PSIHOLOGIJA, 2013, Vol. 46 (3), 331–347
DOI: 10.2298/PSI130801008Z

UDC 159.9.072.59

Impact of different conditions on accuracy of five rules for principal components retention

Aleksandar Zorić and Goran Opačić

Department of Psychology, Faculty of Philosophy, University of Belgrade, Serbia

Polemics about criteria for nontrivial principal components are still present in the literature. Finding of a lot of papers, is that the most frequently used Guttman Kaiser's criterion has very poor performance. In the last three years some new criteria were proposed. In this Monte Carlo experiment we aimed to investigate the impact that sample size, number of analyzed variables, number of supposed factors and proportion of error variance have on the accuracy of analyzed criteria for principal components retention. We compared the following criteria: Bartlett's χ^2 test, Horn's Parallel Analysis, Guttman-Kaiser's eigenvalue over one, Velicer's MAP and CHull originally proposed by Ceulemans & Kiers. Factors were systematically combined resulting in 690 different combinations. A total of 138,000 simulations were performed. Novelty in this research is systematic variation of the error variance. Performed simulations showed that, in favorable research conditions, all analyzed criteria work properly. Bartlett's and Horns criterion expressed the robustness in most of analyzed situations. Velicer's MAP had the best accuracy in situations with small number of subjects and high number of variables. Results confirm earlier findings of Guttman-Kaiser's criterion having the worse performance.

Key words: *Principal component analysis, Criterion for extraction, Factor retention*

Exploratory factor analysis (EFA) is *de facto* psychological method, not just because of its origin, but because it is among the most popular methods in psychology. The idea of identification of the structures underlying measured variables is very close to everyday psychological problems in which phenomena of interest cannot be measured directly, but have to be derived from the direct measures of behavior. Principal components analysis (PCA), in a broader sense one of EFA techniques for factor extraction, is the mostly used one. Reviews of its usage in psychological journals (Conway & Huffcutt, 2003; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Ford, MacCallum, & Tait, 1986) show that the popularity of EFA and PCA, in particular, still holds. After the misconceptions that exploratory is subordinated to confirmatory analysis have been rejected (for example Tukey, 1980; Velicer & Jackson, 1990), the main critique is formed around insufficient preciseness and objectiveness as results

Corresponding author: azoric@gmail.com

are highly influenced by the researcher's choices. The one of the key questions in EFA is how many factors/components are needed in order to successfully represent the data in the space with smaller number of dimensions. Some authors call these rules determining the number of components *stopping rules.* The process for factor extraction could be seen as iterative procedure of components extraction, as originally proposed by Hotelling (see in Mulaik, 1971), which is stopped when the optimal number of factors is extracted. As there are many of these stopping rules it is usually up to researcher to select the one which result is closest to researcher's theoretical model and expectations based on it, and as well as interpretability of that factor solution. Determination of correct i.e. appropriate number of factors has strong theoretical implications in psychology, especially in the fields of personality and cognitive abilities.

In the classical factor analysis true scores are postulated as uncorrelated to unique scores. Moreover, the unique scores are not inter-correlated among themselves. In other words, true scores are responsible for correlations between variables, while unique scores are only responsible for the explanation of variance of variables which remained after the partialization of the true scores. These classical postulates do not lead to unique solution of scores, implicating that true and unique scores can be only estimated.

In order to make the model solvable in the terms of linear algebra Guttman in his image theory relaxed the classical postulates, allowing: a) non-zero correlations among true scores of one variable (*image* scores) and unique scores (*anti image*) of the other measured variables, and b) non-zero correlations among unique scores of different variables (Guttman, 1953). By this relaxation the model leads to solvable unique solution of two parts of each variable. The less desirable consequence of the model was that all random error variables correlate in finite samples.

METHODS FOR DETECTION OF NONTRIVIAL COMPONENTS

In more than hundred years long history of PCA used as EFA technique, a long list of different stopping rules was suggested. Only the most popular rules will be presented here in chronological order.

Bartlett's χ^2 *test*

One of the first presented stopping rules, from 1950, was the statistical test developed by Bartlett (Bartlett, 1950). The idea was to detect when the variance of remaining components do not differ statistically. After all true components are extracted, only error components remain. These error components have variances that only fluctuate by chance. To detect this situation, Bartlett basically computed the ratio of geometric and arithmetic mean of variances i.e. eigenvalues of the remaining components. If all eigenvalues are the same, these means are equal and the ratio has value that is close to one, while its logarithm is close to zero (Horn & Engstrom, 1979). If at least one[1] eigenvalue would be substantially

1 As usual eigenvalues are sorted in descending order, where the first one is the most important.

greater, that would make greater impact on the arithmetic mean than on the geometric mean, and value of logarithm will jump up. Multiplication of this logarithm with a constant defined by the number of observations, variables and the number of already significant components would result with χ^2 distributed statistic.

