Ecology, 84(9), 2003, pp. 2347–2363

GIVING MEANINGFUL INTERPRETATION TO ORDINATION AXES: ASSESSING LOADING SIGNIFICANCE IN PRINCIPAL COMPONENT ANALYSIS

PEDRO R. PERES-NETO,[1] DONALD A. JACKSON, AND KEITH M. SOMERS

Department of Zoology, University of Toronto, Toronto, Ontario, Canada M5S 3G5

Abstract. Principal component analysis (PCA) is one of the most commonly used tools in the analysis of ecological data. This method reduces the effective dimensionality of a multivariate data set by producing linear combinations of the original variables (i.e., components) that summarize the predominant patterns in the data. In order to provide meaningful interpretations for principal components, it is important to determine which variables are associated with particular components. Some data analysts incorrectly test the statistical significance of the correlation between original variables and multivariate scores using standard statistical tables. Others interpret eigenvector coefficients larger than an arbitrary absolute value (e.g., 0.50). Resampling, randomization techniques, and parallel analysis have been applied in a few cases. In this study, we compared the performance of a variety of approaches for assessing the significance of eigenvector coefficients in terms of type I error rates and power. Two novel approaches based on the broken-stick model were also evaluated. We used a variety of simulated scenarios to examine the influence of the number of real dimensions in the data; unique versus complex variables; the magnitude of eigenvector coefficients; and the number of variables associated with a particular dimension. Our results revealed that bootstrap confidence intervals and a modified bootstrap confidence interval for the broken-stick model proved to be the most reliable techniques.

Key words: bootstrap; eigenvector coefficients; multivariate analysis; numerical ecology; power; randomization.

INTRODUCTION

The use of ordination methods to summarize and describe patterns in multivariate data sets is a long-standing approach employed by ecologists. These methods are useful in reducing the effective dimensionality of large data sets by generating combinations of variables showing common trends of variation. The ability to identify relationships and minimize the effect of random variation may contribute substantially to the recognition of meaningful patterns in the data. Ordination results are customarily interpreted on the following basis: (1) eigenvalues quantifying the amount of variation from the original data summarized by ordination axes, and (2) eigenvectors that contain the coefficients that relate the original variables (hereafter referred to as loadings) to the ordination axes. A common approach for examining ordination results is to establish the number of axes to be inspected, and then determine which variables are related to each nontrivial axis. Once these decisions are made, multivariate scores position the observations (e.g., sites, species) along ordination axes.

Manuscript received 10 October 2000; revised 9 December 2002; accepted 10 December 2002; final version received 14 January 2003. Corresponding Editor: P. Legendre.

[1] Present address: Département de Chimie-Biologie, Université du Québec à Trois-Rivières, C.P. 500, Trois-Rivières, Québec, Canada G9A 5H7. E-mail: pperes@zoo.utoronto.ca

Determining the number of interpretable (i.e., nontrivial) ordination axes is perhaps the greatest challenge in both multivariate ecological and statistical literature (e.g., Pimentel 1971, Karr and Martin 1981, Stauffer et al. 1985, Jolliffe 1986, Zwick and Velicer 1986, Grossman et al. 1991, D. Jackson 1991, J. Jackson 1993, Franklin et al. 1995) and has received far more attention than the problem of deciding which variables contribute to nontrivial axes. In part, this is due to the fact that either relevant information can be lost or random variation can be included if the correct number of ordination axes is not selected. If the researcher is interested only in using ordination axes as a means of reducing dimensionality when summarizing patterns of variation in the data (e.g., Ricklefs and Miles 1994, Knight and Morris 1996, Arita 1997, Weiher et al. 1998, Diniz-Filho et al. 1998), then the issue of determining the number of nontrivial axes is the most relevant. Nevertheless, most ecological applications attempt to provide interpretations of ordination axes (see Legendre and Legendre 1998 for a review).

The interpretation of ordination axes is subjective (Kendall 1980, Cadima and Jolliffe 1995). This is due, in part, to the fact that ordination tools are used mainly as means of data exploration rather than hypothesis testing. Once the number of eigenvalues is established, the interpretation of the relative contribution of variables to each ordination axis is rarely questioned. As Jollifee (1986:51) stated: ''It is remarkable how often

it seems to be possible to interpret the first few principal components, though it is probable that some interpretations owe a lot to the analyst's ingenuity or imagination.'' Although a variety of heuristical and statistical methods have been proposed to assess the degree of association of variables with ordination axis, most researchers rely on ''rules of thumb'' for assessing the magnitude of loadings (e.g., Chatfield and Collins 1980, Richman 1988). From an ecologist's view, Karr (in Karr and Martin 1981) concluded, ''I am confident that most biologists could generate plausible post-facto explanations for high loadings after randomly assigning habitat names to the variables in random number tables.'' Alternatively, there are several statistical approaches such as jackknife (Gibson et al. 1984, Peres-Neto and Bizerril 1994), bootstrap confidence intervals (Jackson 1993), and randomization methods (Faith and Norris 1989) to test if loadings differ significantly from zero.

In this study, we compare the performance of a variety of approaches for assessing the significance of loadings in terms of type I error rates and power. In addition, we develop and compare two novel approaches. We use a variety of simulated scenarios to examine the influence of the number of dimensions in the data, unique variables (i.e., those associated with only one component) versus complex variables (i.e., those associated with two or more components), magnitude of loadings and the number of variables associated with a particular dimension on each method's performance. We restrict this study to principal component analysis (PCA) because it is one of the most commonly used tools in the description of ecological data (James and McCulloch 1990).

Statistical Tests and Criteria for Eigenvector Assessment

In the present study, we restricted the analysis to the case of PCA based on correlation matrices. In this case, variables are standardized (i.e., variables with mean = 0.0 and standard deviation = 1.0) and simulated scenarios are simpler to design because only the magnitude of the association between pairs of variables is considered. In the case of covariance matrices (i.e., variables are only centered), one has to consider also the variance of each variable, which would increase dramatically the number of possible scenarios. Nonetheless, when considering most fields, the majority of PCA applications are performed on correlation matrices (Jackson 1991:80). In the specific case of ecological studies, correlation matrices are applied most frequently when conducting PCA on habitat or abiotic variables.

There are several alternatives for scaling eigenvectors (K. Hope 1968, Jackson 1991, Peres-Neto and Jackson 2001*b*. We use one that, when applied to correlation matrices, produces loadings equivalent to the Pearson product-moment correlation between the PC scores and the individual variables. This scale is given by multiplying unit-length eigenvectors by the square root of their associated eigenvalues (Jackson 1991:68). Eigenvectors scaled in this fashion are known as V vectors (Jackson 1991:16). All methods presented in this paper use this scaling, with the exception of two methods that are based on the squared values of these coefficients (i.e., V^2 vectors). For simplicity, when describing each method we will refer to them as either V vectors or V^2 vectors. Our scaling choice was based on the fact it is the one appropriate for assessing loadings based on certain methods applied here (see correlation critical values and methods based on the broken-stick distribution). Note, however, that any transformation that preserves a monotonic function between V vectors and other scaling procedures (e.g., the ones based on the square root of their associated eigenvalues; see Peres-Neto and Jackson 2001*b* for examples) would provide similar results in terms of test assessment.

When one is interested in evaluating the patterns related to each principal component separately (i.e., interpret the positioning of observations along each axis), it is important to evaluate which variables are associated with the component in question. However, one can see things also with respect to the variable, i.e., which components summarize variation related to any particular variable. If two or more variables are summarized by the same or different components, it indicates whether they share similar or different patterns of covariation, thereby leading to a better interpretation (e.g., structural modeling). Under the null hypothesis that the correlation between a particular variable and any given axis is 0, the square of a significant loading will indicate how much of its variation is being summarized by the particular component. Here we describe the different methods for assessing this null hypothesis:

*1) Cutoff rules (*V *vectors).*—This method regards loadings as significant when their absolute value is larger than a certain pre-established arbitrary value. We have considered values that were used in published studies: 0.25 (Chatfield and Collins 1980), 0.30, and 0.50 (Richman 1988).

2) Broken-stick criterion (V^2 *vectors).*—The broken-stick distribution was originally applied in PCA to assess the significance of eigenvalues (Frontier 1976, Jackson 1993, Legendre and Legendre 1998). However, the same concept can be applied to the squared loadings of any given variable across axes in a V vector matrix. When squaring entire rows of a V vector matrix based on a correlation matrix, the sum of squares of any row (i.e., variable) in a V vector matrix is unity (K. Hope 1968:48). In this way, we obtain the relative proportion of the total variance of a variable that is accounted for by a particular component (K. Hope 1968:49). Assuming that the total variance of any particular variable is

TABLE 1. Steps involved in the broken-stick criterion for assessing significance of loadings.

