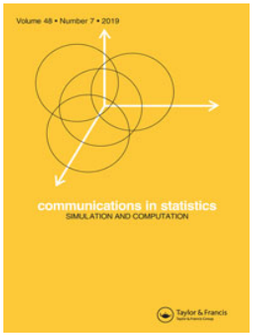

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Minimum average partial correlation and parallel analysis: The influence of oblique structures

P.-O. Caron

To cite this article: P.-O. Caron (2019) Minimum average partial correlation and parallel analysis: The influence of oblique structures, Communications in Statistics - Simulation and Computation, 48:7, 2110-2117, DOI: 10.1080/03610918.2018.1433843

To link to this article: https://doi.org/10.1080/03610918.2018.1433843

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
2019, VOL. 48, NO. 7, 2110–2117
https://doi.org/10.1080/03610918.2018.1433843

Minimum average partial correlation and parallel analysis: The influence of oblique structures

P.-O. Caron

Département des Sciences humaines, Lettres et Communications, Télé-Université, Montréal, Québec, Canada

ABSTRACT

Parallel analysis (Horn 1965) and the minimum average partial correlation (MAP; Velicer 1976) have been widely spread as optimal solutions to identify the correct number of axes in principal component analysis. Previous results showed, however, that they become inefficient when variables belonging to different components strongly correlate. Simulations are used to assess their power to detect the dimensionality of data sets with oblique structures. Overall, MAP had the best performances as it was more powerful and accurate than PA when the component structure was modestly oblique. However, both stopping rules performed poorly in the presence of highly oblique factors.

ARTICLE HISTORY
Received 17 October 2017
Accepted 18 January 2018

KEYWORDS
Monte Carlo simulations; minimum average partial correlation; parallel analysis; principal component analysis; stopping rules

MATHEMATICS SUBJECT CLASSIFICATION
62H25; 11K45

1. Introduction

Principal component analysis (PCA) is generally used to reduce the dimensionality of data sets, because it tends to summarise the meaningful information in the early axes whereas errors and noise remain in the later trivial ones. However, determining the correct number of components has remained a long-standing challenge. Over the last fifty years, thorough evaluations lead researchers, particularly in the biological and psychological literature, to adopt and spread the use of parallel analysis (PA; Horn 1965) and minimum average partial correlation (MAP; Velicer 1976). There is a growing consensus that both stopping rules are optimal solutions to identify the correct number of components (Peres-Neto, Jackson, and Somers 2005; Zwick and Velicer 1982, 1986; Velicer, Eaton, and Fava 2000) and their algorithms are readily implemented in SPSS, SAS, R and MATLAB (O'Connor 2000; Courtney 2013; Dinno 2009).

Despite the extensive effort to evaluate the two methods' performances, some issues have not been fully investigated. Until now, MAP and PA have been reported as accurate with artificial correlation matrices in which loadings were similar between components, axes were orthogonal to each other, all factors were at least modestly salient (loadings over 0.50), and in which components contained relatively the same number of variables (Zwick and Velicer 1986; Peres-Neto, Jackson, and Somers 2005; Garrido, Abad, and Ponsoda 2011; Velicer 1976; Zwick and Velicer 1982; Guadagnoli and Velicer 1988; Velicer, Eaton, and Fava 2000). In other words, they were often tested with components structures for which identifying the correct dimensionality is evident using almost any stopping rule.

CONTACT P.-O. Caron pocaron19@gmail.com Département des Sciences humaines, Lettres et Communications, Télé-Université, Montreal, Québec, Canada.

Even though MAP and PA seem adequate, evidence show that they are generally inaccurate with oblique structures (Beauducel 2001; Peres-Neto, Jackson, and Somers 2005; Garrido, Abad, and Ponsoda 2011). An oblique structure refers to a situation in which a correlation between some factors exists, that is, factors are not orthogonal to each other. They are also seemingly more natural and found in several real-life data sets. Previous studies have also shown that both stopping rules were prone to underestimate the number of components (except when unique variables exist), therefore missing some meaningful components. An explanation lies in that, as the component structure becomes more oblique, variance is attributed to the earliest axes. While the latter dimensions should remain meaningful, this decreases their eigenvalues, making them harder to detect.