By using notation presented in Mulaik (1971), and if we designate with m the number of directly measured variables, with n the sample size, with λ_i the eigenvalues obtained from correlation matrix of m measured variables, sorted in non-ascending order and with p ($p = 0, ..., m$) the number of components already declared as significantly different from the others then:

$$Q = \frac{\prod_{i=p+1}^{m} \lambda_i}{\left(\sum_{i=p+1}^{m} \frac{\lambda_i}{m-p}\right)^{(m-p)}}, \quad K = n - 1 - \frac{(2m+5)}{6} - \frac{2p}{3} \quad \text{and} \quad \chi^2 = -K \log(Q)$$

Resulting χ^2 statistic will have $(m - p - 1)(m - p + 2)/2$ degrees of freedom. In the first iteration ($p = 0$) tested hypothesis is in fact that all population eigenvalues are equal. After rejecting this hypothesis one would then, in the next iteration, test the remaining $m - 1$ eigenvalues. The first test with non-significant outcome would mean that all of the remaining eigenvalues are just error variances, and that the optimal number of components has been detected. According to some authors, Bartlett's χ^2 test was often found to overestimate the number of nontrivial components, especially with conventional significance levels at either 0.05 or 0.01 (Gorsuch, 1973; Horn & Engstrom, 1979; Hubbard & Allen, 1987). Beside, Gorsuch (1973) noticed that this trend increases with the sample size, making the smaller differences become significant in cases of larger samples. It is important to notice that this test is possible to detect structures with no factors at all, when the test is performed for $p = 0$.

Guttman-Kaiser's rule

Describing the conditions necessary for common-factor analysis, Guttman noted that after the removal of unique variances, in a way that resulting matrix is still Gramian, its minimal rank must be the number of eigenvalues greater than one. At the same time, this number is also the minimum number of common factors which should be postulated as truly existing ones (Guttman, 1954). Later, Kaiser noted that in order to have a positive reliability a component must have an eigenvalue greater than one. Another interpretation of this criterion could be that there is no much sense to declare something as component if it carries less information then the original (standardized) variable. And this conclusion is something that almost everyone will agree upon, but the other way around – proclaiming all components with eigenvalues greater than one to be important, is questionable. It is well known that this rule overestimates the number of components (see for example Lorenzo-Seva, Timmerman, & Kieres, 2011), which is probably a consequence of treating the correlation matrix obtained on the sample as a population parameter and not as its estimate. Nevertheless this rule is maybe the mostly used one (Fabrigar et al., 1999; Conway & Huffcutt, 2003).

Horn's Parallel Analysis Test

This test is based on the comparison of eigenvalues obtained from the analyzed matrix and eigenvalues from randomly generated data. All the eigenvalues that are greater from corresponding eigenvalues obtained on random data are designated as important and retained (Horn, 1965). In order to apply this criterion eigenvalues, from the correlation matrix obtained from random data matrix with the same size as the original data matrix, should be calculated. In original paper Horn is simulating just one random matrix, but trough the time method evolved into calculation of larger set of random matrices and complementary eigenvalues. From these eigenvalue distributions for each *k*-th eigenvalue, reference value is calculated for comparison with the tested eigenvalue.

Finding of Zwick & Velicer (1986) that PA overestimates the number of components actualized the question of choice of the critical value obtained on random data.

In order to make criteria more conservative Buja & Eyuboglu (1992) suggested increase of the threshold to 95th percentile. Press-Neto et al. (2005) found that in some cases originally proposed limit has better performances, but popularity of 95th percentile continued (Lorenzo-Seva et al., 2011). On the end some findings (Raîche, Walls, Magis, Riopel, & Blais, 2013) suggested that it is not the increase but the decrease of original threshold, to 5th percentile, that is improving performance of this criterion.

Cattell's Scree Test

This test is based on the graph i.e. scree plot, of eigenvalues on the ordinate and their ordinal values on the abscissa (Cattell, 1966). The idea is that the last eigenvalues, that will be discarded, are just fluctuation of error variance, and therefore make the linear trend. This trend in a similar fashion to Horn's algorithm can be used to detect the first eigenvalue that is above this threshold line. The major critique of this test is on its subjectivity, taking into the account that elbow where curve diverges from the linear trend is not something that is always easily detectable (Horn & Engstrom, 1979). Horn and Engstrom (1979) noted the resemblance of Bartlett's testing, that all remaining eigenvalues are the same, and Cattell's search of the point that diverges from the linear trend formed from error eigenvalues, concluding that these two test are based on the similar idea. In the same time this is also underlying idea of Horn's criteria. It is just that Horn is making the reference values from random data (which do make linear trend also), or by modeling the data just with the error variance, while Cattell is constructing this trend from the error variances that can be found in the data itself.

Velicer's Minimum Average Partial Test

Inter-correlation matrix could be decomposed like,

$$R = \sum_{i=1}^{m} x_i x^t \lambda_i$$

where λ_i is eigenvalue and $\boldsymbol{x}_i$ is corresponding eigenvector of inter-correlation matrix $\boldsymbol{R}$. In that case estimation of this matrix $\boldsymbol{R}_p$ based on the last $m - p$, ($p = 0, ..., m$) eigenvalues and eigenvectors is

$$R_p = D_p^{-1} C_p D_p^{-1}$$

where

$$C_p = \sum_{i=p}^{m} x_i x^t \lambda_i$$

and

$$D_p = diag\left(C_p\right)^{1/2}$$

In other words $\boldsymbol{R}_p$ is matrix of partial correlations among variables when the effect of the first p components is partialized. In this notation Velicer's criteria can be formulated as the number p for which average of squared off diagonal elements of matrix $\boldsymbol{R}_p$ is minimal (Velicer, 1976). Rationale for this is that after all true components are partialized, inter-correlations between variables would be explained, and resulting matrix of partial correlations, which will in that case represent just inter-correlations of error variables, will be in statistical sense the identity matrix. Continuing to partialize remaining, error components, would only increase partial correlations, as these off diagonal values are now the sum of smaller number of error/random components. As Bartlett's test, this test is also capable to detect structures with no important factors.