Parameter	Component 1	2	3	4	5
Step 1: obtain original loadings					
Mean width	−0.458	0.788	0.162	0.374	0.045
Mean depth	0.741	0.559	0.231	−0.238	−0.169
Current velocity	−0.786	−0.003	0.510	−0.338	0.087
Conductivity	0.931	0.252	−0.049	−0.118	0.231
Suspended matter	0.645	−0.448	0.542	0.297	−0.002
Step 2: square loadings					
Mean width	0.210	0.621	0.026	0.140	0.002
Mean depth	0.549	0.312	0.053	0.057	0.029
Current velocity	0.618	0.000	0.260	0.114	0.008
Conductivity	0.867	0.064	0.002	0.014	0.053
Suspended matter	0.416	0.201	0.294	0.088	0.000
Step 3: obtain expected values under broken-stick model					
Expected values	0.457	0.257	0.157	0.090	0.040
Step 4: rank squared loadings and assess their significance according to the broken-stick model					
Mean width	**0.621** (2)	0.210 (1)	0.140 (4)	0.026 (3)	0.002 (5)
Mean depth	**0.549** (1)	**0.313** (2)	0.057 (4)	0.053 (3)	0.028 (5)
Current velocity	**0.618** (1)	**0.261** (3)	0.114 (4)	0.007 (5)	0.000 (2)
Conductivity	**0.867** (1)	0.063 (2)	0.053 (5)	0.014 (4)	0.002 (3)
Suspended matter	0.417 (1)	**0.294** (3)	**0.201** (2)	0.088 (4)	0.000 (5)
Step 5: reorder loadings according to their original axes					
Mean width	−0.458	**0.788**	0.162	0.374	0.045
Mean depth	**0.741**	**0.559**	0.231	−0.238	−0.169
Current velocity	**−0.786**	−0.003	**0.510**	−0.338	0.087
Conductivity	**0.931**	0.252	−0.049	−0.118	0.231
Suspended matter	0.645	**−0.448**	**0.542**	0.297	−0.002

Notes: The table presents original loadings (V vectors) for the stream data set, squared loadings, ranked squared loadings (V^2 vectors), and expected proportion of variance under the broken-stick model. Values in parentheses are the original axis of a particular loading before ranking so that they can be reordered accordingly after significance assessment. Proportions in step 4 larger than the expected values under the broken-stick model in step 3 indicate association of a variable with a particular component (values in bold). For instance, mean width, mean depth, and suspended matter are associated with the second principal component.

divided randomly among all axes, it can be expected that the distribution under the null hypothesis of the squared loadings for a given variable along axes will follow a broken-stick distribution. The solution for estimating the broken-stick distribution for any particular variable based on a correlation matrix is simply

$$b_k = \frac{1}{p}\sum_{i=k}^{p}\frac{1}{i}$$

where p is the number of components (or variables) and b_k is the expected proportion of variance that the kth component summarizes for any particular variable under the broken-stick model. The conceptual framework of the model is that if a stick is randomly broken a large number of times into p pieces by placing $(p - 1)$ random points along its length, b_1 would be the mean or expected size of the largest piece in each set of broken sticks, b_2 would the mean size of the second largest piece, and so on. Because the approach is new, we show an example of the procedure using a data set representing five environmental variables (stream mean width, mean depth, current speed, conductivity, and particulate matter) sampled across 30 sites of an Eastern Brazilian stream (P. Peres-Neto, *unpublished data*). Table 1 presents the original loadings, the squared loadings, and the steps involved in applying a broken stick model to evaluate the association of a variable with a particular component. Since expected proportions under the broken-stick model are obtained in descending order across axes, it is also necessary to rank the observed squared loadings (i.e., V^2 vectors) in the same fashion. For each variable, loadings are considered to be different from random expectation if they exceed the values generated by the broken-stick model. Once the process is finished, loadings and their status of rejection (yes/no) should be reported according to their original axes as in step 1 (step 5, Table 1). Note that values generated by the broken-stick model are fixed for a particular number of variables and therefore do not vary with sample size.

*3) Correlation critical values (*V *vectors).*—This method simply tests loadings against the critical values for parametric correlation from standard statistical tables. The use of standard critical values in PCA has been criticized for two main reasons. (1) Because ordination axes are a composite of the original variables,

the principal components and original variables are not independent and thus their correlations cannot be tested using standard statistical tables (Jackson 1993). (2) Given that the sum of squares of eigenvectors is a function of their respective eigenvalues and that sample eigenvalues decrease along the set of axes (even from spherical populations, i.e., based on an identity matrix, see Buja and Eyuboglu 1992), loadings for latter nontrivial axes would require lower magnitudes to attain significance than early ones. For instance, a loading value of 0.60 may be significant on the second axis, but not on the first. Because standard table correlation values do not take into account the number of variables, which is influential on the distribution of loadings, these critical values are not suitable for assessing loading significance. Appendix A contains a table showing how 95% quantiles for sample loadings change across components, demonstrating that standard tabled values are in fact inappropriate.

*4) Parallel analysis (*V *vectors).*—This method was suggested initially by Horn (1965) and reviewed in a variety of studies (e.g., Zwick and Velicer 1986, Buja and Eyuboglu 1992, Franklin et al. 1995). It applies a Monte Carlo approach to generate critical values for loadings expected under a multivariate normal population having a spherical correlation structure. In this population all variables are uncorrelated, and thus all eigenvalues equal to unity, where each axis has only one loading equal to 1.0 and the remaining ones equal to 0.0. Note, however, that although under normality sample eigenvalues and eigenvectors are maximum likelihood estimators (Anderson 1984), sampling solutions are quite distinct from the population values (Buja and Eyuboglu 1992; see also *Discussion*). The Monte Carlo protocol used here is as follows: (1) generate random normally distributed variables $N(0,1)$ consistent with the original dimensions of the data matrix (i.e., number of observations by number of variables) being analyzed. Note that Buja and Eyuboglu (1992) found that the marginal distribution of the random variables does not influence substantially the critical values generated. (2) Perform a PCA using the matrix generated in step 1, retaining the absolute value of one loading for each axis based on a randomly chosen variable. Given that the sum of squares of eigenvectors along a particular row or column is fixed (i.e., a loss of one degree of freedom), it may be argued that selecting loadings at random from each axis may generate less biased estimates from a spherical population than pooling all loadings across variables or using the values of a fixed variable across axes. However, based on additional simulation results (not presented here), we found no difference between the three procedures. (3) Perform steps 1 and 2 a total of 10 000 times; and (4) calculate for each axis the percentile intervals (e.g., 95% for $\alpha = 0.05$) based on absolute values of loadings, which are then used as critical values. If observed values exceed the critical value, then we reject the null hypothesis according to the pre-established confidence level.

*5) Randomized eigenvector (*V *vectors).*—The randomization protocol was conducted as follows: (1) randomize the values within each variable independently in the data matrix; (2) conduct a PCA on the permuted data matrix, verifying whether the absolute value of each loading from the randomized PCA is greater than the absolute value of corresponding loading of the original data (i.e., the same variable from the same axis). If the random value is greater than the observed, increment the corresponding counter; and (3) repeat steps 1 and 2 a total of 999 times. The P value is then estimated as the probability of obtaining a loading as large as the observed, i.e., P = (number of random loadings equal to or larger than the observed + 1)/(number of randomizations + 1). The observed value is included as one possible value of the randomized distribution, hence the addition of 1 in the numerator and denominator (A. C. A. Hope 1968). The random values, plus the observed value, form the distribution of loadings under the null hypothesis.

*6) Bootstrapped eigenvector (*V *vectors).*—Bootstrap confidence intervals for loadings (Jackson 1993) were based on resampling entire rows from the original data with replacement so that the bootstrapped sample is consistent with the original dimensions of the data matrix. One thousand bootstrapped samples were drawn and a PCA was conducted on each of them. The P value is then estimated as the number of bootstrapped loadings equal to or smaller than zero for loadings that in the original matrix were positive, or alternatively equal to or larger than zero for loadings that originally were negative, divided by the number of bootstrap samples. Two major drawbacks when estimating bootstrap confidence intervals are: (1) axis reflection, which is the arbitrary change in the sign of the eigenvectors of any particular axis (Mehlman et al. 1994, Jackson 1995); and (2) axis reordering (Knox and Peet 1989, Jackson 1995) where two or more axes have very similar eigenvalues. In the latter case, eigenvectors obtained from PCA bootstrap samples may come out altered in their order relative to that found from the original sample. Under either condition, inappropriate bootstrap values in relation to the observed coefficients are used for estimating confidence intervals, which can provide an incorrect estimate of the probability of rejection. In order to address these shortcomings, we applied the following procedure to each bootstrap sample: (1) calculate correlations between the PCA scores for the original data matrix and the PCA scores for the bootstrap sample; and (2) examine whether the highest absolute correlation is between the corresponding axis for the original and bootstrapped samples. Whenever that was not the case, the eigenvectors were reordered. For example, in the case where the correlation between the first original axis and the second bootstrapped axis was the largest correlation, then the loadings from the

second bootstrapped axis are used to estimate the confidence interval for the original first PC axis. This procedure is equivalent to performing orthogonal rotations and correcting for reversals in the axis ordering (Milan and Whittaker 1995). To avoid axis reflections, once reversals were resolved, the signs of the correlations were inspected. A negative correlation between an original axis and a bootstrapped axis indicates a reflection and the coefficients were converted by multiplying them by -1.

7) Bootstrapped broken-stick (V^2 *vectors).*—The procedure is as described for the bootstrapped eigenvector; however in the present case, we determined whether the confidence limits generated included the values expected under the broken-stick model instead of zero. The estimated *P* value was calculated as the number of bootstrapped samples equal to or larger than the appropriate broken-stick value, instead of zero as in the original procedure. Axes reflections and reversals were corrected in all bootstrapped samples using the procedure described above. Note that by coupling a resampling technique with the broken-stick criterion, unlike the broken-stick criterion, critical values become dependent on sample size, possibly providing a more reliable test.

