Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 49 (2005) 974–997
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

How many principal components? stopping rules for determining the number of non-trivial axes revisited

Pedro R. Peres-Neto*, Donald A. Jackson, Keith M. Somers

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

Received 18 June 2004
Available online 14 July 2004

Abstract

Principal component analysis is one of the most widely applied tools in order to summarize common patterns of variation among variables. Several studies have investigated the ability of individual methods, or compared the performance of a number of methods, in determining the number of components describing common variance of simulated data sets. We identify a number of shortcomings related to these studies and conduct an extensive simulation study where we compare a larger number of rules available and develop some new methods. In total we compare 20 stopping rules and propose a two-step approach that appears to be highly effective. First, a Bartlett's test is used to test the significance of the first principal component, indicating whether or not at least two variables share common variation in the entire data set. If significant, a number of different rules can be applied to estimate the number of non-trivial components to be retained. However, the relative merits of these methods depend on whether data contain strongly correlated or uncorrelated variables. We also estimate the number of non-trivial components for a number of field data sets so that we can evaluate the applicability of our conclusions based on simulated data.

Keywords: Principal component analysis; Stopping rules; Monte Carlo simulations

1. Introduction

Principal component analysis (PCA) is one of the most common methods used by data analysts to provide a condensed description and describe patterns of variation in multi-

* Corresponding author. Département des Sciences Biologiques, Université de Montréal, C.P. 6128 succ. A, Montréal, Québec, Canada H3C 3J7. Tel.: +514-343-6111 ext. 1233; fax: +514-3436-111.
E-mail addresses: pedro.peres.neto@umontreal.ca (P.R. Peres-Neto), jackson@zoo.utoronto.ca (D.A. Jackson), somers@zoo.utoronto.ca (K.M. Somers).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.06.015

variate data sets. Moreover, PCA is also able to retain meaningful information in the early axes whereas variation associated to experimental error, measurement inaccuracy, and/or rounding is summarized in later axes (Gauch, 1982). The ability to identify relationships by generating linear combinations of variables showing common trends of variation can contribute substantially to the recognition of patterns in the data. However, the issue of determining whether or not a given axis summarizes meaningful variation (i.e., non-trivial versus trivial components) remains unclear in many cases. There are important considerations about the issue of retaining the correct number of axes. One pitfall is that when the correct number of non-trivial principal components is not retained for subsequent analysis, either relevant information is lost (underestimation) or noise is included (overestimation), causing a distortion in underlying patterns of variation/covariation (Férre, 1995; see Lawrence and Hancock, 1999 for a discussion). Determining the number of non-trivial principal components remains one of the greatest challenges in providing a meaningful interpretation of multivariate data and has been a long-standing issue in both biological and statistical literature (e.g., Pimentel, 1979; Karr and Martin, 1981; Stauffer et al., 1985; Zwick and Velicer, 1986; Jackson, 1991; Grossman et al., 1991; Jackson, 1993; Férre, 1995; Franklin et al., 1995; Jolliffe, 2002), though the issue of providing meaningful interpretation to each axis is also important (Peres-Neto et al., 2003).

A variety of stopping rules to estimate the number of non-trivial axes has been proposed (Jackson, 1991; Jolliffe, 2002). Several studies have investigated the ability of individual methods or compared the performance of a number of methods in determining the number of non-trivial components of simulated data sets (Reddon, 1985; Zwick and Vellicer, 1986; Grossman et al., 1991; Buja and Eyuboglu, 1992; Jackson, 1993; Férre, 1995). Generally, these studies established statistical populations that followed particular correlation structures where the number of non-trivial components was known; subsequently samples were drawn from each population and the reliability of sample tests in estimating the known number of non-trivial components was determined. Despite all these efforts, we identify that certain elements still need to be addressed: (1) the robustness of several rules (e.g., randomization methods) has not been investigated or compared; (2) usually only a small number of correlation scenarios was considered; (3) the number of samples per correlation scenario was usually extremely small (in most cases 5 or less) so that differences found between methods could be due largely to sampling error in simulation experiments; (4) the influence of uncorrelated data (i.e., variables having zero correlations with the others) was not investigated; and, (5) the marginal distribution considered was usually normal, producing a limited assessment of the performance against departure from normality. Herein we conduct an extensive simulation study where we compare a large number of rules available in the literature, and develop some new methods. We also estimate the number of non-trivial components for a number of biological data sets so that we can evaluate how applicable our conclusions based on the simulated data may be.

2. Presentation of methods

For all rules, components were assessed individually and sequentially from the largest eigenvalue to the smallest; once a particular axis was considered as trivial (i.e., non-

significant), all other subsequent components were considered to be trivial as well; hence the term "stopping rule". Acronyms are provided for each rule and are used throughout the text. We restricted the analysis to the case of PCA based on correlation matrices given that variables are standardized (i.e., mean $= 0.0$ and variance $= 1.0$) and hence simulated scenarios are simpler to design. Note that in most fields, the majority of PCA applications are performed on correlation matrices (Jackson, 1991, p. 80). In addition, the definition of non-trivial dimensions applied here only applies to correlation matrices (see Section 3 for further discussion). Note that we did not consider cross validation (Jolliffe, 2002) given the large number of variations around this method including types of measures and methods of testing. This class of methods would probably deserve a separate study. The Scree plot (Jolliffe, 2002, p. 117) was also not directly considered here given its subjective nature. However, a variant based on the a permutation percentile interval (Jolliffe, 2002; see Randomization methods based on eigenvalues below) was considered.

2.1. Stopping rules based on confidence intervals

Maximum-likelihood hypothesis tests have been developed for principal component eigenvalues (e.g., sphericity test); however they are quite complicated and sensitive to departure from multivariate normality (Seber, 1984, p. 197) and to large sample sizes (Crawford, 1975). Here we consider three methods based on quasi-inferential procedures, namely parallel analysis, randomization test and bootstrap resampling methods.

2.1.1. Parallel analysis

This method involves a Monte Carlo approach to generate a large number of eigenvalues based on simulated data sets that are equivalent in size to the observed data set of interest, but comprise independent normally distributed variables (i.e., spherical population). These eigenvalues are then used to build confidence intervals. This method was initially suggested by Horn (1965) and has been reviewed in ecological and statistical studies (e.g., Zwick and Velicer, 1986; Buja and Eyuboglu, 1992; Franklin et al., 1995). The Monte Carlo protocol used here was: (1) generate independent normally distributed variables N (0,1) respecting the original dimensions of the data matrix being analyzed; (2) perform a PCA on the correlation matrix of the data matrix generated in step 1, retaining the eigenvalue for each axis; (3) repeat steps 1 and 2 a total of 1000 times; and (4) calculate for each axis the percentile intervals (e.g., 95% for an $\alpha = 0.05$), which are then used as standard critical values for assessing eigenvalues. If observed values exceed the critical value, then we reject the null hypothesis according to the pre-established significance level (acronym: **PA**). Although Buja and Eyuboglu (1992) noted that the method is robust against normality, Parallel analysis is basically parametric by definition given that is based on independent normally distributed data. Here, we consider two classes of distribution-free tests (i.e., randomization and bootstrap procedures) that may provide a more robust assessment when the normality assumption is violated.

