Research & Reviews: Journal of Statistics and Mathematical Sciences

Determining Dimensionality of Binary Variables: A Monte Carlo Simulation Study

Ting Dai[1]*, Adam Davey[2]

[1]Department of Educational Psychology, University of Illinois, Chicago, USA
[2]Department of Behavioral Health and Nutrition, University of Delaware, Newark, USA

Research Article

Received: 10-Dec-2022, Manuscript No. JSMS-22-83052; **Editor assigned:** 12-Dec-2022, PreQC No. JSMS-22-83052 (PQ); **Reviewed:** 26-Dec-2022, QC No. JSMS-22-83052; **Revised:** 27-Feb-2023, Manuscript No. JSMS-22-83052 (R); **Published:** 08-Mar-2023, DOI: 10.4172/ JSMS.9.2.011

***For Correspondence :** Ting Dai, Department of Educational Psychology, University of Illinois, Chicago, USA;

Email: tdai@uic.edu

Citation: Dai T, et al. Determining Dimensionality of Binary Variables: A Monte Carlo Simulation Study. RRJ Stats Math Sci. 2023;9:0011.

ABSTRACT

Objective: The present study aimed to evaluate four criteria—Kaiser, empirical Kaiser, parallel analysis, and profile likelihood for determining the dimensionality of binary variables.

Methods: A large scale Monte Carlo simulation was conducted to evaluate these criteria across combinations of correlation matrices (Pearson r or tetra choric ρ) and analysis methods (principal component analysis or exploratory factor analysis), and combinations of study characteristics sample sizes (100, 250, 1000), variable splits (10%/90%, 25%/75%, 50%/50%), dimension (1, 3, 5, 10), and items per dimension (3, 5, 10).

Results: Parallel analysis performed best out of the four criteria, recovering dimensionality in 87.9% of replications when using principal component analysis with Pearson correlations.

Conclusion: Our findings suggested that dimensionality of a binary variable data matrix is best determined by parallel analysis using the combination of principal component analysis with a correlation matrix based on Pearson r. We provided recommendations for selecting criteria in different study conditions.

Keywords: Dimensionality determination; Binary variable; Dichotomous variable; Principal component analysis; Parallel analysis; Factor analysis

INTRODUCTION

Growing consensus establishes a set of criteria for determining dimensionality of continuous variables under a variety of conditions, such as Kaiser's criterion and parallel analysis. New methods, e.g., empirical Kaiser criterion, are also being presented in the literature that may outperform benchmark methods in some circumstances. Despite their ubiquity, criteria for determining dimensionality from binary/dichotomous indicators are not yet established. Although some previous research suggests the best approaches for determining the dimensionality of continuous indicators may not work as well with binary indicators, these studies have typically considered only one analysis method (e.g., factor analysis), a selected few criteria, and/or a limited number of study conditions. Furthermore, prior research has not subjected the same data sets to direct comparisons by different combinations of analysis method and data matrix. It raises questions about whether these findings generalize across the range of conditions typical for research in the behavioral, health, social, and educational sciences [1].

Binary data and matrices

Binary (or dichotomous) data arise across nearly every discipline and can correspond with many different types of commonly encountered types of data (e.g., true/false, yes/no, correct/incorrect, agree/disagree, present/absent, observed/missing, positive/negative, dropout/retained etc.). Such data are often conceptualized as resulting from an underlying continuous (e.g., Gaussian) distribution where the observation's value is determined in relation to some underlying cut point or threshold, equatable with a particular split in the data (e.g., 50%/50% or 10%/90%; or the difficulty parameter). The data matrix for each would be identical for comparable criteria, regardless of how the underlying data were conceptualized or generated [2].

Although analysis of binary indicators as though they were continuous is commonplace in disciplines like economics, conventional wisdom, backed with some previous research, leads to an expectation that tetrachoric correlation coefficients outperform Pearson r for binary data in some applications, such as data reduction. Except under conditions reflecting relatively extreme splits in the data, we should generally expect differences to be small based on the extent to which Gaussian, logit, and probit distributions have been shown largely to overlap once standardized for differences in means and variances. In the current study we considered conditions for which these comparisons may not have been systematically evaluated by previous research, we evaluate all methods using both Pearson r and tetrachoric ρ coefficients [3].

Dimension reduction

Researchers across a wide range of disciplines frequently need to perform data reduction on binary data matrices. We focus on two of the most widely used analyses that involve dimensionality determination principal component analysis and factor analysis (or principal axis factoring or exploratory factor analysis), since other approaches appropriate to binary data such as singular value decomposition and non-negative matrix factorization are not yet widely encountered in most areas of the behavioral, educational, health, social sciences. Likewise, we do not consider confirmatory approaches here, such as Confirmatory Factor Analysis (CFA), where dimensionality is specified a priori, or latent class analysis, which relies on very different approaches to determining dimensionality from the methods considered here and can place considerable demands on sample sizes [4].

Principal component analysis

Principal Component Analysis (PCA) seeks to reduce the dimensionality of multivariate data containing multiple inter correlated variables in a way that retains maximal variation present in the data set. This is accomplished by a standardized linear projection which maximizes the variance in the projected space and also minimizes the squared reconstruction error. As such, the first few principal components retain most of the variation present in the original set of variables. In other words, weights are applied to form linear combinations of the original variables, such that the first component has the largest inertia (*i.e.,* variance), the second component is computed such that it is orthogonal to the first component and has the largest remaining variance, and so forth until the same number of components as the observed variables are computed. A key decision for the researcher then is to determine the number of principal components to retain to accomplish the goal of sufficiently extracting the information with a parsimonious structure a reduced set of principal components, sufficient for the aims of data reduction task (e.g., ≥ % variance recovered, identifying the most important signals in the data, denoising, etc.).

Important goals of principal component analysis are to (1) extract the most important information from the multivariate data matrix, (2) reduce the data matrix by retaining only the orthogonal principal components with maximal variance of the original data matrix, and thereby (3) clarify the structure of the observations and the variables in the multivariate data matrix (4) to denoise a dataset. Therefore, for data reduction PCA's emphasis is on accounting for maximal variance, rather than capturing maximal covariance or explaining inter-relations between observed variables in the original data matrix. Mathematically, the principal components are empirical aggregates of the inter correlated observed variables, without much underlying theory about which variables should be associated

with which components [5].

