



Testing the number of required dimensions in exploratory factor analysis

André Achim^a, ^aUniversité du Québec à Montréal

Abstract ■ While maximum likelihood exploratory factor analysis (EFA) provides a statistical test that k dimensions are sufficient to account for the observed correlations among a set of variables, determining the required number of factors in least-squares based EFA has essentially relied on heuristic procedures. Two methods, Revised Parallel Analysis (R-PA) and Comparison Data (CD), were recently proposed that generate surrogate data based on an increasing number of principal axis factors in order to compare their sequence of eigenvalues with that from the data. The latter should be unremarkable among the former if enough dimensions are included. While CD looks for a balance between efficiency and parsimony, R-PA strictly test that k dimensions are sufficient by ranking the next eigenvalue, i.e. at rank $k + 1$, of the actual data among those from the surrogate data. Importing two features of CD into R-PA defines four variants that are here collectively termed Next Eigenvalue Sufficiency Tests (NESTs). Simulations implementing 144 sets of parameters, including correlated factors and presence of a doublet factor, show that all four NESTs largely outperform CD, the standard Parallel Analysis, the Mean Average Partial method and even the maximum likelihood approach, in identifying the correct number of common factors. The recommended, most successful NEST variant is also the only one that never overestimates the correct number of dimensions beyond its nominal α level. This variant is made available as R and MATLAB code as well as a complement incorporated in a Microsoft Excel file.

Keywords ■ Exploratory Factor Analysis, Monte Carlo Comparisons, Number of Factors.

 achim.andre@uqam.ca AA: 0000-0003-2159-6010 [10.20982/tqmp.13.1.p064](https://doi.org/10.20982/tqmp.13.1.p064)

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

Introduction

The present article addresses the question of how many factors, in exploratory factor analysis (EFA), are necessary to explain the data. Although this question may appear outdated from a historical perspective, it remains that true data modeling approaches, rather than heuristics, have only been proposed relatively recently for least squares EFA approaches (Green, Levy, Thompson, Lu, & Lo, 2012; Ruscio & Roche, 2012). Here, several model-based methods meant to identify the number of factors will be applied to simulated data with known numbers of factors across a range of study parameters. One method will clearly emerge as the one to be recommended. For ease of use, this best performing procedure is provided, besides its code in

R and MATLAB, as a programmed complement in an Excel file. All there is to do with the latter is to provide the data as contiguous columns in a separate Excel file and to identify the relevant columns as input to the procedure.

Factor Space Dimensionality

In teaching exploratory data analysis, what should we recommend to assess the number of reliable dimensions in the dataset? The answer must depend on the purpose. For instance, one may seek a reasonable description that accounts for a substantial part of the variance involved, where simplicity and avoidance of minor factors are balanced against completeness (Fabrigar, Wegener, MacCallum, & Strahan, 1999). The present investigation, however, focuses on the relative merit of methods meant to iden-



tify the full factor space dimensionality, i.e. how many dimensions are actually expressed in the data. In functional terms, this question reduces to how many dimensions appear required to completely account for the pattern of correlations among the variables. For this purpose, methods that explicitly model the factor space can be expected to stand the best chances of solving the problem correctly, but the standard parallel analysis (PA: Horn, 1965) is also considered for comparison purposes.

While there are many contexts in which it may be recommended to exclude minor factors, the question addressed here requires not ignoring doublet factors (i.e., factors that affect only two observed variables), or factors with eigenvalues less than unity that nevertheless affect the data structure. Such dimensions could later be consciously discarded to simplify the picture of the phenomenon studied, but this should preferably be done after becoming aware of their existence.

Assumptions

The basic assumption of the present work is the additivity of factor contributions to the observed variables. Situations where, for instance, a variable reflects a weighted sum of factor scores while another would reflect a product of factor scores are excluded a priori. Another, although less crucial assumption is the normal distribution of the several sources of variance. Only normal distributions will be involved, but due to the central limit theorem, the results should be expected to generalize well to skewed distributions. It is, however, generally a better practice to transform variables whose distribution departs markedly from normality before submitting them to analysis. This may actually favor compliance to the additivity assumption, since additive combinations of independent effects tend to produce normal distributions.

Simulations

Candidate methods.

The problem addressed here is actually already solved, in principle, with maximum likelihood factor analysis (MLFA: Lawley, 1940) and, in practice, based on Jöreskog's (1967) implementation. In this approach, factor models with an increasing number of dimensions are developed as long as the hypothesis that the residuals are independent is rejected. Model convergence failures, in the form of Heywood cases (i.e. models that would reproduce more than 100% of the variance for a variable), are treated by removing the variable involved but counting it as one factor, assuming that such variable perfectly represents one underlying factor (Jöreskog, 1967). If a Heywood case indeed reflected a doublet factor, the other variable reflecting the

same doublet factor would then become a singlet factor and would be disregarded as unique variance in the sequel of the procedure.