CHull method

In the essence, this method comes like numerical operationalization of the Cattell's scree test. Originally it was suggested for the detection of the optimal number of components in the three way data matrices (Ceulemans & Kiers, 2006), but lately it has been popularized for the usage in common-factor analysis (Lorenzo-Seva et al., 2011) and in principal components analysis (Wilderjans, Ceulemans, & Meers, 2013). Basic idea is the same as in Cattell's scree, to identify a point on the graph where the curve makes the elbow.

The graph in general represents the relation of the number of free parameters (fp_i) in evaluated model and some measure of the goodness of fit (f_i) for that model. In that case the algorithm could be described trough following steps:

- sort the points by the free parameters value (fp_i)
- exclude all (fp_i, f_i) points where $fp_j < fp_i$ and in the same time $f_{j>} f_{i,}$ i.e. where goodness of fit measure (f_i) is not in the same order (ascending/descending) as the number of free parameters (fp_i)
- check all triplets of adjacent points and exclude all middle points that are located on or below the line that is connecting its neighboring points i.e. if

$$f_i \leq f_{i-1} + (fp_i - fp_{i-1}) \frac{f_{i+1} - f_{i-1}}{fp_{i+1} - fp_{i-1}} \quad *$$

– find the point (fp_i, f_i) where the function

$$\left(\frac{f_i - f_{i-1}}{fp_i - fp_{i-1}}\right) / \left(\frac{f_{i+1} - f_i}{fp_{i+1} - fp}\right)$$

reaches the maximum. At that point the elbow of the curve is located and the optimal model is detected.

The problem with this criterion emerges in the cases of the first and the last eigenvalue, with former naturally being of much greater importance. As the function is not defined on endpoints, it makes impossible that these points are going to be designated as optimal. In common-factor analysis the null model can be used for this purpose (Lorenzo-Seva et al., 2011), as it is simpler it allows evaluation of model with one common factor. In case of PCA as Wilderjans et al. (2013) suggested that the fit function could be the proportion of explained variance by extracted components i.e. the cumulative sum of eigenvalues. By this application of CHull, the formulation of the criterion can be simplified to the search for the maximum of ratios of succeeding eigenvalues, from second eigenvalue onward.

In order to make solution with one component possible the same authors proposed that the null value should be defined as simple zero point, representing the case when none of components is extracted causing that none of the total variance is explained. Authors also made the notion that this is not the ideal solution as the addition of this "virtual" zero point overestimates the function for the first component. Rationale for this could be that the error variance has smaller impact on the size of the first eigenvalue then on the second. The same logic should hold for all succeeding pairs of eigenvalues, which would on the end result with tendency of this criterion to underestimate the number of true components.

We would like to note that this criterion, in this formulation, is similar to Rnd-Ratio criterion (Peres-Neto et al., 2005), just that in their version the ratio was bootstrapped and its significance was estimated.

AIM OF THE STUDY

Aim of this paper is to explore the impact of different research conditions on accuracy of five rules for component retention. The criteria that have been compared are: Guttman-Kaiser's rule of eigenvalue over one, Bartlett's test, Horn's parallel analysis, Velicer's MAP and CHull method.

---

* Even that usually convex function is defined as the function where for any $t \in [0, 1]$ condition $f(t\,x1 + (1\text{-}t)x2) \leq t\,f(x1) + (1\text{-}t)\,f(x2)$ holds true, convexity here is defined like in cited paper

The goal was to compare accuracy of those five rules for different number of postulated components, different sample sizes and number of variables, and also different proportions of error variance.

SIMULATION

In order to evaluate different stopping rules simulated data were generated with predefined structures. First, structure matrix is defined for the given number of important components (*k*) and the given number of variables (*m*) only allowing the structures where each factor has high loadings with at least three variables. The actual number of high loadings of variables per factor was defined by random integer from 3 to *m*/*k,* with the exception for the last factor where high loadings were designated for all remaining variables. For example, if the number of variables was 12, and the specified number of factors 3, the distribution of variables among factors could be 4,4,4 as well as 3,3,6.

All coefficients in structure matrix (**F**) were filled with random numbers uniformly distributed from 0 to 0.2, and after that, on all coefficients for all variables with high loadings, uniformly distributed random number from 0 to 0.7 was added.