2.1.2. Randomization methods based on eigenvalues

The randomization protocol was conducted as follows: (1) randomize the values within variables in the data matrix; (2) conduct a PCA on the reshuffled data matrix; and (3) repeat steps 1 and 2 a total of 999 times. In each randomization, we evaluated four test statistics based on the eigenvalues: (1) the observed eigenvalue for axis k (i.e., λ_k; acronym: **Rnd-Lambda**; e.g., ter Braak, 1988); (2) a Pseudo-F-ratio calculated as each eigenvalue divided by the sum of the remaining (or smaller) eigenvalues (i.e., $\lambda_k / \sum_{j=k+1}^{p} \lambda_j$), where p is the number of variables; acronym: **Rnd-F**; see ter Braak, 1990); (3) the ratio between an eigenvalue and the next adjacent eigenvalue (Jackson, 1993) (i.e., $\lambda_k / \lambda_{k+1}$; acronym: **Rnd-Ratio**); and (4) the difference between an eigenvalue and the next adjacent eigenvalue (i.e., $\lambda_k - \lambda_{k+1}$; acronym: **Rnd-Delta**). The *P*-value for each axis and each test statistic is then estimated as: (number of random values equal to or larger than the observed +1)/1000. Note that the observed value is included as one possible value of the randomized distribution; hence the addition of 1 in the numerator and denominator.

2.1.3. Randomization methods based on eigenvectors

An implicit assumption in PCA is that non-trivial axes have at least two variables associated with them (Zwick and Velicer, 1986). Therefore, estimating the number of significant eigenvector loadings in each axis can provide a rule for component retention (acronym: **Rnd-EigV**). Here eigenvectors were scaled to produce loadings whose values are equivalent to the Pearson product-moment correlation between the PC scores and the individual variables. Eigenvectors scaled in this fashion are known as V-vectors (Jackson, 1991: p. 16; Peres-Neto et al., 2003). The randomization protocol was conducted as follows: (1) randomize the values within variables in the data matrix; (2) conduct a PCA on the randomized data matrix; and (3) repeat steps 1 and 2 a total of 999 times. Each absolute loading for the original data matrix is compared with the corresponding absolute loading from the randomized PCAs (i.e., the same variable from the same axis) to generate a probability associated with the null hypothesis. The *P*-value is then estimated as for the previous randomization method. A particular axis was retained if at least two eigenvector loadings were significantly correlated to it.

2.1.4. Bootstrap methods based on eigenvalues

Bootstrap confidence intervals for eigenvalues (Jackson, 1993) were based on resampling the original data with replacement so that the bootstrapped sample is consistent with the original dimensions of the data matrix. 1000 bootstrapped samples were drawn and a PCA was conducted on each of them. There are a number of methods for estimating confidence intervals (Manly, 1997) and we applied the percentile method (Efron, 1979; Manly, 1997, p. 39). The *P*-value for each axis was estimated as the number of bootstrapped samples equal to or smaller than a particular value representing the expectation under the null hypothesis, divided by the number of bootstrap samples. Given that the average values of all eigenvalues from a spherical population are different than the expected value of unity, we considered four different values under the null hypothesis: (1) eigenvalues > 1.0 (Lambert et al., 1990; acronym: **Bt-KG**), (2) Broken-stick values (acronym: **Bt-BS;** see stopping rules based on average test statistic values for a complete explanation on the Broken-stick model), (3)

average eigenvalue from the Parallel Analysis (acronym: **Bt-PA**), and (4) average eigenvalue from the randomization procedure (acronym: **Bt-RndAvg**).

2.1.5. Bootstrap methods based on eigenvectors

Here bootstrap confidence intervals for loadings (Jackson, 1993; Peres-Neto et al., 2003) instead of eigenvalues were estimated using the same method described above. *P*-values were estimated as the number of bootstrapped samples equal to or smaller than zero for loadings which in the original matrix were positive, or alternatively equal to or larger than zero for loadings which originally were negative, divided by the number of bootstrap samples (acronym: **Bt-Eig**). The same scaling process for the eigenvectors for **Rnd-EigV** was used. Any axis with at least two significant eigenvector loadings was retained. Two major drawbacks when estimating bootstrap confidence intervals for loadings are: (1) axis reflection, which is the arbitrary change in the sign of the eigenvectors of any particular axis (Mehlman et al., 1995; Jackson, 1995); and (2) axis reordering (Knox and Peet, 1989; Jackson, 1995) where two or more axes have very similar eigenvalues. Under either condition, inappropriate values in relation to the observed coefficients are used for estimating confidence intervals. To overcome this problem, we applied a procedure described in Peres-Neto et al. (2003) to each bootstrap sample as follows: (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.

2.1.6. Correlation critical values for eigenvectors

This method simply tests loadings against the critical values for parametric correlation from standard statistical tables (acronym: **r-EigV**). Any particular axis with at least two "significant" eigenvector loadings was retained. The same scaling process for the eigenvectors for **Rnd-EigV** was used (i.e., V-vectors).

2.1.7. Test of sphericity

Once all meaningful variation in the data has been summarized, the remaining variation can be described as a multidimensional sphere where axis orientation is positioned arbitrarily. This point is reached when there is no common correlation among variables. The sphericity test (acronym: **Sphere**) assesses when this arbitrary positioning is reached by finding the axis where all remaining components have similar eigenvalues. The original test was developed by Bartlett (1950) and received several modifications (Jackson, 1991). Here we applied the form presented by Pimentel (1979), which is relevant to correlation

matrices:

$$\chi_k^2 = \left[n - k - \frac{2(p-k) + 7 + 2/(p-k)}{6} + \sum_{j=1}^{k} \left(\frac{\overline{\lambda}}{\lambda_j - \overline{\lambda}} \right)^2 \right] \times \left[-\ln \prod_{j=k+1}^{p} \lambda_j + (p-k) \ln \overline{\lambda} \right],$$

where p denotes the number of components or variables, λ_k represents the eigenvalue for the kth component, $\overline{\lambda} = \sum_{j=k+1}^{p} \lambda_j / (p-k)$ and n is the number of observations in the sample. χ_k^2 is approximately χ^2 distributed with 0.5(p-k-1)(p-k+2) degrees of freedom. For each axis calculate its correspondent χ^2 value.

2.1.8. Bartlett's test for the first principal component

Bartlett (1954) extended his original test of sphericity to evaluate whether the eigenvalue for the first principal component is significantly different from the remaining eigenvalues (acronym: **Bartlett**). The test also received some modifications, and the following version applied by Jackson (1993) was used:

$$\chi^2 = -\left[n - \frac{1}{6}(2p + 11) \right] \ln |R|,$$

where $|R|$ is the determinant of the correlation matrix, n is the sample size and p the number of variables. χ^2 is approximately χ^2 distributed with $p(p\text{-}1)/2$ degrees of freedom. The null hypothesis here is all variables are uncorrelated. If the test is non-significant, one should not look for additional structure in the data.

2.1.9. Lawley's test for the second principal component

Lawley (1956) developed a test that evaluates whether the eigenvalue for the second principal component is significantly different from the remaining set of eigenvalues (acronym: **Lawley**). Thus, the null hypothesis here is that at least two variables are correlated (i.e., first axis significant) and the second eigenvalue is not significantly different from the remaining ones. The test statistic is as follows:

$$\chi^2 = \frac{n-1}{1-\overline{r}} \sum_{\substack{i=1 \\ i \neq j}}^{p} \sum_{\substack{j=1 \\ i \neq j}}^{p} (r_{ij} - \overline{r})^2 - \mu \sum_{k=1}^{p} (\overline{r}_k - \overline{r})^2,$$

where r_{ij} is the correlation between variable i and j and

$$\overline{r} = \frac{2}{p(p-1)} \sum_{i=k+1}^{p} \sum_{j=1}^{p} r_{ij}, \quad \mu = \frac{(p-1)^2 (1 - (1-\overline{r})^2)}{p - (p-2)(1-\overline{r})^2},$$

$$\overline{r}_k = \frac{1}{(p-1)} \sum_{\substack{i=1 \\ i \neq k}}^{p} r_{ik}.$$

χ^2 is approximately χ^2 distributed with $(p+1)(p-2)/2$ degrees of freedom. Although this test only provides an assessment of the first two components, the criterion can be valuable in cases where they summarize a large proportion of the variation in the data.

