Supplementary materials for this article are available at https://doi.org/10.1007/s13253-019-00355-5.

Hypothesis Tests for Principal Component Analysis When Variables are Standardized

Johannes FORKMAN, Julie JOSSE, and Hans-Peter PIEPHO

In principal component analysis (PCA), the first few principal components possibly reveal interesting systematic patterns in the data, whereas the last may reflect random noise. The researcher may wonder how many principal components are statistically significant. Many methods have been proposed for determining how many principal components to retain in the model, but most of these assume non-standardized data. In agricultural, biological and environmental applications, however, standardization is often required. This article proposes parametric bootstrap methods for hypothesis testing of principal components when variables are standardized. Unlike previously proposed methods, the proposed parametric bootstrap methods do not rely on any asymptotic results requiring large dimensions. In a simulation study, the proposed parametric bootstrap methods for standardized data were compared with parallel analysis for PCA and methods using the Tracy–Widom distribution. Parallel analysis performed well when testing the first principal component, but was much too conservative when testing higher-order principal components not reflecting random noise. When variables are standardized, the Tracy–Widom distribution may not approximate the distribution of the largest eigenvalue. The proposed parametric bootstrap methods maintained the level of significance approximately and were up to twice as powerful as the methods using the Tracy–Widom distribution. SAS and R computer code is provided for the recommended methods.

Supplementary materials accompanying this paper appear online

Key Words: Dimensionality reduction; GGE; Parallel analysis; Parametric bootstrap; Principal component analysis; Tracy–Widom distribution.

1. INTRODUCTION

Principal component analysis (PCA) is used extensively in many areas of research, including agriculture, biology and environmental sciences. A question of crucial importance for

Johannes Forkman, (✉) Department of Crop Production Ecology, Swedish University of Agricultural Sciences, PO Box 7043, 750 07 Uppsala, Sweden (E-mail: *johannes.forkman@slu.se*). Julie Josse, CMAP UMR 7641 École Polytechnique INRIA-XPOP CNRS, Route de Saclay, 91128 Palaiseau Cedex, France (E-mail: *julie.josse@polytechnique.edu*). Hans-Peter Piepho, Institute of Crop Science, University of Hohenheim, 70 593 Stuttgart, Germany (E-mail: *hans-peter.piepho@uni-hohenheim.de*).

© 2019 The Author(s)
Journal of Agricultural, Biological, and Environmental Statistics, Volume 24, Number 2, Pages 289–308
https://doi.org/10.1007/s13253-019-00355-5

interpretation of results is how many principal components should be utilized. The first components are often interesting, since these typically account for a large proportion of the total variation, but the last components are usually discarded, since these may reflect noise rather than systematic pattern.

Several procedures (Jolliffe 2002) have been proposed for determining how many principal components to retain: the popular Kaiser (1960) rule and the scree plot (Cattell 1966), resampling methods (Peres-Neto et al. 2005), cross-validation methods (Bro et al. 2008; Josse and Husson 2011), Bayesian procedures (Hoff 2007; Perez-Elizalde et al. 2012; Sobczyk et al. 2017) and statistical tests (Muirhead 1982, p. 409; Johnstone 2001; Choi et al. 2017). However, these statistical tests assume that variables are not standardized before carrying out the PCA. This is in stark contrast to the very common practice and need (Yeater et al. 2015) of standardizing variables in order to remove scale differences between them. The main purpose of this article, therefore, is to fill this yawning gap and provide a solution on how to statistically test significance of principal components when variables are standardized. Parametric bootstrap tests will be proposed for PCA of standardized data.

In agriculture, PCA is a main tool for analysis of genotype-by-environment interaction. When the dataset is an $n \times p$ matrix of column-centred observations from n genotypes grown in p environments, PCA is equivalent to fitting the *genotype main effects and genotype-by-environment interaction effects* (GGE) model (Yan and Kang 2003), which is also known as the *sites regression* model (Crossa et al. 2004). In microarray datasets, with n genes and p treatments, the same model is known as the *treatment regression* model (Crossa et al. 2005). If also rows are centred before conducting the PCA, then the model is the *additive main effects and multiplicative interaction* (AMMI) model (Gauch 1992). The question of how many principal components to retain in the model is a key issue in GGE and AMMI analysis (Yang et al. 2009).

