ethology international journal of behavioural biology

Ethology

Using Principal Components and Factor Analysis in Animal Behaviour Research: Caveats and Guidelines

Sergey V. Budaev

School of Life Sciences, University of Sussex, Brighton, UK

Correspondence
Dr Sergey V. Budaev, Centre for
Neuroscience, School of Life Sciences,
University of Sussex, Brighton BN1 9QG, UK.
E-mail: sbudaev@gmail.com

Received: February 1, 2010
Initial acceptance: February 5, 2010
Final acceptance: February 5, 2010
(J. Kotiaho)

doi: 10.1111/j.1439-0310.2010.01758.x

Abstract

Principal component (PCA) and factor analysis (FA) are widely used in animal behaviour research. However, many authors automatically follow questionable practices implemented by default in general-purpose statistical software. Worse still, the results of such analyses in research reports typically omit many crucial details which may hamper their evaluation. This article provides simple non-technical guidelines for PCA and FA. A standard for reporting the results of these analyses is suggested. Studies using PCA and FA must report: (1) whether the correlation or covariance matrix was used; (2) sample size, preferably as a footnote to the table of factor loadings; (3) indices of sampling adequacy; (4) how the number of factors was assessed; (5) communalities when sample size is small; (6) details of factor rotation; (7) if factor scores are computed, present determinacy indices; (8) preferably they should publish the original correlation matrix.

Introduction

Principal components analysis (PCA) is frequently used in animal behaviour research. It aims to reduce numerous measures to a small set of the most important summary scores (Nunnally 1978; Gorsuch 1983; Tabachnick & Fidell 1996). First, correlations (or variances and covariances if the scale is crucial; correlations are virtually always used in animal behaviour research) between the original behavioural measures are calculated. Second, the correlation matrix is subjected to specific transformations, resulting in a new set of linear combinations of the original measures (principal components) that are orthogonal to each other and account, each, for decreasing proportion of the total variance. Third, loadings of the original measures on these principal components are calculated, which represent correlations between the original measure and the principal components. Then, principal components can substitute for the more numerous original variables. Correlations between components and other external variables can be calculated, average component scores between various treatment groups can be compared, etc. The use of a few orthogonal principal components may help to reduce such problems as multicollinearity of the original variables and, when testing hypothesis, reduce the number of statistical tests. Additionally, the position of the data points in the coordinate space of the principal components reveals patterns and clusters in the data.

Researchers often try to interpret principal components to obtain a better understanding of the patterns of correlations between the original variables, emphasising similarities of this method with factor analysis (FA). Typically, variables loading on the same principal component (i.e., having high correlations with it) share a significant proportion of common variance and are thought to imply a common cause (e.g., behavioural mechanism). For example, various measures of risk taking are often found to be correlated, pointing to an underlying trait of boldness (Budaev & Zworykin 2002; Sih et al. 2004; Reale et al. 2007).

Although PCA and FA are often considered ‘well known’ and relatively ‘elementary’, there exist several potential caveats (Short & Horn 1984; Fabrigar et al. 1999; Henson & Roberts 2006). Here, I provide simple non-technical guidelines for the use of PCA and FA (see Gorsuch 1983; Tabachnick & Fidell 1996; Fabrigar et al. 1999; Henson & Roberts 2006 for more technical discussions). I also review how PCA and FA are currently reported in published research on animal behaviour. Finally, a standard for reporting results of principal components and FA is suggested.

An Artificial Example: Boldness in Fish

Let us consider a typical study using such an approach in the growing field of animal personality (Budaev & Zworykin 2002; Sih et al. 2004; Reale et al. 2007). A sample of fish from two populations with high- and low-predation pressure was tested in five behavioural tests, involving (1) locomotor activity; (2) aggression; (3) exploration of novel environment; (4) exploration of novel food; (5) exploration of a novel object.[1] Two separate principal component analyses were conducted in these two populations, resulting in loading matrices presented in Table 1. In both cases, the first two principal components accounted for considerable proportion of the total variance (64% and 62%). The patterns of factor loadings can be interpreted easily in terms of a general dimension of boldness. In both populations, general activity, aggressiveness and response to novel environment were correlated. Notably, response to novel food was largely unrelated to boldness. There is also an interesting population difference: responses to novel object were linked with boldness in the high, but not in the low-predation population. Overall, these results would agree with the existing literature documenting both consistency and population differences in boldness (e.g., Sih et al. 2004; Reale et al. 2007).