Since MAP and PA are widely spread as optimal stopping rules despite having been systematically evaluated with oblique component structures, it is essential to verify the extent to which they remain accurate as the factors become correlated, especially since this would impact conclusions drawn from studies in the fields of psychology and psychometrics, but also biology and genetics. For instance, Shriner (2011, 2012) found that MAP is suitable for genome-wide genotype data, has smaller bias and smaller variance in estimating the number of components to retain, and is better than the Tracy-Widom distribution (Tracy and Widom 1993) developed form random matrix theory (Wigner 1955). Thus, the purpose of the present simulations is to evaluate the power of MAP and PA and their accuracy to identify the correct number of components in data sets containing oblique structures.

2. Method

We will first describe the two stopping rules. Thereafter, the Monte Carlo simulations used to investigate their ability to detect the correct number of components will be described.

2.1. Parallel analysis

PA was developed by Horn (Horn 1965) as an alternative to Kaiser's eigenvalues-greater-than-one rule (Kaiser 1960). The purpose of PA is to account for sampling error. Since PCA optimizes the variance in a given dimension, the first eigenvalues should catch some error variance. PA is the computation of the average eigenvalues over 100 to 1000 repetitions of multivariate random data for which the correlation matrices are an identity matrix. The first few eigenvalues higher than the average eigenvalues at their respective ranks are considered to contain meaningful information. Several variations of PA exists, but only the original (i.e., Horn 1965) will be consider herein.

2.2. Minimum average partial correlation

MAP was developed specifically for principal component analyses (Velicer 1976). It is computed by successively partialling out $\mathbf{A}_{\mathrm{m}}$, the component loading matrix that contains the first 1 to m components, from $\mathbf{R}$, the original correlation matrix, to obtain $\mathbf{C}_m$, the partial covariance matrix;

$$\mathrm{C}_m = \mathrm{R} - \mathrm{A}_m\mathrm{A}_m^T, \tag{1}$$

The partial correlation matrix, $\mathbf{R}_m{}^*$, is given by;

$$\mathrm{R}_m^* = \mathrm{D}^{-\frac{1}{2}}\mathrm{C}_m\mathrm{D}^{-\frac{1}{2}}, \tag{2}$$

where

$$\mathbf{D} = \text{diag}(\mathbf{C}_m). \tag{3}$$

The MAP criterion is given by the average square of the partial correlations in $\mathbf{R}_m{}^*$;

$$\text{MAP}_m = \sum_{i=1}^{p} \sum_{\substack{j=1 \\ i \neq j}}^{p} \frac{r_{ijm}^2}{p(p-1)}, \tag{4}$$

where p is the number of variables. This procedure ends when $p - 1$ components have been partialled out. To test the first component, Velicer (1976) suggested to compare MAP_1 to MAP_0 which is obtained by;

$$\text{MAP}_0 = \sum_{i=1}^{p} \sum_{\substack{j=1 \\ i \neq j}}^{p} \frac{r_{ij}^2}{p(p-1)}. \tag{5}$$

If $\text{MAP}_0 < \text{MAP}_1$, no component should be retained.

The number of components to retain, m, is the component where the average squared partial correlations reached is minimum, i.e., when m^{th} component $= \min(\text{MAP}_m)$. It is worthy to note that several variations of MAP exist (Velicer, Eaton, and Fava 2000; Garrido, Abad, and Ponsoda 2011). Only the original will be used throughout the current study (i.e.,Velicer 1976), since the differences between outcomes are negligible.

Although many studies suggest one or the other, MAP is considered more theoretically sound than PA (Velicer, Eaton, and Fava 2000). MAP involves assessing to what extent each successive $\mathbf{R}_m{}^*$ is similar to the identity matrix of size p, i.e., how likely it is that all meaningful correlations have been partialed out of $\mathbf{R}$. As such, MAP assesses the effect of removing successive eigenvalues, whereas PA only implements a null model. Therefore, PA is adequate to test the first eigenvalue but its validity is compromised for later ones (Turner 1998; Achim 2017).