2.2. Stopping rules based on average test statistic values

Rules based on average values are assessing whether an observed test statistic based on eigenvalues or eigenvectors is larger than the average value expected under the null hypothesis of no association between variables. One may also see some of these rules as based on 50% confidence intervals; however given that there are a number of rules based on this criteria we have decided to make a distinction between the two types of rules.

2.2.1. Kaiser-Guttman

When using correlation matrices, population components having eigenvalues larger than 1.0 summarize shared variation and should be retained (Guttman, 1954; acronym: **KG**). However, due to sampling variation, approximately one-half of sample eigenvalues from random data will exceed 1.0. Despite being severely criticized for this reason (e.g., Karr and Martin, 1981; Jackson, 1993), this method is still very popular among data analysts.

2.2.2. Broken-stick

If one assumes that the total variance in a multivariate data set is divided at random amongst all components, the expected distribution of the eigenvalues can be assumed to follow a broken-stick distribution (Frontier, 1976; Legendre and Legendre, 1998). The idea underlying the model is that if a stick is randomly broken into p pieces, b_1 would be the average size of the largest piece in each set of broken sticks, b_2 would be the average size of the second largest piece, and so on. In the case of correlation matrices (i.e., standardized variables), p equals the number of components and the total amount of variation across all components. The proportion of total variance associated with the eigenvalue for the kth component under the broken-stick model is obtained by:

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

where p is the number of variables. If the kth component has an eigenvalue larger than b_k, then the component should be retained (acronym: **BS**). Jackson (1993) showed that this method performs well, at least when variables are highly correlated.

2.2.3. Random average under permutation

This rule is based on the average eigenvalue obtained under a randomization of the data matrix, acquired from step (3) of the randomization test based on eigenvalues. If the observed exceeds the average random value, that particular axis is considered to be non-trivial (acronym: **Avg-Rnd**).

2.2.4. Random average under parallel analysis

The rule is based on the average eigenvalue acquired from step (3) of the parallel analysis rule. If the observed eigenvalue values exceed the average value, that particular axis is considered as non-trivial (acronym: **Avg-PA**).

2.2.5. Random average under randomization versus average under bootstrap

Given that the average eigenvalue under a bootstrap protocol may represent a better estimate of the population eigenvalue, we developed this rule such that if the average bootstrapped eigenvalue is larger than the average from the randomization procedure, then the null hypothesis should be rejected (acronym: **BtAvg-RndAvg**). Average bootstrapped eigenvalues are calculated from the 1000 bootstrap samples generated in bootstrap methods based on eigenvalues, and random average eigenvalues are calculated as in the rule "random average under permutation" above.

2.2.6. Minimum average partial correlation

Velicer (1976) suggested a criterion based on the average partial correlation between variables after removing the effect of the first k principal components, which is:

$$\overline{f_k} = \sum_{\substack{i=1 \\ i\neq j}}^{p} \sum_{\substack{j=1 \\ i\neq j}}^{p} (r_{ij.k})^2 / p(p-1),$$

where p is the number of variables or components, $r_{ij.k}$ is the partial correlation between variables i and j when the variation related to the first k components have been removed. The numerical value of $\overline{f_k}$ will initially decrease as k increases, and then increase. Velicer's stopping rule is based on the number of components k that provides the smallest $\overline{f}$. Velicer (1976) also proposed a measure based on the average values for the original correlations to determine whether any component should be retained at all

$$\overline{f_0} = \sum_{\substack{i=1 \\ i\neq j}}^{p} \sum_{\substack{j=1 \\ i\neq j}}^{p} r_{ij}^2 / p(p-1).$$

If $\overline{f_1} > \overline{f_0}$, then no components should be retained (acronym: **Part(f_0)**). In initial simulation trials, we noted that the **Part(f_0)** can be conservative (i.e., extracting a smaller number of non-trivial components than it should) and therefore, we also applied Velicer's method (i.e., Part(f_0)) without using the $\overline{f_0}$ criterion (acronym: **Part**). Zwick and Velicer (1986) found that the original method (i.e., Part(f_0)) is quite accurate in identifying components where many variables load in a particular component, but relatively small non-trivial components are usually not detected.

3. Assessing rule performance

We followed a Monte Carlo protocol equivalent to the one used by Zwick and Velicer (1986) and Jackson (1993) for assessing the robustness of stopping rules in estimating

the number of non-trivial components using PCA. The first step was to design correlation matrices (Fig. 1). All matrices were produced with 9 or 18 variables and divided into groups having different numbers of variables to generate non-trivial components (e.g., matrix 1 has three non-trivial principal components where the first one summarizes the covariation of 4 out of 9 variables or 8 out of 18 variables, Fig. 1). As common to many other studies (Reddon, 1985; Zwick and Vellicer, 1986; Buja and Eyuboglu, 1992; Jackson, 1993), trivial components are degenerate and therefore carry only noise. Conversely, non-trivial components are considered as non-degenerate (i.e., present eigenvalues differing among components). Here we apply the notion that in PCA based on correlation matrices, only population axes having eigenvalues larger than unit are worth interpreting. PCA solutions for axes having population eigenvalues smaller or equal than unity are extremely unstable under sampling variation due to their proximity, presenting themselves nearly degenerate. Therefore, robust stopping rules should be able to the detect at which point eigenvalues become nearly degenerate, hence hindering interpretation. Fava and Velicer (1992) and Lawrence and Hancock (1999) showed that when axes considered as trivial (under the same definition applied here) are included in the analysis, the results appear distorted. These studies conclude that, when comparing sample multivariate solutions for the non-trivial axes with the non-trivial axes for the population solutions, the results are equivalent. Once, trivial axes are compared under the same criterion (i.e., sample versus population solutions), the results are quite different, i.e., distorted. In addition, in a simulation study (unpublished results) we compare Jackson's (1993) eigenvalue bootstrap method and a form of Parallel Analysis for eigenvalues for comparing adjacent eigenvalues. We have found that for axes equal or smaller than one, both tests presented extremely low power ($< 5\%$ in most cases) thus becoming impossible to discern between degenerate and non-degenerate solutions for axes having eigenvalues equal or smaller than unit.

Between-group (either 0.5, 0.3, 0.2, 0.1 or 0.0) and within-group (either 0.8, 0.5 or 0.3) correlations of variables were fixed to a particular uniform value. Four additional matrices containing uniform cross-correlations were also considered. Matrix 15 contained either 9 or 18 variables where all cross-correlation values were 0.8. Matrix 16 contained all cross-correlations at 0.5 and matrix 17 all values at 0.3. Matrix 18 was designed to represent pure uncorrelated data (i.e., a spherical population). These matrices were designed to incorporate various combinations of the following factors: number of non-trivial components; eigenvalue distributions along axes; loading magnitude; number of variables per component; unique variables which load with only one principal component (e.g., Fig. 1, matrices 1 and 4) or complex variables which load with more than one component (e.g., Fig. 1, matrices 2 and 3); and the influence of uncorrelated variation (i.e., Fig. 1 matrices 10–14). The second step was to associate the correlation structures with a particular marginal distribution for their variables. Following Anderson and Legendre (1999), we considered the normal, exponential and exponential^3 distributions. Exponential deviates were generated as the negative logarithm of a uniformly distributed deviate. Those deviates were then cubed to generate exponential^3 deviates. We fixed sample sizes as 30 and 50 observations for populations containing 9 variables and 60 and 100 observations for populations with 18 variables for all simulations performed throughout this study. To draw samples from a population following any particular correlation matrix we have used the following steps (Barr and Slezak, 1972; see also Peres-Neto and Jackson, 2001): (1) generate a matrix composed by the appropriate