Factor analysis

Factor Analysis (FA), in contrast, hypothesizes a set of underlying latent common factors that potentially explain the associations among observed variables. In a common factor model, the factors influence the observed variables (*i.e.,* factor indicators), as modeled in the linear regression functions of observed variables which are dependent on the latent factors. The factors (*i.e.,* predictors in the linear regressions) are shared among the observed variables with different regression coefficients (*i.e.,* factor loadings), whereas the regression residuals, representing the variances unrelated to the common factors, are unique to the observed variables. As such, the total variance of an observed variable is partitioned into the variance contributed by the common factors (*i.e.,* communality) and the variance unrelated to the common factor (*i.e.,* uniqueness). Factor analysis estimates communalities to minimize unique and error variance from the observed variables. This is a key difference from principal component analysis, which provides a mathematically determined empirical solution with all variances included [6-10]. The key to estimating the varying factor loadings for different observed variables lies in the covariance (or correlation) structure among the original observed variables and the common factor model implied covariance structure. Unlike principal component analysis, principal factor analysis is an analysis of the covariances among the observed variables in the original data matrix, and its purpose is primarily to seek the underlying theory (*i.e.,* the common factors) about why the observed variables correlate to the extent they do [11].

Despite the distinctions in goal of analysis and extraction technique, factor analysis is similar to principal component analysis in its utility in reducing a large number of observed variables down to a few latent factors/components (*i.e.,* the underlying dimensions). Factor analysis achieves a reduction in dimensions by invoking a common-factor model relating observed variables to a smaller number of latent common factors [12].

As noted above, PCA and FA differ in terms of their goals, objectives, purpose, and the kinds of applications to which they are best suited, but are similar in terms of an essential step determining dimensionality based on eigenvalue decompositions. For this reason, we compared the two analyses with the primary emphasis of correctly identifying the dimensionality of a set of binary variables using different determination criteria.

Determining dimensionality

There is a vast literature on criteria for determining dimensionality (or retaining underlying factors/components) of multivariate variables that can be applied to different approaches to data reduction, e.g., principal component analysis and factor analysis. One of the most widely applied is the Kaiser criterion (also called the Kaiser-Guttman rule, which retains all dimensions with eigenvalues greater than one (*i.e.,* the dimensions that explain at least as much variance as a typical standardized item). The Kaiser criterion is the default approach in most statistical packages (e.g., SPSS), although some studies have suggested this approach may lead to over extraction of components in many applications [13].

Kaiser's method ignores the fact that eigenvalues are sorted from largest to smallest, thus capitalizing on chance differences associated with sampling variance. To address this issue, parallel analysis uses data simulated under an independence assumption to subtract out this sampling error variance, and solutions can be evaluated using several different criteria (e.g., mean, median, 95%ile). A wide variety of simulation studies have suggested that parallel analysis provides an unbiased estimate of the number of underlying components and works well in practice, but its effectiveness is less clear with binary indicators such as in the current paper [14].

Quite recently, parallel analysis has been extended to consider data generated under data structures with varying numbers of components in the Comparison Data method, with additional components being added to the simulated data for as long as they produce better agreement with the structure of the original data. We do not consider this method here because it: (a) can be quite computationally intensive and (b) generally agrees quite well with results from parallel analysis; however, it represents a potentially important addition to the methods available for determining the number of components to be retained [15].

In empirical contexts, many researchers rely on visual scree plots in order to determine the number of components to retain in an analysis. A plot of eigenvalues against the eigenvalue number is used to identify an "elbow" or "large gap" in the data at the point where the useful "signal" degenerates into noise, or "scree." However, this method provides no definite quantitative cutoffs, and hence is difficult to use for empirical evaluation. The method of profile likelihood attempts to address this shortcoming by quantifying the number of retained components that maximize the observed data likelihood, thus providing an empirical (and automatic) method for determining the point at which a "gap" or "elbow" occurs within a scree plot.

Finally, an Empirical Kaiser Criterion (EKC) has recently been presented in the literature. This approach is grounded in statistical theory and accounts for the serial nature of eigenvalues. In a Monte Carlo study, the EKC approach generally performed at least as well as parallel analysis, particularly with larger sample sizes and a smaller number of variables [16].

There has not been consensus, however, on which criteria may perform well to determine dimensionality of binary variables in various circumstances. A recent simulation study indicated that, while parallel analysis generally performed well, it was not as effective in factor analysis with dichotomous indicators. Adapted from traditional use of

confirmatory factor analysis, Finch recommended using the root mean square error of approximation to determine the number of factors. However, the results also suggested that lack of convergence may often be expected under the kinds of circumstances faced by applied researchers (e.g., smaller sample sizes and larger number of variables), which would make it considerably less robust in the applied research.

Similarly to Finch's approach, the hull method also utilizes fit indices (e.g., comparative fit index) together with degree of freedom, which are traditionally for confirmatory factor analysis, to assist in determining the number of underlying factors to extract for an exploratory factor analysis. Although a heuristic approach, the performance was largely dependent on the choice of a fit index to yield the hull variant, which is suited for specific kind(s) of model estimator. We do not consider the hull method here due to the additional within method conditions one needs to consider, which are not shared by dimension determination approaches such as Kaiser, parallel analysis, and EKC.

To our knowledge, research effort on binary data dimensionality criteria has focused on parallel analysis, but the findings were not conclusive. In an earlier simulation study, the researchers investigated parallel analysis as an approach to dimensionality determination for unidimensional binary data. They found that with the 95th and 99th percentiles of random data eigenvalues as criteria, parallel analysis was accurate in identifying the unidimensionality in the simulated binary variables. On the other hand, in a similar study that evaluated parallel analysis in determining binary variables sharing a single underlying dimension, the results did not agree. The researchers concluded that parallel analysis generally did not perform well with FA in determining unidimensionality among a set of binary variables when a Pearson correlation matrix was analyzed, with the tetrachoric correlation matrix and parallel analysis providing somewhat better, but still unsatisfactory, results. Their findings also pointed to other influencing factors on the performance of parallel analysis, including sample size, and factor loading [17].

Study overview

The purpose of this study was to evaluate the performance of different methods for correctly determining the dimensionality underlying a set of binary indicators across a range of conditions typical for multivariate social, behavioral, and educational research. Specifically, we conducted a large-scale simulation to evaluate several criteria (Kaiser's criterion, empirical Kaiser's criterion, parallel analysis, and profile likelihood) for determining the dimensionality of binary variables given combinations of methods (principal component analysis or exploratory factor analysis) and matrices (Pearson r or tetrachoric ρ), sample sizes (100, 250, 1000), variable splits (10%/90%, 25%/75%, 50%/50%), underlying dimensions (1, 3, 5, 10), and items per dimension (3, 5, 10) with 1000 replications per condition. We focused on determining dimensionality since it is established in identical fashion for both PCA and PAF.