A variable is standardized to zero mean and unit variance in three steps: (1) Based on the n observations of the variable, compute the sample mean and the sample standard deviation. (2) Subtract the sample mean from each observation. (3) Divide these mean-centred observations with the sample standard deviation. This standardization procedure is common practice in PCA (Jolliffe and Cadima 2016). For some examples, see Hoyos-Villegas et al. (2016), Kollah et al. (2017) and Yan and Frgeau-Reid (2018).

The eigenvalues, $\hat{\lambda}_k$, $k = 1, 2, \ldots$, of the sample covariance or correlation matrix reflect the relative importance of the principal components. In the theoretical case that all observations are independent and standard normally distributed, the distribution of the largest eigenvalue of the sample covariance matrix can be approximated by a Tracy–Widom distribution (Johnstone 2001). As an example, the dashed curve in Fig. 1 illustrates the distribution of the largest eigenvalue, $\hat{\lambda}_1$, when a 30×20 matrix of independent standard normally distributed observations were randomly generated 100,000 times. The solid curve is the density of the Tracy–Widom distribution, when scaled as proposed by Johnstone (2007). Since the solid curve approximates the dashed, the Tracy–Widom distribution can indeed be used for inference about the first principal component when all 600 observations are independent standard normally distributed. However, if each of the 20 columns of random standard normally distributed observations is initially standardized to zero mean and unit variance, then the distribution of the first eigenvalue becomes much different, as shown by the dotted curve

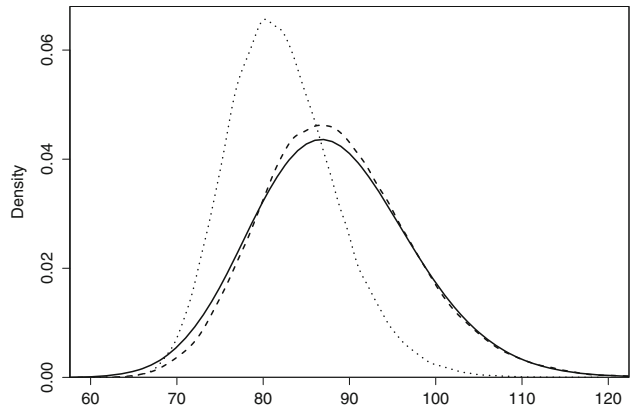

Figure 1. A matrix with 600 standard normally distributed values arranged in $n = 30$ rows and $p = 20$ columns was randomly generated 100,000 times. The figure shows the distribution of the largest eigenvalue (dashed curve), the distribution of the largest eigenvalue computed after initial standardization of each column to zero mean and unit standard deviation (dotted curve) and the Johnstone (2007) scaled Tracy–Widom distribution (solid curve).

in Fig. 1. Clearly, different methods for inference on principal components may be needed, depending on whether columns are standardized or not.

When variables express different quantities (e.g. height and weight), an argument for using scaled data is that without scaling, results will depend on the choice of unit of measurement (e.g. whether height is measured in metres or centimetres). Underhill (1990) proposed scaling variables by dividing with means instead of standard deviations. This method also makes results independent of the choice of unit. As an example of the usefulness of such scaling, Underhill (1990) analysed a matrix with data on areas grown with different vegetable crops in the UK between the years 1968 and 1987. Scaling using means was useful for study of the relative variability of the areas over time, considering that some crops, e.g. peas and cabbage, were grown on much larger areas than other crops, such as celery and rhubarb. The present article considers both these options of scaling, i.e. scaling using means or standard deviations, as well as the option of using non-scaled data. Models are proposed for each of the three options, and it is shown how principal components can be tested in these models.

We focus on small datasets, which are still commonly encountered in applications, especially in analysis of genotype-by-environment interaction. When data are non-scaled, the *simple parametric bootstrap* method (Forkman and Piepho 2014) can be used for testing principal components in PCA (Forkman 2015), but, as Sect. 5 will show, this method does not work well when variables have been scaled to unit variance.

Franklin et al. (1995) and Peres-Neto et al. (2005) recommended parallel analysis (Horn 1965) for selecting the number of components in PCA. According to this method, simulated random matrices of normally distributed values are subjected to PCA. Glorfeld (1995) proposed using upper percentiles, instead of means, of simulated eigenvalues, in order to decrease Type I error. Parallel analysis is legitimate for evaluation of the first principal component, but questionable for evaluating remaining principal components (Crawford

et al. 2010). The present article builds on the basic idea of parallel analysis, but suggests that the reference distribution needs to be simulated differently, depending on the principal component to be tested and whether variables are standardized or not.