We decided to compare the parallel analysis based on 10 000 samples, whereas for all the other methods also based on confidence interval estimations we used 1000 samples. This decision was based on the computational time constraints that would be generated if 10 000 samples were to be used for all randomization and resampling methods. Given that critical values for the parallel analysis are the same for all samples, the constraint can be relaxed for this method. There were two points that we considered when making this decision. First, given that the critical values for the parallel analysis is generated only once (i.e., prior to testing), by chance, it could be argued that their critical values could be more influenced by sampling variation than the ones evaluated at each sample test (i.e., randomization and bootstrap based methods). Thus, we expect that the use of larger sample sizes for estimating the critical values based on the parallel analysis would compensate for this chance. Note, however, that at 10 000 observations, we found that sampling variation around the critical values was minimal. Second, we also felt that this decision resembles the one made when comparing methods based on standard tabled critical values (analytical), which are based on infinite number of samples, with methods based on randomization or resampling (empirical), based on a restricted number of samples.

Examining Type I Error Rates and Power

In this study, we follow standard Monte Carlo protocols for estimating probabilities of type I error and power for all methods described above (e.g., Manly 1997, Anderson and Legendre 1999, Peres-Neto and Olden 2001). In this case, one simulates population correlation matrices and manipulates them in order to introduce a desirable effect size (i.e., loading magnitude). Following this simulation, a large number of samples are taken and the test is conducted each time. If the effect size is manipulated to be zero (i.e., the null hypothesis is true), the probability of committing a type I error is estimated as the fraction of tests that erroneously rejected the null hypothesis. If the effect size is set different from zero, the proportion of cases in which the null hypothesis was correctly rejected is used as an estimate of statistical power.

The first step was to design population correlation matrices containing loadings suitable to estimate type I error (i.e., $\rho = 0.0$) and power (i.e., $\rho \neq 0.0$), where ρ denotes the off-diagonal correlation (Fig. 1). All matrices were produced with nine or 18 variables and they were divided into groups of varying number of variables to generate data dimensionality (e.g., matrix 1 has three principal components where the first one summarizes the covariation of four out of nine variables or eight out of 18 variables, Fig. 1). Note that the between- and within-dimensions correlations were fixed to a particular uniform value. Within-group correlations were equal to 0.8, 0.5, or 0.3, whereas the between-group correlations were equal to 0.5, 0.3, 0.2, 0.1, or 0.0. These matrices were designed to account for various combinations of the following factors: number of dimensions in the data set (1, 2, 3), loading magnitude, number of variables per component, unique variables which load on only one principal component (e.g., Fig. 1, matrices 1 and 4), complex variables which load on more than one component (e.g., Fig. 1, matrices 2 and 3), and influence of uncorrelated variables (matrices 10 to 14). Unique variables were generated by setting between-group correlations to zero, whereas complex variables were produced by setting between-group correlations at a level different from zero. Note that the between- and within-groups correlations were fixed to a particular value. For instance, matrix 5 (Fig. 1) had three dimensions and each had within-group correlations fixed at $\rho = 0.5$ and between-group correlation at $\rho = 0.2$.

The second step was to generate data sets based on the established correlation structures which, associated with a particular marginal distribution for their variables, were considered as statistical populations. Following Anderson and Legendre (1999), we considered the normal, exponential, and $(\text{exponential})^3$ distributions. We used sample sizes of 30, 40, and 50 observations for populations containing nine variables and 60, 80, and 100 observations for populations with 18 variables for all simulations performed throughout this study. These sample sizes provide a ratio larger than 3:1 of the number of observations relative to variables. The ratio 3:1 or greater was shown to provide stable solutions in PCA (Grossman et al. 1991). To draw sam-

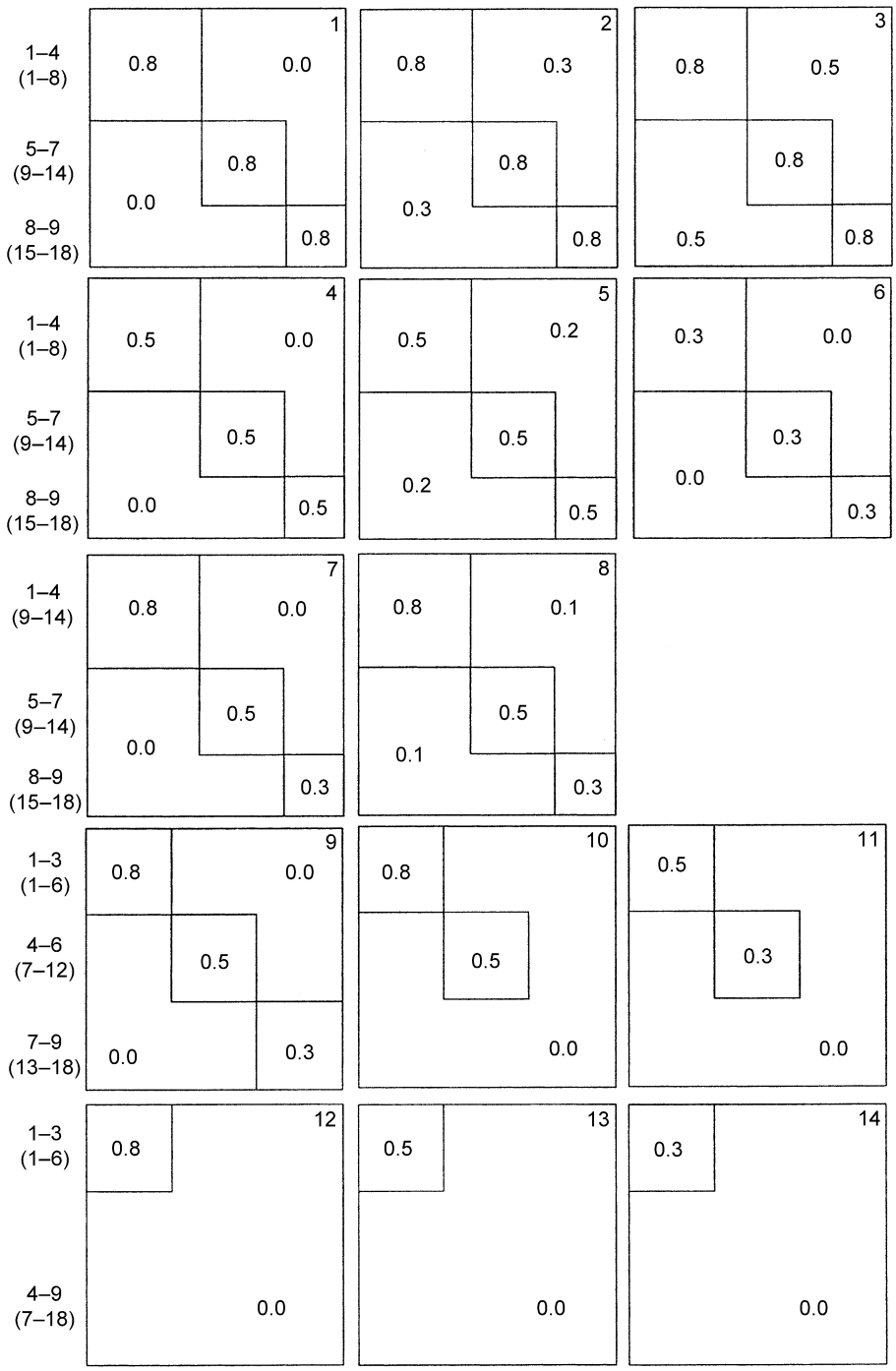

FIG. 1. Correlation matrices considered in this study. The values presented are the off-diagonal intervariable correlations. For example, in matrix 1, variables 1–4 (9 variables) or 1–8 (18 variables) were correlated with one another at $\rho = 0.8$. However, correlations between these and all other variables were equal to 0.

ples from a population following any particular correlation matrix, we have used the following steps (see also Peres-Neto and Jackson 2001*a*): (1) generate a matrix composed by the appropriate number of observations and number of variables containing random deviates with mean = 0 and variance = 1 following one of the three distributions considered [i.e., normal, exponential, and (exponential)^3]; given that the variance of deviates from an (exponential)^3 distribution is rather different than unity, it was necessary to standardize the columns of the generated matrix afterward (Legendre 2000). (2) decompose the correlation matrix using Cho-

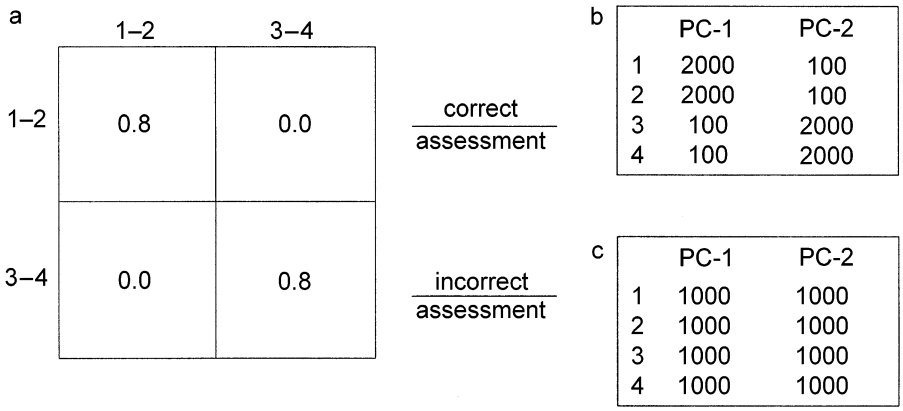

FIG. 2. An example illustrating the rationale behind correcting for sample PCA axis reversals in relation to the PCA population solution: (a) a correlation matrix with four variables that presents equal amount of variation in the first two principal component axes, so that the probability of sample reversals is 50%; (b) expected type I error rates and power estimates for loadings on the first two components for a hypothetical indefectible method ($\alpha = 0.05$); (c) the same estimates if a correction to record the number of rejections according to its target population axis is not performed.

lesky decomposition; and (3) postmultiply the upper triangular matrix resulting from the matrix factorization of step 2 by the matrix of step 1. Note that the resultant sample data matrix in step 3 follows a multivariate distribution according to the marginal distribution and particular correlation matrix chosen (Fig. 1).

