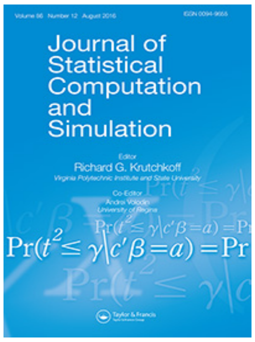
Journal of Statistical Computation and Simulation
Editor
Richard G. Krutchkoff
Virginia Polytechnic Institute and State University
Co-Editor
Andrei Volodin
University of Regina
$\Pr(t^2 \leq \gamma|c'\beta = a) = \Pr$
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

A Monte Carlo examination of the broken-stick distribution to identify components to retain in principal component analysis

Pier-Olivier Caron

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2016
VOL. 86, NO. 12, 2405–2410
http://dx.doi.org/10.1080/00949655.2015.1112390

Taylor & Francis
Taylor & Francis Group

A Monte Carlo examination of the broken-stick distribution to identify components to retain in principal component analysis

Pier-Olivier Caron

Laboratoire des sciences appliquées du comportement, Département de Psychologie, Université du Québec à Montréal, Montréal, Québec, Canada

ABSTRACT

The broken-stick (BS) is a popular stopping rule in ecology to determine the number of meaningful components of principal component analysis. However, its properties have not been systematically investigated. The purpose of the current study is to evaluate its ability to detect the correct dimensionality in a data set and whether it tends to over- or underestimate it. A Monte Carlo protocol was carried out. Two main correlation matrices deemed usual in practice were used with three levels of correlation (0, 0.10 and 0.30) between components (generating oblique structure) and with different sample sizes. Analyses of the population correlation matrices indicated that, for extremely large sample sizes, the BS method could be correct for only one of the six simulated structure. It actually failed to identify the correct dimensionality half the time with orthogonal structures and did even worse with some oblique ones. In harder conditions, results show that the power of the BS decreases as sample size increases: weakening its usefulness in practice. Since the BS method seems unlikely to identify the underlying dimensionality of the data, and given that better stopping rules exist it appears as a poor choice when carrying principal component analysis.

ARTICLE HISTORY
Received 23 September 2014
Accepted 21 October 2015

KEYWORDS
Principal component analysis; stopping rules; Monte Carlo simulations; broken-stick

AMS SUBJECT CLASSIFICATION
62H25; 11K45

1. Introduction

The broken-stick (BS) [1] method is one of many stopping rules to determine the number of meaningful components of principal component analysis (PCA). This model came from studies concerned by the abundance of species among habitats and whether their distribution is structured or random.[2,3] Because the BS model is deeply rooted in the ecological literature and has already been implemented in many R packages, it is widely used by ecologists and biologists.

To understand the BS method, consider that if one breaks a stick into p pieces (after randomly selecting $p-1$ breaking points), and sorts them in decreasing length, the expected length, b_k, of the kth longest piece is $b_k = \sum_{i=k}^{p}(1/i)$. This equation defines the BS distribution. Frontier [1] suggested that these decreasing theoretical means, b_k, can be used as critical values to decide how many principal components contain meaningful information. It has to be used for PCA on a correlation matrix in which p is the number of variables. According to this rule, the components to retain are the first ones with eigenvalues all higher than the corresponding b_k.

The BS has previously been studied in comparison to other stopping rules. It was found to be accurate for salient components (high component saturation) in combination with orthogonal

CONTACT Pier-Olivier Caron pocaron19@gmail.com

structures.[4,5] That is, component structures for which identifying the correct number of meaningful axes is easy for almost any rule. In order to avoid doubts on their validity, stopping rules should be assessed against simulations with hard conditions, in which they could theoretically fail, rather than mere simple ones in which they are likely to succeed. However, no systematic analysis of BS has been carried on complex correlation structures. Since it is widely used, it is necessary to further verify the extent to which this stopping rule is efficient. Thus the purpose of the current study is to evaluate whether the BS tends to over- or underestimate the number of components (accuracy) and its ability to detect the correct number of components (power) regarding sample size and correlations between components.