Factor analysis uses a model with random factors (Johnson and Wichern 2007), whereas our proposed models consist of a fixed systematic part and a single random normally distributed error part. Several methods exist for determining the number of components in factor analysis (Bai and Ng 2002; Kritchman and Nadler 2008; Onatski 2009; Owen and Wang 2016; Passimier et al. 2017). The "revised parallel analysis" (Green et al. 2012) and the "comparison data" method (Ruscio and Roche 2012) use resampling and are related to our work.

Section 2 presents three motivating examples. Section 3 specifies models and proposes bootstrap methods. Section 4 describes other methods for hypothesis testing: the parallel method for PCA and methods using the Tracy–Widom distribution. Section 5 presents a simulation study comparing the methods. Section 6 analyses the three examples using the proposed bootstrap methods. Section 7 discusses the results. Computer code is provided in Supplementary materials.

2. MOTIVATING EXAMPLES

This article uses three examples:

(a) The peanut dataset consists of observations of yield from $n = 10$ peanut genotypes grown in $p = 15$ environments (E01–E15). The dataset, kang.peanut, is included in the R package agridat and published in Kang et al. (2004).

(b) The Bumpus (1899) female sparrows dataset contains observations of $p = 5$ body measurements on $n = 49$ female sparrows. The variables are L1: total length, L2: alar extent, L3: length of beak and head, L4: length of humerus and L5: length of keel of sternum. Manly (1986) includes the dataset, but it is also published on the Internet (North Dakota State University 1997).

(c) The fish dataset comprises mass fractions of $p = 7$ chemicals (C1–C7) measured in $n = 10$ samples of fish collected in the Bay of Seine (Galgani et al. 1991). Zitko (1994) used this dataset for promoting an increased use of PCA for evaluation of environmental data. The chemicals are C1: PCB, C2: DDE, C3: DDD, C4: DDT, C5: α-HCH, C6: γ-HCH and C7: PAH.

Table 1 lists means, standard deviations and coefficients of variation for the variables of the three datasets.

The peanut dataset is a typical dataset for GGE analysis. Figure 2a, b shows biplots when variables, i.e. environments, are scaled to zero means and unit variance, as recommended by Yan and Kang (2003, p. 56). Figure 2a shows the first two principal components (PC1 and PC2), and Fig. 2b shows the third and fourth principal components (PC3 and PC4). Commonly in scientific reports, only the first two principal components are presented, but here we display also the third and the fourth.

Table 1. Means, standard deviations (SD) and coefficients of variation (CV) for the variables of the three example datasets.

Variable	Mean	SD	CV
(a) The peanut dataset			
E01	1.10	0.20	0.186
E02	1.65	0.78	0.475
E03	2.36	0.17	0.070
E04	2.05	0.32	0.157
E05	1.43	0.37	0.257
E06	2.80	0.30	0.105
E07	2.93	0.24	0.082
E08	1.91	0.53	0.278
E09	2.90	0.41	0.142
E10	3.98	0.43	0.108
E11	1.99	0.26	0.133
E12	4.80	0.60	0.126
E13	1.68	0.44	0.260
E14	4.17	0.33	0.079
E15	3.06	0.52	0.171
(b) The sparrows dataset			
L1	158	3.65	0.023
L2	241	5.07	0.021
L3	31.5	0.79	0.025
L4	18.5	0.56	0.031
L5	20.8	0.99	0.048
(c) The fish dataset			
C1	5110	3595	0.704
C2	69.80	46.74	0.670
C3	22.40	10.09	0.450
C4	30.86	25.28	0.819
C5	1.690	0.624	0.369
C6	13.47	5.20	0.386
C7	14.19	13.34	0.940

Analysis of genotype-by-environment interaction aims to answer many questions, among them which genotype performs the best, i.e. gives the highest yield in each environment (Yan and Tinker 2006; Josse et al. 2014). The answer to this question may depend on how many principal components the researcher chooses to retain in the model. In the peanut example, this choice of the number of principal components is particularly important for the conclusions for environment E09. Figure 2a suggests that genotype G03 (manf393) is the top performing genotype for environment E09. However, if only the first principal component is considered important, then genotype G06 (mf480), which is the genotype located furthest to the left on the PC1 axis of Fig. 2a, is estimated to be the best genotype for environment E09. The decisive factor is whether the variation along the PC2 axis should be considered as random or not. In the extreme case, when the biplot just shows random noise, there are no differences at all between the genotypes.