The final step was to estimate empirically type I error and power by drawing a large number of random samples from each population and then applying all the criteria and tests described previously to each sample. Type I error rates were measured using the correlation matrices from which their PCA generated loadings equal to zero (i.e., matrices 1, 4, 6, 7, 9, 10, 11, 12, 13, and 14, Fig. 1). All 14 matrices (Fig. 1) generated PC loadings different from zero (i.e., $\rho \neq 0.0$) and thus were suitable for assessing power. As a result of combinations of each correlation matrix and each associated marginal distribution, a total of 42 populations were considered. For samples based on 30 (nine variables) and 60 (18 variables) observations, 2000 random samples were drawn. Due to computational time constraints, and expected smaller sampling variation, for samples based on 40, 50 (nine variables), 80, and 100 (18 variables), tests were evaluated on the basis of 1000 random samples.

For sake of simplicity, only eigenvectors from the nontrivial dimensions in relation to the population correlation matrices were assessed. For example, with populations based on correlation matrix 1 having nine variables (Fig. 1), a total of 27 loadings distributed along the first three axes (nine suitable for estimating power and 18 for type I error) were assessed by each method, whereas for population 10 a total of 18 loadings (six suitable for estimating power and 12 for type I error) were tested. For matrices containing 18 variables, these values are doubled. Because the percentage of variation contained in each dimension is the same when contrasting the same matrix but with different number of variables, the effects of increasing the number of variables can be better evaluated. For methods based on a statistical criterion, sample tests were conducted with and without applying the sequential Bonferroni correction developed by Holm (1979; see also Peres-Neto 1999) in order to verify the efficiency of this correction in minimizing type I error probabilities.

In order to minimize differences related to sampling variation in Monte Carlo simulations, all methods were applied to the same samples for a specific correlation matrix population. A significance level of $\alpha = 0.05$ was used for evaluating the significance of all statistical tests. Type I error rates were estimated as the proportion of sample tests out of 2000 that rejected the null hypothesis when the null hypothesis was true (i.e., $\rho = 0.0$ for matrices in Fig. 1); whereas power was calculated as the proportion of sample tests that correctly rejected the null hypothesis for loadings different from zero (i.e., $\rho \neq 0.0$ for matrices in Fig. 1).

The same concerns related to axis reversals apply to the estimation of power and type I error. Due to sampling variability in PCA parameters (e.g., Cadima and Jolliffe 1995), the order of sample axes can appear reversed in relation to the PCA solution for the population, and by consequence type I error and power may not be estimated properly. For instance, consider the correlation matrix with four variables and two dimensions in Fig. 2a. After 2000 sample tests, an appropriate method (i.e., power = 100% and type I error rates = 5%) should present the number of rejections per loading as shown in Fig. 2b. Nevertheless, because the first two dimensions have equal amounts of variation (i.e., eigenvalues), the probability of sample axis reversal is 50%. If corrections are not implemented in order to record the number of rejections of each test according to its population axis, power and type I error estimates would be more likely as presented in Fig. 2c.

In the extreme case, for populations with similar or equal eigenvalues, sample loadings can vary widely (Buja and Eyuboglu 1992). When comparing the performance of different methods, a correction for this problem appears necessary in order to estimate type I error rate and power appropriately. Moreover, in practical situations there is no interest in knowing if any particular axis is at the same position relative to its population axis. The important issue is how efficient are the various methods in assessing structural combinations of variables, as found in the original population. Therefore, the position (i.e., axis) in which the combination of particular variables appears in the sample when related to the population is somewhat irrelevant provided that they are retained and interpreted. We adopted a similar solution as provided for the bootstrap samples in relation to the original sample in order to minimize this problem. We applied the following procedure for each sample: (1) calculate PCA scores for the sample by multiplying the standardized sample data matrix by the sample eigenvectors; (2) calculate another set of PCA scores by multiplying the same standardized data matrix by the reference population eigenvectors, generating a set of scores that better approximates the population scores. Then, as in the bootstrap case, we compared the correlations between sample and ''population'' multivariate scores to evaluate possible reversals. After possible axis reversals were identified and altered, the status of rejection for the tests based on the coefficients were saved. For all cases, we have designed correlation matrices (Fig. 1) such that their eigenvalues are quite distinct among dimensions, lessening possible axis reversals of samples.

We estimated 95% confidence intervals for both power and type I error rates. Because we conducted a large number of simulations, we applied a normal approximation for a binomial confidence interval with mean p and variance $p(1 - p)/N$ (Manly 1997). In this case, a 95% confidence interval was estimated as $p \pm 1.96 \sqrt{p(1 - p)/N}$, where p is the proportion of rejections and N is the number of sample trials. All simulations were carried out using a computer routine in Borland Pascal 7.0 developed by P. Peres-Neto.

Results

Probability estimates of type I error were based on correlation scenarios that provided null loadings (i.e., variables with $\rho = 0.0$, Fig. 1). Although the sequential Bonferroni correction resulted in type I error rates for some tests that were closer to 5%, the correction resulted in an extreme reduction in power (in most cases a loss of 50%), and therefore outcomes based on Bonferroni corrections are not reported. Although probabilities of a type I error were not equal within and between groups of variables (i.e., some dimensions were more prone to slightly higher type I errors), we averaged type I error estimates for all loadings as we do not see them being influential in the differences in terms of power between methods. For instance, correlation matrix 4 with nine variables provides 18 estimates for type I error rates distributed along three dimensions, with the average error rate for the normal distribution being 0.04 (Table 2). Results are only presented for samples based on 30 (nine variables) and 60 (18 variables) observations, as estimates of type I error rates for the other sample sizes were comparable across all scenarios (Tables 2 and 3, respectively). Note that the expected type I error for the statistical tests is 0.05, whereas the expected type I error for assessments based on the broken-stick criterion and cutoff values criterion should be zero because they were not compared against a confidence interval. For simplicity, in the case of type I error assessment, we established a fixed confidence interval around the alpha level as 0.04–0.06. In this way, any estimated type I error rate smaller than 0.04 or larger than 0.06 can be considered as significantly different from the expected value. No method presented type I errors comparable to the nominal alpha across all scenarios. Assessments based on cutoff values, broken-stick criterion and the correlation critical values resulted in excessively high probabilities. Type I errors from the bootstrapped broken-stick, randomized eigenvector and parallel analysis were typically smaller than expected. The bootstrapped eigenvector was inconsistent being either slightly lower or higher in a number of correlation scenarios.

Because all methods that were not computer intensive (i.e., correlation critical values, broken-stick criteria, and cutoff values) presented large type I error rates, with the exception of the cutoff values of 0.50 for 18 variables (Table 3), we will not report the results in terms of power for these methods as they can be considered as invalid tests (sensu Edgington 1995). In addition, parallel analysis and the randomization procedure, with hardly any exceptions, always exhibited significantly lower power when compared to the bootstrapped eigenvector and bootstrapped broken-stick methods. Note, however, that the relative differences between the two group of methods decreased in certain cases as sample size and number of variables increased (see Appendix B). Their results are shown only for normal populations based on 18 variables so that the general behavior of the two methods can be observed. In addition, the pattern based on these populations is relatively consistent across the other distributions and for matrices containing nine variables.

Power was defined as the proportion of rejections out of the appropriate number of sample tests (2000 for 30 and 60 observations, and 1000 for the other sample sizes) based on population loadings different from 0 (i.e., the null hypothesis was false). Because loadings within any particular dimension in the population matrices were uniformly set, and thus differences in power estimates are due just to chance, the estimates were averaged within dimensions. Mean estimates of power were examined according to their as-

TABLE 2. Type I error estimates for all methods considered in this study based on correlation matrices containing nine variables based on 30 observations.