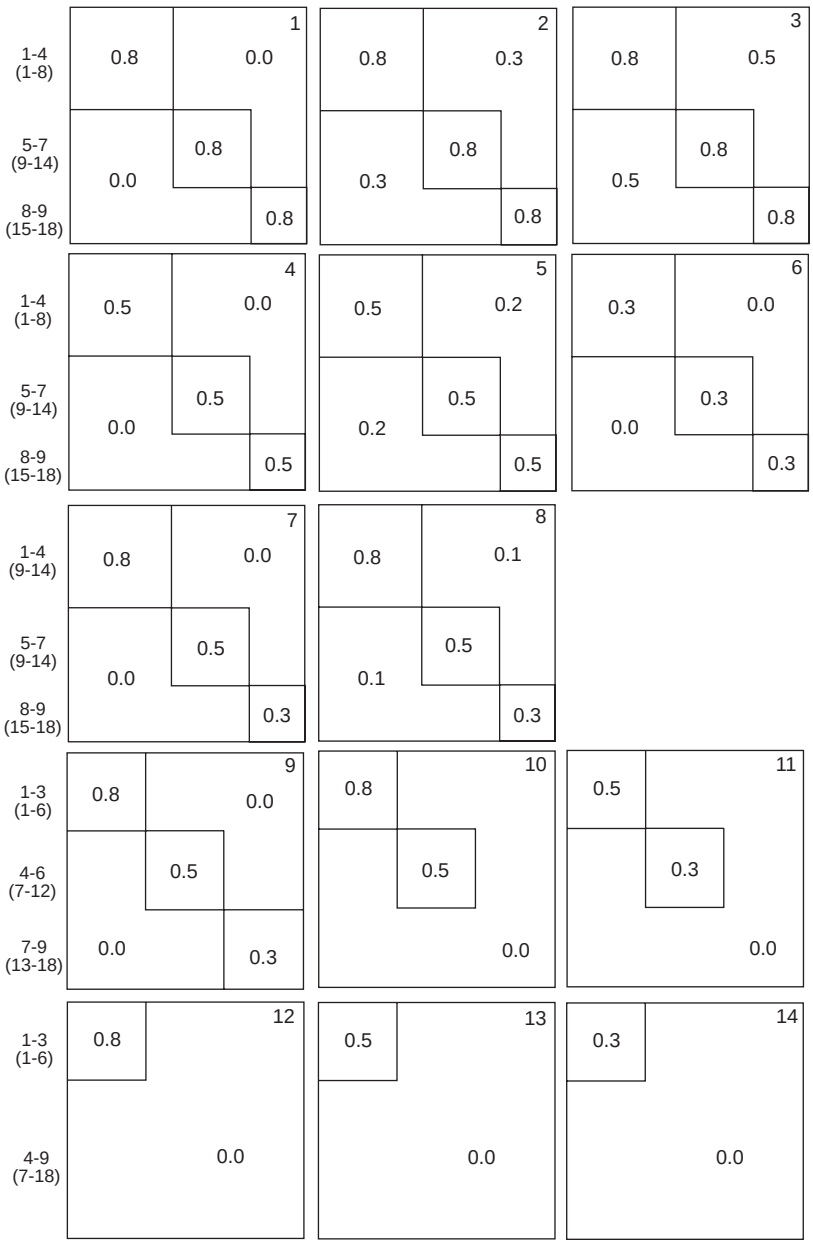

Fig. 1. Correlation matrices considered in this study. The values presented are the off-diagonal intervariable correlations. For example in matrix 1, variables 1–4 were correlated with one another at $r = 0.8$. However, correlations between these and all other variables were equal to 0.

number of observations (30 or 60) and number of variables (9 or 18) 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 by some method of matrix triangular factorization (in this case we used Cholesky decomposition), and (3) post-multiply 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 (30 observations by 9 variables, or 60 observations by 18 variables) in step 3 follows a multivariate distribution according to the marginal distribution and particular correlation matrix chosen (Fig. 1).

The last step was to estimate empirically the accuracy of each method in assessing the number of non-trivial components of the simulated populations. For each population containing 9 variables (i.e., a total of 54 populations as a result of combinations of each correlation matrix and each associated marginal distribution), 1000 random samples were drawn. Due to smaller sampling variation, populations containing 18 variables were evaluated on the basis of 500 random samples. Based on initial simulation trials, we determined that most rules tended to underestimate the number of non-trivial components; therefore we did not attempt to provide any correction for test probabilities for preventing inflated type I errors (Peres-Neto, 1999). We adopted a paired-test protocol where all rules were applied to the same samples for all populations, thus minimizing differences related to sampling variation in Monte Carlo simulations. A significance level of $\alpha = 0.05$ was used for evaluating the significance of procedures based on statistical tests, though equivalent results were found for $\alpha = 0.01$.

4. Analyzing simulation results

For each population (i.e., correlation structure associated with a distributional type), sample size and rule, we calculated the average of the absolute difference between the estimated and expected values across all estimates available. These averages estimated by how many components the rules for each population tended to be in error, and served as a measure of the overall quality of the estimation for each rule. For instance, a value of 0.0 would indicate perfect assessment; a value of 1.0 would indicate that the particular combination tended to over or underestimate the actual value by one axis. Note, however, that this average value does not assess the relative merits of over or underestimating the number of non-trivial components. To reduce the vast amount of results, this assessment was done subsequent to identifying the more accurate methods. For each combination of sample size and distributional type, a table containing the average of the absolute differences for each method (rows) and for each correlation matrix (columns) across all sample tests was built. An additional row containing zeros (i.e., after removing the number of non-trivial components by calculating absolute differences) was included so that methods could be contrasted to the correct number of non-trivial components. In order to assess the similarities among methods and their relative proximity to the correct number of non-trivial components, for each of these tables a matrix of Euclidean distances between all possible pairs of methods and the

correct number of non-trivial components (i.e., between rows of the tables described above) was calculated and a Principal Coordinate Analysis (PcoA; Legendre and Legendre, 1998) was conducted. Two of the stopping rules were analyzed separately from the PCoA (i.e., Bartlett's and Lawley's tests) as they only assessed the significance of the first and second eigenvalues, respectively.

5. Examining real data sets

The use of simulated data is preferable because the number of non-trivial components is known and conditions of interest can be manipulated. However, there is always the question of whether simulated data represent plausible biological scenarios and how many of the conclusions are directly applicable in real situations. We selected four data sets representing a gradient from weak to strong correlation structures as in our simulated scenarios (Fig. 1). Data set **MORPH1** represents 7 morphological variables for 38 bird species from Burgundy (France) and California (USA) studied by Blondel et al. (1984). Data set **MORPH2** comprises 6 morphological variables for 13 species of West Indian *Anolis* lizards (Losos, 1990). Data set **BEHAVIOR** represents 6 behavioral variables for the same 13 lizard species in **MORPH2.** Raw data for **MORPH2** and **BEHAVIOR** were both $\log_{10}$ transformed. Finally, data set **ENVIRON** represents 8 environmental variables for 42 lakes sampled by Robinson and Tonn (1989) in the Athabasca River basin. Lakes having missing data were deleted from the analysis. All variables, with the exception of pH, were $\log_{10}$ transformed.