Other modeling approaches have been attempted. The mean average partial (MAP: Velicer, 1976) can be considered a model based approach, since principal components (data model) are successively removed from the correlation matrix, leaving partial correlations. MAP assesses dimensionality as the number of removed components for which the average partial correlation, raised to exponent 2 or 4, is minimal.

PA only implements a null model (Turner, 1998) and, in terms of data modeling, is theoretically adequate only to decide on the need for at least one factor, i.e. to negate the hypothesis of no common factor. Beyond this, PA is strictly heuristic. PA and MAP fared reasonably well in simulations (Peres-Neto, Jackson, & Somers, 2005; Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986) and are the main methods recommended, even quite recently, by journal editors (e.g., Ziegler & Hagemann, 2015) despite promising newer approaches (Green et al., 2012; Ruscio & Roche, 2012).

In recent developments, Green et al. (2012) proposed two forms of revised parallel analysis (R-PA), where either principal components or principal axis factors are used as data models meant to account for the observed pattern of correlations using an increasing number of factors. The logic of the test is that if k dimensions are not sufficient to account for the data, then the data eigenvalue at rank $k + 1$ should be larger than those for most of parallel datasets generated with the k -factor model. For this test, the authors used eigenvalues from the same technique that provided the factor models, ranking the critical eigenvalue from a principal component analysis (PCA) model among PCA eigenvalues, and those from a principal axis factor analysis (PAFA) model among PAFA eigenvalues. They further rejected PAFA models with negative eigenvalues, which is compatible with the present purpose, but their a priori rejection of PCA models with eigenvalues less than unity is not. Their principal axis factoring model simply used the squared multiple correlation of each variable as communalities, without iterative improvement. The same authors (Green, Thompson, Levy, & Lo, 2015) later compared their own preferred version of R-PA, that based on PAFA, with the MLFA approach, reporting advantage for their own method.

Contemporary to the development of R-PA, Ruscio and Roche (2012) elaborated an approach named comparison data (CD) that also uses increasing number of PAFA components to model the data, but uses the PCA eigenvalues to identify the sufficient number of dimensions. CD applies a criterion widely different from R-PA, based on the



stabilization of the full pattern of surrogate data eigenvalues around the actual data eigenvalues, once enough factors are included in the model. To our knowledge, no direct comparison of CD and R-PA has yet been published, although Green et al. (2015), after becoming aware of the CD method, expressed that “it would be interesting in the future to compare the revised PA and the Ruscio–Roche methods to understand how and why they differ”.

Although PA and MAP are included, the present work mainly aimed at comparing approaches that explicitly model the data space, namely the MLFA approach of Jöreskog (1967) and the PAFA modeling approaches of Green et al. (2012) and of Ruscio and Roche (2012). Three related approaches based of PAFA models are also considered which incorporate elements of the CD and R-PA methods, namely CD’s iteration of commonality estimates and its use of PCA eigenvalues to assess models derived from PAFA, and R-PA’s ranking of eigenvalues at rank $k + 1$ to decide on the sufficiency of the k dimension model.

The use of PCA for modeling the data, as in Green et al. (2012), is excluded a priori, because mathematical analysis indicates that data produced from the first k principal components plus a random variable completing the variances to unity must systematically have their first k PCA eigenvalues larger than those of the original data, simply because the unique variances of the p original variables span the complete p dimensional space and therefore necessarily contributes to all principal components. PCA based models would thus assign too much signal variance to the factors, underestimating the unique variances required for expected unit variance of the surrogate variables. This can be verified by ranking the original PCA eigenvalues among those of corresponding rank from the PCA based surrogate data. PCA based models are thus rejected here on the principle that an acceptable model for the first k data dimensions should generate surrogate data that correctly reproduce the first k eigenvalues of the original data.

Test Conditions.

The synthetic problems on which the methods are to be compared must be inherently difficult, since for easy situations, nearly all methods should work well. For instance, if all eigenvalues to be retained are above 2.0 and the communalities of all the variables are similar, then Kaiser’s eigenvalue > 1 rule or Cattell’s Scree test will unambiguously provide the correct answer.

Difficult conditions include situations in which at least one factor has low weights on all the variables in which it is expressed. For correlated factors, even if they have reasonably large loadings on their variables, the resulting factor space can have a small eigenvalue (and therefore an eigenvector with only small loadings). Ignoring such dimension

does not amount to ignoring an independent minor factor. For two correlated factors, for instance, a common meta-factor accounts for their shared variance while a second factor, possibly with an eigenvalue less than unity, essentially accounts for what distinguishes the two actual factors. Ignoring minor factors would impede correct understanding of the underlying phenomenon. Another type of difficult situation is the presence of a doublet factor, especially when its weights are rather modest. Such situations are included here.