Method and distribution	Correlation matrix									
	1	4	6	7	9	10	11	12	13	14
Bootstrapped eigenvector										
Normal	0.07	0.04	0.01	0.06	0.06	0.08	0.04	0.08	0.06	0.02
Exponential	0.08	0.04	0.02	0.06	0.06	0.07	0.04	0.07	0.07	0.03
(Exponential)^3	0.08	0.04	0.02	0.04	0.03	0.07	0.04	0.09	0.05	0.03
Bootstrapped broken-stick										
Normal	0.04	0.05	0.04	0.03	0.03	0.03	0.03	0.02	0.02	0.04
Exponential	0.04	0.05	0.03	0.04	0.04	0.04	0.02	0.03	0.02	0.05
(Exponential)^3	0.03	0.05	0.03	0.05	0.04	0.03	0.05	0.03	0.02	0.04
Randomized eigenvector										
Normal	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.01
Exponential	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.01
(Exponential)^3	0.02	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01
Parallel analysis										
Normal	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Exponential	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.01
(Exponential)^3	0.03	0.02	0.03	0.02	0.03	0.03	0.02	0.02	0.02	0.03
Correlation critical values										
Normal	0.27	0.32	0.35	0.15	0.30	0.34	0.37	0.21	0.32	0.40
Exponential	0.28	0.31	0.33	0.16	0.31	0.34	0.36	0.22	0.32	0.39
(Exponential)^3	0.27	0.29	0.32	0.17	0.29	0.32	0.35	0.21	0.33	0.37
Broken-stick criteria										
Normal	0.18	0.26	0.31	0.12	0.24	0.26	0.31	0.21	0.29	0.36
Exponential	0.19	0.26	0.30	0.12	0.25	0.27	0.31	0.21	0.29	0.35
(Exponential)^3	0.20	0.25	0.29	0.14	0.23	0.25	0.29	0.22	0.30	0.34
Cutoff value (0.25)										
Normal	0.47	0.50	0.52	0.31	0.49	0.54	0.55	0.41	0.52	0.58
Exponential	0.47	0.49	0.52	0.31	0.50	0.53	0.55	0.41	0.51	0.57
(Exponential)^3	0.48	0.47	0.51	0.32	0.47	0.51	0.53	0.42	0.49	0.56
Cutoff value (0.30)										
Normal	0.37	0.42	0.44	0.23	0.40	0.44	0.47	0.31	0.43	0.50
Exponential	0.38	0.41	0.43	0.23	0.41	0.44	0.46	0.32	0.42	0.49
(Exponential)^3	0.36	0.43	0.41	0.24	0.38	0.45	0.45	0.33	0.41	0.48
Cutoff value (0.50)										
Normal	0.12	0.15	0.17	0.05	0.13	0.14	0.18	0.07	0.13	0.20
Exponential	0.12	0.15	0.16	0.06	0.14	0.15	0.17	0.07	0.13	0.20
(Exponential)^3	0.14	0.12	0.15	0.07	0.15	0.14	0.15	0.08	0.12	0.18

Notes: Estimates are based on the mean proportion of rejections ($\alpha = 0.05$) per 2000 tests for all null loadings, for normal [first row within each method], exponential [second row], and (exponential)^3 [third row] populations. Confidence limits for estimates based on $\alpha = 0.05$ are 0.04–0.06.

sociated loading in the corresponding population correlation matrix (Fig. 1). For the sake of simplicity, results are only presented for tests based on sample sizes of 30 (nine variables) and 60 (18 variables) observations. Results for the other sample sizes are presented in Appendix B. Figs. 3 and 4 show these mean estimates for normal and (exponential)^3 distributions based on nine and 18 variables, respectively. Results for the exponential distribution were omitted, as their power was comparable to the normal distribution across all scenarios. Results for the randomized eigenvector and parallel analysis based on 18 variables are shown in Fig. 5. Examining these results lead to several conclusions: (1) As expected, greater power was achieved for loadings associated with larger eigenvalues for all methods. This result was consistent for all correlation matrices. Thus, variables related to the first principal component are more likely to be deemed significant than variables related to the second component and so on. (2) Increasing sample size and the number of variables resulted in a large increase in power for all matrices. (3) The bootstrapped broken-stick method showed higher power largely for correlation structures exclusively composed of unique variables (but see next result) that load on only one component (e.g., correlation matrices 1, 4, 6, 7, 9, 10, 11), whereas the bootstrapped eigenvector method showed higher power for correlation structures composed of complex variables which load multiple components (i.e., correlation matrices 2, 3, 5, and 8). Randomized eigenvector and par-

TABLE 3. Type I error estimates for all methods considered in this study based on correlation matrices containing 18 variables based on 50 observations.

Method and distribution	Correlation matrix									
	1	4	6	7	9	10	11	12	13	14
Bootstrapped eigenvector										
Normal	0.09	0.03	0.01	0.06	0.06	0.10	0.05	0.11	0.10	0.04
Exponential	0.11	0.03	0.02	0.07	0.06	0.10	0.06	0.12	0.10	0.05
(Exponential)^3	0.08	0.02	0.02	0.06	0.05	0.11	0.06	0.11	0.09	0.04
Bootstrapped broken-stick										
Normal	0.01	0.02	0.03	0.02	0.02	0.01	0.02	0.00	0.00	0.01
Exponential	0.01	0.03	0.05	0.00	0.02	0.01	0.02	0.00	0.00	0.01
(Exponential)^3	0.01	0.02	0.04	0.03	0.04	0.02	0.03	0.02	0.01	0.01
Randomized eigenvector										
Normal	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Exponential	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00
(Exponential)^3	0.02	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.01
Parallel analysis										
Normal	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Exponential	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(Exponential)^3	0.02	0.01	0.00	0.01	0.02	0.02	0.01	0.00	0.01	0.02
Correlation critical values										
Normal	0.31	0.32	0.37	0.26	0.26	0.24	0.26	0.12	0.16	0.25
Exponential	0.32	0.33	0.35	0.27	0.27	0.25	0.28	0.11	0.17	0.24
(Exponential)^3	0.14	0.28	0.32	0.30	0.31	0.23	0.27	0.09	0.18	0.23
Broken-stick criteria										
Normal	0.14	0.19	0.13	0.14	0.14	0.11	0.16	0.08	0.12	0.18
Exponential	0.16	0.20	0.25	0.16	0.16	0.13	0.17	0.08	0.12	0.18
(Exponential)^3	0.19	0.21	0.27	0.18	0.19	0.16	0.21	0.07	0.08	0.13
Cutoff value (0.25)										
Normal	0.32	0.33	0.18	0.27	0.27	0.24	0.27	0.12	0.17	0.26
Exponential	0.33	0.34	0.16	0.28	0.28	0.26	0.29	0.12	0.17	0.25
(Exponential)^3	0.27	0.31	0.17	0.29	0.33	0.27	0.26	0.10	0.12	0.25
Cutoff value (0.30)										
Normal	0.23	0.24	0.13	0.18	0.18	0.16	0.18	0.06	0.01	0.17
Exponential	0.24	0.25	0.16	0.19	0.19	0.17	0.20	0.06	0.00	0.16
(Exponential)^3	0.21	0.23	0.14	0.18	0.21	0.15	0.21	0.05	0.03	0.13
Cutoff value (0.50)										
Normal	0.04	0.04	0.02	0.02	0.02	0.01	0.02	0.00	0.00	0.01
Exponential	0.05	0.05	0.04	0.03	0.03	0.02	0.00	0.00	0.00	0.01
(Exponential)^3	0.02	0.03	0.03	0.02	0.04	0.04	0.04	0.01	0.01	0.04

Notes: Estimates are based on the mean proportion of rejections ($\alpha = 0.05$) per 2000 tests for all null loadings, for normal [first row within each method], exponential [second row], and (exponential)^3 [third row] populations. Confidence limits for estimates based on $\alpha = 0.05$ are 0.04–0.06.

allel analysis were also more sensitive to complex variables (Fig. 5). Interestingly, it seems that the bootstrapped broken-stick presented larger power in certain matrices containing complex variables for later dimensions (i.e., beyond the first component). For example, with matrices 5 and 8 the bootstrapped broken-stick method was more powerful in detecting a loading related to the last two dimensions (Tables 2 and 3). This was mainly the case for the most important loadings within a particular axis. (4) For correlation structures exclusively composed of unique variables, but containing uncorrelated variables, (i.e., matrices 10 to 14), the bootstrapped eigenvector method presented in some instances larger power (matrices with nine variables and sample sizes 40 and 50, see Appendix B) than the bootstrapped broken-stick method. Note that this result is more evident for matrices containing a larger number of uncorrelated variables (i.e., matrices 12, 13, and 14) and also for larger sample sizes (see results in Appendix B). (5) Comparable power was found for normal and exponential data, whereas (exponential)^3 data presented lower levels especially for latter nontrivial dimensions. For the (exponential)^3 distribution, which generates variables with radically non-normal distribution, the bootstrapped broken-stick was less affected than bootstrapped eigenvector. (6) The randomized eigenvector and parallel analysis provided very similar power, especially for larger sample sizes (Appendix B). The bootstrapped eigenvector and the bootstrapped broken-stick methods are more powerful than the ran-