However, for one important reason all the above results and conclusions are meaningless: namely that the 'original data' were random uncorrelated and normally distributed numbers generated by the computer. The variable labels were randomly chosen from a set frequently used in animal personality literature. While performing the analysis, I used the default options built in most general-purpose statistical packages (principal component analysis, extract factors with eigenvalues > 1, Varimax rotation). With the easy availability of statistical software using simple graphical user interface, one could produce thoughtless, ignorant and sometimes overtly wrong 'automatic' data analysis.

Table 1: Principal components analysis of boldness-related scores from two populations

	High predation		Low predation	
Behavioural measure	PC1	PC2	PC1	PC2
Locomotion	**0.89**	−0.34	**−0.82**	0.01
Aggression	**−0.49**	**−0.64**	**0.51**	**−0.77**
Novel environment	**−0.82**	−0.08	**−0.77**	−0.11
Novel food	−0.12	**0.88**	−0.03	−0.26
Novel object	**−0.40**	−0.07	−0.35	**−0.87**
Eigenvalue	1.9	1.3	1.7	1.4

Interpretable factor loadings are in bold. The original data are artificial: random normal variates generated by the computer.

This example shows that an arbitrary data can be 'meaningfully' interpreted and placed into a broader context involving other published research (see also Armstrong & Soelberg 1968). The above 'study' could pass the usual peer-review process: many referees and readers would be happy with the statistical analysis, although some aspects of the interpretation may be disputed. How could one make sure that the results of such analyses are not attributed to just sampling error?

Guidelines for Principal Components and Factor Analysis

Both FA and PCA are examples of exploratory analysis. They are used to *summarise* the data and *generate* hypotheses (see Haig 2006 for more discussion). Neither method usually involves explicit testing of specific hypotheses. When conducting PCA or FA, the researcher is concerned with several important questions: (1) whether the PCA or common FA should be used; (2) minimum sample size; (3) the optimal number of the original measures and their sampling characteristics (e.g., the overall level of correlations); (4) what is the optimal number of components/factors to extract; (5) whether factor rotation is necessary and which method should be used; (6) which loadings should be considered for the interpretation of the factors; (7) how factor scores are calculated; (8) how repeated measures should be treated. As noted above, many general-purpose software packages silently provide default analysis options, which are often not optimal and could lead to wrong results.

Principal Components or Factor Analysis?

Both historically and theoretically, PCA and FA represent different data analysis methods (Gorsuch 1983). There has been a long debate about the

[1] The data are artificial, generated by the computer.

relationship between PCA and FA (see Velicer & Jackson 1990 and other papers in the same issue). Often they are considered simply as two types of 'factor analysis'. Although their results are sometimes very similar (Velicer & Jackson 1990), PCA and FA represent conceptually distinct models (see Gorsuch 1983, 1990, 1997). PCA is a dimensionality reduction method, whereas the purpose of FA is to measure an unobservable latent construct that accounts for correlations between variables. PCA assumes that all variability in the data is accounted for by the PCs, FA provides a two-component model including both latent common factors and an error term (unique components specific to each variable).

Generally, PCA is most appropriate when the main objective is just to reduce the number of dimensions. For example, the researcher may wish to avoid multicollinearity of independent variables in a regression analysis, use a single composite index instead of several available measures of body size or combine several related (but not mechanistically linked, see Short & Horn 1984) behavioural measures into a single score. FA is most appropriate when the main aim is to determine and assess unobservable behavioural constructs. In particular, most studies seeking to identify the dimensions of animal personality should naturally use the FA. The habitual use of PCA instead of FA is an historical computational compromise: PCA involves very few simple matrix operations, whereas FA requires complex iterative calculations. Although computational difficulty is no longer an issue (even a feeble desktop PC now outperforms old mainframes), the historical practices are still followed.

Principal component analysis does not include an error term and tends to inflate factor loadings. For example, low and non-significant correlations can easily produce high PCA loadings, which is unlikely in FA (Gorsuch 1983, 1997). The above fictitious study indicates that PCA of a random correlation matrix easily produces factors accounting for a significant proportion of the (very low) common variance with high loadings. FA of this matrix is computationally impossible, no factors can be extracted because all variance goes to the error term. Nonetheless, if the data are well conditioned – all variables have high communalities (communality is the proportion of variance due to common factors), several variables load highly on the same factor and the number of factors is correctly specified (model error is low) – FA and PCA yield almost identical results (Schneeweiss 1997).