If three principal components are retained, then genotype G08 is estimated to be the best choice for environment E09. This is a consequence of genotype G08 (mf485) and environment E09 both having large negative scores, whereas genotypes G03 and G06 have positive

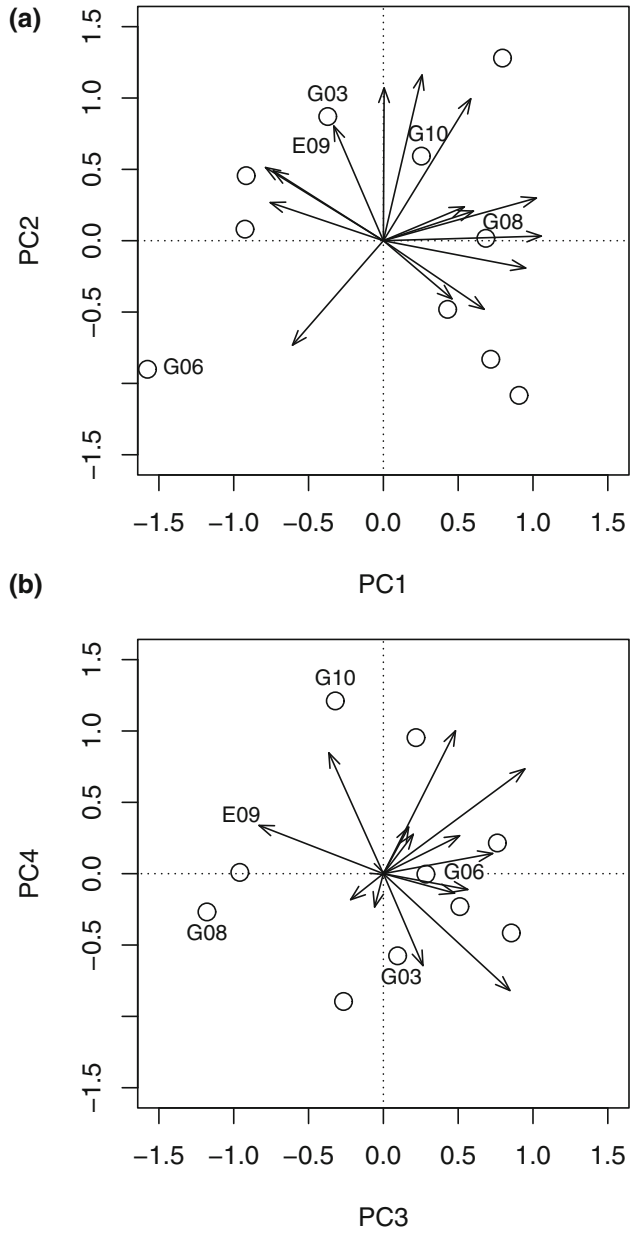

Figure 2. Biplots showing **a** the first and second , and **b** the third and fourth principal components of the peanut dataset. Arrows indicate environments and open circles genotypes. Before analysis, environments were standardized to zero mean and unit variance.

scores, on the third principal component axis (Fig. 2b). If four components are retained, then genotype G10 (mf489) can be recommended for environment E09. The problem studied in this article is how to determine the number of significant principal components when the data are standardized.

3. MODELS AND BOOTSTRAP METHODS

3.1. MODELS

Assume n observations have been made on p variables, and let $\mathbf{Y}$ denote the $n \times p$ matrix of observations. Let y_{ij} denote the observation in the ith row and jth column of $\mathbf{Y}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$. Before PCA, column sample means are subtracted from all observations. The PCA could be performed directly on these mean-centred data or after division with column sample means or after division with column sample standard deviations. Let $\mathbf{X}$ denote the $n \times p$ matrix used for PCA, i.e.

$$\mathbf{X} = \left\{ y_{ij} - \sum_{i=1}^{n} y_{ij}/n \right\}, \tag{1}$$

$$\mathbf{X} = \left\{ \frac{y_{ij} - \sum_i y_{ij}/n}{\sum_i y_{ij}/n} \right\}, \tag{2}$$

$$\mathbf{X} = \left\{ \frac{(y_{ij} - \sum_i y_{ij}/n)}{s_j} \right\}, \tag{3}$$

for non-scaled data, data scaled by means and data scaled by standard deviations, respectively, where $s_j = (\sum_i (y_{ij} - \sum_i y_{ij}/n)^2/(n-1))^{1/2}$. The eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}/(n-1)$ are $\hat{\lambda}_k = \hat{\tau}_k^2/(n-1)$, for $k = 1, 2, \ldots, M$, where $\hat{\tau}_k^2$ is the square of the kth singular value of $\mathbf{X}$.