MATERIALS AND METHODS

Design

In order to consider a range of conditions typical of many real world applications with binary variables (e.g., 0/1, incorrect/correct, observed/missing) we used a factorial design. Between subjects factors, which represent various characteristics of study data, included: (1) the number of underlying dimensions-1, 3, 5, or 10; (2) the number of items per dimension-3, 5, or 10; (3) sample size 100, 250, or 1000; and (4) binary variable splits 90%/10%, 75%/25%, or 50%/50%. Within subjects factors, which represent researcher selected analytic approaches, were: (i) correlation matrix Pearson r or tetrachoric ρ; and (ii) analysis method PCA or PAF.

Outcomes

We evaluated four different criteria for determining dimensionality the number of underlying dimensions. These included: 1) Empirical Kaiser Criterion (EKC), 2) Kaiser criterion, 3) parallel analysis with 95%ile criterion, and 4) profile likelihood. We considered a criterion performed successfully in a specific condition if it recovered the correct number of underlying dimensions in at least 95% of replications [18].

Data and procedure

Between subject: We constructed 1000 replications in each condition with distinct between subject factors, for example, 1000 replications with (1) one underlying dimension, (2) three items per dimension, (3) N=250, and (4) a 10%/90% split. Population correlations were set at 70 within variables on the same dimension and 30 between variables on different dimensions, and all data were drawn from multivariate normal distributions. Consistent with our hypothesized latent variable model, data were dichotomized *via* a probit link function. Variables were dichotomized based on whether observed values of the function exceeded the threshold associated with the population %ile cut point for that condition [19].

Within subject: For each replication, the binary data matrix was converted to correlation matrices (Pearson r and tetrachoric ρ) and analyzed by Principal Component Analysis (PCA) and principal axis Factoring (FA). The number of

dimensions indicated by each criterion was determined for each of the four combinations of correlation matrix and analysis method.
Whether the dimensionality was correctly recovered was determined for each criterion under each combination of method and matrix. A nominal indicator (1=parallel analysis, 2=empirical Kaiser, 3=Kaiser, and 4=profile likelihood) was constructed with the value referring to the criterion performed best in a condition.

Statistical analysis: Descriptive statistics and factorial Analysis of Variance (ANOVA) were used to examine the association of each between and within subject factor with recovery of the correct number of dimensions. Factorial ANOVA was in preference to logistic regression owing to the large number of conditions where some criteria performed perfectly, which can lead to estimation difficulties with likelihood based techniques. We also conducted a recursive partitioning analysis (or, Classification and Regression Trees (CART)) to classify the simulation results, with the tree pruned based on the minimum value of Mallow's Cp. The R statistical package was used for all data simulation and analyses. All code and data generated are available from an online data repository.

RESULTS

Convergence

Convergence was achieved for 99.98% of data sets, regardless of matrix-analysis combination. Failed convergence occurred only for N=100 and when the split was 10%/90%. The lowest convergence rate (98.6%) was when N=100, with 1 underlying dimension, 3 items per dimension, and a 10%/90% split, likely due to variables having insufficient variance/covariance for analysis.

Criterion performances

Parallel analysis was the best performing criterion of the four (all >86%; Table 1) in all combinations of matrix and analysis, except in PAF/ρ where EKC was the criterion that correctly determined dimensionality for the most replications (77.3%). Note, however, the performances of dimension determination criteria not only varied across the four matrix analysis combinations as shown above, but also differed by the between-subject factors. This calls for a detailed examination of performances by sample size, the number of underlying dimensions, the number of items per dimension, and the variable split, within a combination of matrix and analysis type, which we elaborate below.
Among the four matrix analysis combinations, PCA/r had the highest average percentage of correctly recovered replications (77.0%; Table 1), and in this combination the best performing criterion parallel analysis correctly determined the highest percentage of replications (87.9%) compared to any criterion in any other matrix analysis combination. Given these results, we present the detailed evaluation of criteria only for the combination of PCA with Pearson r (Supplementary Table 1). Full results are available as supplementary material for researchers whose application suggests a different combination.

Table 1. Percentage of replications with dimensionality correctly determined by the four criteria given method and correlation matrix.

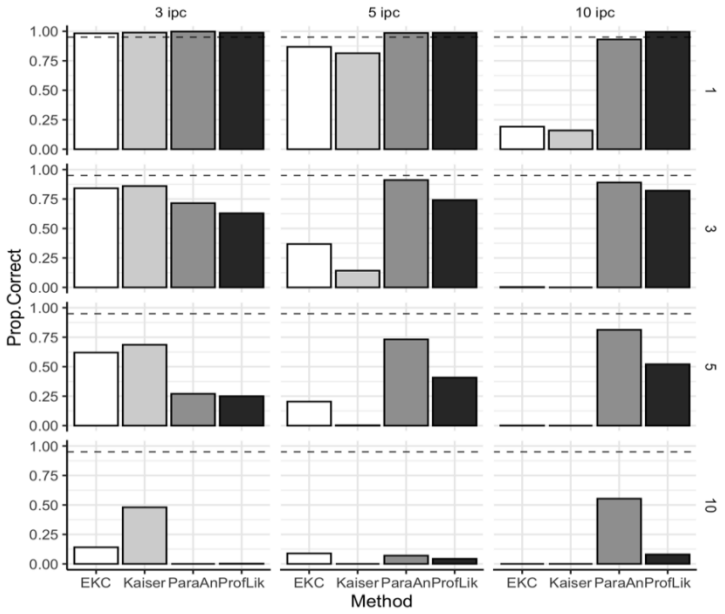
Analysis	Correlation	Criteria averaged	EKC	Kaiser	ParaAn	ProfLik
PCA	Pearson	77.00%	79.30%	73.40%	87.90%	67.40%
PCA	Tetrachoric	71.10%	73.70%	73.70%	86.20%	50.80%
FA	Pearson	67.50%	58.90%	63.90%	86.10%	61.10%
FA	Tetrachoric	61.40%	77.30%	76.70%	44.10%	47.40%

Note: Boldfaced indicates the method/criterion that performed the best with each combination. PCA: Principal Component Analysis; FA: Factor Analysis.