6. Results

Summaries of the average absolute difference between each rule and the correct number of non-trivial components across all correlation matrices for each sample size and distributional type are presented in Table 1. Irrespective of matrix size and type of distribution, **Rnd-Lambda**, **Rnd-F**, **Avg-Rnd, PA**, **Avg-PA** and **Part** were the most accurate rules overall. As expected, accuracy is better for populations with larger number of variables and sample sizes. We removed **KG**, **Sphere**, **Bt-BS**, and **r-EigV** from further inspection due to their poor performance (Table 1).

The first two PCoA provided a good summary of differences in performance between rules (Fig. 2) given that they summarized larger amounts of variation. Only results for the normal distribution are presented as the other distributional types led to equivalent interpretations. For correlation matrices containing the same number of variables, PCoAs were extremely concordant (Fig. 2), suggesting that the relative performance of these rules is highly consistent irrespective of sample size and type of distribution. However, the results were not consistent across matrices having different number of variables. Because the amount of variation of equivalent axes was similar among scenarios, Euclidian distances among loadings representing a particular method are comparable between PCoAs. Overall tests based on 50 observations were more accurate than samples based on 30 observations for matrices with 9 variables. Similar findings resulted when comparing samples sizes of 100 and 60 observations for matrices with 18 variables.

Table 1
Values of the average absolute difference between sample test estimates and the known dimensionality across correlation matrices for each rule according to number of variables, sample size and distributional type

Rules	*9 variables*						*18 variables*					
	Normal		*E ponential*		*E ponential*^{3}		Normal		Exponential		Exponential^3	
	($n = 30$)	($n = 50$)	($n = 30$)	($n = 50$)	($n = 30$)	($n = 50$)	($n = 60$)	($n = 100$)	($n = 60$)	($n = 100$)	($n = 60$)	($n = 100$)
KG	1.13	1.07	1.14	1.07	1.1	1.05	2.46	2.21	2.47	2.22	3.74	3.68
BS	1.03	1.11	1.02	1.09	1.06	1.15	0.47	0.51	0.45	0.49	1.42	1.47
Rnd-Lambda	0.77	0.54	0.77	0.54	0.86	0.64	0.19	0.09	0.19	0.08	1.15	1.32
Rnd-F	0.77	0.54	0.77	0.54	1.88	5.6	0.19	0.09	0.19	0.08	4.51	5.05
Rnd-Ratio	0.91	0.73	0.91	0.73	1.02	0.82	0.43	0.32	0.43	0.31	0.97	1.09
Rnd-Delta	0.91	0.61	0.95	0.65	1.04	0.78	0.25	0.14	0.22	0.16	1.4	1.43
Avg-Rnd	0.66	0.54	0.65	0.53	0.63	0.51	0.29	0.20	0.29	0.23	1.61	1.83
Rnd-EigV	1.14	0.92	1.15	0.94	1.33	1.17	0.58	0.49	0.52	0.49	1.46	1.61
r-EigV	2.30	4.10	2.33	4.12	2.44	4.16	9.41	11.17	9.29	11.65	11.02	12.62
PA	0.75	0.54	0.75	0.54	0.81	0.62	0.25	0.14	0.23	0.14	1.31	1.55
Avg-PA	0.65	0.55	0.65	0.53	0.62	0.49	0.29	0.20	0.29	0.24	1.46	1.67
Bt-KG	1.00	0.91	0.99	0.92	1.03	1.00	1.90	1.87	2.01	1.88	2.88	2.98
Bt-BS	1.48	1.51	1.49	1.51	1.62	1.65	0.80	0.76	0.79	0.76	1.4	1.51
Bt-PA	0.86	0.72	0.86	0.71	0.93	0.78	0.71	0.61	0.68	0.59	0.92	1.09
Bt-RndAvg	0.86	0.72	0.86	0.71	0.91	0.75	0.71	0.61	0.69	0.60	1.06	1.24
BtAvg-RndAvg	0.93	0.88	0.93	0.88	0.92	0.83	1.18	1.15	1.17	1.13	2.32	2.55
Bt-Eigv	0.88	0.64	0.96	0.71	1.27	1.12	0.34	0.21	0.31	0.24	1.35	1.32
Sphere	1.45	1.46	1.80	1.78	2.43	2.41	4.81	3.76	3.29	4.04	6.78	6.72
Part(fo)	1.02	0.93	1.09	0.98	1.08	0.95	0.22	0.12	0.21	0.14	1.43	1.47
Part	0.68	0.63	0.70	0.64	0.73	0.66	0.26	0.17	0.20	0.17	0.92	0.9

The results represent averaged values from all simulated matrices. Methods in bold represent the smaller means across all scenarios. Acronyms are defined in the te t. Methods are in order of appearance in the methods section.

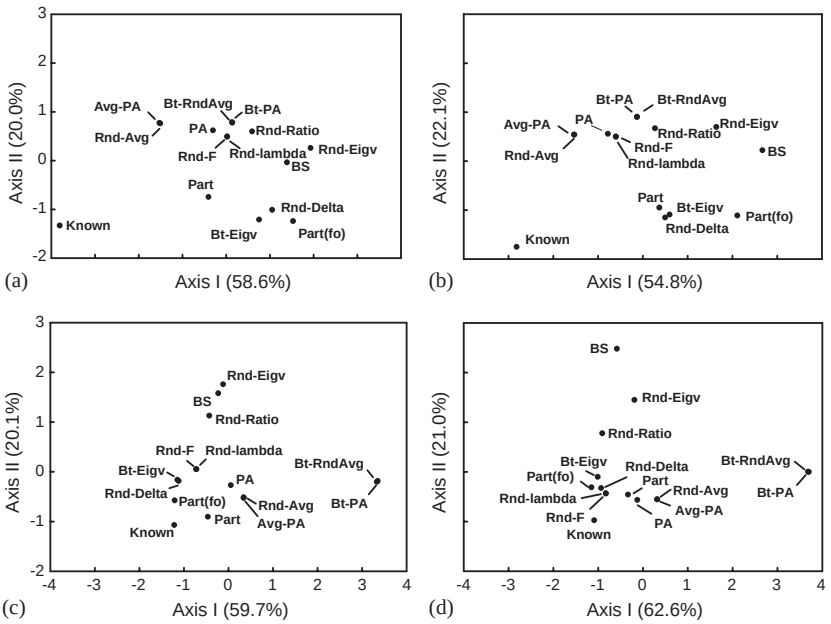

Fig. 2. First two axes from a principal coordinates analysis of the known dimensionality and that obtained from each stopping rule and the known dimensionality for normally distributed samples from matrices containing 9 variables based on (a) 30 observations, (b) 50 observations, and from matrix containing 18 variables based on (c) 60 observations and (d) 100 observations. Acronyms are defined in the text.

In order to acquire more specific details on the performance of the most reliable rules (i.e., **Rnd-Lambda**, **Rnd-F**, **Avg-Rnd**, **PA**, **Avg-PA** and **Part**; Table 1), for each rule we calculated the percentage of sample estimates that deviated a set amount from the correct number of non-trivial components for each correlation matrix. For simplicity of results, we only consider samples with 50 (Table 2) and 100 (Table 3) observations from the normal distribution. Although numerical performance differs when considering the other scenarios, the relative performance among rules did not (Fig. 2). All methods tended to underestimate the correct number of non-trivial axes for populations in which all variables were correlated at some level (i.e., matrices 1–9), whereas for matrices with uncorrelated variables, the rules tended to overestimate the number of non-trivial components (Tables 3 and 4). All methods were very precise in extracting the number of correct components from uniform matrices (matrices 15–18). Overall, **PA**, **Rnd-Lambda**, **Rnd-F** and **Part** were more precise in the presence of uncorrelated variance (i.e., matrices 10–14), whereas **Avg-PA** and **Avg-Rnd** performed better against data without uncorrelated variance (i.e., matrices 1–9). Rules based on confidence interval (**Rnd-Lambda, Rnd-F** and **PA**) were less sensitive to the amount of uncorrelated variance in the data, but did not perform as well for highly correlated data as was the case for **Avg-PA** and **Avg-Rnd** that are based on average values. **Part** underestimated the number of non-trivial components more than any other rule. Tests based on samples from the largest matrix with 18 variables (Table 3) provided greater accuracy.