2. Simulation method

A Monte Carlo protocol was carried out. To evaluate the influence of the component structure, two initial correlation matrices were borrowed from Peres-Neto et al. [5] (their matrices 7 and 9 corresponding to matrices 2 and 1 herein), which they considered as the representative of studies in ecology. Figure 1 depicts their component structure. Centre areas (0.80, 0.50, and 0.30) represent the correlation between variables within components. Off-diagonal areas correspond to the level of correlations between components (values of 0, 0.10 and 0.30, but left empty here). As such, there are three different settings for each matrix giving a total six component structure. Each matrix is composed of nine variables. Matrix 1 is composed of three components each expressed in three variables. Matrix 2 is composed of three components, respectively, expressed in four, three and two variables. Matrices 1 and 2 both contained three components and are used to assess the ability of the BS to detect their correct number and, when it errs, whether it tends to over- or underestimate. The amount of variables was not varied and remained nine across all conditions, because increasing the number of variables in a correlation matrix while maintaining the same number of components decreases the difficulty of finding the correct number of axes by cumulating more variance for each component.

In order to increase the difficulty of the matrices, correlations between variables depending on different underlying components were added (0.10 and 0.30) to create oblique structures that transfer variance from the later to the earlier underlying components.[5] Obliqueness was implemented by changing outside component areas (off-diagonal areas of the matrices shown in Figure 1) to the specified level of correlations. Finally, to evaluate the influence of sample size, we systematically varied the number of objects (or subjects) to specified values (8, 16, 32, 64, 128 256, 512, 1024, and 2048) chosen to represent a variety of fields and studies in ecology and psychology (approximation of the current values have been used by Peres-Neto et al. [5] and Beauducel [6]).

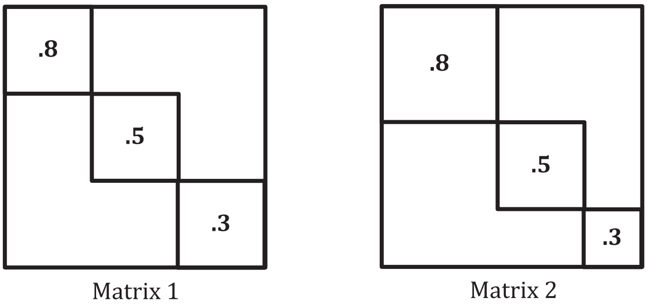

Figure 1. Visual representation of correlation matrices. Centre areas (0.80, 0.50 and 0.30) represent the correlation between variables within components (variables sharing the same axe). The size of squares correspond to the proportion of variables per component, that is, three variables by component for matrix 1, and four, three and two variables for matrix 2. Off-diagonal areas correspond to the level of correlations between components (values of 0, 0.10 and 0.30, but left empty here). There is an implicit diagonal of unities in matrices 1 and 2. Inspired by Peres-Neto et al. [5].

Simulations were carried out in Matlab (R2012b) on a Window® 7 operating system. The computer had 2.53 GHz Intel® core processor, 4 GB of RAM, and 450 GB of hard disk space. The numerical simulation was as follow:

(1) Generate, with the function *mvnrnd*, an artificial data set from a multivariate normal distribution with expected null means and expected covariance specified by the population correlation matrix and with a specified sample size.
(2) Carry a PCA on the data sets and record the eigenvalues.
(3) Compare the consecutive eigenvalues to the corresponding BS critical values, stopping at k, just before the first eigenvalue lower than the criterion.
(4) Record k as the identified number of components.
(5) Repeat steps 1, 2, 3 and 4 a total of 10,000 times.
(6) Compute the average difference between the number of detected and true components (accuracy) and the frequency with which the correct number was identified (power).
(7) Repeat the simulations for the two correlation matrices, the three levels of correlations between components and the nine levels of sample size.

3. Results

3.1. Population analysis

Table 1 gives the population component eigenvalues (i.e. for the specified ideal correlation matrix) along with the BS critical values for nine variables. Apart from the sampling variation that depends on sample size, the level of difficulty in estimating the number of dimensions depends on whether the first three population eigenvalues are all higher (easier situation) or some are lower (harder situation) than the critical values specified by the BS. Population component eigenvalues below the BS criterion (making the correct dimensionality harder to identify) are in bold face.

One can see from Table 1 that the dimensionality of matrix 1 with an orthogonal and an oblique structure of 0.30 will actually be harder to identify than the oblique structure of 0.10. In the former case, the first eigenvalue is expected to be smaller than its corresponding BS criterion and, in the latter case, both the second and third ones are so. All structures of matrix 2 produce expected eigenvalues below the BS criterion for the second and third components. It is worth noting that, as the correlations between components increase, the first eigenvalue accounts more variance whereas latter ones tend to decrease.