Criterion performances given PCA with Pearson r

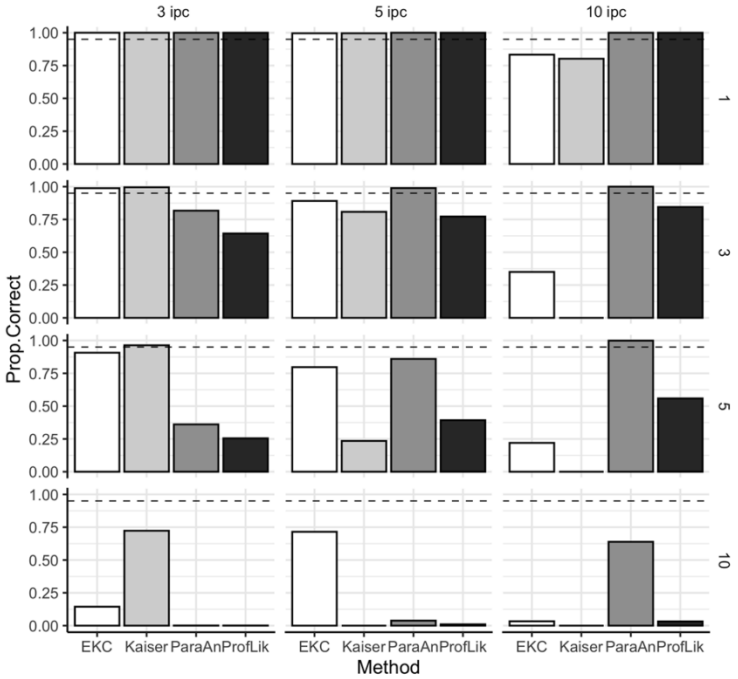
Our design included four between subject factors 3 sample sizes × 4 numbers of underlying dimensions × 3 numbers of items per dimension × 3 splits of variable values, which created a total of 108 conditions given the combination of PCA and Pearson r (and 432 conditions overall). Due to the large number of replications, ANOVAs indicated that all model effects were statistically significant using α=.05, including the 5 way interaction. For this reason, we summarize the key findings and influences for each criterion (Figures 1-9).

Figure 1. Proportion of rep. recovering the correct dimensionality (N=100, 10%/90% split).



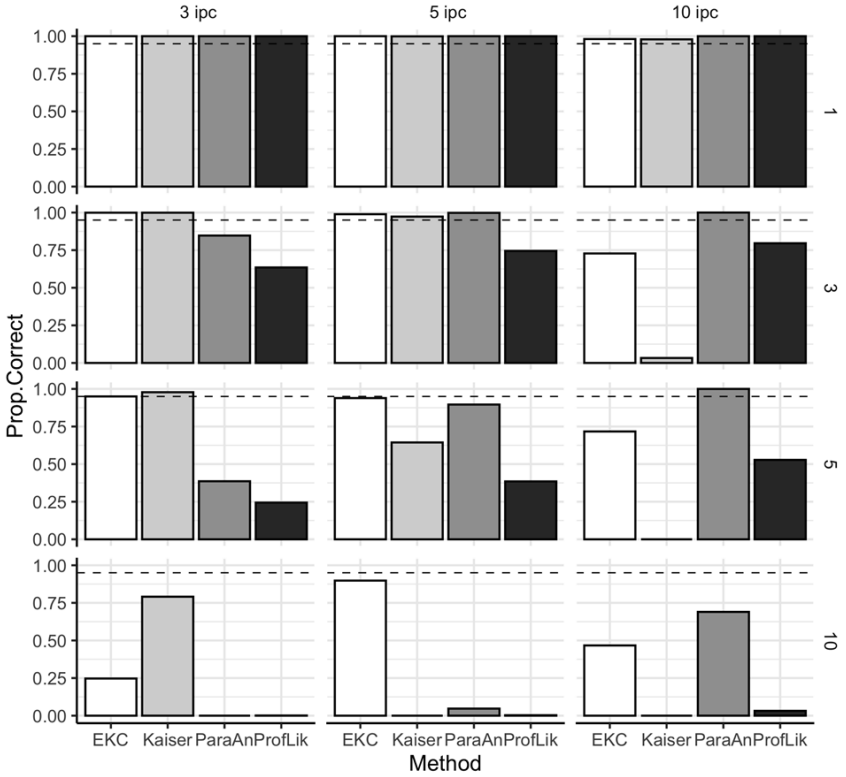
Note: The rows of the bar graph matrix represent numbers of underlying dimensions/components and the columns represent numbers of Items Per Component (IPC). The X axis represents the 4 criteria/method, the Y axis represents the proportion of replications that recover the correct dimensionality, and the dotted line is the 95% reference line. EKC: Empirical Kaiser Criterion; Kaiser: Eigenvalue>1.0; ParaAn: Parallel Analysis; and ProfLik: Profile Likelihood.

Figure 2. Proportion of rep. recovering the correct dimensionality (N=100, 25%/75% split).



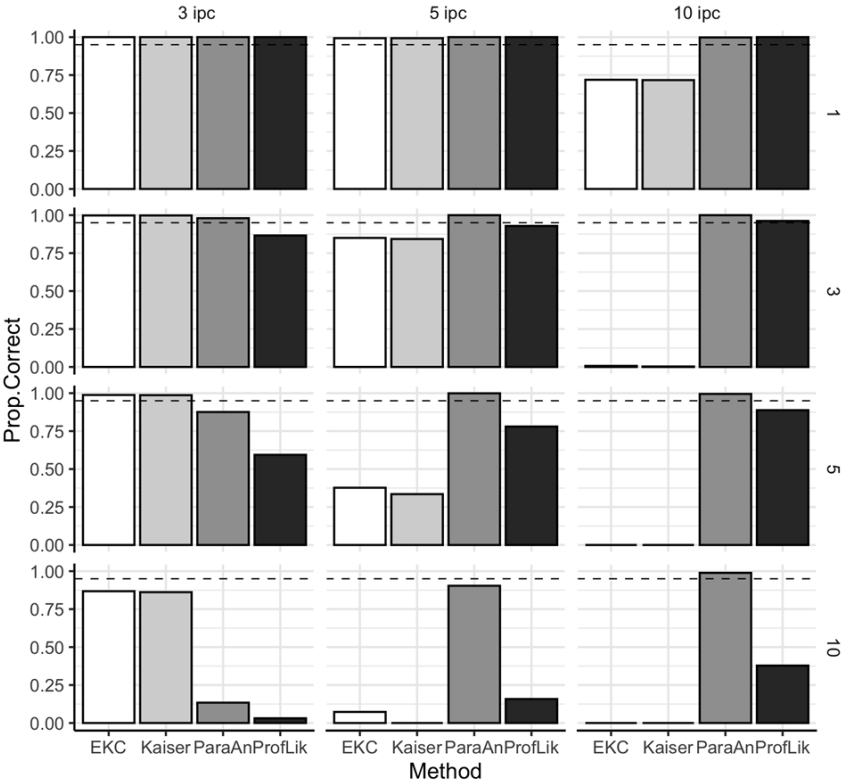
Note: The rows of the bar graph matrix represent numbers of underlying dimensions/components and the columns represent numbers of Items Per Component (IPC). The X axis represents the 4 criteria/method, the Y axis represents the proportion of replications that recover the correct dimensionality, and the dotted line is the 95% reference line. EKC: Empirical Kaiser Criterion; Kaiser: Eigenvalue>1.0; ParaAn: Parallel Analysis; and ProfLik: Profile Likelihood.