2.3. Simulation method

A series of Monte Carlo simulations was carried out. Five correlation matrices were chosen to assess the influence of the component structure. Figure 1 provides a visual representation of the correlation matrices which was inspired by Peres-Neto et al. (2005). Each matrix is composed of nine descriptors and three components. Matrices 1, 2, and 3 contain three components expressed in three variables each. Matrices 4 and 5 contain three components expressed in four, three and two variables respectively. It is worthy to note that the number of descriptors was not varied, because increasing their number while maintaining the same number of components decreases the difficulty of finding the correct amount of components by cumulating more non-trivial signals in the earlier axes.

In order to generate oblique structures, correlations between variables belonging to different underlying components were added (0.10 and 0.30). Finally, to evaluate the influence of sample size, the sample size was systematically varied to some specified values, $n = 16, 32, 64, 128, 256$, and 512.

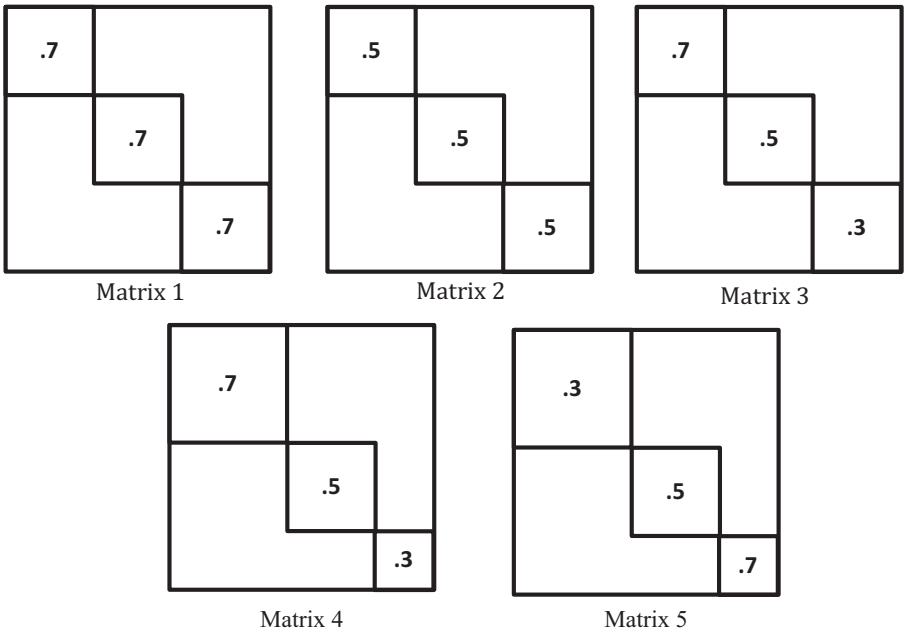

Figure 1. Visual representation correlation matrices of the population. The size of squares correspond to the proportion of variables per component (see the text for more information). Outside squares correspond to the level of correlations between components. Inspired by Peres-Neto et al. (2005).

2.4. Data generation

Simulations were carried out in Matlab (2012a). The steps were as follow: (1) generate an artificial data set from a multivariate normal distribution with a specified population correlation matrix and a specified sample size; (2) carry a PCA on the data sets; (3) apply PA and MAP and record the number of detected components; (4) compute the agreement or the disagreement between both stopping rules; (5) repeat steps 1, 2, 3 and 4 a total of 1000 times; (6) sum frequency with which the correct number of components was identified divided by the number of replications (power). Finally, (7) repeat previous steps for the five correlation matrices, the three levels of correlation between components and the six levels of sample size.