Methods

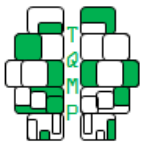
General Consideration.

This work was carried using MATLAB release R2015a (The MathWorks, Inc., Natick, Massachusetts, United States), using the default pseudo random number generation algorithm, the Mersenne Twister (Matsumoto & Nishimura, 1998).

Factor Structures.

The combinations of parameters used to generate test data were inspired from Velicer et al. (2000), although with some variation, including the introduction of correlated factors and the presence of doublet factors. The present investigation involved data structures with complete crossing of number of factors (3 or 6), of variables per factor (4 or 8, each associated with only one factor), of factor loadings (.6 or .8), of factor correlation (0, .3, or .6), and of sample sizes (75, 150, or 300). Besides these 72 combinations, a similar set of simulations was performed in which one of the factors was expressed in only two variable (creating a doublet factor), for a total of 144 simulation conditions. For each of these, 100 datasets were generated to which each candidate method was applied to assess the number of dimensions.

For comparison purposes, the study by Velicer et al. (2000) used the same sample sizes, but had only uncorrelated factors and no doublet condition. It involved simulations with 3, 6, or 9 factors, each expressed, as done here, in 4 or 8 different variables. Their maximum number of variables was thus 72 compared to 48 in the present study without 9 factor conditions. In their study, the factor loadings were all fixed to .4, to .6, or to .8 in different conditions. In the present study, .4 loadings were not used but small expected eigenvalues are rather produced by correlating the factors or by the presence of a doublet factor. Another important difference is that while the preceding authors generated only five datasets for each condition, here 100 datasets per condition are used to also allow within condition statistical comparisons of the dimensionality assessment methods.



When all of k dimensions are to be identified, the difficulty of a given problem should depend largely on its k th eigenvalue. Based on the population eigenvalues, 6 of the 48 data structures have their critical (third or sixth) eigenvalue less than 1.0; all are for a correlation of .6 between the factors and including a doublet factor. Of the remaining data structures, 18 have their most critical eigenvalue between 1.0 and 1.5; 8 others have it between 1.5 and 2.0, and the remaining 16 data structures have it above 2.0. The comparisons of candidate methods will be carried separately for these four ranges of critical population eigenvalues.

Algorithms.

Dimensionality assessment methods.

The eight factor space dimensionality assessment methods compared are PA, MAP, MLFA, CD, the original R-PA and three hybrid versions incorporating elements of CD and of R-PA. The latter and the three new versions all rank eigenvalue $k + 1$ of the actual data among those of surrogate data, rejecting the k factor model if the former ranks high, indicating that shared variance is still present in the data beyond the first k dimensions. They are thus collectively called Next Eigenvalue Sufficiency Test (NEST). They differ among themselves in two ways, model generation and type of eigenvalue ranked, as will be described shortly.

MAP was implemented as its version that performed best in the study by Velicer et al. (2000), namely with the off-diagonal elements of the partial correlation matrix raised to the fourth power before averaging. MLFA used MATLAB's `factoran` function embedded in a custom function that, following the suggestion of Jöreskog (1967), removes any variable causing a Heywood case, counting each such variable as an extra factor added to the lowest number of factors for which the residuals of the remaining data qualified as independent by an α level of .05.

The remaining six methods all compare eigenvalues from either the full or a reduced correlation matrix of the original data to those of surrogate data. The latter are 500 independent synthetic datasets having the same number of variables and cases as the test dataset, each generated by what would be the unrotated factor structure of the test dataset according to PAFA, assuming a given number of factors. Except for PA, for which the model consists of only independent variables, two related ways are used to produce the model with which the surrogate datasets are generated. These are two versions of PAFA differing in their communality estimate method for the reduced correlation matrix. The original R-PA method simply estimates the communalities as the values of R^2 obtained from the multiple regressions predicting each variable from the re-

maining ones. The CD method rather adapts its communality estimates to the number k of factors to model; it iteratively updates the diagonal of the correlation matrix by the variances reproduced by the first k dimensions based on the current reduced matrix estimate.