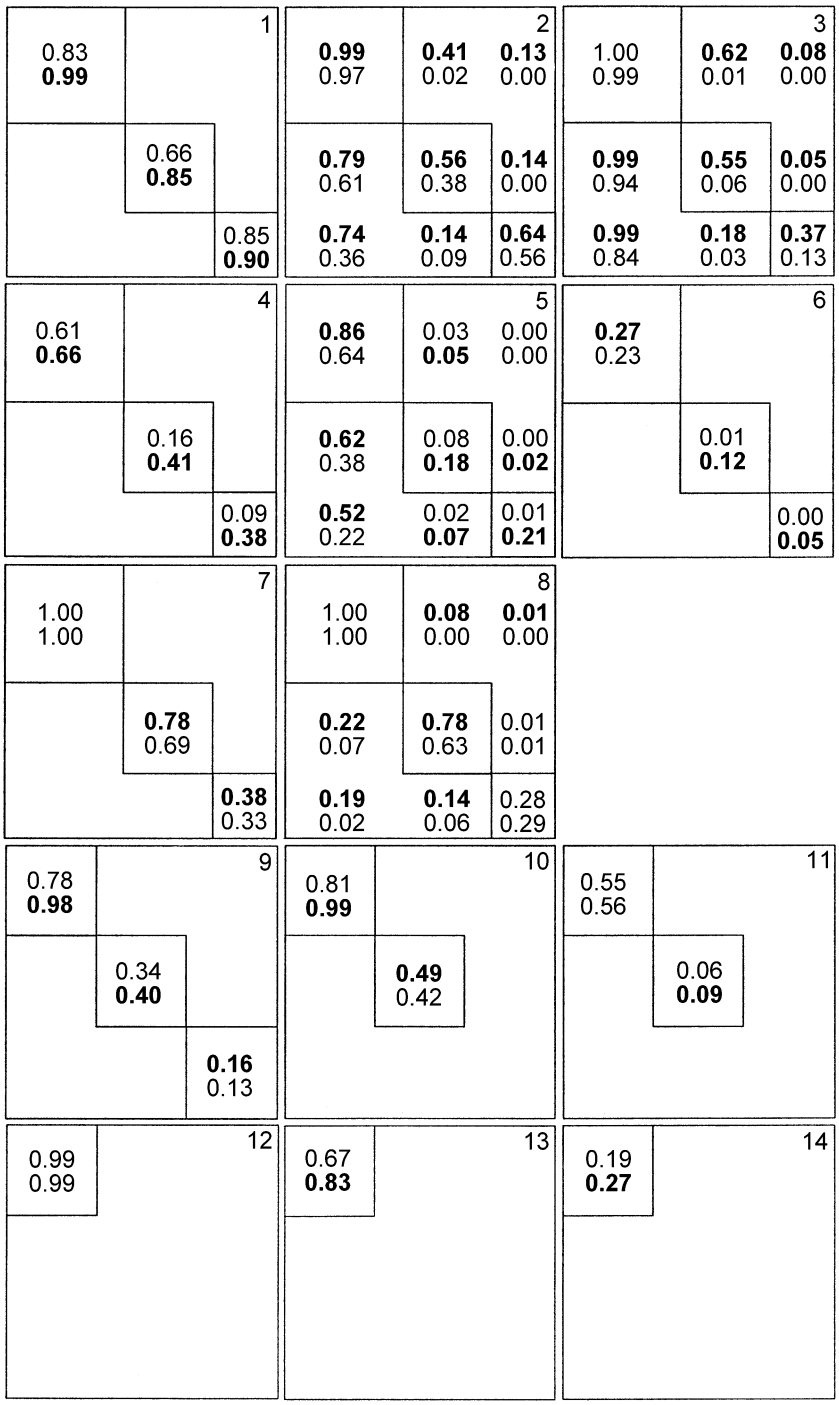

FIG. 3. Mean power estimates for each correlation structure considered containing nine variables (Fig. 1) based on samples containing 30 observations, measured as the mean proportion of rejections ($\alpha = 0.05$) per 2000 tests, for normal and $(\text{exponential})^3$ populations. Estimates are presented according to their associated loading in the corresponding population correlation matrix. Upper values within each block of loadings represent estimates for the bootstrapped eigenvector, whereas lower values represent values for the bootstrapped broken-stick model. For example, for correlation matrix 2 the loadings for variables 5–7 are the same within the first dimension; therefore their power estimates were averaged out. Values in bold indicate that a particular method showed significantly larger power based on the confidence interval for the estimate (see *Examining type I error rates and power* for details).

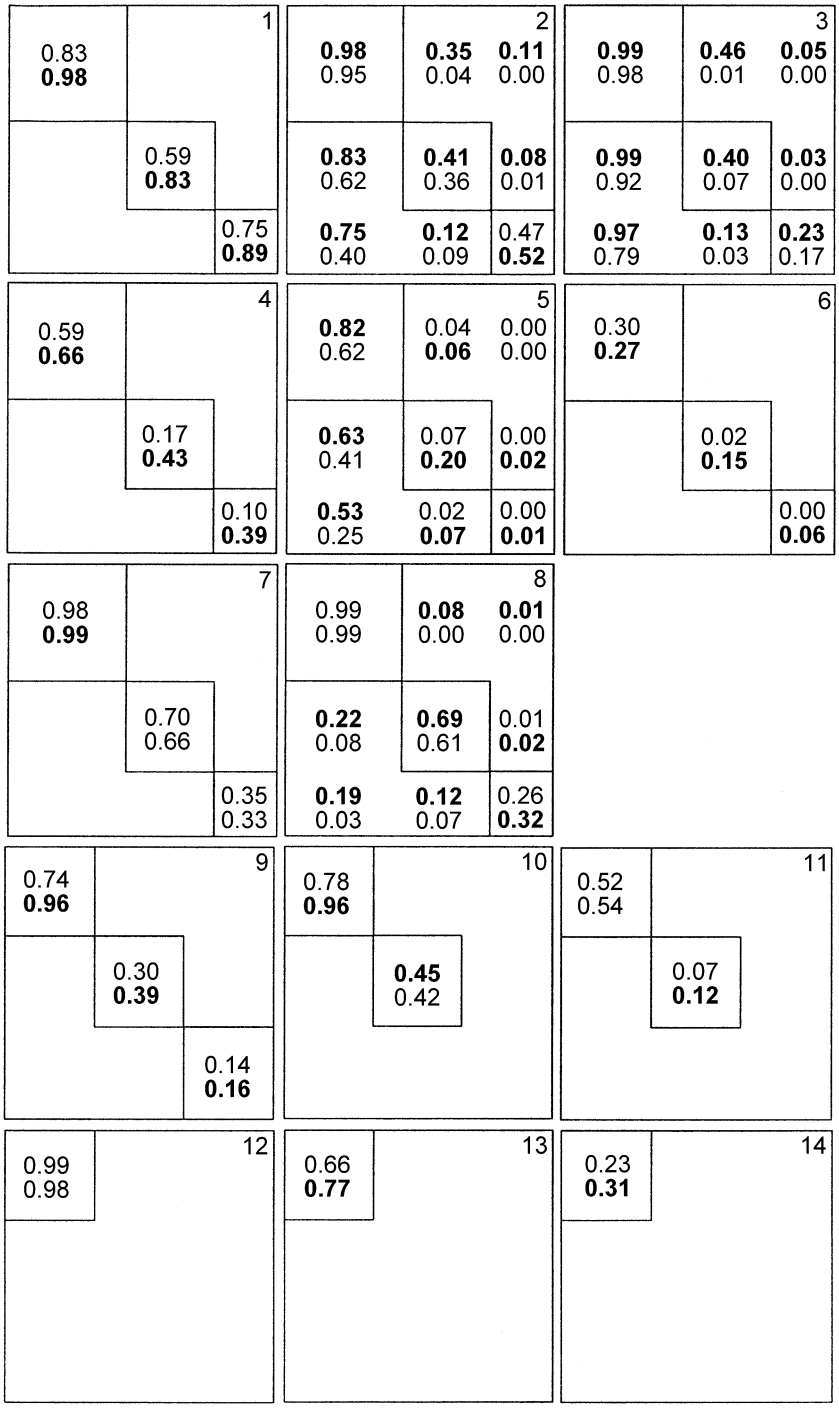

FIG. 4. Mean power estimates for each correlation structure considered containing 18 variables (Fig. 1) based on samples containing 50 observations, measured as the mean proportion of rejections ($\alpha = 0.05$) per 2000 tests, for normal and $(\text{exponential})^3$ populations. Estimates are presented according to their associated loading in the corresponding population correlation matrix. Upper values within each block of loadings represent estimates for the bootstrapped eigenvector, whereas lower values represent values for the bootstrapped broken-stick model. For example, for correlation matrix 2 the loadings for variables 9–14 are the same within the first dimension; therefore their power estimates were averaged out. Values in bold indicate that a particular method showed significantly larger power based on the confidence interval for the estimate (see *Examining type I error rates and power* for details).