3. Results

An overall assessment of MAP and PA is presented in Table 1. The results show that MAP was the better of the two stopping rules. It retained the correct number of axes 62% of the time, across all simulated conditions. When PA was correct, MAP was also right 89% of the time. However, when MAP was correct, PA was also correct only 41% of the time. MAP was also more likely than PA to overestimate the number of components whereas PA has a stronger tendency to underestimate. This result is in agreement with past methodological literature showing that PA is more likely to underestimate the number of components with correlated components (Turner 1998; Beauducel 2001). Both methods will now be discussed independently.

Table 2 shows the performance of PA to determine the correct dimensionality of data sets. Regarding results from matrix 1, the method performed well when sample size was higher than 30 and when each component was salient (loadings over 0.70), regardless of the correlation between axes. However, for all other matrices, PA was barely able to detect the correct

Table 1. Overall performance of parallel analysis (PA) and the minimum average partial correlation (MAP) to determine the correct number of components.

	MAP					
PA	0	1	2	**3**	4	Total
1	8	*9255*	3001	**463**	6	12733
2	566	5224	*12318*	**2317**	55	20480
3	**11147**	**6443**	**14706**	***22980***	**72**	**55348**
4	723	252	113	**82**	*1*	1171
5	166	49	12	**4**		231
6	22	8	1	**1**		32
7	3	1	1			5
Total	12635	21232	30152	**25847**	134	90000

Note. Values of zero are left blank. Boldface represents the identification of the correct number of components (3). Italic represents agreement between MAP and PA.

amount of components. Worse, in matrix 3, 4, and 5, its best performances were below 15% of correct identification. With the exception of matrix 1, it failed to find the appropriate solution less than half the time in several cases. These results clearly show that when using harder component structures (less salient component saturation, loading below 0.50, or obliqueness, correlation between components) stopping rules lose their reliability. Finally, the dimensionality determined by PA was underestimated 37% of the time. That means that, if PA finds meaningful axes, they are likely to be real ones. However, PA will systematically fail to find less, but still meaningful, components.

Table 3 shows the performance of MAP in identifying the correct number of meaningful axes. MAP had its best performances with highly saturated components (matrix 1) regardless of the oblique structure. MAP had a poor performance, when the oblique structure was 0.30. Moreover, it was likely to underestimate the number of axes. Notwithstanding these results, Table 3 shows that the power of MAP was high when sample size was at least of 128 and when the correlation between components was below not exceed 0.30. However, its efficiency, restricted to specific sets of circumstances, does not redeem the loss of meaningful information. Yet, except for the easiest conditions (i.e., correlation matrix 1), MAP did have a better overall performance than PA.

Table 2. Power (proportion of correct identification) of parallel analysis (PA) according to correlation matrices, correlations between components (corr. comp.), and sample size.

		Sample size					
Matrix	corr. comp.	16	32	64	128	256	512
1	0.00	.70	.98	1	1	1	1
	0.10	.71	.98	1	1	1	1
	0.30	.55	.89	.99	1	1	1
2	0.00	.17	.30	.45	.54	.49	.52
	0.10	.13	.28	.46	.65	.77	.94
	0.30	.05	.03	.01			
3	0.00	.14	.19	.21	.10	.02	
	0.10	.13	.17	.18	.11	.04	.01
	0.30	.07	.02				
4	0.00	.11	.23	.17	.09	.02	
	0.10	.12	20	.18	.12	.05	
	0.30	.05	.02				
5	0.00	.07	.09	.06	.01		
	0.10	.08	.10	.06	.04	.01	
	0.30	.03					

Note. Values of zeros are left blank.

Table 3. Power (proportion of correct identification) of the minimum average partial correlation (MAP) according to correlation matrices, correlation between components (corr. comp.), and sample size.