Use of one or the other communality estimate for the data model will be denoted by appending 'NEST' with 'r' (for regression) or 'i' (for iterative). In assessing whether k factors are sufficient to account for the data, ranking the original data eigenvalue at rank $k + 1$ among those of the surrogate data can be based on either PCA or PAFA eigenvalues. This distinction in eigenvalue used to assess the sufficiency of k dimensions will be denoted by appending 'p' or 'f' to 'NESTr' or 'NESTi'. This defines the four related methods: NESTrf, NESTrp, NESTif and NESTip, where the former is the original R-PA method. When it comes to ranking PAFA eigenvalues, i.e. in the NEST methods with a final 'f', the only realistic choice is to reduce the correlation matrices with the regression approach, without further iteration.

Data generation.

In CD and all NEST versions, a k factor model consists of the first k components of the reduced matrix, weighted by the square root of their respective eigenvalues, with the addition of independent random numbers for the unique variance bringing the expected variance of each variable to unity. Both for building all models and to test them with PAFA eigenvalues, component extraction from a reduced correlation matrix is done with MATLAB's `svd` function embedded in a function that reassesses the signs of the eigenvalues; this is required since `svd` itself returns all eigenvalues as positive. This is achieved by post-multiplying the reduced matrix by the eigenvectors and comparing any row of the result with the corresponding row of the eigenvectors. The signs change for eigenvalues that should be negative. The eigenvalues and corresponding eigenvectors are then reordered to ensure decreasing eigenvalues. For k factors expressed on p variables in c cases, each surrogate dataset is produced by applying a matrix of dimensions $(c, k + p)$ of independent random numbers to the generating model.

In generating the test datasets for conditions in which the factors are specified as correlated, a matrix (produced by the MATLAB `chol` function) transforming k independent sources into k variables that are pairwise inter-correlated with the value r is applied to the k sets of factor loadings before complementation with the diagonal matrix implementing the unique variance that complements to unity the expected variance of each variable.

For the methods relying on surrogate data, namely PA, CD, and all NEST methods, 500 random datasets are used.



This number was selected because the CD criterion has been tuned for that number of comparison data. The same surrogate data are used for all methods based on the same model; more explicitly, CD, NESTif and NESTip use the same surrogate data based on the iterated model, while NESTrf and NESTrp use the same surrogate data based on the non-iterated PAFA model.

Implementation Details.

PA was implemented by generating 500 sets of (c, p) random numbers from a normal distribution and using the means of the resulting 500 sets of eigenvalues as thresholds at each rank. The mean, rather than the 95th percentile as suggested by Glorfeld (1995), is used because the PA reputation of tending to overestimate the number of dimensions, on which this suggestion is based, essentially comes from simulations including singlet factors (i.e. variables depending on none of the common factors). The 95th percentile, however, is the relevant cut-off value to reject the hypothesis of no common factor. This criterion also applies in all NEST methods in their assessing the rank of the data eigenvalue $k + 1$ among those of the surrogate data. A model with k dimensions is thus rejected when no more than 25 out of 500 surrogate datasets produce their relevant eigenvalue equal to or larger than that of the actual data.

The CD method was implemented essentially as Ruscio and Roche (2012) described it, except that all data generation are here based on independent normally distributed random numbers, instead of random numbers reflecting the observed distributions. Although this is not explicitly described in Ruscio and Roche (2012), the original implementation of CD likely used the PAFA algorithm described in Ruscio and Kaczetow (2008) to develop the surrogate data generating model. It iterates the communalities as the variances explained by the first k eigenvalues of the initially full, then reduced correlation matrix until convergence. The CD decision rule is based on PCA eigenvalues (i.e. of full correlations matrices), more specifically on the root mean square (rms) difference between the original set of eigenvalues and each set of eigenvalues of the surrogate data. A set of 500 rms difference values is thus available for each hypothesis of k factors to retain. To decide on a significant reduction of the rms differences due to the inclusion of one extra factor in the model, a rank test is applied to pairs of rms sets that differ by one factor. When the Mann-Whitney rank-sum test between sets for k and $k + 1$ factors becomes not significant by a liberal criterion, the number of factors is set to k , the lower of the two model dimensions involved in the comparison. The authors determined that operating the rank test at $\alpha = .3$ provides a good balance between tendencies to over- and to under-

estimate the number of data dimensions.

Comparison of methods.

The methods are mainly to be compared in terms of their success rates. Since the dimensionality assessments produced by the various methods are paired within test dataset, the correct chi-squared test excludes equality cases. In the present context of pairwise comparison of methods, this means comparing the counts of datasets correctly solved by only one or by only the other of the two methods.

Besides pooling the trials across conditions that share the same range of critical eigenvalue, the methods may be compared pairwise on the number of conditions in which one method significantly outperforms the other. For this purpose, a liberal α level of .10 is applied within each of the 144 test conditions, but the usual .05 criterion remains used for deciding whether two methods differ on the number of conditions in which each outperforms the other by the liberal .10 criterion.

