_b\mathbf{C}_2>0}$ the number of data sets in

which there is at least one false significant result at a Bonferroni corrected significance level of $0.05/m$, and R the number of replications. Finally, for the FDR corrected results, we computed the Type I error rate as: $FDR = sig_f\mathbf{C}_2/(sig_f\mathbf{C}_2 + sig_f\mathbf{C}_1)$, that is, the number of false significant results within the total number of significant results at an FDR corrected significance level ($r/t \times \alpha$). The FDR is averaged over replications. Note that the number of significant results in the block $\mathbf{C}_2$ differs for each type of significance level condition, as the significance level used depends upon the correction method.

The Type II error rate for the strong and moderate two-dimensional data structures is given by the proportion of times that a variable contributing to the two-dimensional data structure (i.e., a variable in $\mathbf{C}_1$) does not come up with a significant VAF. The proportion of Type II error is computed as $TypeII = insig\mathbf{C}_1/m_1$, with $insig\mathbf{C}_1$ the number of times a variable in $\mathbf{C}_1$ is falsely found insignificant (at the significance level corresponding to the correction method used), and m_1 the number of tests with a possible false insignificant outcome. Data sets without a principal components structure contain no variables that contribute to a two-dimensional structure (i.e., have no possiblity of rendering false insignificant results), and thus for such data, Type II errors do not exist.

4.3.4 Choosing the number of Monte Carlo samples

The Type I and Type II error rates are sample estimates that are expected to deviate from the population values by some error margin. We used confidence intervals for proportions to decide how many Monte Carlo replications are needed to obtain an acceptable margin of error. Because we did not wish the total confidence interval to be larger than 1%, we chose 0.005 as an acceptable margin of error. We used the Wilson estimate (Wilson, 1927) to avoid the margin of error of the confidence intervals becoming 0 (also see Agresti & Coull, 1998). In Appendix E, the exact procedure we used to calculate the number of Monte Carlo replications R with these criteria is given. The results from this procedure show that it is sufficient to use 1500 Monte Carlo replications for data sets with 20 variables, and 750 Monte Carlo replications for data sets with 40 variables. In the next section, we will discuss the results for the different data conditions, in terms of proportions of Type I and Type II errors for both permutation strategies.

4.4 Results

The main question of this study is which combination of permutation strategy (PermD or PermV) and significance level condition (UNC, BF, or FDR) performs best for different sizes and structures of data sets. In the following sections, we will first focus on the most useful results for answering that question, that is, the data sets with 20 variables, as these show higher error rates than the data sets with 40 variables. The relatively high error rates for data sets with 20 variables result from the fact that we kept the dominance of the first two principal components constant across the two data size conditions. Consequently, the principal components structure is more evident in data sets with 40 variables than in data sets with 20 variables. In a moderately structured data set with 40 variables, $\mathbf{C}_1$ contains 30 variables, of which 30% of the variance is accounted for by the first component, and 10% by the second component. This leads to eigenvalues of approximately $0.30 \times 30 = 9$ for the first component, and $0.10 \times 30 = 3$ for the second component. The remaining components are all approximately equally unimportant and have eigenvalues close to 1. For a similarly structured data set with 20 variables, with $\mathbf{C}_1$ containing 15 variables, the eigenvalues will be approximately $0.30 \times 15 = 4.5$ for the first, $0.10 \times 15 = 1.5$ for the second, and 1 for the other components. In other words, when the proportion of VAF by the first two components is constant, a data set with 40 variables will have a relatively stronger principal components structure compared to a data set with 20 variables.

4.4.1 Permutation strategies: Overall comparison

First, we will focus on the general comparison between the two permutation strategies (permutation of the entire data set, PermD, and permutation of the separate variables, PermV). Table 4.1 displays a selection of some general results of the study, showing the proportions of Type I and Type II error for these two different strategies. The error proportions have been determined with an uncorrected 5% significance level (UNC). These general results show that proportions of Type I error are smaller with PermD, while proportions of Type II error are considerably smaller with PermV. The mean proportion of Type I error over numbers of objects is 0.005 for PermD and 0.063 for PermV, and the mean proportion of Type II error is 0.271 for PermD and 0.043 for PermV. The differences between the two strategies after correcting the significance level for multiple testing (BF and FDR) will be discussed in detail below. These results will be shown to point in the same direction, and to be even more pronounced for the proportions of Type II error.