After this the matrix of normally distributed random numbers with *k* columns and *n* rows (Z_r) that represented the matrix of factor scores, was multiplied with the structure matrix in order to obtain the matrix of true scores of variables.

$$Z_t = Z_r F^t$$

These true scores were standardized to have the variances of (1-*e*) where *e* is the specified proportion of the error in the data. After this the matrix of error scores (Z_e) was summed with this true scores matrix in order to obtain the data matrix (Z) used for evaluation of different criteria. This error *n* by *m* matrix (Z_e) was also filled with random numbers from normal distribution and columns scores were standardized to have the variance of *e*.

$$Z = Z_t + Z_e$$

This method for error integration into data is not unknown (Josse & Husson, 2012), but we have to note that it is not popular as the Tucker's method with mayor and minor components and unique error variances (Tucker, Koopman, & Linn, 1969).

After construction of the data matrix the correlation matrix and its eigenvalues were calculated and selected stopping criteria were applied.

The simulation was done in R software version 2.15.2 (R Core Team, 2013) and the source code of the procedure can be found on www.kal.rs/simulation.

The choice of the number of variables, sample size and percent of error variance was based on a few meta-analytic studies of EFA practices in psychology and social sciences (Fabrigar et al., 1999; Conway & Huffcutt, 2003; Cangelosi & Goriely, 2007; Costello & Osborne, 2005).

Tested number of postulated factors (*k*) was limited to values of 1, 2, 3, 5, 8 and 10, and from these values and the rule of at least three variables with high loading per factor, number of variables (*m*) was derived to be: 9, 15, 22, 35 and 40. Following the same rule, some numbers of components became inapplicable for the some number of variables. For example, for numbers of 35 and 40 variables all combinations of number of components were tested, but for 15 variables only situations with 1, 2, 3 and 5 components were applicable. This resulted with total of 23 different combinations.

Sample size (*n*) was limited to values of 50, 100, 200, 300 and 600, where 50 and 100 were representing fairly small sample sizes and 600 the large ones. These small sample sizes, especially in situations with large number of variables 35 and 40 are not conditions that are commonly advisable for application of factor analysis, but in clinical psychology, for example, as well as in other scientific fields, these sample sizes are not uncommon, so these situations were as well included in simulation.

The percentage of error variance (*err*) was limited to 30, 40, 50, 60, 70 and 80.

Inclusion of error variance in the model makes the model of the data much more realistic, as some amount of error variance is inevitably part of the measurement. In similar Monte Carlo studies, the amount of the error was almost never directly systematically controlled as a factor that is producing the impact on the size of eigenvalues. Underlining premise based on classical measurement model was that correlation between error and true scores should be zero, as well as correlations between errors of different variables. Guttmann (1953) demonstrated that the second hypothesis is true only if measurement was done on the universe of variables, but because of proposed method for decomposition of true and error variances, he was forced to keep the first premise. We can easily agree that correlation between true and error scores should be zero on infinite sample, but sample sizes in psychological researches can hardly be approximated with this model. Cumulatively, those statistically insignificant correlations could build up the size of eigenvalues. After all, on this assumption Horn's parallel criterion is based.

By the suggested model for integration of error variance in the data that was applied in this paper, correlations among all true and error scores as well as inter-correlations among error scores were allowed.

In earlier papers (Jackson, 1993; Peres-Neto et al., 2005; Zwick & Velicer, 1986; Zwick & Velicer, 1982) error factors were introduced by minor loadings or by addition of unique variance on diagonal of reproduced correlation matrix in order to achieve standardized variance. Much more realistic approach was suggested by Tucker et al. (1969). Their model consist of three parts: common-factors (major factors), minor factors and unique factors, where minor and

unique factors represent the error in the model. But that model didn't allowed correlations between major (true) components and error components. The model suggested by Hong (1999) resolved that problem which existed in all preceding simulations (Jackson, 1993; Peres-Neto et al., 2005; Zwick & Velicer, 1986; Zwick & Velicer, 1982) allowing the modeling of these inter-correlations. Hong's algorithm as not being based on score matrices (Z_r, Z_t and Z_e), is insensitive on the sample size.

So for 23 different plausible combinations of number of variables to number of components, all five sample sizes and all six error levels were tested, which resulted in 690 different combinations. For each combination process of data generation was repeated 200 times. On the end 138,000 different data matrices were analyzed.

Bartlett's χ^2 test was applied with two standard significance levels 0.05 (**BAR5**) and 0.01 (**BAR1**). In addition test was applied in two similar versions. The first, which is something like the standard interpretation of the test, stops the extraction after the first non-significant component is detected. The second version aims to find the last significant component, not taking into account if there were one or more non significant outcomes before. These two versions produced almost the same results, even the later was something better performing, and only its results are presented in the paper.

Standard Guttman-Kaiser rule (**GK**) of eigenvalue larger than one was the third option tested.

In the case of the parallel analysis comparison of different critical values was performed. The tested values were 50th (**PA50**) and 95th (**PA95**) percentiles of eigenvalues obtained from random data. The 5th percentile (Raîche et al, 2013) was also tested, but as its performance, in overall comparison, was worse than of PA50 its results are not presented.

Minimum average partial correlation (**MAP**) test was also included in this comparison.