One potential caveat is that there exist several approaches to factor extraction in FA (see Gorsuch 1983; Tabachnick & Fidell 1996). The maximum likelihood FA has an advantage of various goodness of fit indices (e.g., CFI, RMSEA) and statistical significance tests (Chi-squared test). However, its use in animal behaviour research is extremely limited by high sensitivity to deviations from the normality assumption and requirements of very large sample size (see Fabrigar et al. 1999). Fabrigar et al. recommend the use of the principal axis factoring in all other cases, making it the only feasible choice in animal behaviour research. A simple rule of thumb, therefore, is to perform PCA if factors are not interpreted and used only to reduce the dimensionality, principal axis FA model is better when measurement is involved, factors are interpreted and meaningfully labelled.

Sample Size, the Number of Variables and Their Sampling Characteristics

Most texts on FA state that quite a large sample size is needed to ensure stable assessment of the raw correlation coefficients. Gorsuch (1983) recommends a minimum sample size of 100, others often require even larger minimum sample size (see MacCallum et al. 1999 for a review). Further, many texts point out that the variable to subject ratio is more crucial than absolute sample size (see Velicer & Fava 1998; MacCallum et al. 1999 for reviews). For example, Nunnally (1978) suggested that the sample size should be at least ten times the number of variables, other researchers proposed a less stringent rules of 5:1 (Gorsuch 1983) or 3:1 (Gorsuch 1997). Recent investigations, however, show that such rigid rules may be too simplistic: good recovery of the true population factor structure depends more on communalities of the variables and the number of variables per factor ('overdetermination', minimum 3, Velicer & Fava 1998; MacCallum et al. 1999, 2001). If the data are well conditioned, then FA can be legitimately conducted even on much smaller sample size (MacCallum et al. 1999, 2001; Preacher & MacCallum 2002). For example, Preacher & MacCallum (2001) found that in the nearly ideal case of highly reliable variables (e.g., in behavioural genetics, representing stable strain means rather than individual scores), sample size approx. 20–30 may be adequate. This is supported by de Winter et al. (2009) who gave evidence that in such cases FA can provide reliable results with n = 25 and sometimes even less. Notably, the number of variables in well-conditioned data may even exceed the sample size, a highly unusual situation explicitly forbidden by most textbooks (e.g., Gorsuch 1983).

The fact that FA can be used with small samples, however, does not provide a license for its indiscriminate applications. Well-conditioned data are likely to be relatively infrequent in animal behaviour studies. Possible exceptions include highly stable and reliable morphological measures or aggregated or composite behavioural indices in animal personality research. Another notable exception is highly reliable measures representing group averages (e.g., strains or phenotypes in behavioural genetics research) rather than individual values (Preacher & MacCallum 2002). As a general rule, PCA and FA usually require large sample ($n > 100$), however, if the data are well conditioned (original measures are highly reliable, there are a few well-defined factors, communalities are all high) it can be adequately conducted with such small samples as $n = 25$.

Measures of Sampling Adequacy

Two measures have been developed to determine the sampling adequacy of the correlation matrix for FA: the Bartlett sphericity test and the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy. The former tests whether all correlations are zero, whereas the second compares the observed correlations and partial correlations among the original variables.

The power of the Bartlett's test is relatively high (see Dziuban & Shirkey 1974; Fouladi & Steiger 1993), however it depends on the chi-squared approximation and assumes that the measures are normally distributed. In practical applications, it sometimes rejects the null hypothesis even when the correlation matrix is ill conditioned (Gorsuch 1983). Dziuban & Shirkey (1974) recommend that Bartlett's test 'may be used as a lower bound to the quality of the matrix. This is, if one fails to reject the independence hypothesis, the matrix need be subjected to no further analysis. On the other hand, rejection of the independence hypothesis on the Bartlett test is not a clear indication that the matrix is psychometrically sound (p. 360).' There is also a less known alternative to the Bartlett's test, proposed by Steiger (see Fouladi & Steiger 1993) which has higher power, especially with small samples.

Dziuban & Shirkey (1974) provided a review of the Bartlett's test and KMO. Simulation study by Dziuban et al. (1979) investigated the behaviour of KMO in relation to the sample size, communality and other factors. It was shown that KMO increases with increase of the number of variables and the correlation coefficients between them but does not much depend on the sample size. Specifically, correlation matrices with KMO < 0.5 are entirely inappropriate whereas those with KMO below 0.6–0.7 must be treated with caution.