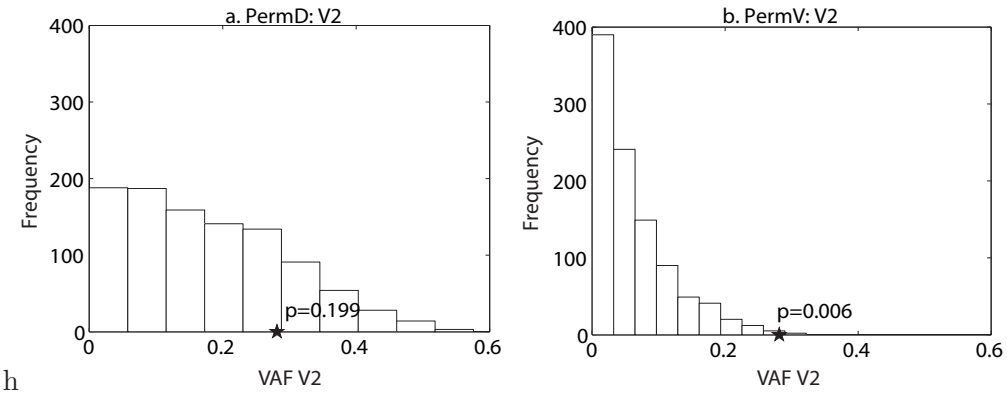
These general differences between the two strategies can be explained as

Table 4.1: *Proportions of Type I and Type II errors, and the total proportion of errors for permutation of the entire data set (PermD) and permutation of separate variables (PermV). Results are based on 1500 replications on data sets with a moderately strong principal components structure and 20 variables. Uncorrected 5% significance levels were used.*

	Type I		Type II		Total	
Nr. of objects	PermD	PermV	PermD	PermV	PermD	PermV
100	0.015	0.063	0.538	0.159	0.553	0.222
200	0.004	0.064	0.303	0.011	0.307	0.075
300	0.001	0.062	0.177	0.001	0.178	0.063
500	0.000	0.064	0.067	0.000	0.067	0.064
mean	0.005	0.063	0.271	0.043	0.276	0.106

follows, using the permutation distributions from one permutation study as an illustration (see Figure 4.3). In the PermD condition, the permuted data set will have a random structure. In such a structure, variables may sometimes by chance obtain a relatively high component loading. Therefore, the permutation distributions for the variables will show quite some spread. Consequently, the observed VAF values will sometimes be close to or even within the range of the permutation distribution, and may therefore less frequently be marked significant. Thus, the proportion of Type I error will be quite small, but the proportion of Type II error will be large. In Figure 4.3, examples of the permutation distributions for the VAF of one variable (V2) in a particular simulated data set are shown. Panel a contains the distribution obtained by PermD. The original VAF values are displayed by stars. The distribution for PermD is quite widely spread, and the observed VAF value lies within the permutation distribution, with a corresponding p-value of 0.199, which is not significant.

If, alternatively, the PermV strategy is used, where only one variable is permuted, the permuted data set will still have a principal components structure as determined by the other $\mathbf{C}_1$ variables. The chance that the permuted variable will obtain a permutation distribution containing relatively large VAF values is very small. In other words, the permutation distribution for the VAF for that variable will show relatively low values with a small spread, and the observed value will be farther from the center of the distribution. This is illustrated in panel b of Figure 4.3, in which the permutation distribution for V2 is displayed as obtained with the PermV strategy. This permutation distribution is much more narrow than that for PermD in panel a of Figure 4.3, and



h

Figure 4.3: *Examples of permutation distributions of the VAF in variable V2 of a particular Monte Carlo data set, after permutation of the entire data set (PermD) and after permutation of separate variables (PermV). Results for a moderately structured data set with 20 variables and 100 objects. The original VAF in V2 is indicated by the star.*

the original VAF (indicated by a star) is close to the tail of the distribution, with a p-value of 0.006, which is significant.

Thus, with the PermV strategy, the proportion of Type I error will be larger compared to the PermD strategy, whereas the proportion of Type II error will be smaller. In exploratory research, Type II error is often considered more serious than Type I error, because in such research it is important to find any effect that might be present in the data. When an effect is discovered in an exploratory study, new studies might be conducted to confirm this result. However, when an exploratory study fails to find an effect that is present in the population (i.e., a Type II error is made), new studies on this effect might never be conducted. This emphasis on avoiding Type II error implies the higher suitability of PermV in exploratory contexts. In addition, from the sum of the error proportions in Table 4.1, we conclude that the total proportion of errors is always smaller for PermV, which also indicates PermV as the most favorable strategy of the two.