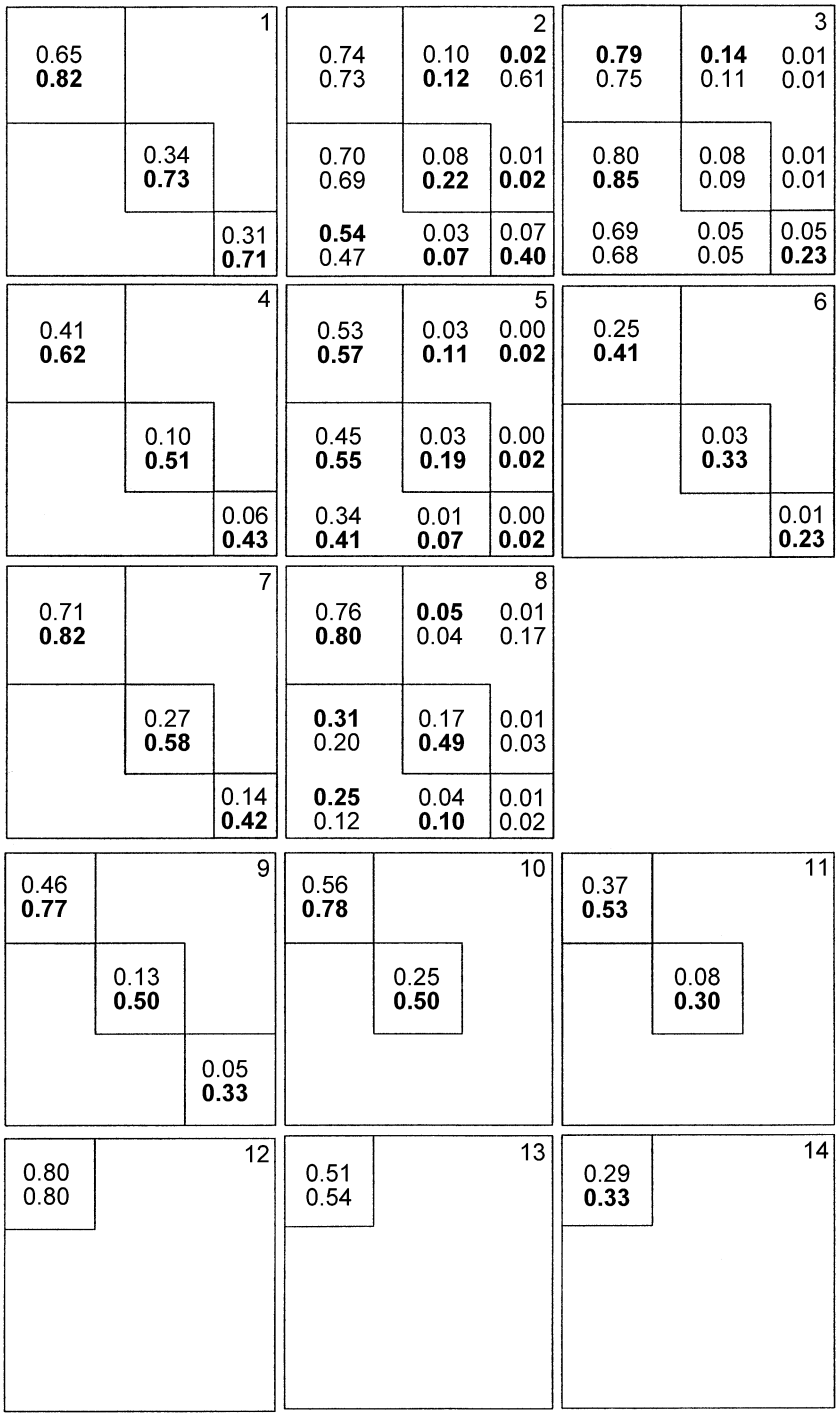

FIG. 5. Mean power estimates for each correlation structure considered containing 18 variables (Fig. 1) based on samples containing 50 observations, measured as the mean proportion of rejections ($\alpha = 0.05$) per 2000 tests, for normal populations. Estimates are presented according to their associated loading in the corresponding population correlation matrix. Upper values within each block of loadings represent estimates for the randomized eigenvector, whereas lower values represent values for the parallel analysis. For example, for correlation matrix 2 the loadings for variables 9–14 are the same within the first dimension; therefore their power estimates were averaged out. Values in bold indicate that a particular method showed significantly larger power based on the confidence interval for the estimate (see *Examining type I error rates and power* for details).

domized eigenvector and parallel analysis methods, though the difference is moderate with increased sample size and number of variables. It is interesting to note that the latter methods are extremely low in power when matrices contain complex variables. In these cases, significant power was only achieved for either variables associated with the first axis or for the most important variables within each axis. However, given that the randomized eigenvector and parallel analysis are similar in this respect to the bootstrapped broken-stick method, the latter method can substitute the formers even in cases of matrices containing complex variables.

Discussion

When analyzing the effectiveness of different decision methods there is always the question of whether simulated data, rather than real data, represent plausible ecological scenarios and how applicable the conclusions are. The use of simulated data is preferable because their characteristics are known (e.g., dimensionality, correlation structure, underlying distribution) and can be kept simple in order to understand the main features of the tests being evaluated (Fava and Velicer 1992). One can argue that sample correlation matrices from ecological studies can be used as input in simulation protocols in order to mimic more relevant ecological scenarios. However, sample correlation matrices from ecological variables never present features such as known dimensionality and correlation values at $\rho = 0$, so type I error rates can not be evaluated. Therefore, the use of simulated data, rather than ecological data, is the only option that permits an examination of the properties and robustness of the different methods compared here. Thus, we assume that if any particular test demonstrates reasonable performance in a large number of scenarios, one can consider that it will exhibit similar abilities when applied to data of interest. It is important to reiterate that our results may be applicable only to standardized data (i.e., mean = 0.0 and variance = 1.0) because we only used correlation matrices. However we believe that this case covers a large number of cases in ecological applications, especially the ones involving environmental data. Future simulation studies should consider also the case of covariance matrices (i.e., for variables centered at mean = 0.0), especially when considering their use in direct gradient analysis (Legendre and Legendre 1998: 582).

Our simulation results revealed that the bootstrapped eigenvector and bootstrapped broken-stick methods are most suitable for assessing the significance of loadings. Because the performances of the two methods are dependent on the correlation structure of the population from which the sample data were drawn, the decision about which method to apply is not straightforward. The bootstrapped eigenvector method was preferable in the presence of complex variables whose variation is expressed on more than one component, and often in cases of matrices with nine variables containing a larger number of independent variables (e.g., matrices 12 to 14, Fig. 1). On the other hand, the bootstrapped broken-stick method was more appropriate where data structure was based on unique variables in the absence of uncorrelated variables. Note, however, that the latter often presented larger power than the bootstrapped eigenvector for the highest loadings within a dimension for complex matrices. This contrast between these two methods in terms of power is clearly not related to their differences in type I error probabilities (Tables 2 and 3). The choice between the two methods is not as simple as it will depend on prior knowledge about the ecological complexity of variables. Since we did not have any expectation about differences in performance of these two methods, we have only considered scenarios where either all variables were complex (Fig. 1, matrices 2, 3, 5, and 8) or unique (Fig. 1, matrices 1, 4, 6, 7, 9, 10, 11, 12, 13, and 14). Nevertheless, in applied situations, principal components containing both unique and complex variables are to be expected (Tucker et al. 1969). Future simulation studies should take into account a mix of these two types of variables. This could be accomplished by constructing population matrices based on samples from ecological studies, where lower values of correlation (e.g., ≤ 0.20) can be converted to 0 in order to generate nontrivial axes. The loss in power of the bootstrapped broken-stick model for scenarios where there are complex variables is understandable. Values under the broken-stick model are calculated assuming that variation associated with any particular variable is partitioned independently (i.e., at random) among multivariate dimensions, which is clearly not the case for complex variables that share variation with more than one principal component axis. However, when analyzing five data sets in order to observe the performance of methods in real ecological situations, both methods largely agreed (Appendix C), indicating that perhaps in real situations, where there is a mix of unique and complex variables, the two methods may be similarly efficient.

It seems that the bootstrapped eigenvector and bootstrapped broken-stick methods were quite robust against departure from normality as they were not affected greatly by an exponential marginal distribution of the variables. Nevertheless, they showed significant reduction in power for the $(\text{exponential})^3$ distribution, which is extremely skewed. In this case, the bootstrapped broken-stick was less sensitive even for data sets composed of complex variables and its use appears suitable where data depart from normality. Another important finding was that the sequential Bonferroni correction for controlling inflated probabilities of type I error was not appropriate for statistical tests conducted on PCA loadings. However, because the dimensions of our correlation matrices were not too large, it might be advisable to determine whether some control over fam-

ily-wise type I error is necessary when larger matrices are used. Another consideration may be to discard a number of redundant variables (Krzanowski 1987, King and Jackson 1999) so that a smaller number of tests are conducted. In addition, it may be desirable to eliminate variables that only load in one component as their variation may be better evaluated separately. In the latter case, the procedures presented here may be used as well.

Although the parallel analysis and the randomization eigenvector approach presented lower power when compared to the two methods discussed above, it is worthwhile to explore some possible reasons for their behavior. There are two possible, and complementary, explanations. One possibility is that, under the null hypothesis, all components are independent and only one variable is related to each eigenvector, whose absolute loading is maximum at 1.0. Due to random axis reorderings during the process of estimating the null distribution for both methods, every variable has the same probability of being related to any axis. Therefore, it can be expected that the null distribution for each coefficient should contain a percentage of values, proportional to the number of variables, that are expected to be larger than the observed value under the alternative hypothesis. As a consequence, both methods become too conservative in rejecting the null hypothesis. In fact, they showed lower type I error when compared to the predetermined alpha significance level and lower power (Tables 2 and 3, Fig. 5). Even if we were able to solve the problem of reorderings, another aspect contributing to the inflation of the confidence intervals of these two methods is that samples from nearly identical or degenerate (equal) population eigenvalues suffer from great sampling variability (Cliff and Hamburguer 1967). Spherical populations have only degenerate eigenvalues (i.e., equal to 1.0) so that their eigenvectors can assume any direction and by consequence, their loadings are arbitrary (note, however, that their sums of squares are kept as a function of their eigenvalues and orthogonality is maintained). An inevitable consequence of this behavior is that any particular variable could have high loadings for more than one axis in any given sample solution. In that case, confidence intervals are inflated, thereby decreasing power. See Seber (1984:198) for further discussion on the subject of sampling from degenerate eigenvalues. However, the two methods were not always so conservative that the null hypothesis was never rejected. This is because accuracy and sampling variation of loadings seem to be inversely related to their eigenvalues (P. Peres-Neto, *unpublished data*) so that confidence intervals based on random matrices will also experience the greatest bias. For instance, 10 000 samples based on 30 observations from an identity correlation matrix with nine variables provided a mean absolute loading for the first variable on PC-1 of 0.41 (standard deviation of 0.22), whereas the expected is 1.0. On the other hand, the mean corresponding loading obtained from 10 000 PCA samples from correlation matrix 1 (Fig. 1) was 0.85 (standard deviation of 0.07) whereas the expected is 0.92. Thus, although confidence intervals based on random matrices are conservative, due to the large bias in their estimation, the null hypothesis is in fact rejected in many cases.