Figure 3. Proportion of rep. recovering the correct dimensionality (N=100, 50%/50% split).



Note: The rows of the bar graph matrix represent numbers of underlying dimensions/components and the columns represent numbers of Items Per Component (IPC). The X axis represents the 4 criteria/method, the Y axis represents the proportion of replications that recover the correct dimensionality, and the dotted line is the 95% reference line. EKC: Empirical Kaiser Criterion; Kaiser: Eigenvalue>1.0; ParaAn: Parallel Analysis; and ProfLik: Profile Likelihood.

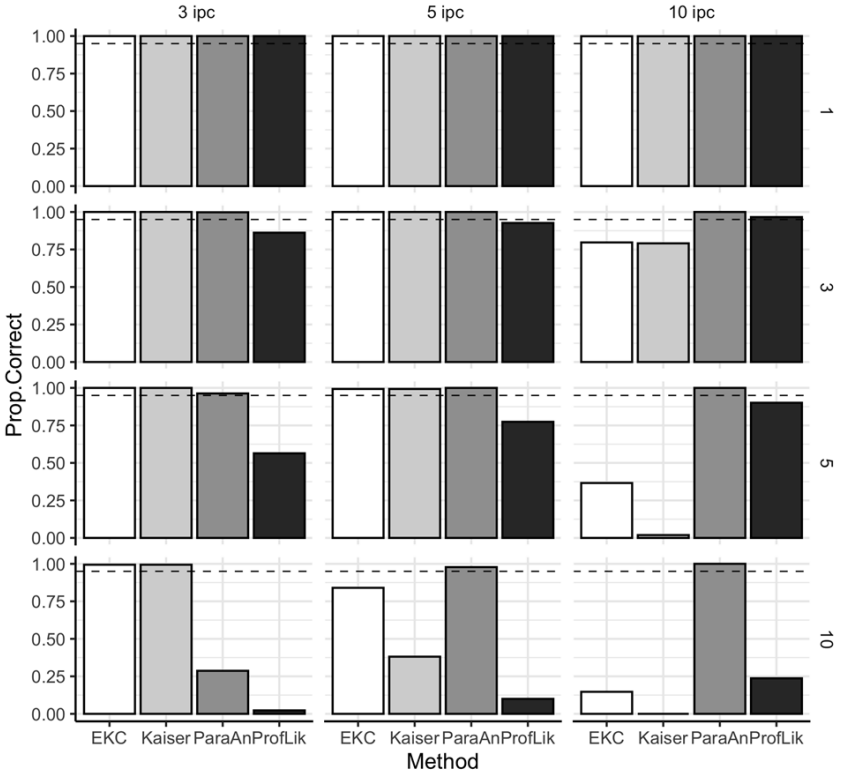
Figure 4. Proportion of rep. recovering the correct dimensionality (N=250, 10%/90% split).



Note: The rows of the bar graph matrix represent numbers of underlying dimensions/components and the columns

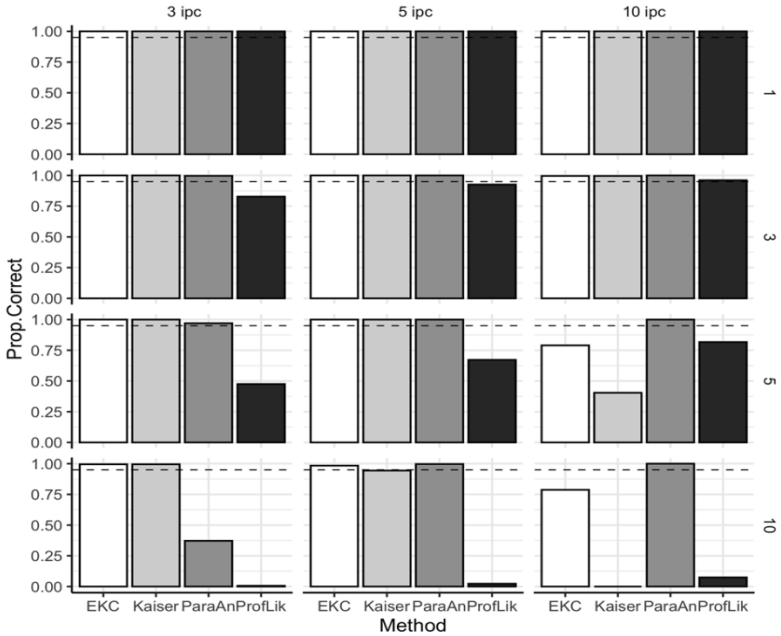
represent numbers of Items Per Component (IPC). The X axis represents the 4 criteria/method, the Y axis represents the proportion of replications that recover the correct dimensionality, and the dotted line is the 95% reference line. EKC: Empirical Kaiser Criterion; Kaiser: Eigenvalue>1.0; ParaAn: Parallel Analysis; and ProfLik: Profile Likelihood.

Figure 5. Proportion of rep. recovering the correct dimensionality (N=250, 25%/75% split).



Note: The rows of the bar graph matrix represent numbers of underlying dimensions/components and the columns represent numbers of Items Per Component (IPC). The X axis represents the 4 criteria/method, the Y axis represents the proportion of replications that recover the correct dimensionality, and the dotted line is the 95% reference line. EKC: Empirical Kaiser Criterion; Kaiser: Eigenvalue>1.0; ParaAn: Parallel Analysis; and ProfLik: Profile Likelihood.

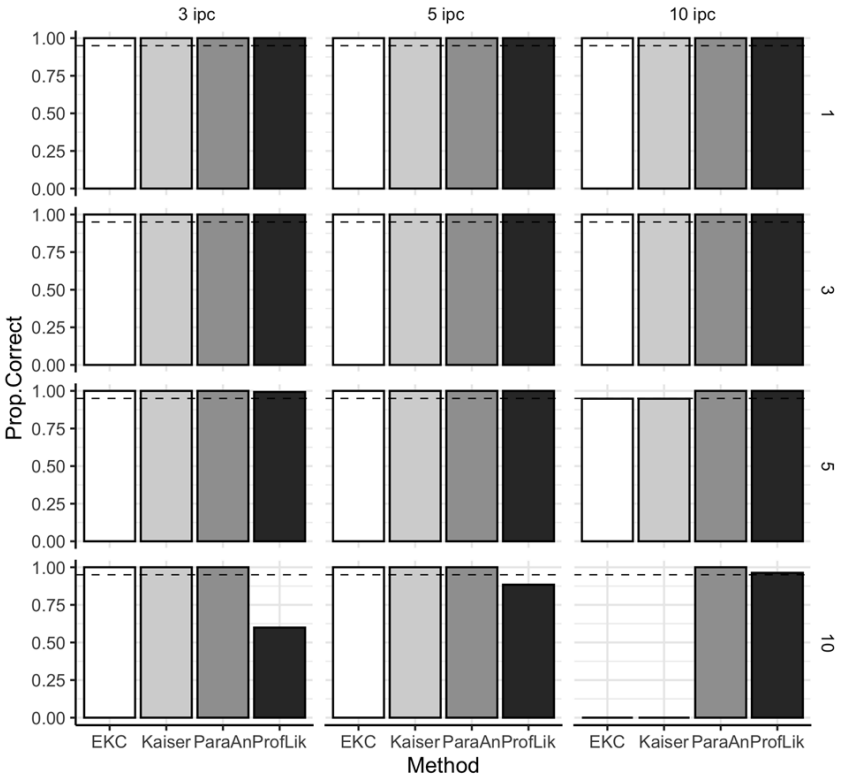
Figure 6. Proportion of rep. recovering the correct dimensionality (N=250, 50%/50% split).