4.4.2 Permutation strategies combined with different confidence level conditions

In this section, we combine the two permutation strategies with the three different confidence level conditions: the uncorrected condition (UNC) with 5% significance level, the BF corrected, and the FDR corrected condition.

Table 4.2: *Proportions of Type I and Type II error based on 1500 replications (with PermD as well as PermV) on data sets with strong, moderately strong, or no distinct principal components structure. m=**20**. UNC= uncorrected 5% significance level; FDR = FDR corrected significance level; BF= Bonferroni corrected significance level. Type I error rates have been computed in accordance with each type of control.*

			Strong structure		Moderate struct.		No struct.
Nr. objects	Permutation	Sign.	Type I	Type II	Type I	Type II	Type I
100	PermD	UNC	0.000	0.000	0.015	0.538	0.052
		FDR	0.000	0.000	0.004	0.878	0.049
		BF	0.000	0.000	0.003	0.961	0.048
	PermV	UNC	0.048	0.000	0.063	0.159	0.060
		FDR	0.012	0.000	0.015	0.261	0.065
		BF	0.008	0.000	0.017	0.624	0.058
200	PermD	UNC	0.000	0.000	0.004	0.303	0.052
		FDR	0.000	0.000	0.001	0.600	0.055
		BF	0.000	0.000	0.001	0.907	0.055
	PermV	UNC	0.048	0.000	0.064	0.011	0.077
		FDR	0.012	0.000	0.017	0.024	0.087
		BF	0.010	0.000	0.019	0.167	0.079
300	PermD	UNC	0.000	0.000	0.001	0.177	0.054
		FDR	0.000	0.000	0.000	0.381	0.057
		BF	0.000	0.000	0.001	0.855	0.057
	PermV	UNC	0.049	0.000	0.062	0.001	0.091
		FDR	0.013	0.000	0.016	0.002	0.123
		BF	0.009	0.000	0.018	0.035	0.110
500	PermD	UNC	0.000	0.000	0.000	0.067	0.057
		FDR	0.000	0.000	0.000	0.149	0.070
		BF	0.000	0.000	0.000	0.783	0.065
	PermV	UNC	0.054	0.000	0.064	0.000	0.125
		FDR	0.014	0.000	0.017	0.000	0.215
		BF	0.012	0.000	0.015	0.001	0.193

Table 4.2 shows the Type I error rates for data sets with 20 variables and with three different types of structure. These error rates have been computed in accordance with the rates that are supposed to be controlled (see the paragraph above on the computation of error rates). When we compare the two permutation strategies, for all confidence level conditions, PermD gives smaller proportions of Type I error than PermV, showing that the conclusion for the uncorrected confidence level in Table 4.1 also applies to the BF and

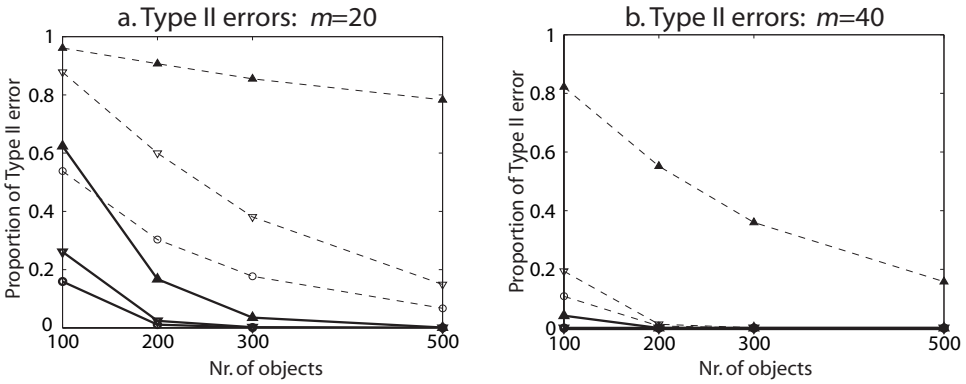


Figure 4.4: *Proportions of Type II error for data sets with a moderately strong principal components structure, calculated after R replications of the permutation test. ($R = 1500$ when $m = 20$, and $R = 750$ when $m = 40$.) Results after PermD are indicated by dashed lines; results after PermV are indicated by solid lines. Circles indicate results with an uncorrected 5% significance level, downward-pointing triangles indicate results after FDR correction, and upward-pointing triangles after BF correction.*

FDR conditions.