We are unaware of any other simulation study that attempted to evaluate and contrast different statistical and heuristical methods for assessing the significance of loadings in principal component analysis. Our main goal was to conduct a comparative study so that differences in performance and behavior of available and novel methods could be revealed. The most promising approaches were based on bootstrap techniques. For variables sharing variation with different principal component axes, the bootstrapped broken-stick method is preferable; whereas variables whose variation is partitioned among different components, should be analyzed by the bootstrapped eigenvector method. Because this knowledge is not known a priori and in our real ecological applications these two approaches provided similar outcomes, the choice between approaches may become arbitrary. Ultimately, the ability to assess the loading significance is essential in the process of interpreting the association among ecological variables and their contribution to each non-trivial component, thereby leading to the separation of meaningful patterns (i.e., variables that covary) from sources of random variation (i.e., independent variables). Consequently, the use of more appropriate analytical tools in this assessment provides a more rigorous analysis, aiding in the interpretation of the possible processes involved in pattern definition.

Acknowledgments

We thank Pierre Legendre and two anonymous reviewers for comments on the manuscript and our simulation protocols. Funding for this project was provided by a CNPq Doctoral Fellowship to P. R. Peres-Neto, and an NSERC operating grant to D. A. Jackson.

Literature Cited

Anderson, M. J., and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficient in a linear model. Journal of Statistical and Computational Simulation **62**:271–303.

Anderson, T. W. 1984. An introduction to multivariate statistical analysis. Second edition. Wiley, New York, New York, USA.

Arita, H. T. 1997. Species composition and morphological structure of the bat fauna of Yucatan, Mexico. Journal of Animal Ecology **66**:83–97.

Buja, A., and N. Eyuboglu. 1992. Remarks on parallel analysis. Multivariate Behavioral Research **27**:509–540.

Cadima, J., and I. T. Jolliffe. 1995. Loadings and correlations in the interpretation of principal components. Journal of Applied Statistics **22**:203–214.

Chatfield, C., and A. J. Collins. 1980. Introduction to multivariate analysis. Chapman and Hall, London, UK.

Cliff, N., and C. D. Hamburger. 1976. The study of sampling errors in factor analysis by means of artificial experiments. Psychological Bulletin **68**:430–445.

Diniz-Filho, J. A. F., C. E. R. Sant'Ana, and L. M. Bini. 1998. An eigenvector method for estimating phylogenetic inertia. Evolution **52**:1247–1262.
Edgington, E. S. 1995. Randomization tests. Third edition. Marcel Decker, New York, New York, USA.
Faith, D. P., and R. H. Norris. 1989. Correlation of environmental variables with patterns of distribution and abundance of common and rare freshwater macroinvertebrates. Biological Conservation **50**:77–98.
Fava, J. L., and W. F. Velicer. 1992. The effects of overextraction on factor and component analysis. Multivariate Behavioral Research **27**:387–415.
Franklin, S. B., D. J. Gibson, P. A. Robertson, J. T. Pohlmann, and J. S. Fralish. 1995. Parallel analysis: a method for determining significant principal components. Journal of Vegetation Science **6**:9–106.
Frontier, S. 1976. Étude de la décroissance des valeurs propres dans une analyse en composantes principales: comparaison avec le modèle du bâton brisé. Journal of Experimental Marine Biology and Ecology **25**:67–75.
Gibson, A. R., A. J. Baker, and A. Moeed. 1984. Morphometric variation in introduced populations of the common myna (*Acridotheres tristis*): an application of the jackknife to principal component analysis. Systematic Zoology **33**: 408–421.
Grossman, G. D., D. M. Nickerson, and M. C. Freeman. 1991. Principal component analysis of assemblage structure data: utility of tests based on eigenvalues. Ecology **72**:41–347.
Holm, S. 1979. A simple sequential rejective multiple test procedure. Scandinavian Journal of Statistics **6**:65–70.
Hope, A. C. A. 1968. A simplified Monte Carlo test procedure. Journal of the Royal Statistical Society B **30**:582–598.
Hope, K. 1968. Methods of multivariate analysis. University of London Press, London, UK.
Horn, J. L. 1965. A rationale and test for the number of factors in factor analysis. Psychometrika **30**:79–185.
Jackson, D. A. 1993. Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches. Ecology **74**:204–2214.
Jackson, D. A. 1995. Bootstrapped principal component analysis—reply to Mehlman et al. Ecology **76**:644–645.
Jackson, J. E. 1991. A user's guide to principal components. John Wiley and Sons, New York, New York, USA.
James, F. C., and C. E. McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea or pandora's box? Annual Review in Ecology and Systematics **21**:129–166.
Jolliffe, I. T. 1986. Principal component analysis. Springer-Verlag, New York, New York, USA.
Karr, R. J., and T. E. Martin. 1981. Random numbers and principal components: further searches for the unicorn. Pages 2–24 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. United States Forest Service General Technical Report RM-87.
Kendall, M. 1980. Multivariate analysis. Second edition. Charles Griffin, London, UK.
King, J. R., and D. A. Jackson. 1999. Variable selection in large environmental data sets using principal component analysis. Environmetrics **10**:67–77.
Knight, T. W., and D. W. Morris. 1996. How many habitats do landscapes contain? Ecology **77**:1756–1764.
Knox, R. G., and R. K. Peet. 1989. Bootstrapped ordination: a method for estimating sampling effects in indirect gradient analysis. Vegetation **80**:153–165.
Krzanowski, W. J. 1987. Selection of variables to preserve multivariate structure, using principal components. Applied Statistics **36**:22–33.
Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. Journal of Statistical and Computational Simulation **67**:37–73.
Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier Science BV, Amsterdam, The Netherlands.
Manly, B. J. F. 1997. Randomization, bootstrap and monte carlo methods in biology. Second edition. Chapman and Hall, London, UK.
Mehlman, D. W., U. L. Shepherd, and D. A. Kelt. 1995. Bootstrapping principal component analysis—a comment. Ecology **76**:640–643.
Milan, L., and J. Whittaker. 1995. Application of the parametric bootstrap to models that incorporate a singular value decomposition. Applied Statistics **44**:31–49.
Peres-Neto, P. R. 1999. How many statistical tests are too many? The problem of conducting multiple ecological inferences revisited. Marine Ecology Progress Series **176**: 303–306.
Peres-Neto, P. R., and C. R. S. F. Bizerril. 1994. The jackknifing of multivariate allometric coefficient (Jolicouer 1963): a case study on allometry and morphometric variation in *Corydoras barbatus* (Quoy & Gaimard, 1824) (Siluriformes, Callichthydae). Archives of Biology and Technology **37**:449–454.
Peres-Neto, P. R., and D. A. Jackson. 2001*a*. How well do multivariate data sets match? The robustness and flexibility of a Procrustean superimposition approach over the Mantel test. Oecologia **129**:169–178.
Peres-Neto, P. R., and D. A. Jackson. 2001*b*. The importance of scaling of multivariate analysis in ecological studies. Écoscience **8**:522–526.
Peres-Neto, P. R., and J. D. Olden. 2001. Assessing the robustness of randomization tests: examples from behavioural studies. Animal Behaviour **61**:79–86.
Pimentel, R. A. 1979. Morphometrics: the multivariate analysis of biological data. Kendall-Hunt, Dubuque, Iowa, USA.
Richman, M. B. 1988. A cautionary note concerning a commonly applied eigenanalysis procedure. Tellus **40B**:50–58.
Ricklefs, R. E., and D. B. Miles. 1994. Ecological and evolutionary inferences from morphology: an ecological perspective. Pages 13–41 *in* P. C. Wainwright and S. M. Reilly, editors. Ecological morphology: integrative organismal biology. The University of Chicago Press, Chicago, Illinois, USA.
Seber, G. A. F. 1984. Multivariate observations. John Wiley and Sons, New York, New York, USA.
Stauffer, D. F., E. O. Garton, and R. K. Steinhorst. 1985. A comparison of principal components from real and random data. Ecology **66**:1693–1698.
Tucker, L. B., R. F. Koopman, and R. L. Linn. 1969. Evaluating of factors analytic research procedures by means of simulated correlation matrices. Psychometrika **34**:421.
Weiher, E., G. D. Clarke, and P. A. Keddy. 1998. Community assembly rules, morphological dispersion, and the coexistence of plant species. Oikos **81**:309–322.
Zwick, R. W., and W. F. Velicer. 1986. Comparison of five rules for determining the number of components to retain. Psychological Bulletin **99**:432–442.

APPENDIX A

Values of the 95% quantiles for absolute loadings of the first variable in the first five principal components for each correlation structure considered in this study and a spherical correlation matrix are available in ESA's Electronic Data Archive: *Ecological Archives* E084-056-A1.

APPENDIX B

Power estimates for larger sample size are available in ESA's Electronic Data Archive: *Ecological Archives* E084-056-A2.

APPENDIX C

An examination of real ecological data using these methods is available in ESA's Electronic Data Archive: *Ecological Archives* E084-056-A3.