A few general-purpose statistical software packages (e.g., SAS [SAS Institute Inc., Cary, NC, USA] and SPSS [SPSS Inc., Chicago, IL, USA]) allow calculation of the Bartlett's test and KMO. There is also an R function cortest.bartlett (package psych, R Development Core Team 2008) and a free stand-alone Windows utility available on the internet (Budaev 1997). Steiger's test can be calculated using the cortest.mat function in R (package psych). It is highly recommended to calculate the Bartlett's or Steiger's tests and especially KMO prior to PCA or FA: the correlation matrix is appropriate if the hypothesis of all zero correlations is rejected and KMO significantly exceeds 0.5.

How Many Factors Should be Extracted?

This is one of the most important issues in PCA and especially FA; what will occur if the number of factors is wrongly determined? Gorsuch (1983) discussed several studies that tried to extract successive factors from the same data with known factor structure. He found that the factor pattern may be wrong when too few factors are extracted ('underfactoring', several sources of variability are confounded), it stabilises around the true number of factors and does not change much with addition of a few more factors ('overfactoring', the spare factors will be random and non-reproducible). In case of a doubt, it is better to extract more factors and simply drop those which are not theoretically interpretable or reproducible across studies (also see Gorsuch 1983; Fabrigar et al. 1999).

Several techniques are used to determine the optimal number of factors inherent in the data. The most widely used rule is to extract factors with eigenvalues > 1. Although this remains the default method in most general-purpose statistical packages, many studies have shown that it performs extremely poorly in most cases (see Revelle & Rocklin 1979; Zwick & Velicer 1986; Fabrigar et al. 1999). Another popular technique, the Cattell's scree-test, involves visual examination of the eigenvalue plot and is therefore relatively subjective, requires some experience and often involves running several factor analyses with different number of factors (see Gorsuch 1983 for more detail). Though, a non-graphical procedure nScree is implemented in R (package nFactors, R Development Core Team 2008).

More computationally complicated methods for the assessment of the optimal number of factors have been developed: Very simple structure (VSS, Revelle & Rocklin 1979), minimum average partial (MAP, Velicer 1976), parallel analysis (PA, Horn 1965) and a few others. Modelling studies show that PA and MAP consistently give the most reliable results in realistic situations (Zwick & Velicer 1986). PA is based on comparison of eigenvalues of the observed correlation matrix with average eigenvalues of random uncorrelated correlation matrices with the same number of variables and observations. MAP involves calculation of the average squared partial correlation after removing successive factors until a minimum is reached (theoretically, it must decrease as shared variance is extracted but increases again once only error variance is left). However, these procedures are still unavailable in general-purpose statistical packages. PA, MAP and VSS are nonetheless implemented in R (libraries paran, psych and nFactor, R Development Core Team 2008), Factor (Lorenzo-Seva & Ferrando 2005). SAS and SPSS syntax scripts allowing to calculate MAP and PA have been published (O'Connor 2000), there is also a stand-alone utility for PA (Budaev 1997).

Because FA is a special case of a more general structural equation modelling (SEM), certain SEM procedures and indices (e.g., CFI, RMSEA, chi-squared test) for model selection can also be used to assess the number of factors (see Fabrigar et al. 1999). However, they are much more complex, require normal distribution, large sample size and not easily available. Due to its reliable results, relative simplicity and availability, PA and MAP are highly recommended for animal behaviour researchers.

Factor Rotation

Factors can be arbitrarily rotated in the multidimensional space, although certain positions corresponding to a 'simple structure' allow their easy interpretation. Unrotated factor solution is typically biased towards the first general factor that confounds many variables and does not replicate well, rotation is necessary in most cases (Gorsuch 1983, 1997). Numerous rotation methods have been developed. Some (e.g., Varimax) maintain factors that are orthogonal whereas other (e.g., Promax, Oblimin) allow final factors to be correlated (Nunnally 1978; Gorsuch 1983; Tabachnick & Fidell 1996). Oblique rotations (e.g., Oblimin and Promax) are usually recommended because they allow orthogonality as a special case. However, this recommendation applies to classical FA with large sample size. The best rotation strategy for a small sample size is unknown. Because orthogonal rotations are mathematically simpler and involve estimation of fewer parameters (Nunnally 1978) they would probably provide more stable and replicable results in case of small sample size. Sample estimates of inter-factor correlations in such cases would be of little value due to very large standard errors. Thus, if factors are interpreted, rotation is usually required. Oblique rotations (e.g., Oblimin and Promax) are optimal in most cases, although orthogonal (Varimax) may be a better choice when the sample size is very small.