The results considering Type II error for the moderately structured data sets with 20 variables are also displayed in Table 4.2. (The results for data sets with 40 variables will be discussed shortly.) To show the differences among the confidence level conditions more clearly, we have additionally displayed the Type II errors in Figure 4.4. Panel a shows results for data sets with 20 variables, and Panel b for data sets with 40 variables. The results are averaged over the Monte Carlo replications (1500 for the 20 variables condition, and 750 for the 40 variables condition; the validation of this number of Monte Carlo replications can be found in Appendix E.) Estimates for proportions of Type II error are displayed for the UNC condition (marked with circles), the BF condition (upward-pointing triangles), and FDR condition (downward-pointing triangles). The dashed lines indicate the results from the PermD strategy, and the solid lines those from the PermV strategy.

Considering the overall comparison of the two permutation strategies, the results for Type II error are completely reversed compared to the results for the Type I error: PermD has much larger proportions of Type II error than PermV, specifically for the BF and the FDR condition. The proportions of Type II error in the UNC condition are smaller than those with FDR, which are smaller than those with BF. Over different numbers of objects, for PermD, proportions of Type II error range from 0.07 to 0.54 for the UNC

condition, from 0.15 to 0.88 for the FDR condition, and from 0.78 to 0.96 for the BF condition. For PermV, these ranges are 0.00 to 0.16 (UNC), 0.00 to 0.26 (FDR), and 0.00 to 0.63 (BF). If we consider a power of .80 or higher acceptable, the PermD strategy never gives acceptable results (except for $n = 500$, with UNC and FDR), while PermV is acceptable under all conditions (except for $n = 100$, with BF and FDR). For both PermD and PermV, the Bonferroni correction leads to the highest loss of power, thus we conclude that this correction is much too conservative.

Permutation strategies with different confidence level conditions for different data structures

The results in Table 4.2 show that, for all confidence level conditions, the Type I error rates are smaller when the data structure is stronger. In unstructured data sets, the Type I error rates are much larger than for structured data sets, which can be explained as follows. In data sets with a strong or moderately strong principal components structure, the variables that do not contribute to that structure ($\mathbf{C}_2$ variables) will have small component loadings compared to the variables that do contribute ($\mathbf{C}_1$ variables). The probability that one of the $\mathbf{C}_2$ variables will turn up significant in the presence of $\mathbf{C}_1$ variables is small. However, when an observed (or generated) data set is *unstructured* (has a random structure), $\mathbf{C}_1$ and $\mathbf{C}_2$ variables are equivalent, and each variable may coincidentally obtain a component loading that is relatively high, which can be compensated by relatively small loadings in other variables. If the data set is permuted, a corresponding high VAF value may (incorrectly) be marked significant. As a result, Type I errors are more likely to occur for unstructured than for structured data sets.

The above reasoning applies to the traditional Type I error rate as well as the FWE (controlled by the BF correction) and the FDR (controlled by the FDR correction). If the data have no component structure, each significant result is a false significant. Then, if one or more significant results are found, both the FWE and the FDR are equal to 1 (thus, when the same correction procedure is applied, the FDR and FWE are exactly equal). Consequently, we would expect the average FWE and FDR over replications to also become inflated. In the worst case for unstructured data (PermV for data sets with 20 variables and 500 objects), the FDR and FWE are around 0.20, meaning that 20% of the data sets contained at least one significant result (which was false because the data had no structure). However, the traditional Type I error rates were still controlled by FDR and BF correction: The proportion of significant results out of all significance tests performed on the unstructured data ranged between 0.001 and 0.022 with FDR correction, and between 0.001

and 0.012 with BF correction over all data sizes and permutation strategies. As a comparison, in the UNC condition, the FDR and FWE were dramatically inflated, ranging from 0.631 to 0.952. (These latter results were computed separately, and are not displayed in the table for conciseness.) Therefore, correction of the significance level is still worthwhile. Fortunately, data sets without any correlational structure (except for a method effect) are not very likely to be analyzed in practice. Therefore, it seems more sensible to focus on Type I errors in the structured data sets.