We propose using the models

$$\mathbf{Y} = \mathbf{A} + \boldsymbol{\Theta}_m + \mathbf{E}, \tag{4}$$

$$\mathbf{Y} = \mathbf{A} + (\boldsymbol{\Theta}_m + \mathbf{E})\boldsymbol{\Delta}, \tag{5}$$

$$\mathbf{Y} = \mathbf{A} + (\boldsymbol{\Theta}_m + \mathbf{E})\boldsymbol{\Sigma}, \tag{6}$$

for non-scaled data, data scaled by means and data scaled by standard deviations, respectively. In these models, $\mathbf{A} = \mathbf{1}_n \boldsymbol{\mu}^{\mathrm{T}}$ is a matrix of intercepts for the p variables, $\boldsymbol{\mu}^{\mathrm{T}} = (\mu_1, \mu_2, \ldots, \mu_p)$, $\mathbf{E}$ is an $n \times p$ matrix of independent $\mathrm{N}(0, \sigma^2)$ distributed errors, $\boldsymbol{\Delta}$ is a diagonal matrix with elements $\mu_1, \mu_2, \ldots, \mu_p$ in the diagonal, and $\boldsymbol{\Sigma}$ is a diagonal matrix with p unknown standard deviations in the diagonal. The matrix $\boldsymbol{\Theta}_m$ can, through singular value decomposition, be written as $\boldsymbol{\Theta}_m = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U}$ is a $p \times m$ matrix of left singular vectors, $\mathbf{V}$ is an $n \times m$ matrix of right singular vectors and $\mathbf{S}$ is a diagonal matrix of positive singular values $\tau_1, \tau_2, \ldots, \tau_m$, sorted in descending order. The rank of $\boldsymbol{\Theta}_m + \mathbf{E}$ is M, where $M = \min(n-1, p)$, but the rank of $\boldsymbol{\Theta}_m$ is m, where $m < M$. We are interested in the unknown parameter m, i.e. the true order of the model.

3.2. NULL HYPOTHESES AND TEST STATISTIC

The null hypotheses are $H_0 : m = K$, where $K \in \{0, 1, \ldots, M-2\}$, with corresponding alternative hypotheses $H_1 : m > K$, where K is the candidate order of the model. Forkman and Piepho (2014) proposed testing these null hypotheses sequentially, starting with $K = 0$

and continuing with $K = 1, 2, \ldots$, until a non-significant result is obtained or $K = M - 2$. As long as $K < M - 2$, the statistic

$$T = \frac{\hat{\tau}_{K+1}^2}{\sum_{k=K+1}^{M} \hat{\tau}_k^2} \tag{7}$$

can be used as a test statistic for the null hypothesis $H_0 : m = K$, as proposed by Yochmowitz and Cornell (1978). The sequential testing procedure ensures the level of significance conditionally on the null model and protects against overfitting (Forkman and Piepho 2014).

3.3. BOOTSTRAP METHODS

For non-scaled data, the simple parametric bootstrap method (Forkman and Piepho 2014) can be used for testing $H_0 : m = K$:

For $b = 1, 2, \ldots, B$, where B is large, do the following:

1. Generate an $(n - 1 - K) \times (p - K)$ matrix $\mathbf{Z}_b = \{z_{ij}\}$, where z_{ij} are independent N(0, 1).
2. Compute the $M - K$ singular values $t_1, t_2, \ldots, t_{M-K}$ of $\mathbf{Z}_b$, and let $T_b = t_1^2 / \sum_{k=1}^{M-K} t_k^2$.

The estimate of the p value is the frequency of T_b larger than the observed test statistic T.