Which Loadings are Interpretable?

There seems to be no universal agreement as to what minimum factor loadings (absolute value) should be considered when interpreting factors. Most textbooks (e.g., Nunnally 1978; Gorsuch 1983; Tabachnick & Fidell 1996) recommend 0.3 or 0.4 as a minimum cutoff. However, this classical recommendation is based on the assumption of large sample size and no rules have been developed for small samples because of complex relations between SEs of factor loadings, sample size and other factors (Cudeck & O'Dell 1994). Although it is possible to compute confidence intervals of factor loadings and test statistical significance of loadings (Cudeck & O'Dell 1994), such a statistical approach has not been much used and is not available in general-purpose software. Thus, when sample size is large (>100) it is possible to follow the classical heuristics (loadings > 0.4), when it is small, minimum interpretable loadings should be higher (higher than 0.5 or even 0.7).

How Factor Scores are Calculated?

In many animal behaviour research factors themselves are rarely considered in isolation. The researcher is usually more interested in correlations between the factors and other measures or in differences in the levels of the factors across various experimental groups. Therefore, factor scores for each individual are frequently calculated. While calculation of such scores in PCA is straightforward, it is a difficult problem in FA (there may be an infinite number of scores consistent with the same pattern of factor loadings) and several solutions have been proposed, the regression method being most popular and implemented in virtually all statistical software (see Nunnally 1978; Gorsuch 1983; Tabachnick & Fidell 1996; Grice 2001). Approaches have also been proposed to assess the validity of factor score estimates (Grice 2001). Griece recommends calculation of the maximum proportion of determinacy ρ^2 (squared multiple correlation between each factor and the original variables, which must significantly exceed 0.5), and $2\rho^2 - 1$ (must be a high positive value). Some general-purpose software (e.g., SAS and SPSS) calculate ρ for the regression based factor scores. Still, even though the factor score indeterminancy problem has not been solved, most common factor score estimation methods (including regression) well reproduce the same covariance matrix, so 'the impact of differences between score estimates on research results may not have been very large' (Beauducel 2007, p. 441). Thus, commonly used methods for the assessment of factor scores, especially regression, can be legitimately used, but indices of determinacy (ρ, ρ^2, $2\rho^2 - 1$) should be computed to assess their adequacy.

How Repeated Measures Data Should be Treated?

Animal behaviour research often involve repeated measurements of the same behaviours or indices. However, common PCA and FA methods work only with a 'flat' two-mode data matrix (variables by individuals). Using multiple measures from the same individuals as independent while computing the correlation matrix is pseudoreplication and is incorrect. For example, such pseudoreplication is involved when individuals are tested repeatedly and then, when calculating correlations, repeated tests enter the raw data matrix as independent 'rows'. Fortunately, several methods of three-mode PCA and FA and longitudinal latent variable models were developed (see Gorsuch 1983; Law et al. 1984; Smilde et al. 2004). The PARAFAC model is characterised by simplicity and software availability. It is implemented in R (package PTAk, R Development Core Team 2008) and as a MATLAB module (Bro 2009). There is at least one example of its application in animal behaviour research (Ossenkopp et al. 1994). Thus, PARAFAC model can be recommended for repeated measures.

Reassessing the 'Fish Boldness Study'

It may be informative to consider the above fictitious 'fish boldness example' to show how these guidelines may help in assessment of the analysis validity. First, analysis of the sampling adequacy indices indicates that the correlation matrices are inappropriate (population 1: Bartlett's sphericity test $\chi^2 = 8.47$, df = 10, p = 0.58; KMO = 0.35; population 2: Bartlett's sphericity test $\chi^2 = 5.75$, df = 10, p = 0.84; KMO = 0.37). Second, as already noted, FA of random matrices cannot be calculated (most software packages will result in computation errors). Third, assessment of the number of factors using parallel analysis reveals that optimal number of factors in both cases is just zero. Without any information about how the data matrices have been produced, it becomes obvious that they are not appropriate for FA or PCA and all further analysis and interpretations are meaningless.

Applicability of Structural Equations: The Sample Size Issue

Factor analysis is a special case of a more general framework: SEM. SEM models directional and non-directional linear relationships between multiple variables: both manifest and latent (represented by manifest variables). Confirmatory factor analysis (CFA) is a special case of SEM in which a hypothesised factor structure is explicitly tested. Although SEM have long been used in the social sciences, they are very rare in animal behaviour research. Nonetheless, recent examples of SEM application in this field include CFA of mice temperament (Wall & Messier 2000; Ibanez et al. 2007) and testing alternative models of behavioural syndromes (Dochtermann & Jenkins 2007).