Considering Type II error rates, we can conclude that these rates are also smaller for data sets with a strong structure than with a moderate structure, as (of course) the power is much higher when effect sizes are high. For unstructured data sets, Type II errors cannot be computed.

Permutation strategies with different confidence level conditions for different numbers of objects

For structured data sets, the Type I error rates are not dependent on the number of objects in the data set (see Table 4.2). However, for unstructured data sets, Type I error rates are higher for data sets containing more objects, which reflects the fact that significant results are more easily found for larger samples. For the unstructured data sets, this effect surfaces, because each significant result is a false significant.

From Figure 4.4, we conclude that the proportion of Type II error decreases when the number of objects in the data set increases. Figure 4.4a shows that this conclusion holds specifically for the PermV condition, where the error drops considerably from 100 to 200 objects.

Permutation strategies with different confidence level conditions for different numbers of variables

The results for the data sets with 40 variables are displayed in Table 4.3. All the results described for the 20 variables condition are confirmed, and as expected, there is an overall drop in proportion of errors compared to the 20 variables condition. In the structured data sets, for PermV with the UNC condition, the Type I error rate is around the desired 0.05 in the whole range of $n = 100$ up to $n = 500$; the proportions of Type I error in the FDR condition are slightly smaller than expected (they vary slightly around 0.04); and the Type I error rates in the BF condition are close to 0. For PermD with structured data sets, the proportion of Type I error is always close to zero. For the unstructured data sets, the Type I error rates are much higher, specifically for PermV with large samples.

Table 4.3: *Proportions of Type I and Type II error based on 750 replications (with PermD as well as PermV) on data sets with strong, moderately strong, or no distinct principal components structure. m=**40**. UNC= uncorrected 5% significance level; FDR = FDR corrected significance level; BF= Bonferroni corrected significance level. Type I error rates have been computed in accordance with each type of control.*

			Strong structure		Moderate struct.		No struct.
Nr. objects	Permutation	Sign.	Type I	Type II	Type I	Type II	Type I
100	PermD	UNC	0.000	0.000	0.000	0.108	0.051
		FDR	0.000	0.000	0.000	0.194	0.037
		BF	0.000	0.000	0.000	0.822	0.037
	PermV	UNC	0.048	0.000	0.048	0.000	0.059
		FDR	0.012	0.000	0.012	0.000	0.057
		BF	0.004	0.000	0.007	0.042	0.057
200	PermD	UNC	0.000	0.000	0.000	0.007	0.051
		FDR	0.000	0.000	0.000	0.013	0.040
		BF	0.000	0.000	0.000	0.552	0.040
	PermV	UNC	0.045	0.000	0.050	0.000	0.068
		FDR	0.011	0.000	0.013	0.000	0.079
		BF	0.008	0.000	0.009	0.000	0.077
300	PermD	UNC	0.000	0.000	0.000	0.000	0.054
		FDR	0.000	0.000	0.000	0.001	0.041
		BF	0.000	0.000	0.000	0.360	0.040
	PermV	UNC	0.052	0.000	0.048	0.000	0.081
		FDR	0.013	0.000	0.012	0.000	0.105
		BF	0.012	0.000	0.019	0.000	0.101
500	PermD	UNC	0.000	0.000	0.000	0.000	0.055
		FDR	0.000	0.000	0.000	0.000	0.055
		BF	0.000	0.000	0.000	0.158	0.055
	PermV	UNC	0.048	0.000	0.050	0.000	0.100
		FDR	0.012	0.000	0.012	0.000	0.179
		BF	0.007	0.000	0.009	0.000	0.165

The results for the proportion of Type II error are even more clear than for the 20 variables condition: For PermD, the error proportions in the BF condition are too large if $n < 500$. Results for the other conditions range from acceptable (for $n = 100$) to excellent.

In summary, the number of objects and variables in the data set is primarily important for the proportion of Type II error. If the entire data set is permuted, more objects and variables are needed to obtain enough power compared to the permutation of one variable at a time. In addition, for small data sets, the Bonferroni correction leads to an enormous loss of power for both permutation strategies, especially for PermD. Considering the size of structured data sets, acceptable results for both Type I and Type II error rates were obtained for PermV with FDR and BF, for data sets with 20 variables and between 100 and 200 objects, or with 40 variables and at least 100 objects. With FDR correction, a higher level of power was reached than with BF correction. For data sets without any structure, Type I error rates were severely inflated. However, traditional Type I error rates were much smaller with than without correction for multiple testing.