		Sample size					
Matrix	corr. comp.	16	32	64	128	256	512
1	0.00	.61	.95	1	1	1	1
	0.10	.53	.89	1	1	1	1
	0.30	.21	.37	.65	.92	1	1
2	0.00	.45	.81	.99	1	1	1
	0.10	.40	.66	.93	1	1	1
	0.30	.12	.05	.06	.07	.11	.22
3	0.00	.38	.69	.94	1	1	1
	0.10	.32	.53	.77	.95	1	1
	0.30	.08	.03	.01			
4	0.00	.39	.67	.93	1	1	1
	0.10	.32	.52	.79	.95	1	1
	0.30	.11	.02	.01	.01		
5	0.00	.40	.73	.96	1	1	1
	0.10	.37	.55	.83	.98	1	1
	0.30	.13	.03	.01			

Note. Values of zeros are left blank.

It is worthy to note that, in some oblique case (especially .3), increasing the sample size decreased the power of both stopping rules. This is likely due to the fact that empirical eigenvalues were getting closer to the population eigenvalues and that the third eigenvalue of matrices 3, 4 and 5 were under unity, see Table 4.

4. Discussion

The purpose of the current study was to evaluate the ability of MAP and PA to detect the correct number of meaningful axes from a PCA when the correlation matrices contain an oblique structure. Their failures are generally reported, but largely undiscussed in the methodological literature (Garrido, Abad, and Ponsoda 2011; Peres-Neto, Jackson, and Somers 2005). The current study used Monte Carlo simulations to investigate the performance of MAP and PA.

Table 4. The first five eigenvalues of the population according to correlation matrices, correlations between components.

		Components				
Matrix	corr. comp.	I	II	III	IV	V
1	0.00	2.40	2.40	2.40	0.30	0.30
	0.10	3.00	2.10	2.10	0.30	0.30
	0.30	4.20	1.50	1.50	0.30	0.30
2	0.00	2.00	2.00	2.00	0.50	0.50
	0.10	2.60	1.70	1.70	0.50	0.50
	0.30	3.80	1.10	1.10	0.50	0.50
3	0.00	2.40	2.00	1.60	0.70	0.70
	0.10	2.71	1.87	1.42	0.70	0.70
	0.30	3.84	1.31	0.85	0.70	0.70
4	0.00	3.10	2.00	1.30	0.70	0.50
	0.10	3.26	1.93	1.21	0.70	0.50
	0.30	4.16	1.43	0.81	0.70	0.50
5	0.00	2.00	1.90	1.70	0.70	0.70
	0.10	2.48	1.62	1.50	0.70	0.70
	0.30	3.64	1.10	0.86	0.70	0.70

Overall, MAP had a better overall performance than PA. Both had their best performances with highly saturated components (matrix 1) regardless of the oblique structure and both had poor performance, when the oblique structure was 0.30. somewhat similar. PA had poor performances in most of the conditions and failed to identify the correct number of components half the times (except with matrix 1, where PA had better performance than MAP). MAP was more likely to give correct results when the correlation matrix was modestly oblique (0.10), but both stopping rules had poor performance when the structure was highly oblique (0.30). This poor performance is explained by population eigenvalues of matrices 3, 4, and 5 being under unity (0.85, 0.81 and 0.86 respectively) when there was a correlation between components. Moreover, as sample size increases, empirical eigenvalues were getting closer to the population eigenvalues. PCA cannot take advantage of the sample variation, and smaller eigenvalues become increasingly harder to identify. In the current cases, the obliqueness of the third component is taken into account by the first ones, reducing its own eigenvalue.

Two other results of the current study to the methodological literature are worth noting. First, a sample size of at least 100 was sufficient for MAP to identify the correct dimensionality when there were no oblique structures. In agreement with previous studies, MAP and PA were likely to underestimate the number of component. PA errors nearly always consisted in an underestimate of the number of axes whereas MAP produced a non-negligible proportion of overestimates. Second, low ratios of variables by components are generally considered hard conditions, leading to poor performances in counting components, but were not particularly determinant in the current study. It can be stated that 4:1 is just over the recommended minimum of variables by component (Velicer et al. 2000), however, three variables is generally seen as sufficient to consider an axe as meaningful (Fabrigar et al. 1999) whereas two might lead to unreliable estimations (see doublet factors; Mulaik 2009). Yet, given that the recommendation goes for both stopping rules, and that MAP had good performances, this only brings it stronger support.