On the end two versions of convex hull criterion were applied, one version (**CHull.CFI**) on a problem of number of factors in common-factor analysis where the value of comparative fit index (Bentler, 1990) of a model is plotted against its degrees of freedom (Lorenzo-Seva et al., 2011), and the other suggested for PCA (Wilderjans et al., 2013) as a plot of cumulative values of eigenvalues and their ordinal number in situation with (**CHull.PCA0**) and without (**CHull. PCA1)** inclusion of zero point. On both methods that were based on PCA upper limit of possible solutions was limited to $m/2$ i.e. half of the number of variables analyzed.

Classical eigenvalue decomposition was performed in all cases of all criteria, except for CHull based on CFI where ML common-factor model was used. In this case appropriate procedure from R package "psych" was applied (Rawelle, 2013).

RESULTS AND DISCUSSION

As we can see from Table 1 and 2 and Figure 1 the obtained results are mostly in line with the findings in literature. As the overall comparison is influenced by the selected sets of parameters, and as there is difference among criteria under different parameter values, the straightforward estimation of criteria performances is not possible.

Table 1. *Percent of correct detections of number of components by different criteria*

		BAR5	**BAR1**	**GK**	**PA 50**	**PA 95**	**MAP**	**CHull. CFI**	**CHull. PCA0**	**CHull. PCA1**	Count in 000
n	50	50.5	48.4	31.6	55.2	51.3	57.4	43.2	45.7	42.7	27.6
	100	68.0	64.0	47.4	69.3	63.3	67.9	52.9	52.1	53.1	27.6
	200	80.6	78.3	62.6	79.9	75.4	70.5	64.4	57.6	60.9	27.6
	300	86.1	84.7	69.7	85.8	82.2	71.2	71.1	61.2	64.8	27.6
	600	93.0	92.6	81.5	93.3	91.2	71.5	80.6	65.5	69.1	27.6
m	9	87.2	85.8	85.7	94.2	90.1	67.8	91.4	86.2	61.5	18
	15	82.4	80.4	73.0	87.6	82.6	67.2	76.5	68.9	64.5	24
	22	82.4	80.0	59.8	87.1	82.5	74.3	67.3	63.3	62.2	24
	35	68.5	66.0	48.6	64.6	61.0	64.1	49.3	42.5	53.7	36
	40	68.0	66.3	44.6	65.8	62.5	67.1	48.4	42.5	53.9	36
k	1	96.0	98.5	57.1	98.0	99.9	98.1	99.8	100.0	0.0	30
	2	87.7	86.9	59.6	94.8	93.7	80.4	70.3	68.7	97.7	30
	3	78.3	75.2	62.4	87.7	81.3	64.8	62.4	57.3	81.0	30
	5	63.6	58.9	60.2	65.5	54.7	48.9	45.1	33.6	67.3	24
	8	47.7	42.1	51.9	33.2	26.6	41.0	24.8	12.6	46.7	12
	10	40.0	34.9	53.6	16.7	12.4	31.3	21.5	3.8	40.5	12
err	30	93.7	94.9	97.4	86.5	83.9	97.5	85.7	83.3	75.8	23
	40	90.1	90.1	89.3	84.3	81.3	94.4	77.6	70.6	71.6	23
	50	85.0	83.3	75.5	81.1	77.4	79.0	67.9	59.4	65.3	23
	60	75.9	72.9	52.9	76.7	72.6	56.1	57.2	49.7	56.2	23
	70	63.2	59.0	27.0	71.1	65.6	44.8	47.9	41.4	45.6	23
	80	45.9	41.3	9.2	60.5	55.2	34.3	38.2	34.1	34.1	23
n/m	<2	39.0	39.8	17.6	41.1	40.6	50.1	28.5	32.4	35.6	14.4
	2 – 4.9	62.4	57.5	38.3	63.1	57.2	67.4	45.2	46.7	49.9	28.8
	5 – 9.9	77.4	74.6	57.6	76.0	70.6	70.1	59.9	52.3	59.5	42
	10 – 19.9	88.5	87.3	73.5	90.7	86.8	71.2	75.7	62.2	66.6	27.6
	> 20	94.7	94.7	90.4	98.5	96.7	70.2	91.1	81.6	68.8	25.2
total		75.6	73.6	58.6	76.7	72.7	67.7	62.4	56.4	58.1	138

Table 2. *Average deviation from correct number of components in situations when criteria produced incorrect number of components by different criteria*