Given the expected (population) eigenvalues presented in Table 1, one should anticipate that the BS method will poorly detect the correct dimensionality in most cases. Good BS performance will depend on 'favourable' sampling error, that is, the opportunistic characteristic of PCA to summarize

Table 1. The critical values of the BS distribution followed by the first five theoretical eigenvalues of the correlation matrices as a function of the correlations between components (corr. comp.).

		Components				
		I	II	III	IV	V
BS critical values:		2.83	1.83	1.33	1.00	0.75
Matrix	corr. comp.					
1	0.00	**2.60**	2.00	1.60	0.70	0.70
	0.10	2.85	1.93	1.42	0.70	0.70
	0.30	3.93	**1.41**	**0.85**	0.70	0.70
2	0.00	3.40	**1.79**	**1.30**	0.70	0.50
	0.10	3.52	**1.78**	**1.20**	0.70	0.50
	0.30	4.31	**1.45**	**0.80**	0.70	0.50

Note: Population component eigenvalues below the BS criterion are in bold face.

meaningful information as well as sampling errors (accidently correlated) in the earlier axes. Nevertheless, the main expectation is that, as sample variation decreases, the BS should be more likely to miss meaningful axes.

3.2. Simulation analysis

Figure 2 depicts the relation between the ability of the BS to detect the correct number of component and sample size levels. It shows that the BS generally did not find the correct number of components even half the time. For all five situations where at least one of the first three population eigenvalues was below the critical BS level, success decreases as sample size increased (i.e. as sampling error decreased). The BS success for the remaining condition at first decreased with increasing sample size up to 64 and then increased, reaching only 65% at $N = 2048$. Compared to the conventionally required power of 0.80, the BS had, at its best, a weak ability to identify the correct number of components. For the easiest condition, Caron and Achim [7] have found that the BS reached a power of 0.80 (inside a confidence interval of 95%) at $N = 5605$.

As the population analysis pointed out for harder conditions, the power decreases as sample size becomes large. The exception is the second matrix with correlations of 0.10 between components. This is the only case where the first three theoretical eigenvalues are higher than the expected BS distribution. Adding a small oblique structure leads the first component to be easier to identify compared to the orthogonal structure, while the next two expected eigenvalues decreased but remained above their respective criterion. In all other cases, increasing sample size did decrease the ability to detect the correct number of components.

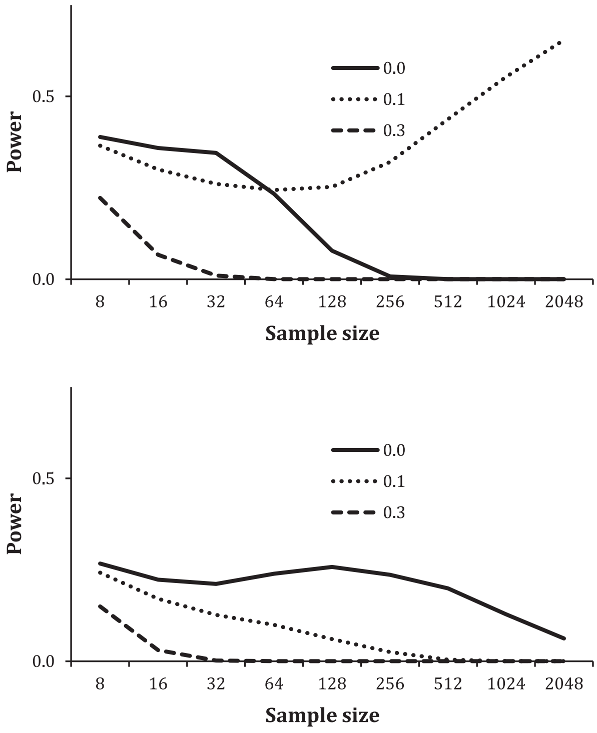

Figure 2. The ability of the BS to identify the correct number of components (power) according to sample size (abscissa) and correlations between components (lines shape). Upper and bottom panels depict the results for matrices 1 and 2, respectively.