The current results show that a visual inspection of the correlation matrix is necessary to verify if any oblique structure exists. In these circumstances, researchers are likely to miss the real dimensionality of the data sets; they should avoid PA and MAP in order to prevent loss of information. The results also call on the necessity to systematically investigate the influence of oblique structures (especially when meaningful axes' eigenvalue are below unity) and on the need to develop new stopping rules taking obliqueness into account, and being less dependent on eigenvalues (i.e., by taking loadings into account) or less vulnerable when meaningful eigenvalues fall below unity. Present results show that stopping rules may lose their reliability when they face harder-to-detect component structures.

Acknowledgement

The author would like to thank André Achim and Anne-Josée Piazza for their helpful comments on the manuscript.

ORCID

P.-O. Caron http://orcid.org/0000-0001-6346-5583

References

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

Beauducel, André. 2001. Problems with parallel analysis in data sets with oblique simple structure. *Methods of Psychological Research Online* 6 (2):141–157.

Courtney, Matthew Gordon Ray. 2013. Determining the number of factors to retain in EFA: using the SPSS R-menu v2.0 to make more judicious estimations. *Practical Assessment, Research & Evaluation* 18 (8):1–14.

Dinno, Alexis. 2009. Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research* 44 (3):362–388. doi:10.1080/00273170902938969. PMID:20234802

Fabrigar, Leandre R., Duane T. Wegener, Robert C. MacCallum, and Erin J. Strahan. 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.

Garrido, L. E., F. J. Abad, and V. Ponsoda. 2011. Performance of Velicer's minimum average partial factor retention method with categorical data. *Educational and Psychological Measurement* 71 (3):551–570. doi:10.1177/0013164410389489.

Guadagnoli, Edward, and Wayne F. Velicer. 1988. Relation of sample size to the stability of component patterns. *Psychological Bulletin* 103 (2):265–275. doi:10.1037/0033-2909.103.2.265. PMID:3363047

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

Kaiser, H. F. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 20:141–151. doi:10.1177/001316446002000116.

Mulaik, S. A. 2009. *Foundations of factor analysis*. Boca Raton (FL): Chapman and Hall/CRC.

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, Instruments, & Computers* 32:396–402. doi:10.3758/BF03200807.

Peres-Neto, Pedro R., Donald A. Jackson, and Keith Mé Somers. 2005. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis* 49:974–997. doi:10.1016/j.csda.2004.06.015.

Shriner, D. 2011. Investigating population stratification and admixture using eigenanalysis of dense genotypes. *Heredity* 107:413–420. doi:10.1038/hdy.2011.26. PMID:21448230

Shriner, D. 2012. Improved eigenanalysis of discrete subpopulations and admixture using the minimum average partial test. *Human Heredity* 73:73–83. doi:10.1159/000335899. PMID:22441298

Tracy, Craig A., and Harold Widom. 1993. Level-spacing distributions and the Airy Kernel. *Physics Letters B* 305:115–118. doi:10.1016/0370-2693(93)91114-3.

Turner, Nigel E. 1998. The effect of common variance and structure pattern on random data eigenvalues: Implications for the accuracy of parallel analysis. *Educational and Psychological Measurement* 58:541–568. doi:10.1177/0013164498058004001.

Velicer, Wayne F. 1976. Determining the number of components from the matrix of partial correlations. *Psychometrika* 41:321–327. doi:10.1007/BF02293557.

Velicer, Wayne F., C. A. Eaton, and J. L. Fava. 2000. Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In *Problems and solutions in human assessment: honoring Douglas N. Jackson at seventy*, edited by R. D. Goffin and E. Helmes, 41–71. Boston: Kluwer.

Wigner, E. 1955. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics* 62 (3):548–564. doi:10.2307/1970079.

Zwick, William R., and Wayne F. Velicer. 1982. Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research* 17:253–269. doi:10.1207/s15327906mbr1702_5. PMID:26810950

Zwick, William R., and Wayne F. Velicer. 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.