Table 2
Percentage of number of non-trivial components estimated by **Rnd-Lambda**, **Avg-PA**, **PA** and **Bt-EiV** deviating from a set amount from the known dimensionality for each correlation matrix based on normally distributed samples for 9 variables and 50 observations

Matri	*Avg-PA (and Avg-Rnd)*							*PA*						
	$<=-3$	-2	-1	0	1	2	$>=3$	$<=-3$	-2	-1	0	1	2	$>=3$
1				0.9	99.1					3.8	96.2			
2			3.6	50.6	45.8				9.6	64.8	25.6			
3			70.6	28.2	1.2				84.8	15.0	0.2			
4			0.2	13.8	86.0				1.0	36.7	62.3			
5			18.2	56.0	25.8				41.0	52.3	6.7			
6	0.4		7.6	36.1	49.8	5.3	0.7	0.1	3.6	32.0	52.8	11.5	0.1		
7				12.3	87.6	0.1				0.5	31.6	67.9			
8			0.6	66.9	32.5					4.2	85.2	10.6			
9			0.1	55.2	44.7					1.4	82.7	15.9			
10				0.5	72.8	24.7	2.0				2.5	92.6	4.9		
11				8.8	51.6	31.2	8.0	0.4		0.9	33.7	59.1	6.2	0.1	
12					50.6	30.9	15.9	2.6				86.4	12.8	0.8	
13				0.2	42.8	33.5	16.7	6.8			4.6	76.5	17.5	1.3	0.1
14				12.9	35.7	25.5	17.2	8.7			42.9	47.0	9.5	0.6	
15					100.0							100.0			
16					100.0							100.0			
17					98.1	1.7	0.2					99.8	0.2		
18					50.7	20.5	13.8	15.0				69.0	25.4	5.1	0.5

Table 2. (*Continued*)

Matrix	Rnd-Lambda (and Rnd-F)							Part						
	$<=-3$	-2	-1	0	1	2	$>=3$	$<=-3$	-2	-1	0	1	2	$>=3$
1			3.8	96.2					3.4	24.7	71.9			
2		10.1	64.8	25.1					6.6	5.3	88.1			
3		85.4	14.4	0.2					10.6	28.8	60.6			
4		2.2	37.8	60.0					39.6	58.6	1.8			
5		43.4	49.9	6.7					45.9	53.0	1.1			
6	11.0	32.8	44.1	12.0	0.1				93.3	6.7				
7		1.4	31.1	67.5					28.0	53.3	18.7			
8		4.7	85.1	10.2					18.4	81.3	0.3			
9		1.4	82.7	15.9					27.5	72.3	0.2			
10			2.6	92.2	5.2					15.6	83.8	0.6		
11		3.5	35.4	55.8	5.3					82.1	17.9			
12			0.0	88.0	11.1	0.9	0.0				97.7	2.3		
13			6.8	80.0	11.8	1.3	0.1				99.6	0.4		
14			59.3	35.1	5.3	0.3	0.0				100.0			
15				100.0							100.0			
16				100.0							100.0			
17				99.8	0.2						99.9	0.1		
18				95.6	3.5	0.8	0.1					100.0		

Avg-Rnd provided extremely similar results to **Avg-PA**, and the same with **Rnd-Lambda** and **Rnd-F** and thus their results were not tabulated, but for convenience they are indicated in the table with the rule that provided similar outcome. Differences of zero indicate perfect assessment. Deviations of more than 3 components were combined and those less than 3 components were treated similarly. Acronyms are defined in the text.

Table 3
Percentage of number of non-trivial components estimated by **Rnd-Lambda**, **Avg-PA**, **PA** and **Bt-EiV** deviating from a set amount from the known dimensionality for each correlation matrix based on normally distributed samples for 18 variables and 100 observations

Matrix	Avg-PA (and Avg-Rnd)							PA						
	$<=-3$	-2	-1	0	1	2	$>=3$	$<=-3$	-2	-1	0	1	2	$>=3$
1				100.0							100.0			
2				100.0							100.0			
3			49.0	51.0					2.0	64.0	34.0			
4				100.0						0.0	100.0			
5			3.0	97.0						11.0	89.0			
6			1.0	98.0	1.0					4.0	96.0			
7			0	100.0							100.0			
8			12.0	88.0						26.0	74.0			
9			2.0	98.0						5.0	95.0			
10				94.0	6.0						100.0			
11				77.0	22.0	1.0					94.0	6.0		
12				68.0	26.0	6.0					89.0	11.0		
13				60.0	30.0	7.0	3.0				88.0	12.0		
14				46.0	37.0	12.0	5.0				88.0	9.0	3.0	
15				100.0							100.0			
16				100.0							100.0			
17				100.0							100.0			
18				53.0	25.0	7.0	15.0				34.0	46.0	16.0	4.0

Table 3. (*Continued*)

Matrix	Rnd-Lambda (and Rnd-F)							Part						
	$<=-3$	-2	-1	0	1	2	$>=3$	$<=-3$	-2	-1	0	1	2	$>=3$
1				100.0							100.0			
2				100.0					1.0		99.0			
3		2.0	64.0	34.0					2.0		98.0			
4				100.0							100.0			
5				88.0						1.0	99.0			
6			7.0	93.0					5.0	59.0	36.0			
7			0.0	100.0							100.0			
8			26.0	74.0						70.0	30.0			
9			5.0	95.0						58.0	42.0			
10				99.0	1.0						100.0			
11				96.0	4.0					3.0	97.0			
12				91.0	9.0						100.0			
13				90.0	10.0						100.0			
14				90.0	7.0	3.0					100.0			
15				100.0							100.0			
16				100.0							100.0			
17				100.0							100.0			
18				94.0	4.0	2.0					0.0	100.0		

Avg-Rnd provided extremely similar results to **Avg-PA**, and the same with **Rnd-Lambda** and **Rnd-F** and thus their results were not tabulated, but for convenience they are indicated in the table with the rule that provided similar outcome. Differences of zero indicate perfect assessment. Deviations of more than 3 components were combined and similarly those less than 3 were combined. Acronyms are defined in the text.

Table 4
Percentage of significant results for the Bartlett and Lawley tests according to sample size and distributional type

Matri	*Bartlett*								*Lawley*							
	9 variables				*18 variable*				*9 variables*				*18 variable*			
	Normal		*E ponential*[3]		Normal		Exponential[3]		Normal		Exponential[3]		Normal		Exponential[3]	
	($n = 30$)	($n = 50$)	($n = 30$)	($n = 50$)	($n = 60$)	($n = 100$)	($n = 60$)	($n = 100$)	($n = 30$)	($n = 50$)	($n = 30$)	($n = 50$)	($n = 60$)	($n = 100$)	($n = 60$)	($n = 100$)
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	0.61	0.91	0.60	0.75	1.00	1.00	0.91	0.99	0.97	0.99	0.99	1.00	1.00	1.00	0.99	1.00
7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
11	0.72	0.97	0.68	0.89	0.98	1.00	0.96	1.00	0.95	0.98	0.96	1.00	0.98	1.00	0.99	1.00
12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.98	1.00	1.00	1.00	1.00	1.00
13	0.57	0.88	0.54	0.72	1.00	1.00	0.88	0.99	0.85	0.92	0.89	0.95	1.00	1.00	0.95	0.99
14	0.18	0.34	0.16	0.29	0.72	0.99	0.53	0.64	0.76	0.81	0.75	0.82	1.00	1.00	0.82	0.90
15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
17	0.98	1.00	0.80	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
18	0.06	0.05	0.31	0.30	0.07	0.05	0.39	0.35	0.46	0.47	0.44	0.45	0.37	0.50	0.44	0.44

The e ponential distribution is not shown as it presents results similar to the normal distribution.