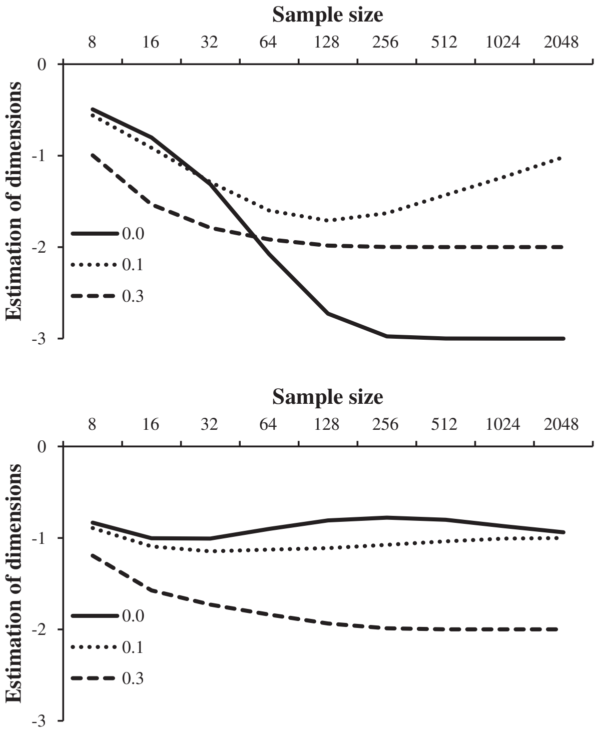

Figure 3. The accuracy (average difference between the number of detected and true components) of the BS according to sample size (abscissa) and correlations between components (lines shape). Upper and bottom panels depict the results for the matrices 1 and 2, respectively.

Figure 3 shows that, at every sample size level, the BS underestimates the number of components for all six component structures. It should be noted that a sample size of 8 is lower than the number of variables. At this point, one should already be concerned by the variables-to-sample-size ratio regardless of the stopping rule. Still, it is the level at which the BS, in five out six conditions, had its most accurate performances, the exception being the simpler case pointed out by the population analysis.

4. Discussion

The purpose of the current study was to assess whether the BS method tends to over- or underestimate the number of components and its ability to detect the correct number of components, especially for representative but not so clear-cut situations. According to the correlation matrices used herein, the current results show that the BS will miss the component structure, if there is such in the data set, more than half the times at best. In harder cases, it will fail to identify the right number of components and will underestimate their amount. The population eigenvalue analysis showed that as sample size increases, the power and accuracy of BS is expected to decrease when an eigenvalue is below the criterion. This was merely due to the decreasing influence of sampling error, that is, empirical eigenvalues getting closer to population eigenvalues, which lead them to be harder to identify.

One might suggest that the number of conditions is actually limited. However, expanding the number of settings (and more specifically hard ones) would not change the current conclusions. Knowing population eigenvalues a priori, and whether they are over or below the BS criterion, will always lead to the expected outcome. Either true eigenvalues are over the BS criterion, in which case this is an easy setting, or they are below the criterion, which is an unsolvable setting for the BS. If a single eigenvalue of a true component is below the criterion, then whatever the correlation matrix is, as sample size increases, less the stopping rule can detect the correct number of components to retain. In other terms, adding harder scenarios would not change the outcome of the current study: the BS will always lack power if a component's eigenvalue is below the threshold. Because the stopping rule seems unlikely to give correct results in difficult situations and given that better methods already exist,[5] we must recommend to avoid it as a means to identify the number of meaningful components from PCA. Other methods to evaluate the dimensionality of data sets should be considered instead.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

The author is currently subsidized by the Fonds de recherche du Québec – Société et culture.

ORCID

Pier-Olivier Caron http://orcid.org/0000-0001-6346-5583

References

[1] Frontier S. Étude de la décroissance des valeurs propres dans une analyse en composantes principales: comparaison avec le modèle du bâton brisé. J Exp Marine Biol Ecol. 1976;26:67–75.
[2] Barton DE, David FN. The dispersion of a number of species. J R Stat Soc Ser B. 1959;21:190–194.
[3] MacArtur RH. On the relative abundance of bird species. Proc Natl Acad Sci USA. 1957;43:293–295.
[4] Jackson DA. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology. 1993;74:2204–2214.
[5] Peres-Neto PR, Jackson DA. Somers KM. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Comput Stat Data Anal. 2005;49:974–997.
[6] Beauducel A. Problems with parallel analysis in data sets with oblique simple structure. Methods Psychol Res Online. 2001;6:141–157. Available from: http://www.dgps.de/fachgruppen/methoden/mpr-online/issue14/art2/article.html
[7] Caron P-O, Achim A. Nombre de composantes à retenir dans l'analyse en composantes principales: Comparaison de méthodes selon la taille d'échantillon requise pour 80% de succès. Poster presented at the 36th congress of the Société Québécoise de Recherche en Psychologie: Montréal; 2014.