For the methods that appear possibly worth retaining, it is relevant to verify their actual type I error when the correct number of dimensions constitutes the null hypothesis. If a method is prone to overestimation, i.e. it tends to reject the correct model significantly above the nominal 5% α level, then some of the success it achieves in conditions where much power is required to detect the true dimension with a small eigenvalue might be due to this bias rather than to its true sensitivity.

Results

Success rates.

The eight methods are mainly to be compared in terms of percentage correct. These results are presented in Table ?? with the row-wise best, or equal best, performing methods signaled in bold. Although all methods except CD correctly identified the number of dimensions at least 85% of the time when the critical eigenvalue was above 2.0, only the four NEST methods performed relatively well in the more challenging test conditions. NESTrf, the original R-PA method, outperformed the other NEST variants only when the critical population eigenvalue was below 1.0. These are all cases involving a doublet factor and a correlation of .6 between all pairs of factor. Otherwise, NESTif and NESTip outperformed all other candidate methods and did not differ from one another for the next two ranges of critical population eigenvalues, but NESTip surpassed in accuracy all other methods when the critical population eigenvalue was at least 2.0. Although PA compared to the NEST variants for the least difficult problems, only the NEST variants appear worth retaining and further considering for finer



Table 1 ■ Percentage of correct identifications of the number of factors for the eight methods studied by range of the smallest eigenvalue to be retained.

Critical eigenvalue range	Dimensionality assessment method							
	PA	MAP	MLFA	CD	NESTrf	NESTrp	NESTif	NESTip
≤ 1.0	0.4 _a	0.8 _a	14.6	33.8	57.2	55.0 _b	55.8 _b	53.1
1.0-1.5	36.2	6.7	26.9	33.8	73.8 _a	73.9 _a	77.1_b	76.6_b
1.5-2.0	74.0	12.0	41.5 _a	38.5 _a	87.4 _b	88.0 _b	89.4_c	90.0_c
≥ 2.0	95.3 _a	85.0	87.7	45.9	90.6	94.7 _a	92.1	96.3
Unrestricted	57.7	32.9	48.1	38.6	79.6	80.8	81.5_a	82.5_a

Note. On each critical eigenvalue line, percentages sharing a common subscript do not differ at the .05 α level. The best methods, within critical eigenvalue ranges, are marked in bold. As a reminder, NESTrf is the same as R-PA.

analysis.

The NEST methods may also be compared pairwise on the number of test conditions in which one outperforms the other at a liberal level (namely .10) over the 100 test datasets of the condition. In this approach, NESTip always outperformed the other NEST methods. The difference is not significant in its comparison with NESTif (13 against 6 test conditions, $\chi^2(1) = 2.58, p = .108$), but is significant when NESTip is compared to NESTrf (31 against 7, $\chi^2(1) = 15.16, p = .0001$) or to NESTrp (20 against 6, $\chi^2(1) = 7.54, p = .006$). NESTif more often outperformed NESTrf than was outperformed by it (26 against 7, $\chi^2(1) = 10.95, p = .0009$); it also outperformed NESTif in more conditions than the opposite (27 against 13, $\chi^2(1) = 4.90, p = .027$). Finally, NESTrp outperformed NESTrf in 11 conditions with the opposite relationship observed in only 3 test conditions ($\chi^2(1) = 4.57, p = .033$).

Other considerations relevant for selecting among NEST variants.

When NEST methods are applied with the correct number of dimensions as null hypothesis, they would be expected to reject that hypothesis at the 5% nominal rate, provided their surrogate data generating model is faithful. Table ?? lists the false alarm rate of the four NEST variants for the four ranges of critical population eigenvalues. The error rates that differ significantly, at the .05 α level, from the nominal 5% expected error rate are marked in bold and those that differ from the nominal rate at the .001 level are expressed in red. It can be seen that the two NEST variants that rank the eigenvalues of reduced correlation matrices, namely NESTrf and NESTif, express some systematic positive bias to reject the correct null hypothesis which results in overestimating the factor space dimension. The NEST variants that rank the eigenvalues of full correlation matrices did not exceed the nominal 5% rate in the different ranges of critical population eigenvalue; NESTip is actually conservative at all levels of the critical population eigenvalue.

Discussion

Of all the methods tested, NESTif and NESTip were observed to be the most reliable, but NESTif sometimes displays a positive bias, which may account for some of its success. Interestingly, the positive bias of NESTif was observed only for the conditions in which the critical population eigenvalue was above 2.0, conditions for which NESTip significantly outperformed all other methods, including NESTif.