The main feature of the simple parametric bootstrap method is that the dimensions of the matrix of random standard normally distributed values are not $n \times p$, but $(n-1-K) \times (p-K)$. Thus, the numbers of rows and columns are reduced by the number of principal components assumed under the null hypothesis. In addition, the number of rows is reduced by 1, since the columns of $\mathbf{X}$ are centred around zero, which implies a linear relationship between the rows. The idea of reducing the dimensions of the matrix originates from Marasinghe (1985), who used an approximation for the distribution of eigenvalues provided by Muirhead (1978).

The simple parametric bootstrap method is *parametric*, because it assumes a model with a normal distribution of errors, and *simple* since it is based on sampling of random standard normally distributed observations. Thus, no parameters must be estimated. In this regard, the simple parametric bootstrap method differs from the full parametric bootstrap methods that are now proposed for data scaled by means and data scaled by standard deviations.

Let $\hat{\boldsymbol{\mu}} = (\sum_{i=1}^{n} y_{i1}/n, \sum_{i=1}^{n} y_{i2}/n, \ldots, \sum_{i=1}^{n} y_{ip}/n)$, $\hat{\boldsymbol{\Delta}} = \text{diag}(\hat{\boldsymbol{\mu}})$, and $\hat{\mathbf{A}} = \mathbf{1}_n \hat{\boldsymbol{\mu}}^{\mathrm{T}}$. For $K = 0$, let $\hat{\boldsymbol{\Theta}}_K = \mathbf{0}$ (i.e. an $n \times p$ matrix of zeros). For $K \in \{1, 2, \ldots, M - 2\}$, let $\hat{\boldsymbol{\Theta}}_K$ denote the singular value decomposition of $\mathbf{X}$, as specified in Eqs. (2) or (3), with the first K terms retained. Let $\hat{\sigma}_K^2 = \sum_{k=K+1}^{M} \hat{\tau}_k^2 / ((n - 1 - K)(p - K))$, where $\hat{\tau}_k$ is the kth singular value of $\mathbf{X}$.

For $b = 1, 2, \ldots, B$, where B is large, do the following:

1. Generate an $n \times p$ matrix $\mathbf{E}_b$, where the elements are independent N(0, $\hat{\sigma}_K^2$). If data are scaled by means, let $\mathbf{Y}_b = \hat{\mathbf{A}} + (\hat{\boldsymbol{\Theta}}_K + \mathbf{E}_b)\hat{\boldsymbol{\Delta}}$. If data are scaled by standard deviations, let $\mathbf{Y}_b = \hat{\boldsymbol{\Theta}}_K + \mathbf{E}_b$.

2. Let y_{bij} denote the element in the ith row and jth column of $\mathbf{Y}_b$. If data are scaled by means, let $\mathbf{X}_b = \{(y_{bij} - \sum_i y_{bij}/n)/\sum_i y_{bij}/n)\}$. If data are scaled by standard deviations, let $\mathbf{X}_b = \{(y_{bij} - \sum_i y_{bij}/n)/s_{bj}\}$, where $s_{bj} = (\sum_i (y_{bij} - \sum_i y_{bij}/n)^2/(n-1))^{1/2}$.

3. Compute the M singular values $t_1, t_2, \ldots, t_M$ of $\mathbf{X}_b$, and let $T_b = t_{K+1}^2/\sum_{k=K+1}^{M} t_k^2$.

The estimate of the p value is the frequency of T_b larger than the observed test statistic T.

For data scaled by standard deviations, estimates of $\mathbf{A}$ and $\mathbf{\Sigma}$ were not used in Step 1. The obvious parametric bootstrap method would have been to define $\mathbf{Y}_b$ as $\hat{\mathbf{A}} + (\hat{\mathbf{\Theta}}_K + \mathbf{E}_b)\hat{\mathbf{\Sigma}}$, where $\hat{\mathbf{\Sigma}}$ is a diagonal matrix with diagonal elements $s_1, s_2, \ldots, s_p$. However, the standardization of $\mathbf{Y}_b$ in Step 2 gives the same matrix $\mathbf{X}_b$ regardless of whether $\mathbf{Y}_b$ is defined as $\hat{\mathbf{\Theta}}_K + \mathbf{E}_b$ or $\hat{\mathbf{A}} + (\hat{\mathbf{\Theta}}_K + \mathbf{E}_b)\hat{\mathbf{\Sigma}}$. Thus, $\hat{\mathbf{A}}$ and $\hat{\mathbf{\Sigma}}$ were not needed in Step 1. A similar shortcut was not possible for data scaled by means.