Note: The rows of the bar graph matrix represent numbers of underlying dimensions/components and the columns represent numbers of items per Component (IPC). The X axis represents the 4 criteria/method, the Y axis represents

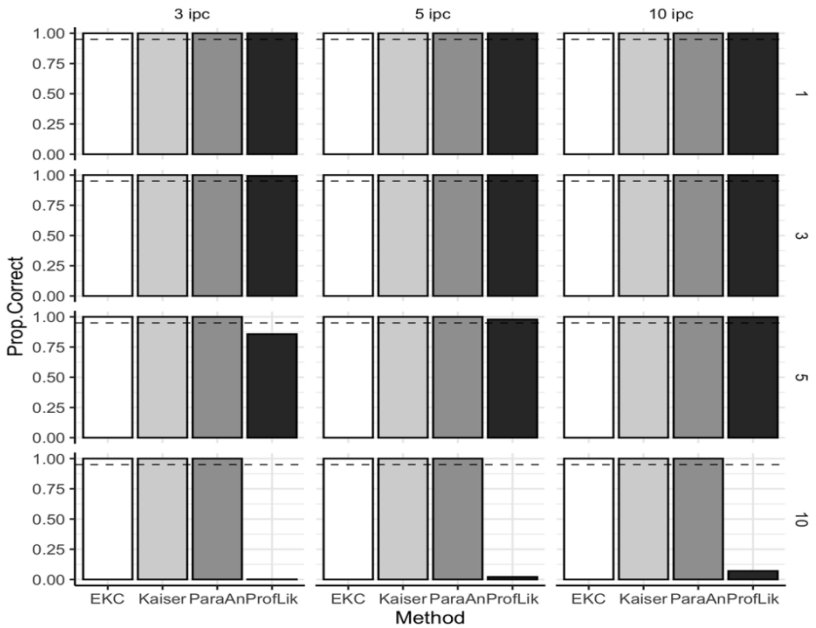
the proportion of replications that recover the correct dimensionality, and the dotted line is the 95% reference line. EKC: Empirical Kaiser Criterion; Kaiser: Eigenvalue>1.0; ParaAn: Parallel Analysis; and ProfLik: Profile Likelihood.

Figure 7. Proportion of rep. recovering the correct dimensionality (N=1000, 10%/90% split).



Note: The rows of the bar graph matrix represent numbers of underlying dimensions/components and the columns represent numbers of Items Per Component (IPC). The X axis represents the 4 criteria/method, the Y axis represents the proportion of replications that recover the correct dimensionality, and the dotted line is the 95% reference line. EKC: Empirical Kaiser Criterion; Kaiser: Eigenvalue>1.0; ParaAn: Parallel Analysis; and ProfLik: Profile Likelihood.

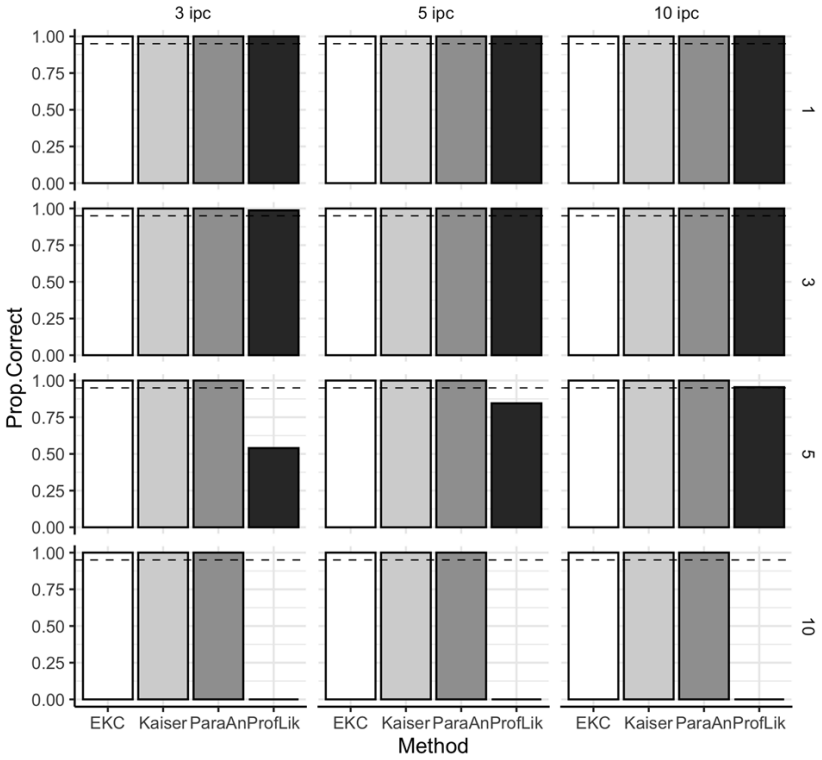
Figure 8. Proportion of rep. recovering the correct dimensionality (N=1000, 25%/75% split).



Note: The rows of the bar graph matrix represent numbers of underlying dimensions/components and the columns represent numbers of Items Per Component (IPC). The X axis represents the 4 criteria/method, the Y axis represents the proportion of replications that recover the correct dimensionality, and the dotted line is the 95% reference line.

EKC: Empirical Kaiser Criterion; Kaiser: Eigenvalue>1.0; ParaAn: Parallel Analysis; and ProfLik: Profile Likelihood.

Figure 9. Proportion of rep. recovering the correct dimensionality (N=1000, 50%/50% split).



Note: The rows of the bar graph matrix represent numbers of underlying dimensions/components and the columns represent numbers of Items Per Component (IPC). The X axis represents the 4 criteria/method, the Y axis represents the proportion of replications that recover the correct dimensionality, and the dotted line is the 95% reference line. EKC: Empirical Kaiser Criterion; Kaiser: Eigenvalue>1.0; ParaAn: Parallel Analysis; and ProfLik: Profile Likelihood.