Table 5
Eigenvalues and relative proportion of variation explained (in brackets) for the first 4 principal components based on the ecological data sets and the number of components retained for each rule

Component	MORPH1	MORPH2	PERFORM	ENVIRON
	Eigenvalues			
1	4.90 (70.0)	4.60 (76.7)	2.55 (42.5)	1.94 (24.2)
2	0.83 (11.9)	0.71 (11.9)	1.69 (28.1)	1.64 (20.5)
3	0.55 (7.9)	0.48 (8.1)	1.03 (17.2)	1.31 (16.3)
4	0.36 (5.1)	0.14 (2.3)	0.45 (7.5)	0.98 (12.3)
Rule	Number of components retained			
BS	1	1	3	0
Rnd-lambda	1	1	1	0
Rnd-F	1	1	1	0
Rnd-Ratio	1	1	1	2
Rnd-Delta	1	1	0	0
Avg-Rnd	1	1	2	3
Rnd-EigV	1	1	0	1
PA	1	1	1	1
Avg-PA	1	1	2	3
Bt-KG	1	1	2	3
Bt-PA	1	1	1	2
Bt-Rnd Avg	1	1	1	2
BtAv-RndAvg	1	1	2	3
Bt-Eigv	4	1	1	1
Part(fo)	1	1	0	0
Part	1	1	1	1
Bartlett	1	1	1	1
Lawley	2	2	0	2

Acronyms are defined in the text.

Performance of the Bartlett's test for the first principal component and Lawley's test for the second principal component were evaluated on the basis of number of significant tests out of the total number of sample tests for each population (Table 4). Bartlett's test was very efficient in detecting the presence of at least one component in most cases, being affected slightly by sample size and by samples from the exponential[3] distribution. Lawley's test did not discriminate between matrices with greater or fewer than two components (e.g., contrast results for matrices 11 and 12) and also provided poor assessment when evaluating the spherical population indicating in some cases that 50% of the samples contained two non-trivial components. The distributional type was influential in both tests.

The results for the real data sets are presented in Table 5. Note that **MORPH1** and **MORPH2** comprise highly correlated variables, such that the first principal component explains 70.0% and 76.7% of the total variation, respectively. For these two data sets, the majority of rules retained only the first component. The data set **BEHAVIOR** represents a situation where the total variation is partitioned more evenly among more components, so that the first one represents only 42.5% of the total variation (Table 5). Most rules tended

to consider between 1 or 2 non-trivial components. Finally, **ENVIRON** is a case where variables are weakly correlated, and thus the first component only represents 24.2% of the total variation. In this last case, the rules showed little agreement, extracting between 0 and 3 components as being interpretable (Table 5). Thus, as indicated by the simulation results, matrices where correlations were larger provided greater agreement between rules.

7. Discussion and concluding remarks

Our intent in using a large number of scenarios was twofold: (1) establish overall rule performance; and (2) capitalize on the chances of establishing general trends in their statistical behavior. Given that rule performance is also dependent on correlation structure (Tables 3 and 4), it seems inappropriate to establish quality of estimation based only on overall performance because conclusions depend on how many of each particular correlation structures were used. For instance, we found that some rules outperform others in the presence of uncorrelated variables or where matrices contain larger number of variables. Moreover, if only highly correlated variables are used, one can have a better chance of finding that some methods perform better or perhaps more similarly. For instance, Zwick and Velicer (1986) found that **Part(fo)** was very accurate for components with large eigenvalues, or where there was an average of eight or more variables per component. Because these features were not in our matrices, this may account for why **Part(fo)** did not perform well.

We attempted to establish matrices that approximate the typical size of ecological and evolutionary data sets (between 9 and 18 variables) and that provided a wide range of eigenvalue distributions along trivial and non-trivial components (Fig. 1). We found that **Avg-PA**, **Avg-Rnd**, **PA**, **Rnd-Lambda**, **Rnd-F** and **Part** are the most promising rules for component evaluation. They were quite accurate, especially considering some of the weakly correlated structures we have employed, and in general, at least 80% of the samples under or overestimated the correct number of non-trivial components by only one component (Tables 2 and 3). However their performance varied across different correlation scenarios. **Avg-PA** and **Avg-Rnd** outperformed the other ones when data did not contain uncorrelated variables. **Rnd-Lambda** and **Rnd-F** were more effective against uncorrelated variance (Tables 2 and 3). It seems that because **Avg-PA** and **Avg-Rnd** are based on less stringent criteria (i.e., average rules) than confidence interval rules, they were too liberal, thus providing good performance for highly correlated data, but not for data containing uncorrelated variables. Perhaps a less conservative interval (e.g., 80%) would be more appropriate, maximizing the chances of correctly estimating the correct number of non-trivial components of highly correlated data, while still minimizing excessive trivial components when data contain uncorrelated variance.

Given the difference in rule performance, it is interesting to evaluate the relative disadvantages of over or underextraction of number of principal components. Fava and Velicer (1992) anticipated that overextraction might not be as serious a problem as underextraction because the amount of variation for each component decreases successively. In fact, their simulation results indicated that, at least for strong correlation structures, overextraction did not degrade the patterns related to multivariate scores, whereas in weakly correlated data structures produced very negative effects. Lawrence and Hancock (1999) were concerned

about the integrity of eigenvector loadings under overextraction and found similar results compared to Fava and Velicer (1992). Under these considerations, if one wants to be conservative and is willing to lose information rather than incorporate noise into the model, the use of **Rnd-Lambda** or **Rnd-F** seems to be more appropriate than **Avg-PA** and **Avg-Rnd** (Tables 2 and 3). Another uncertainty related to the issue of overextraction is that sample axes, especially the later ones, can appear reordered in relation to the PCA solution for the population. Thus although rules may provide a correct assessment, there is no guarantee that the correct components are being retained relative to the population PCA, thus potentially compromising interpretation.

Some methods falsely assigned at least one non-trivial component in spherical populations (matrix 18, Tables 2 and 3). To avoid this problem, one could first apply a particularly efficient method to detect if the matrix contain at least one non-trivial component and then use another method to estimate the number of principal components. This approach is handicapped because most rules also failed in assessing correlation matrices with at least one non-trivial component (Tables 2 and 3). Overall, the Bartlett's test seems the best method for this assessment (Table 4). It maximized the chances of detecting at least one non-trivial component for matrices 1–14, and identified samples from matrix 18 as coming from a spherical population. Therefore, we suggest the Bartlett's test should be applied first as a means to assess whether data contain at least one principal component. Then, if the null hypothesis is rejected, use another rule to quantify the number of non-trivial components.