Structural equation modelling involves modelling, parameter estimation and model selection in a hypothesis-testing framework. Therefore, more stringent sample size requirements are often recommended for SEM than for exploratory FA. The classical recommendation is to have a minimum of 100 (better, 200) observations (Kline 2005). Modelling studies (Bentler & Yuan 1999; Boomsma & Hoogland 2001) indicate that 200 is a reasonable minimum sample size. This led to the recommendation that 'SEM analyses based upon samples of less than 200 should simply be rejected outright for publication' (Barrett 2007, p. 820). In various modelling approaches, sample sizes < 200 too often lead to non-convergence and improper solutions (Boomsma & Hoogland 2001). This would significantly limit the general applicability of SEM in animal behaviour studies.

How PCA and FA are used in Animal Behaviour Studies?

I examined how FA and PCA are used and reported in animal behavioural research. One whole volume of four journals from the year 2008 (*Animal Behaviour* vol. 78; *Applied Animal Behaviour Sciences* vol. 111; *Behavioral Ecology* vol. 19; *Ethology* vol. 114) were searched for the keywords 'principal component', 'factor analysis', 'PCA', yielding 51 papers. Almost all studies used PCA (98%) rather than FA (only one). More than half of the studies (55%) provided some interpretation of the extracted factors and labelled them using conceptual terms, making FA more suitable. Notably, very few if any of these studies provided complete and sufficient information allowing to assess the analysis. Only 35% of studies explicitly stated the sample size used to calculate the correlation matrix (either clearly in the text or in footnote to the factor loadings table).

In most cases, it was difficult to determine the sample size, especially when the analysis involved several different groups and complex design. I therefore tried to infer sample size from the text and other reported statistical analyses (e.g., from the degrees of freedom of other statistical tests), in case of uncertainty using the larger figures (an optimistic approach assuming the researchers followed the well-known recommendation of most textbooks for large sample size). The average sample size was 64 (the minimum, 4 and next 12, are clearly below the normal limits).

Some studies (approx. 10%) may have involved pseudoreplication while calculating the correlation matrix, although the description of the analysis in such cases was typically too vague. In some cases, pseudoreplication was clearer: 'We conducted a principal component (PC) analysis on these 5 parameters.... We then applied a repeated-measures analysis of variance (ANOVA) using the PC scores with 2 levels of the within-subject factor....' (Schmidt et al. 2008, p. 638). Given pseudoreplication has long been anathematised, it is not surprising that the authors would try to not reveal questionable analysis too clearly.

The average ratio of the number of cases to the number of variables was 8.9, with minimum 0.35 (i.e., 34 variables and only 12 cases, which is an anathema in virtually all FA textbooks). The ratio of the number of variables to the number of factors was on average 4.06 (minimum 2). Thus, although the average values are in the acceptable range, the validity of some published papers may be questioned.

Measures of sampling adequacy were calculated only in two studies (KMO > 0.5 in both cases). Only 25% of studies clearly described the criteria used to determine the number of factors. Usually the eigenvalue > 1 rule was used (in one case, parallel analysis) indicating that the default strategy provided by the software was probably silently followed. Given the root-one criterion tends to overfactor (see above), some spurious factors may be extracted and 'meaningfully' interpreted. 31% of studies reported the use of factor rotation (Varimax in 12 cases, one 'orthogonal', one Promax, and two unrotated), but no information (whether rotation was used or not and why) was given in the majority of studies.

Finally, assessment of the three animal behaviour papers involving CFA and SEM (Wall & Messier 2000; Dochtermann & Jenkins 2007; Ibanez et al. 2007) provided a mixed result. The quality of the two CFAs of anxiety (Wall & Messier 2000) and temperament (Ibanez et al. 2007) in mice appear more or less satisfactory (sample size, respectively, 200 and 70, several manifest variables per factor). However, SEM of behavioural syndromes in the kangaroo rat by Dochtermann & Jenkins (2007) is clearly flawed. First, both sample size and the number of variables were unacceptably small (n = 19 with four manifest and one or two latent variables). Analysis of the published correlation matrix (Table 1 in the original paper) revealed that it was not appropriate for SEM. The maximum correlation coefficient was only 0.28 (not significant at $p < 0.05$). I calculated the Bartlett's sphericity test ($\chi^2 = 3.51$, df = 6, p = 0.74) and the Steiger's test ($\chi^2 = 3.07$, df = 6, p = 0.80), clearly indicating that the correlation matrix is random (KMO = 0.40, also unacceptable). The authors' conclusions that 'SEM is a more powerful approach to testing behavioural syndrome hypotheses than is the use of bivariate correlation coefficients' (p. 2347) and especially that they 'detected an underlying covariance pattern which would not be interpretable using probability values' (p. 2348) are wrong. SEM is not a magical tool somehow making non-significant data conclusive. If the original correlation or covariance matrix does not significantly differ from random (or is otherwise ill-conditioned), then any further multivariate analysis is unjustified.