Parallel analysis

Parallel analysis performed the best among the four criteria given the PCA/r combination. Parallel analysis's performance was driven primarily by sample size and the number of underlying dimensions. Parallel analysis correctly recovered the number of underlying dimensions at least 95% of the time under all 36 conditions when N=1000 in Figures 7-9, 31 of 36 conditions when N=250 in Figures 4-5, and 14 of 36 conditions when N=100 in Figures 1-3. PA also performed less well when the number of underlying dimensions was larger (e.g., 10 dimensions). The next influential factor was the number of items per dimension. Most of the failing cases for parallel analysis occurred given fewer items per dimension. The binary value split did not matter for the performances of PA.

Empirical Kaiser criterion

Performance of the empirical Kaiser criterion was driven primarily by sample size and binary value split. Better performances were observed when sample sizes were larger (e.g., 1000; Figures 7-9) and with a more even split (e.g., 50%/50%). The next influential factors were the number of underlying dimensions and items per dimension. EKC performed better when there were fewer underlying dimensions and fewer items per dimension.

Overall, EKC did not outperform parallel analysis as a criterion for determining the dimensionality of binary variables in most conditions examined in the present study. The only conditions where EKC correctly determined the number of underlying dimensions in more replications than parallel analysis was when N=100 or 250, the true number of dimensions was 3, 5, or 10, and given 3 items per dimension.

Kaiser

Performance of the Kaiser criterion was driven primarily by sample size and the number of items per dimension. Better performances were observed given a larger sample size and few items per dimension. The next influential factors were the number of underlying dimensions (favoring fewer), followed by the binary value split (favoring a more even split).

Overall, the Kaiser criterion did not outperform parallel analysis as a criterion for determining the dimensionality of binary variables in most conditions we examined. It only outperformed parallel analysis in the same conditions where EKC also outperformed parallel analysis, in which its performances were not as well as EKC.

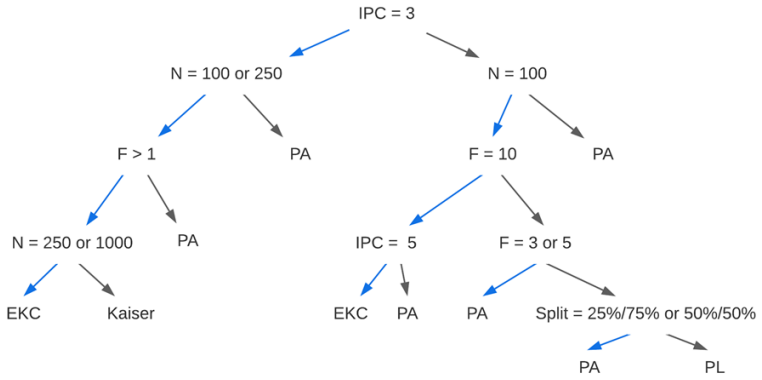
Profile likelihood

Performance of profile likelihood was driven primarily by the number of underlying dimensions. With a greater number of underlying dimensions, profile likelihood performed better. The next influential factor was sample size, favoring a larger sample size. Overall, profile likelihood did not outperform parallel analysis as a criterion for determining the dimensionality of binary variables in almost all conditions we examined in the present study.

Guidance for selecting a criterion

We obtained a classification and regression tree from the PCA/r results. Our data support using parallel analysis given PCA with Pearson correlations, except in four scenarios. First, use EKC when there are 3 items per dimension, >1 underlying dimensions, and N=250. Second, use the Kaiser criterion when there are 3 items per dimension, >1 underlying dimensions, and N=100. Third, use EKC when there are 5 or 10 underlying dimensions, N=100, and 5 items per dimension (Figure 10). Finally, use profile likelihood only when the underlying structure is unidimensional with 10 items per dimension, N=100, and the value split is an extreme 10%/90%. We note that in this case profile likelihood (99.5%) outperformed parallel analysis (93.1%) only by a small degree (Supplemental Figures 1-3 for the CART for the other three matrix-analysis combinations.)

Figure 10. Classification and regression tree for dimension determination criteria given PCA with Pearson r.



Note: Left (blue path) if node statement is true, right (black path) if node statement is false. IPC: number of items per component/dimension; F: number of dimensions; N: sample size; PA: Parallel Analysis with 95%ile; EKC: Empirical Kaiser Criterion; PL: Profile Likelihood.

Although parallel analysis generally outperformed the other methods under most circumstances when PCA was used with Pearson r, researchers should consider the closest scenario to their own from the conditions considered in our simulation. Then, dimensionality should be determined using one or all acceptable criteria for that combination of conditions and, in situations where there is disagreement a sensitivity analysis should be performed to make a final determination.

DISCUSSION

Dimensionality determination is critical across a range of disciplines and applications and represents the first and most critical step in data reduction, but criteria are not well established for binary data. Without clear criteria for first determining the dimensions underlying the large set of binary variables, subsequent steps (e.g., factor/component rotations) cannot be carried out and further analytic goals (e.g., data reduction, obtaining factor scores) cannot be achieved. This large scale simulation study examined the performance of four criteria (*i.e.,* parallel analysis, EKC, Kaiser, and profile likelihood) for determining dimensionality of binary variables under a considerably wider range of conditions than previous research. Specifically, we used a factorial design with 4 between subject factors-sample size, number of underlying dimensions, number of items per dimension, and variable value split. More importantly, this is the first study, to our knowledge, that directly compares criterion performance across combinations of method of analysis (principal component analysis *vs.* factor analysis) and type of matrix analyzed (Pearson correlations *vs.* tetrachoric correlations).

Our findings suggest that dimensionality of a binary variable data matrix is best determined using the combination of PCA with a correlation matrix based on Pearson r, regardless of how the data will ultimately be analyzed. This is important because numerous disciplines (e.g., educational testing, clinical assessment, personality research) 1,6,32,33,37 rely heavily on dichotomously scored measures and computationally intense tetrachoric ρ, which has risks of over estimating linear relations. We also find that parallel analysis is the criterion that most frequently

recovers the correct underlying dimensionality, expanding on prior research on parallel analysis of continuous variables to that of binary or dichotomous variables as a reliable approach to dimensionality determination. We discuss each important finding below.

FA was originally developed to identify common factors among normally distributed continuous variables; however, many assessments and questionnaires are made up of dichotomous or ordered categorical items. Prior literature calls EFA with categorical variables "Item Factor Analysis (IFA)"1. Conceptually, the least squares approach to IFA is based on the assumption that underlying each categorical indicator is a normally distributed continuous latent response variable 5,39. An individual's standing on this latent variable relative to a set of thresholds determines which response category they fall into. For example, for a dichotomous item, if the individual's standing on the latent response variable is below a certain threshold they will endorse a score of 0, whereas if they are above this threshold they will endorse a score of 1.