		BAR5	BAR1	GK	PA 50	PA 95	MAP	CHull. CFI	CHull. PCA0	CHull. PCA1
n	50	-1.00	-2.44	4.51	-1.82	-3.17	-2.74	-3.69	-4.11	0.71
	100	-2.18	-2.66	4.48	-2.37	-2.87	-3.33	-4.24	-4.43	0.27
	200	-1.90	-2.34	4.27	-2.06	-2.42	-3.57	-4.48	-4.65	0.28
	300	-1.60	-2.09	3.92	-1.87	-2.17	-3.68	-4.64	-4.78	0.56
	600	-0.85	-1.59	3.25	-1.57	-1.78	-3.71	-4.80	-4.88	0.91
m	9	-0.92	-1.51	1.26	-0.19	-1.26	-2.05	-0.67	-1.41	1.57
	15	-1.39	-1.97	1.93	-0.84	-1.75	-2.70	-1.58	-2.51	1.79
	22	-1.20	-1.78	3.10	-0.89	-1.58	-2.09	-2.29	-2.58	2.02
	35	-2.19	-2.96	4.62	-2.48	-3.27	-4.45	-5.21	-5.50	-0.52
	40	-1.12	-2.43	5.63	-2.21	-3.04	-3.88	-5.21	-5.48	-0.28
k	1	3.83	1.56	4.84	1.55	0.90	1.03	1.87	1.75	3.99
	2	0.52	-0.79	4.65	0.30	-0.98	-0.93	-0.88	-0.99	3.40
	3	-0.55	-1.31	4.45	-0.45	-1.19	-1.69	-1.68	-1.86	-0.60
	5	-1.46	-2.10	4.03	-1.21	-1.81	-2.90	-3.60	-3.91	-1.84
	8	-2.45	-3.12	3.82	-2.41	-3.17	-4.70	-6.88	-6.97	-4.69
	10	-3.55	-4.20	2.53	-3.47	-4.47	-7.00	-8.86	-8.98	-6.04
err	30	2.45	-0.51	-0.40	-2.20	-2.35	0.42	-4.72	-6.09	3.15
	40	0.75	-1.12	1.68	-2.28	-2.46	-1.27	-4.68	-5.51	2.09
	50	-0.28	-1.66	2.71	-2.33	-2.54	-3.22	-4.69	-4.94	0.98
	60	-1.10	-2.04	3.34	-2.31	-2.67	-3.25	-4.52	-4.48	0.02
	70	-1.90	-2.52	4.21	-2.19	-2.85	-3.47	-4.17	-4.13	-0.44
	80	-2.64	-3.07	5.63	-1.33	-3.04	-3.67	-3.60	-3.81	-0.18
n/m	<2	-0.71	-2.72	5.60	-2.39	-4.00	-3.17	-4.85	-5.16	0.18
	2 – 4.9	-2.10	-2.60	4.36	-2.01	-2.72	-3.18	-3.80	-4.16	0.39
	5 – 9.9	-1.92	-2.30	4.20	-1.85	-2.29	-3.93	-4.59	-4.92	0.00
	10 – 19.9	-1.07	-1.64	2.91	-1.38	-1.65	-3.68	-4.18	-4.63	0.83
	> 20	-0.14	-1.18	1.49	-1.00	-1.25	-2.39	-2.00	-2.46	2.05
total		-1.51	-2.40	4.26	-2.00	-2.73	-3.35	-4.24	-4.53	0.54

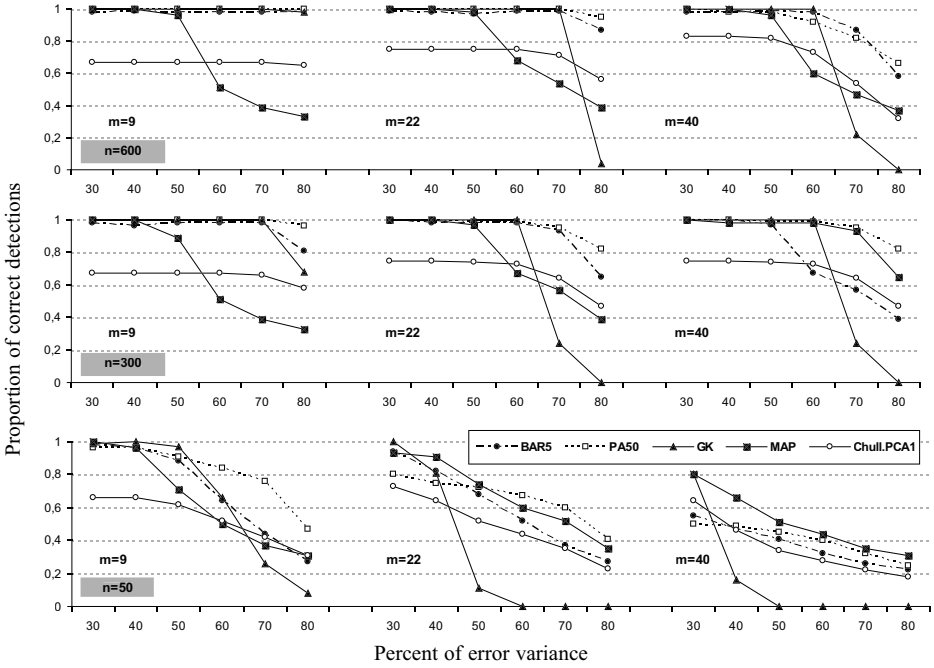

Figure 1. *Proportions of correct detections for selected criteria, for three different sample sizes (n), and three different numbers of variables (m), and sixe tested levels of error variances (err)*

Introduction of systematic variation of amount of error variance also provided useful information. The first conclusion is that all criteria, except HULL PCA, work properly in favorable circumstances like high number of subjects per variable and low level of the error in a system. Inaccuracy of all criteria starts to be evident in situations with small *n* by *m* ratio, high percent of error variance and high number of supposed factors. But some differences between them also exist.