Guidelines for Reporting PCA and FA in Animal Behaviour Research

Obviously, reporting differences between the experimental and control group as 'significant at $p < 0.05$'

without noting what statistical test is used, what is the sample size or df is not acceptable in published research. However, similar omissions in PCA are currently the norm. Therefore, a set of standards for reporting the results of PCA and FA in published research on animal behaviour is needed (see also Henson & Roberts 2006). FA and PCA involve several crucial steps outlined above, all decisions must be reported. The researcher should thence (1) report whether the correlation or covariance matrix was used (although the former is used almost exclusively); (2) clearly state sample size used to calculate the correlation matrix, preferably in a footnote to the table of factor loadings; (3) present indices of sampling adequacy; (4) clarify the assessment of the number of factors; (5) report communalities, especially with small sample size; (6) report whether factor rotation was used and what was the rotation method; (7) present determinacy ρ^2 and $2\rho^2 - 1$ if factor scores are computed; (8) include the original correlation matrix (preferably as an on-line supplementary material), making reanalysis possible.

Acknowledgements

I thank Daniel Osorio, Richard Andrew and referees for valuable comments on earlier drafts of the manuscript and John Reddon for helpful discussions. The author is supported by the EU Sixth Framework Programme grant 'Evolution and Development of Cognitive, Behavioural and Neural Lateralisation.'

Literature Cited

Armstrong, J. S. & Soelberg, P. 1968: On the interpretation of factor analysis. Psychol. Bull. **70**, 361—364.

Barrett, P. 2007: Structural equation modelling: Adjudging model fit. Person. Individ. Diff. **42**, 815—824.

Beauducel, A. 2007: In spite of indeterminacy many common factor score estimates yield an identical reproduced covariance matrix. Psychometrika **72**, 437—441.

Bentler, P. M. & Yuan, K. H. 1999: Structural equation modeling with small samples: Test statistics. Multivar. Behav. Res. **34**, 181—197.

Boomsma, A. & Hoogland, J. J. 2001: The robustness of LISREL modeling revisited. In: Structural Equation Modeling: Present and Future (Cudeck, R., Du Toit, S. & Sorbom, D., eds). SSI Scientific Software, Chicago, pp. 139—168.

Bro, R. 2009: The N-way toolbox. Available at http://www.mathworks.com/matlabcentral/fileexchange/1088.

Budaev, S. V. 1997: Paranal: Parallel analysis and factor adequacy of correlation matrix. Available at: http://sourceforge.net/projects/paranal.

Budaev, S. V. & Zworykin, D. D. 2002: Individuality in fish behavior: Ecology and comparative psychology. J. Ichthyol. **42**(Suppl. 2), S189—S195.

Cudeck, R. & O'Dell, L. L. 1994: Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. Psychol. Bull. **115**, 475—487.

Dochtermann, N. A. & Jenkins, S. H. 2007: Behavioural syndromes in Merriam's kangaroo rats (*Dipodomys merriami*): A test of competing hypotheses. Proc. R. Soc. London B **274**, 2343—2349.

Dziuban, C. D. & Shirkey, E. S. 1974: When is a correlation matrix appropriate for factor analysis? Some decision rules. Psychol. Bull. **81**, 358—361.

Dziuban, C. D., Shirkey, E. S. & Peeples, T. O. 1979: An investigation of some distributional characteristics of the measure of sampling adequacy. Educat. Psychol. Measur. **39**, 543—549.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. 1999: Evaluating the use of exploratory factor analysis in psychological research. Psychol. Meth. **4**, 272—299.

Fouladi, R. T. & Steiger, J. H. 1993: Tests of multivariate independence: A critical analysis of ''A Monte Carlo study of testing the significance of correlation matrices'' by Silver and Dunlap. Educat. Psychol. Measur. **53**, 927—932.

Gorsuch, R. L. 1983: Factor Analysis, 2nd edn. Erlbaum, Hillsdale, New Jersey.