Previous comparisons of criteria for dimensionality determination with binary data have concluded that FA with tetrachoric correlations outperforms FA with Pearson's correlations. Our results found similar performance under these conditions, but further demonstrated that the results for dimensionality determination are poorest under precisely these conditions. Instead, parallel analysis combined with PCA using Pearson's correlations most often correctly determines dimensionality. We note that this does not in any way detract from the application of FA with binary data, which likely should be performed using tetrachoric correlations once dimensionality is established, in order to correctly estimate factor structure.

In several regards, results of this large scale simulation study differ from conventional wisdom, perhaps due to the broader range of conditions we considered, and also perhaps because we compared criterion performance across methods and matrices head-to-head on identical data sets. Most importantly, these results suggest that it is not necessary or beneficial to use tetrachoric correlation matrices in preference to Pearson r for the purposes of dimensionality determination, regardless of the combination of method and matrix ultimately used for analysis. Again, we believe that this should generally be the case except in situations with extreme splits on binary variables or other potentially ill conditioned circumstances. This is important because numerous fields and disciplines rely heavily on dichotomously scored measures. For example, ability testing in educational settings, and symptom assessment in clinical psychology, are often based on questionnaires in which responses can fall into either one of two categories (*i.e.,* correct/incorrect and symptom present/symptom absent), and the computationally intense tetrachoric ρ have often been used for analyses of these binary data, which may cause issues such as difficulty in estimation, over estimation of linear relationships.

Another finding that can be considered surprising is that, for the purposes of dimensionality determination, PCA outperforms FA. This is important because in educational and psychological research, factor analysis is primarily used in developing validity arguments for scales and theory development (e.g., intelligence, personality, executive functioning), because of its focus on identifying the underlying latent constructs that lead to correlations among observed variables. Dimension reduction is both a necessary step toward the research goal and an inevitable byproduct of process. Principal component analysis on the other hand is used to achieve a parsimonious summary of high-dimensional, multivariate data, so, arguably, data reduction is the most important goal of PCA. These results suggest that, although common practice is to apply either PCA or FA to a dataset exclusively for the entire analysis, it might be more effective, particularly in exploratory contexts, to use PCA for dimensionality determination, regardless of whether analyses will ultimately rely on PCA or FA.

A third important finding from the current study is that parallel analysis is the criterion that most frequently recovers the correct dimensionality. Traditional EFA, and the tools used to guide determinations of dimensionality, were developed for use with continuous data, and the application of these techniques to categorical data, especially dichotomous data, can lead to more suspect and difficult to interpret results. For example, parallel analysis might be less effective when scales contain dichotomous items. Parallel analysis can also be effective, but is less accessible for models with categorical indicators, and has also demonstrated a tendency to over factor in certain circumstances or under extract in others. However, contrary to these concerns, our large scale simulation results indicated generally more accurate performance of parallel analysis for binary variable dimensionality determination in various conditions tested.

In specific circumstances, EKC and Kaiser (and, in a very unique condition, profile likelihood), may be the most effective criteria to use. Some researchers consider using the Kaiser criterion to determine dimensions a common mistake with factor analysis. This is somewhat surprising since prior literature has shown that the Kaiser criterion tends to extract too many dimensions in FA with continuous variables. However, as a method to determine the dimensionality of multivariate dichotomous data, the traditional Kaiser criterion and its recent extension showed good performance in some conditions.

CONCLUSION

Our study provided empirical evidence for criteria to use for dimensionality determination the first and most important step in a variety of data reduction techniques. We directly compared the performance of four widely used criteria in a variety of study conditions and given different combinations of analysis method and type of matrix analyzed. Our

result supported using PCA with a Pearson correlation matrix as a preferred approach to determine how many dimensions that underlie a large set of binary variables for a variety of circumstances examined.

Limitations

The current study has some limitations. Although we considered a much wider range of conditions than most previous research addressing this topic, we only considered a single set of correlations among and within dimensions. Future research should consider more complex data structures. These include considering a number of conditions, such as different correlation structures, combinations of oblique and orthogonal dimensions (components or factors), deviations from simple structure, and even ordered categorical latent variables and finite mixture models.

REFERENCES

1. Clark DA, et al. Model fit and item factor analysis: Over factoring, under factoring, and a program to guide interpretation. Multivariate Behav Res. 2018;53:544-558.
2. Braeken J, et al. An empirical Kaiser criterion. Psychol Methods. 2017;22:450-466.
3. Kaiser HF, et al. The application of electronic computers to factor analysis. Educ Psychol Meas. 1960;20:141-151.
4. Horn JL, et al. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965;30:179-185.
5. Wirth RJ, et al. Item factor analysis: Current approaches and future directions. Psychol Methods. 2007;12:58.
6. Finch WH, et al. Using fit statistic differences to determine the optimal number of factors to retain in an exploratory factor analysis. Educ Psychol Meas. 2020;80:217-241.
7. Yang Y, et al. On the number of factors to retain in exploratory factor analysis for ordered categorical data. Behav Res. 2015;47:756-772.
8. Divgi DR, et al. Calculation of the tetrachoric correlation coefficient. Psychometrika. 1979;44:169-172.
9. Owen DB, et al. Tables for computing bivariate normal probabilities. Ann Math Stat. 1956;27:1075-1090.
10. Tallis GM, et al. The maximum likelihood estimation of correlation from contingency tables. Biometrics. 1962;18:342-353.
11. Carreira-Perpinan MA, et al. A review of dimension reduction techniques. Tech Rep. 1997;9:1-69.
12. Fahrmeir L, et al. Multivariate statistical modelling based on generalized linear models. 2nd Edition. Springer, New York. USA. 2010: 425.
13. Hastie T, et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Springer. New York, USA. 2009;536.
14. Collins LM, et al. Latent class and latent transition analysis: With applications in the social behavioral, and health sciences. 2nd Edition. John Wiley and Sons Publisher, USA. 2010;295.
15. Dalmaijer ES, et al. Statistical power for cluster analysis. BMC Bioinformatics. 2022;23:1-28.
16. Tein JY, et al. Statistical power to detect the correct number of classes in latent profile analysis. Struct Equ Modeling. 2013;20:640-657.
17. Hotelling H, et al. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24:417.
18. Guttman L, et al. Some necessary conditions for common factor analysis. Psychometrika. 1954;19:149-161.
19. Flora DB, et al. The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. Can J Behav Sci. 2017;49:78-88.