4.5 Conclusions and Discussion

The main conclusion from this study is that for assessing the significance of the contribution of the variables to the PCA solution, permuting one variable while keeping the others fixed, combined with FDR correction of the significance level, yields the most favorable combination of proportions of Type I and Type II error. The strategy of permuting the entire data set leads to an excessive loss of power, especially when the size of the data sets is small. Permutation of separate variables gives higher, but still acceptable, proportions of Type I error for structured data sets. The Bonferroni correction is much too conservative and leads to a huge loss of power. Regarding the number of objects, we can conclude that permutation studies should preferably be applied to data sets with more than 100 objects. For smaller data sets, the power of the permutation test is somewhat low. This results from the fact that a principal components structure is less manifest when $n \leq 100$.

Based on these results, we can give researchers who wish to apply permutation tests to PCA for assessing the significance of the VAF per variable the following advice: Permute one variable at a time, while keeping the others fixed (PermV). For small samples (with $n \leq 100$), do not apply the Bonferroni correction. If, with such small samples, Type I error is considered more serious than Type II error (which in exploratory research is not very likely), apply the FDR correction, otherwise use an uncorrected significance level. For larger samples ($n > 100$), FDR correction is recommended, not only because it implies higher power than the Bonferroni correction, but also because it theoretically fits the objective of exploratory data analysis (Keselman et al., 1999).

In the structured data sets in this study, the Type I error rates with Bonferroni and FDR correction were lower than we would expect (< 0.05). In these structured data sets, the relative number of true alternative hypotheses (corresponding to variables in $\mathbf{C}_1$) compared to the true null hypotheses (corresponding to variables in $\mathbf{C}_2$) was high: the ratio was 3:1. For other ratios of true alternatives (TA's) and true null (TN's), we may expect other FWE and FDR values. As an illustration, imagine a data set on which 20 hypotheses are tested, resulting in one false significant result. If the data set contained 10 TA's and 10 TN's, the FDR may range from 1/11 (0.09) to 1/1 (1.00); if the data set contained 1 TA and 19 TN's, the FDR may range from 1/2 (0.50) to 1/1 (1.00). Thus, when the relative number of TA's decreases, the FDR becomes high more easily. The dependency of the FDR on the relative number of TA's compared to TN's is not surprising and also not specific for data analyzed by PCA. With the simulation program by Verhoeven et al. (2005) which randomly simulates p-values for true null hypotheses from a uniform distribution, and z-values for true alternative hypotheses from the normal distribution (with added effect size), high FDR is obtained more easily for data containing relatively few TA's. In such cases, the FWE will also become inflated, because the probability of finding a false significant result is higher when relatively few true significant results exist. This overall effect concurs with logic: If we search for a very rare phenomenon (for instance a rare disease), the chance of obtaining a false positive when doing a test on a random individual is large, even when we use a reliable instrument. Based on these results, we (obviously) advise researchers to perform permutation tests on the VAF of variables only if the data are theoretically founded, and are thus expected to be structured. Otherwise, the FDR and BF correction do not control the error rates they are supposed to control. However, both correction methods still keep the traditional proportion of Type I error within the total number of tests performed on a data set quite small (far below 0.05).

For comparing the performance of the permutation strategies in this study, we mainly focused on the VAF of the variables across components, as this VAF measure remains constant over all possible rotations of the solution. However, for interpretation purposes, it could be more interesting to look at the VAF of the variables *per component*. In the next chapter, we will apply the strategy of permuting the variables independently and sequentially (PermV) to an empirical data set, and also pay attention to the VAF of the variables for each component separately.

The simulation studies were performed using Matlab code, and took quite some computation time. The smallest study, for PermD with 1500 Monte Carlo replications, each involving 999 permutations of data sets with 100 objects and 20 variables, took about 3 hours. The largest study, for PermV with 750 Monte Carlo replications with samples containing 500 objects and 40 variables – each replication involving 999×40 permutations – took almost 50 hours. In practice, of course, a researcher will only apply one permutation test (for example, with 999 permutations) to establish the statistical significance for the variables. Such a single test on a data set (comparable in size to the data in our simulation study) will take less than a minute for permutation of the entire data set, and about four minutes for permutation of single variables (Pentium 4, 3.00 GHz).