Results of the accuracy of the Bartlett's χ^2 test are not in concordance with authors who have found that this test overestimates the number of factors especially on large samples (Gorsuch, 1973; Horn & Engstrom, 1979; Hubbard & Allen, 1987; Henson & Roberts, 2006; Raîche et al., 2013). Velicer, Eaton, & Fava (2000) even do not recommend its usage. Our results are much more in line with Ferré, (1995) who have found that accuracy of Bartlett test increase with sample size.

Results in Table 1 suggest that percent of correct detections of number of components for this criterion is the same as the one obtained by using PA. In the same time, in situations with wrong estimations, average error for this criterion is smaller than average error acquired by PA.

Our results imply that this method should be considered as the method of choice except in stations with extremely small samples (see Figure 1). This disagreement between our results and findings of the majority of other authors could be the consequence of the fact that mentioned authors did not used the original Barttlett's formula (Barttlett, 1950) but instead some of its modifications (see Peres-Neto et al., 2005).

A lot of authors reported that Guttman-Kaiser's criterion regularly overestimates the number of components (see for example Lorenzo-Seva at al., 2011; Costello & Osborne, 2005; Josse & Husson, 2012; Raîche et al., 2013; Wilderjans et al., 2013). Despite this it is the mostly used rule in researches in social sciences. The reason for this probably lies in the fact that this criterion is the default criterion in most statistical packages. In our simulation this was the only criterion that was overestimating, while all the others were underestimating the number of postulated factors. But the evaluation of this criterion is not that straightforward as it is accurate as much as other criteria in situations when *n* by *m* ratio is over 10 and when error variance is not above 50 percent. In other, non favorable, situations this criterion is the most sensitive one and very quickly loses its accuracy. When EFA is performed on test items, that often have very low reliability, sometimes even lower than .3, the GK criterion should be avoided.

Horn's Parallel Test was in many studies considered as the most accurate (Franklin et al., 1995; Hayton, Allen, & Scarpello, 2004: Zwick & Velicer, 1986; Ledesma & Valero-Mora, 2007; Velicer et al., 2000). That was a good argument for its recommendation despite pretty inconvenient procedure for its application. In order to overcome alleged tendency of this criterion to overestimate number of components some authors (Zwick & Velicer, 1986; Buja & Eyuboglu, 1992; Lorenzo-Seva et al., 2011) suggested increase of the cutoff value to 95th percentile of eigenvalue distribution. Other authors (Press-Neto et al., 2005) have found that usage of average eigenvalues obtained on random matrices could be more appropriate than usage of 95th percentile. Our findings are in line with this recommendation, as in our simulations in overall comparison median value outperformed the 95th percentile.

There are some new papers that suggest lowering of threshold value to 5th percentile (Raîche et al, 2013). In our study this low threshold level was performing better only in the situations with large sample sizes and low error levels, which can be also described as situations which make high correlations more possible. This finding is in line with Press-Neto et al. (2005). In overall comparison this threshold value under performs PA50. Our findings suggest that PA50 criterion is the most accurate one when the number of subject is small, and the percent of error variance is reasonably high but the number of variables is rather small.

In situations with small sample size, large error variance and large number of variables the best criterion is Velicer's MAP. Practically logic of MAP doesn't include sample size but only number of variables and number of factors, so this result should not be unexpected. Estimation of average partial correlation, as a parameter that should be minimized, is much more stable if a system has more

variables. So in cases of almost square data matrices the MAP should be the preferred criteria.

CHull criterion, except in situations with the only one component, in overall comparison outperforms the other criteria. This criterion in the case of PCA can be simplified as location of the maximum of ratios between all pairs of consecutive eigenvalues (Wilderjans et al., 2013). And in that case at least theoretically it can correctly detect existence of only one component but in reality it is not the case. Inclusion of zero point underestimates the number of components making a strong affinity toward solution with only one important component. This could explain why on the first glance it looks like much worse than other criteria. In situations with more than one component, standard variant of CHull criterion based on cumulative eigenvalues (without zero point) is one of the most accurate. CHull based on CFI suggested by Lorenzo-Seva et al. (2011) wasn't as accurate as authors claimed in their paper, but it should be noted that this version of CHull does not have a problem with detection of solutions with just one factor.

CONCLUSIONS

Our results partially confirm the results of earlier studies. Accuracy of all criteria decreases with decrease of sample size, and with increase of: number of variables, number of supposed factors and proportion of error variance. There is no unambiguous answer which of analyzed criteria has the best performance. In favorable research conditions all criteria have good accuracy. We can recommend usage of Bartlett's χ^2 test and Horn Parallel Analysis test working properly in all conditions except in situations with small number of subjects and relatively high number of variables. In these situations, with small sample size and relatively high number of variables, we can recommend Velicer's MAP criterion. Mostly used criterion Guttman-Kaiser's is the most sensitive on increase of number of variables and increase of proportion of error variance.

Those practitioners that would like to apply recommended criteria (PA, MAP) can find the macros in SPSS and SAS in the article of O'Connor (2000).

By our knowledge this paper is the first one that systematically varied error variance. Besides, non-zero correlations between error scores of different variables, as well as non-zero correlations of error score and true scores were allowed. This fact makes our simulations much more realistic. Results strongly confirm our hypothesis that reliability of analyzed measures (proportion of error variance) has crucial role in determination of number of components. This effect could be neutralized only by increasing the sample size.