Since bias can only be estimated for the correct number of dimensions constituting the null hypothesis, reasoning for situations where the current null hypothesis is one less than the correct number may be hazardous, especially for a method that also displayed a negative bias in some range of critical population eigenvalue. A method that is known to sometimes significantly exceed its nominal type I error rate may nevertheless be suspected of deriving some of its success from this characteristic.

Although the original R-PA, here called NESTrf, was the best performing method when the critical population eigenvalue was below unity, it may be suspected that the differences among the NEST methods for this range of conditions is entirely due to type I errors. Indeed, if the false alarm rates in these conditions are subtracted from the success rates, then all four NEST methods give very close net success values, between 52.4 and 52.6%.

Preferring NESTip over NESTif appears justified based on the observation that the difference in success rate between the two methods either did not differ or favored NESTip, despite NESTip being systematically more conservative than NESTif. This preferred version of the Next Eigenvalue Sufficiency Test differs from the original R-PA by generating its surrogate data from a model of the data factor space based on iterated communality estimates, rather than simply from multiple regression R^2 . Furthermore, to test the sufficiency of k dimensions, the eigenvalues $k + 1$ used for ranking the original data among the surrogates are based, in NESTip, on full, rather than reduced,



Table 2 ■ Percentages of false rejection of the correct null hypothesis when tested at the 5% level by the four NEST variants by range of the smallest eigenvalue to be retained.

Critical eigenvalue range	Dimensionality assessment method			
	NESTrf	NESTrp	NESTif	NESTip
≤ 1.0	4.63	3.60	2.33	0.52
1.0-1.5	5.95	4.04	3.99	2.25
1.5-2.0	6.97	5.04	4.75	2.88
≥ 2.0	8.97	4.84	6.71	3.28
Unrestricted	7.19	4.50	5.03	2.63

Note. Percentages that differ at the .05 α level from the nominal 5% error rate are in bold. Those that differ from 5% at the .001 level are further shown in red.

correlation matrices. These two substitutions mildly but significantly improved performance.

It seems that the reduced precision of deriving the model used to generate the surrogate datasets based on simple estimates of the communalities from multiple regression explains the excess of type I errors when the data structures provide for relatively large critical population eigenvalues. It also appears that ranking eigenvalues from full correlation matrices, rather than from reduced correlation matrices, introduces a negative bias at all levels of the critical eigenvalue. NESTrp would thus tend to balance the tendencies to over and underestimate the factor space dimensionality, while NESTip should, in principle, be handicapped by its negative bias. This handicap does not, however, prevent NESTip to generally perform best.

Clearly, the success rates observed here cannot be taken as those to be expected with actual data. These rates are averages over a range of clearly heterogeneous conditions, all given equal weights. There can be no pretension that these conditions are as uniformly distributed, in any area of investigation using EFA, as they are here. What the present work provides, however, is some rather solid indication of the relative advantage of NESTip, at least in the range of conditions assessed here.

While the other NEST variants exhibited significant excess of type I errors when the correct number of dimensions constituted the null hypothesis, NESTip rather proved systematically conservative, effectively overestimating the factor space dimensionality at a rate significantly less than the nominal 5% α level. Thus, contrary to the other NEST variants, it does not derive any part of its power from a slight propensity to reject the null hypothesis above the nominal α rate when the last dimension to identify would otherwise be missed.

Comparison with earlier studies.

PA and MAP. While MAP rivalled PA in some earlier studies, it was not impressive in the current set of conditions; it did particularly poorly in all conditions having the crit-

ical population eigenvalue below 2.0 and its success rate was 10% below that of PA in the remaining, less demanding conditions. Although PA is a heuristic rather than a data modeling approach, it was a surprise to observe that, in these easier conditions, it actually significantly surpassed some NEST variants, namely NESTrf and NESTif that rank eigenvalues from reduced correlation matrices, doing nearly as well as the best performing NESTip. PA, however, did not stand the comparison with any NEST method in more challenging test conditions, including the range of critical population eigenvalues between 1.5 and 2.0.

As to the question of the alleged tendency of PA to overestimate the number of dimensions, examination of its overestimations rates confirmed this to be incorrect in absence of singlet factors. Its observed overestimation rates are extremely low, below 0.1%, in all four ranges of critical population eigenvalues.

CD. Another surprise concerns the relatively poor performance of the CD method, it being correct only about 38% overall, ranging 34% to 46% across the four levels of critical population eigenvalues. Ruscio and Roche (2012) documented the good performance of CD, relative to PA and MAP, on simulations that randomly selected the number of subjects between 200 and 1000, the number of factors between 1 and 5 and the number of variables between 15 and 60. Factors affecting more than one variable were classified as correlated, although the generating factor scores were not made to correlate. In their study, PA was correct in 76.4% of the simulations, MAP in 59.6% and CD in 87.1%. The present modest performance of CD may not be attributed exclusively to the conditions in which the factor scores were correlated. Examination of its performance separately for $r = 0, .3, \text{ or } .6$, irrespective of the critical population eigenvalue, yields respective success rates of 39.2%, 40.5%, and 36.1%.