With permutation tests, p-values are calculated as $p = (q + 1)/(P + 1)$, with q the number of values as extreme as or more extreme than the observed value, and P the number of permutations. Thus, p-values have a lower bound of $1/(P + 1)$. When applying permutation tests, one should realize that the number of permutations has an effect on the minimum p-value that can be obtained. If too few permutations are used, the minimum p-value will be relatively large. Buja and Eyuboglu (1992) suggested using either 99 or 499 permutations, which would lead to minimum p-values of $p = (0+1)/(99+1) = 0.01$ and $p = (0 + 1)/(499 + 1) = 0.002$, respectively. In the current study, we used 999 permutations (with a minimum p-value of 0.001), which leads to satisfactory results.

In this study, we focused on the contribution of single variables to the PCA solution. It may be conceived that in other studies, the contribution of pairs or sets of variables might be of interest. In that case, two or more variables might be permuted at a time, keeping the other variables fixed, such that the significance of the sum of the variances accounted for by these permuted variables on top of the structure of the others may be assessed. In the most extreme case, all variables would be permuted at the same time for assessing the significance of their total VAF, indicated by the eigenvalues, which equals the Buja and Eyuboglu strategy.

When generating the data, we explicitly decided to make the signal variables independent of the noise variables (the correlations between the data blocks $\mathbf{C}_1$ and $\mathbf{C}_2$ were zero). If we allow for correlations between these two types of variables, it becomes much harder to distinguish method effects from the actual signal. One could argue that assuming all correlations to be zero may not be considered very realistic. An alternative would then be to allow for small correlations between *subsets* of the signal variables and the noise variables, instead of between all variables in the data set. However, the conclusions of the current study should not be considered particularly limited, as the correlations between signal and noise variables were only zero in the population, while they were unequal to zero in the data sets analyzed, due to the sampling effect induced by imposing a very weak one-dimensional structure on the variables in $\mathbf{C}_2$, and by replacing the orthonormal matrix $\mathbf{B}$ with the matrix $\tilde{\mathbf{B}}$ when creating data matrices from correlation matrices (see Appendix D).

In the literature, there has been an ongoing discussion about the validity of null hypothesis significance testing (NHST) in the traditional sense, as proposed by Fisher. The main point brought forward by opponents of NHST is that it is not valid to use the probability of observed data *given that the null hypothesis is true* as an answer to the reversed question, that is, what is the probability of the null hypothesis *given the observed data* (Cohen, 1994; Gliner, Leech, & Morgan, 2002; Killeen, 2005, 2006). Consistent with that line of thought, an alternative to the traditional p-value, called p_{rep} has been proposed (Killeen, 2005), which gives the probability that the direction of a certain effect (positive or negative) can be replicated in another study, under the same circumstances. P_{rep} can be calculated within a parametric framework, under the assumption that the data are normally distributed. In addition, it can be calculated nonparametrically by doing a bootstrap study, and calculating the proportion of bootstrap values for a specific outcome value that point in the same direction as the observed result (Killeen, 2005). This latter approach may also be used in the PCA context.

Buja and Eyuboglu (1992) noted that significance of loadings should not be mistaken for sampling stability. Significance means that loadings of a certain magnitude are unlikely to be due to chance alone. Sampling stability refers to the question of whether the solution of an analysis would be the same if the analysis was performed on a slightly different data set. Stability can be established by resampling techniques, like the bootstrap, but not by permutation tests. In theory, statistically significant loadings can be unstable, whereas insignificant loadings might be quite stable. A stability study on the PCA solution has been reported in Chapter 3.

Nonlinear PCA is an alternative for linear PCA that is useful for data sets that contain variables of different measurement levels (numeric as well as categorical) that may be nonlinearly related to each other (for example, see Chapter 2 of this thesis, and Meulman, Van der Kooij, and Heiser (2004)). There are no standard provisions for establishing inferential statistics for nonlinear PCA, like stability measures and p-values. In Chapter 3, the stability of the nonlinear PCA solution was established, and compared to that of the linear PCA solution. The permutation strategy proposed in the current study may be used in the context of nonlinear PCA as well. Doing so can be considered worthwhile for the application of multivariate categorical data analysis methods in the social and behavioral sciences.