Main limitation of this paper emerges from the limited set of values that were varied trough simulations. This limitation is predominantly related to the number of variables, and components. So, we are not sure that trends observed in our analyses could be generalized on those situations with more than 100 or 200 variables, and situations with more than 10 factors.

Second limitation is related to the shape of distribution. We simulated data only from normal distribution, but it is not rare in psychological researches that authors applied EFA without proving that the distributions are normal.

It should be noted that even that we used one criterion that is applicable for common-factor model, our findings are strongly related just to situations that PCA is used for factor extraction.

REFERENCES

Bartlett, B. M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, *3*(2), 77-85.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246. pmid:2320703. doi:10.1037/0033-2909.107.2.238

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, *27*, 509-540. doi:10.1207/s15327906mbr2704_2

Cangelosi, R., & Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct*, *2*(2). doi:10.1186/1745-6150-2-2

Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245-276.

Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: a numerical convex hull based method. *British journal of mathematical and statistical psychology*, *59*(Pt 1), 133-150. pmid:16709283

Conway, J. M., & Huffcutt, A. I. (2003). A Review and Evaluation of Exploratory Factor Analysis Practices in Organizational Research. *Organizational Research Methods*, *6*(2), 147-168. doi:10.1177/1094428103251541

Costello, A. B., & Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1–9.

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*(3), 272-299. doi:10.1037/1082-989X.4.3.272

Ferré, L. (1995). Selection of components in principal component analysis: A comparison of methods. *Computational Statistics & Data Analysis*, *19*, 669-682.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychology*, *39*, 291-314.

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., Fralish, J. S., Scott, B., et al. (1995). Parallel Analysis: a Method for Determining Significant Principal Components significant principal components. *Journal of Vegetation Science*, *6*, 99-106.

Gorsuch, R. L. (1973). Using Bartlett's Significance Test to Determine the Number of Factors to Extract. *Educational and Psychological Measurement*, *33*(2), 361–364. doi:10.1177/001316447303300216

Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, *18*(4), 277–296. doi:10.1007/BF02289264

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149–161. doi:10.1007/BF02289162

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis. *Organizational Research Methods*, *7*(2), 191–205. doi:10.1177/1094428104263675

Henson, R., & Roberts, J. (2006). Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement*, *66*(3), 393–416. doi:10.1177/0013164405282485

Hubbard, R., & Allen, S. J. (1987). An empirical comparison of alternative methods for principal component extraction. *Journal of Business Research*, *15*(2), 173–190. doi:10.1016/0148-2963(84)90047-X

Hong, S. (1999). Generating correlation matrices with model error for simulation studies in factor analysis: A combination of the Tucker-Koopman-Linn model and Wijsman's algorithm. *Behavior Research Methods, Instruments, & Computers*, *31*(4), Retrieved from http://link.springer.com/article/10.3758/BF03200754

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185. pmid:14306381. doi:10.1007/BF02289447

Horn, J. L., & Engstrom, R. (1979). Cattell's scree test in relation to Bartlett's chi-square test and other observations on the number of factors problem. *Multivariate Behavioral Research*, *14*(3), 283–300. doi:10.1207/s15327906mbr1403_1

Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, *74*(3), 2204–2214. Retrieved from http://www.jstor.org/stable/10.2307/1939574

Josse, J., & Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, *56*(6), 1869–1879. doi:10.1016/j.csda.2011.11.012

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the Number of Factors to Retain in EFA: an easy-to– use computer program for carrying out Parallel Analysis. *Practical Assessment, Research & Evaluation*, *12*(2).

Lorenzo-Seva, U., Timmerman, M. E., & Kieres, H. A. L. (2011). Multivariate Behavioral The Hull Method for Selecting the Number of Common Factors. *Multivariate Behavioral Research*, *46*, 340–364. doi:10.1080/00273171.2011.564527

Mulaik, S. (1971). *The Foundations of Factor Analysis*. New York: McGraw-Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers*, *32*, 396-402.

Peres-Neto, P. R., Jackson, D. a., & Somers, K. M. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, *49*(4), 974–997. doi:10.1016/j.csda.2004.06.015

R Core Team, R. (2013). *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing* (Vol. 1, p. 409). R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J. G. (2013). Non-Graphical Solutions for Cattell's Scree Test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(1), 23–29. doi:10.1027/1614-2241/a000051

Rawelle, W. (2013). psych: Procedures for Personality and Psychological Research. Evanston, Illinois, USA: Northwestern University. Version 1.3.2. Retrieved from http://cran.r-project.org/package=psych

Tucker, L., Koopman, R., & Linn, R. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(4), 421–459. Retrieved from http://link.springer.com/article/10.1007/BF02290601

Tukey, J. W. (1980). We need both exploratory and confirmatory. *American Statistician*, *34*(1), 23-Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321–327.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*(1), 1-28.

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. (pp. 1-31).

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). CHull: a generic convex-hull-based model selection method. *Behavior research methods*, *45*(1), 1–15. doi:10.3758/s13428-012-0238-5

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research, 17*(2), 253–269.

Zwick, W., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442. doi:10.1037/0033-2909.99.3.432