The discrepancy of the present results with those of Ruscio and Roche (2012) could suggest that CD might need specific conditions, like a large number of variables, to fare



well. The poorer performance of CD compared to any NEST version and even to PA for critical population eigenvalues above 1.5 may, however, be attributed to the stopping rule that characterises CD. By looking for a relative stabilisation of the rms difference across the pattern of eigenvalues over increasing number of factors in the model, the approach implicitly looks for a balance between efficiency and parsimony. Further balance is expressed in the liberal $\alpha = .30$ criterion selected for the rank test. The test is bilateral, but since the rms difference of eigenvalues can only decrease, this is functionally a .15 significance level. This α level might be suboptimal for conditions other than those on which it was determined. Further work would be needed to explain the poor success of CD in the present simulating conditions. The general NEST approach is, however, more congruent with the present objective of completeness, regardless of parsimony.

MLFA. Green et al. (2015) compared MLFA and PA to their R-PA version that corresponds to the present NESTrf, and observed that although there was no uniformly superior method, R-PA generally fared better than PA and the latter better than MLFA. Many of their simulations included singlet variables or less than two factors, but the two or three factor conditions yielded this same general pattern. The present study, across all conditions, replicates this order, with success rates of 79.6% for R-PA/NESTrf, 57.7% for PA and 48.1% for MLFA.

Their implementation of MLFA, however, applied Jöreskog's (1967) proposal of removing the offending variables but counting them as representing factors. To assess a posteriori the effect of this addition to the procedure, the set of 144 test conditions was repeated to compare the success rates of MLFA with and without Jöreskog's proposal. With it, the observed success rates were, as expected, within the confidence limits of those reported in Table ?? (13.6%, 26.7%, 48.1%, and 79.6% across the four levels of critical eigenvalues). It turned out however, that MLFA is generally more accurate when the procedure stops whenever a Heywood case is detected, accepting the current hypothesis as the number of dimensions (since the hypothesis of one less dimension was rejected as insufficient). The success rates in the four levels of critical eigenvalues were respectively 32.9%, 49.9%, 70.9%, and 87.3%. The overall success rate of 61% makes this version of MLFA more successful than PA, thus changing, with the current set of simulation conditions, the order observed by Green et al. (2015). The improvements observed for the four levels of critical eigenvalues are, however, not large enough to bring MLFA to levels of success comparable to those of the NEST variants, which all respectively exceeded 50%, 70%, 85%, and 90%.

The comparison of the two implementation forms of

MLFA is also worth examining with respect to the absence or presence of a doublet factor. Without doublet, the two forms of MLFA fared equivalently (78.2% and 78.3%) but the version incorporating Jöreskog's proposal actually did more poorly than the plain version in the presence of a doublet (17.4% versus 49.6%).

Limitations and future work.

As already mentioned in discussing the current performance of CD that does not match that documented by Ruscio and Roche (2012), there is no guarantee that NESTip or any other NEST method should quite generally outperform CD. But given the current poor performance of CD, and in absence of contradicting information using problem structures more favorable to CD, the current study minimally establishes that there is a class of problem complexity for which NEST methods markedly outperform CD as well as MLFA, even without the handling of Heywood cases proposed by Jöreskog (1967).

Reasons were proposed why CD is handicapped for the purpose of identifying the total number of factors contributing to the correlations between the observed variables. This handicap of CD would remain, even for much larger sample sizes than those studied here. MLFA, however, might fare far better than it did here for much larger sample sizes, with larger number of factors and of observed variables, perhaps especially when the factor space includes small eigenvalues due to factor correlations. Would there be conditions sufficiently different from the present ones under which MLFA outperforms NESTip? This is possible but such conditions remain to be identified. If such conditions exist for which MLFA would perform better than NESTip, the other NEST variants should also be verified, because perhaps the widely changed conditions would also change their relative power. More work is thus required either to extend the confidence that NESTip is generally the method to be recommended or to develop rules to adapt the recommendation to particular observable data conditions.

Conclusion

In EFA, data modeling approaches may be used to identify how many dimensions are required to account for the correlations expressed in a data set. Ideally, all dimensions should be identified, including so-called minor factors. If these reflect negligible sources of information, they can be consciously disregarded to simplify the picture. But if they result from the correlation of important factors, their neglect would prevent correct understanding of the underlying dynamics. To identify all the common dimensions expressed in a data set, the Next Eigenvalue Sufficiency Test, implemented in its NESTip version, was shown, over a lim-



ited but representative range of conditions, to be the most reliable of the eight methods tested. This NEST procedure is made available in R and MATLAB code as well as a complement in Excel (see Appendix A). An illustrative application of the latter is provided in Appendix B.