We have contrasted the different methods using real data so that we could evaluate to what degree our conclusions based on the simulated scenarios are congruent with real data sets. Given that we considered an extensive number of scenarios, the real data sets should resemble some of the populations manipulated. In fact, the differences in estimates between methods for the real data sets were in agreement with the differences found in the simulation study. However, one possible avenue to follow could be that the researcher specify correlation matrices that resemble his or her particular data (e.g., sample size, number of variables, correlation structure), conduct a customized simulation study, similar to the one presented here, and select a method that maximizes the chances of correctly retaining the correct number of non-trivial components. Dimensionality may be then introduced by transforming small correlation values (say 0.25 and lower) into zero correlations. Note that attention has to be paid in order to guarantee that the matrix is positive definite. We hope that this process may bring some general knowledge about the behavior of different methods in cases of particular correlation structure of interest.

In conclusion, our main goal was to conduct a comparative study so that differences in performance and behavior of available and new methods could be contrasted and revealed. Sample size and great departure from normality affected rule performance. However, the most important finding was that the performance of the methods was largely dependent on whether data contained uncorrelated variables and on the size of the correlation matrix. Once at least the first component is detected as significant by the Bartlett's test, we suggest that **Avg-PA**, **Avg-Rnd**, **PA**, **Rnd-Lambda**, **Rnd-F** or **Part** can be applied. However, one should keep in mind the impact of different types of correlation structures and the relative merits of under and overestimating the number of non-trivial components in the ordination.

Acknowledgements

We would like to thank Pierre Legendre and two anonymous reviewers for valuable comments regarding our study. Funding was provided by a CNPq Fellowship to PRP-N and an NSERC operating grant to DAJ. Software for running all the analyses conducted here is available from the senior author.

References

Anderson, M.J., Legendre, P., 1999. An empirical comparison of permutation methods for tests of partial regression coefficient in a linear model. J. Statist. Comput Simulation 62, 271–303.

Anderson, T.W., 1984. An Introduction to Multivariate Statistical Analysis. 2nd Edition. Wiley, New York.

Barr, D.R., Slezak, N.L., 1972. Comparison of multivariate normal generators. Comm. Assoc. Comput. Mach. 15, 1048–1972.

Bartlett, M.S., 1950. Tests of significance in factor analysis. British J. Psych. (Statistical Section) 3, 77–85.

Bartlett, M.S., 1954. A note on the multiplying factors for various X^2 approximations. J. Roy. Statist. Soc. Ser B 16, 296–298.

Blondel, J., Vuilleumier, F., Marcus, L.F., Terouanne, E., 1984. Is there ecomorphological convergence among Mediterranean bird communities of Chile, California, and France?. Evol. Biol. 18, 141–213.

Buja, A., Eyuboglu, N., 1992. Remarks on parallel analysis. Mult. Beh. Res. 27, 509–540.

Crawford, C.B., 1975. Determining the number of interpretable factors. Psych. Bul. 82, 226–237.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Statist. 7, 1–26.

Fava, J.L., Velicer, W.F., 1992. The effects of overextraction on factor and component analysis. Mult. Beh. Res. 27, 387–415.

Férre, L., 1995. Selection of components in principal component analysis: a comparison of methods. Comput. Statist. Data Anal. 19, 669–682.

Franklin, S.B., Gibson, D.J., Robertson, P.A., Pohlmann, J.T., Fralish, J.S., 1995. Parallel analysis: a method for determining significant principal components. J. Veg. Sci. 6, 99–106.

Frontier, S., 1976. Étude de la decroissance des valeurs propers dans une analyze en composantes principales: comparison avec le modèle de baton brisé. J. Exp. Mar. Biol. Ecol. 25, 67–75.

Gauch Jr., H.G., 1982. Noise reduction by eigenvector ordination. Ecology 63, 1643–1649.

Grossman, G.D., Nickerson, D.M., Freeman, M.C., 1991. Principal component analysis of assemblage structure data: utility of tests based on eigenvalues. Ecology 72, 341–347.

Guttman, L., 1954. Some necessary conditions for common factor analysis. Psychometrika 19, 149–161.

Horn, J.L., 1965. A rationale and test for the number of factors in factor analysis. Psychometrika 30, 179–185.

Jackson, D.A., 1993. Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches. Ecology 74, 2204–2214.

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. Wiley, New York.

Jolliffe, I.T., 2002. Principal Component Analysis. 2nd Edition. Springer, New York.

Karr, R.J., Martin, T.E., 1981. Random numbers and principal components: further searches for the unicorn. In: Capen, D.E. (Ed.), The Use of Multivariate Statistics in Studies of Wildlife Habitat. United States Forest Service General Technical Report RM-87.

Knox, R.G., Peet, R.K., 1989. Bootstrapped ordination: a method for estimating sampling effects in indirect gradient analysis. Vegetation 80, 153–165.

Lambert, Z.V., Wildt, A.R., Durand, R.M., 1990. Assessing sampling variation relative to number-of-factors criteria. Ed. Psych. Meas. 50, 33–49.

Lawley, D.N., 1956. Tests of the significance for the latent roots of covariance and correlation matrices. Biometrika 43, 128–136.

Lawrence, F.R., Hancock, G.R., 1999. Conditions affecting integrity of a factor solution under varying degrees of overextraction. Ed. Psych. Meas. 59, 549–579.

Legendre, P., 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. J. Statist. Comput. Simulation 67, 37–73.

Legendre, P., Legendre, L., 1998. Numerical Ecology. 2nd English Edition. Elsevier Science BV, Amsterdam.

Losos, J.B., 1990. Ecomorphology, performance capability, and scaling of West Indian Anolis lizards: an evolutionary analysis. Ecol. Monog. 60, 368–388.

Manly, B.J.F., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. 2nd Edition. Chapman and Hall, London.

Mehlman, D.W., Shepherd, U.L., Kelt, D.A., 1995. Bootstrapping principal component analysis—a comment. Ecology 76, 640–643.

Milan, L., Whittaker, J., 1995. Application of the parametric bootstrap to models that incorporate a singular value decomposition. Appl. Statist. 44, 31–49.

Peres-Neto, P.R., 1999. How many statistical tests are too many? The problem of conducting multiple ecological inferences revisited. Mar. Ecol. Prog. Ser. 176, 303–306.

Peres-Neto, P.R., Jackson, D.A., 2001. 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., Olden, J.D., 2001. Assessing the robustness of randomization tests: examples from behavioural studies. An. Beh. 61, 79–86.

Peres-Neto, P.R., Jackson, D.A., Somers, K.M., 2003. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. Ecology 84, 2347–2363.

Pimentel, R.A., 1979. Morphometrics: the Multivariate Analysis of Biological Data. Kendall-Hunt, Dubuque.

Reddon, J.R., 1985. Monte Carlo type one error rates for Velicer's partial correlation test for the number of principal components. Criminometrica 1, 13–23.

Robinson, C.L.K., Tonn, W.M., 1989. Influence of environmental factors and piscivory in structuring fish assemblages of small Alberta lakes. Canad. J. Fish. Aqua. Sci. 46, 81–89.

Seber, G.A.F., 1984. Multivariate Observations. Wiley, New York.

Stauffer, D.F., Garton, E.O., Steinhorst, R.K., 1985. A comparison of principal components from real and random data. Ecology 66, 1693–1698.

ter Braak, C.F.J., 1988. CANOCO—a Fortran program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis (version 2.1), Agricultural Mathematic Group, Report LWA-88-02, Wageningen.

ter Braak, C.F.J., 1990. Update notes: CANOCO (version 3.1). Agricultural Mathematic Group, Wageningen.

Velicer, W.F., 1976. Determining the number of components from the matrix of partial correlations. Psychometrika 41, 321–327.

Zwick, R.W., Velicer, W.F., 1986. Comparison of five rules for determining the number of components to retain. Psych. Bull. 99, 432–442.