Gorsuch, R. L. 1990: Common factor analysis versus component analysis—some well and little known facts. Multivar. Behav. Res. **25**, 33—39.

Gorsuch, R. L. 1997: Exploratory factor analysis: Its role in item analysis. J. Pers. Assess. **68**, 532—560.

Grice, J. W. 2001: Computing and evaluating factor scores. Psychol. Meth. **6**, 430—450.

Haig, B. D. 2006: Exploratory factor analysis, theory generation, and scientific method. Multivar. Behav. Res. **40**, 303—329.

Henson, R. K. & Roberts, J. K. 2006: Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. Educ. Psychol. Meas. **66**, 393—416.

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

Ibanez, M., Avila, C., Ruiperez, M. A., Moro, M. & Ortet, G. 2007: Temperamental traits in mice (I): factor structure. Person. Individ. Diff. **43**, 255—265.

Kline, R. B. 2005: Principles and Practice of Structural Equation Modeling, 2nd edn. Guilford, New York.

Law, H., Snyder, C., Hattie, J. & MacDonald, R. (Eds). 1984: Research Methods for Multimode Data Analysis. Praeger, New York.

Lorenzo-Seva, U. & Ferrando, P.J. 2005: Factor: Exploratory factor analysis program for Windows. Available at: http://psico.fcep.urv.es/utilitats/factor/.

MacCallum, R. C., Widaman, K. F., Zhang, S. & Hong, S. 1999: Sample size in factor analysis. Psychol. Meth. **4**, 84—99.

MacCallum, R. C., Widaman, K. F., Preacher, K. J. & Hong, S. 2001: Sample size in factor analysis: The role of model error. Multivar. Behav. Res. **36**, 611—637.

Nunnally, J. C. 1978: Psychometric Theory, 2nd edn. McGraw-Hill, New York.

O'Connor, B. 2000: SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. Behav. Res. Methods Instrum. Comput. **32**, 396—402.

Ossenkopp, K.-P., Sorenson, L. & Mazmanian, D. S. 1994: Factor analysis of open field behavior in the rat (*Rattus norvegicus*): Application of the three-way PARAFAC model to a longitudinal data set. Behav. Proc. **31**, 129—144.

Preacher, K. J. & MacCallum, R. C. 2002: Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. Behav. Genet. **32**, 153—161.

R Development Core Team 2008: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.r-project.org.

Reale, D., Reader, S. M., Sol, D., McDougall, P. T. & Dingemanse, N. J. 2007: Integrating animal temperament within ecology and evolution, Biol. Rev. **82**, 291—318.

Revelle, W. & Rocklin, T. 1979: Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. Multivar. Behav. Res. **14**, 403—414.

Schmidt, R., Kunc, H. P., Amrhein, V. & Naguiba, M. 2008: Aggressive responses to broadband trills are related to subsequent pairing success in nightingales. Behav. Ecol. **19**, 635—641.

Schneeweiss, H. 1997: Factors and principal components in the near spherical case. Multivar. Behav. Res. **32**, 375—401.

Short, R. & Horn, J. 1984: Some notes on factor analysis of behavioral data. Behaviour **90**, 203—214.

Sih, A., Bell, A. M., Johnson, J. C. & Ziemba, R. E. 2004: Behavioral syndromes: An integrative overview. Q. Rev. Biol. **79**, 241—277.

Smilde, A., Bro, R. & Geladi, P. 2004: Multi-way Analysis: Applications in the Chemical Sciences. Wiley, New York.

Tabachnick, B. G. & Fidell, L. S. 1996: Using Mulitvariate Statistics, 3rd edn. Harper & Row, New York.

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

Velicer, W. F. & Fava, J. L. 1998: Effects of variable and subject sampling on factor pattern recovery. Psychol. Meth. **3**, 231—251.

Velicer, W. F. & Jackson, D. N. 1990: Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. Multivar. Behav. Res. **25**, 1—28.

Wall, P. M. & Messier, C. 2000: Ethological confirmatory factor analysis of anxiety-like behaviour in the murine elevated plus-maze. Behav. Brain Res. **114**, 199—212.

de Winter, J. C. F., Dodou, D. & Wieringa, P. A. 2009: Exploratory factor analysis with small sample sizes. Multivar. Behav. Res. **44**, 147—181.

Zwick, W. R. & Velicer, W. F. 1986: Comparison of five rules for determining the number of components to retain. Psychol. Bull. **99**, 432—442.