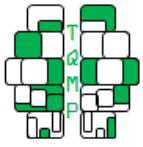
Authors' note

The author thanks Louis Laurencelle for a critical review of this manuscript.

References

- Achim, A. (2012). Detecting outliers in multivariate data while controlling false alarm rate. *Tutorials in Quantitative Methods for Psychology*, 8, 108–121.
- Briggs, S. R., Cheek, J. M., & Buss, A. H. (1980). An analysis of the self-monitoring scale. *Journal of Personality and Social Psychology*, 38, 679–686. doi:10.1037/0022-3514.38.4.679
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272–299. doi:10.1037/1082-989X.4.3.272
- Finch, J. F. & West, S. G. (1997). The investigation of personality structure: statistical models. *Journal of Research in Personality*, 31, 439–485. doi:10.1006/jrpe.1997.2194
- Glorfeld, L. W. (1995). An improvement on horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55, 377–393. doi:10.1177/0013164495055003002
- Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W.-j. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement*, 72, 357–374. doi:10.1177/0013164411422252
- Green, S. B., Thompson, M. S., Levy, R., & Lo, W. J. (2015). Type i and type ii error rates and overall accuracy of the revised parallel analysis method for determining the number of factors. *Educational and Psychological Measurement*, 75, 428–457. doi:10.1177/0013164414546566
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185. doi:10.1007/BF02289447
- Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32, 443–482. doi:10.1007/BF02289658
- Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, 60, 64–82. doi:10.1017/S037016460002006X
- Matsumoto, M. & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. on Modeling and Computer Simulation*, 8, 3–30. doi:10.1145/272991.272995
- Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? *Computational Statistics & Data Analysis*, 49, 974–997. doi:10.1016/j.csda.2004.06.015
- Ruscio, J. & Kaczetow, W. (2008). Simulating multivariate non-normal data using an iterative algorithm. *Multivariate Behavioral Research*, 43, 355–381. doi:10.1080/00273170802285693
- Ruscio, J. & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of a known factorial structure. *Psychological Assessment*, 24, 282–292. doi:10.1037/a0025697
- Turner, N. E. (1998). The effect of common variance and structure pattern on random data eigenvalues: implications for the accuracy of parallel analysis. *Educational and Psychological Measurement*, 58, 541–568. doi:10.1177/0013164498058004001
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlation. *Psychometrika*, 41, 321–327. doi:10.1007/BF02293557
- Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: a review and evaluation of alternative procedures for determining the number of factors or components. In J. L. Fava (Ed.), *Problems and solutions in human assessment* (pp. 41–71). New York: Springer. doi:10.1007/978-1-4615-4397-8_3
- Ziegler, M. & Hagemann, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment*, 31, 231–237. doi:10.1027/1015-5759/a000309
- Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432–442. doi:10.1037/0033-2909.99.3.432

Appendices follow on next page.



ranked first among the 1001 available and all remaining eigenvalues ranked last. The number of factors that PA suggests terminates that line. This is followed by the NEST (preferred NESTip version) for models with increasing number of dimensions, where the rank of the data eigenvalues among the corresponding eigenvalues of the surrogate data is given for each consecutive eigenvalue up to one past the number of factors constituting the current null hypothesis. This last rank is the actual Next Eigenvalue Sufficiency Test of the model; the previous ranks are presented to assure that the PAFA model was successful at approximating the data eigenvalues. For instance, in Figure ??, for “H0: 3 factors”, the ranks are 528, 511, 422, and 12; the first three ranks confirm close approximation of the corresponding data eigenvalues by the model and the last one specifies $p < .012$ (i.e., = 12/1001) for the test of the current null hypothesis. The test ranks for consecutive eigenvalues up to $k + 1$ are followed by the test verdict (model rejected or accepted) using the .05 significance level. All previous ranks should be around 500 since the model was adjusted to match the data eigenvalues.

Open practices

📄 The *Open Material* badge was earned because supplementary material(s) are available on the [journal's web site](#).

Citation

Achim, A. (2017). Testing the number of required dimensions in exploratory factor analysis. *The Quantitative Methods for Psychology*, 13(1), 64–74. doi:[10.20982/tqmp.13.1.p064](https://doi.org/10.20982/tqmp.13.1.p064)

Copyright © 2017, Achim. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 31/10/2016 ~ Accepted: 10/11/2016