4. OTHER METHODS FOR ANALYSIS

4.1. THE PARALLEL METHOD

The simulation study compares the parametric bootstrap methods with the parallel method, which is used for testing principal components when data are scaled by standard deviations (Glorfeld 1995). The parallel method can be performed like this:

For $b = 1, 2, \ldots, B$, where B is large, do the following:

1. Generate an $n \times p$ matrix $\mathbf{Z}_b = \{z_{ij}\}$, where z_{ij} are independent N(0, 1).

2. Standardize the columns of $\mathbf{Z}_b$ using Eq. (3), with z_{ij} substituted for y_{ij}. Let $\mathbf{X}_b$ denote the resulting standardized matrix.

3. Compute the $(K+1)$th singular value t_{K+1} of $\mathbf{X}_b$.

For the null hypothesis $H_0 : m = K$, the estimate of the p value is the frequency of t_{K+1} larger than the observed $(K+1)$th singular value of $\mathbf{X}$, where $\mathbf{X}$ is computed using Eq. (3).

The main difference between the parallel method and the simple parametric bootstrap method is that with the former, the dimensions of the random matrices are $n \times p$, whereas with the latter, the dimensions are $(n-1-K) \times (p-K)$. The two methods also use different test statistics. Furthermore, using the parallel method, the columns of the random matrices are centred and scaled. This is not done with the simple parametric bootstrap method. The parallel method is equivalent to the full parametric bootstrap method for data scaled by standard deviations when testing the significance of the first principal component, i.e. for the hypothesis $H_0 : m = 0$. These two methods differ when testing $H_0 : m = K$, where $K > 0$.

4.2. METHODS USING THE TRACY–WIDOM DISTRIBUTION

Johnstone (2001) studied the distribution of the square of the largest singular value, $\hat{\tau}_1^2$, of an $n \times p$ matrix when all elements of the matrix are standard normally distributed. Specifically, Johnstone (2001) showed that V, defined as

$$V = (\hat{\tau}_1^2 - \mu_{np})/\sigma_{np}\,, \tag{8}$$

approximately has a Tracy–Widom distribution of order 1, where

$$\mu_{np} = (\sqrt{n-1} + \sqrt{p})^{\,2}, \tag{9}$$

$$\sigma_{np} = (\sqrt{n-1} + \sqrt{p})\left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}}\right)^{1/3}. \tag{10}$$

Patterson et al. (2006) recommended V for inference on principal components in genetic data. However, since the theorem of this approximation does not apply when PCA is performed on a correlation matrix, Johnstone (2001) also proposed an "ad hoc" method intended for use in that case, i.e. when columns have been standardized to zero mean and unit variance. According to this ad hoc method, the largest singular value, l_1, of $\mathbf{XR}/(n-1)$ should be computed, where $\mathbf{X}$ is the standardized matrix as specified in Eq. (3) and $\mathbf{R}$ is a diagonal matrix with diagonal elements r_j that are positive roots of independent $\chi^2(n)$ distributed variables, $j = 1, 2, \ldots, p$. Approximately, W, defined as

$$W = (l_1^2 - \mu_{np})/\sigma_{np}\,, \tag{11}$$

follows a Tracy–Widom distribution of order 1.

Later, Johnstone (2007) proposed using

$$\mu_{np} = \left(\sqrt{n-1/2} + \sqrt{p-1/2}\right)^2, \tag{12}$$

$$\sigma_{np} = \left(\sqrt{n-1/2} + \sqrt{p-1/2}\right)\left(\frac{1}{\sqrt{n-1/2}} + \frac{1}{\sqrt{p-1/2}}\right)^{1/3} \tag{13}$$

in place of Eqs. (9) and (10). Thus, either V or W may be computed, using either the Johnstone (2001) Eqs. (9) and (10) or the Johnstone (2007) Eqs. (12) and (13). In the following, these four methods will, with obvious notation, be denoted V-2001, V-2007, W-2001 and W-2007.

In summary, when Model 6 is used and columns are standardized using Eq. (3), the null hypothesis $H_0 : m = 0$, which is used for checking the significance of the first principal component, is readily testable using the Tracy–Widom distribution. The distribution function of the Tracy–Widom distribution is available in the RMTstat package of R.

5. SIMULATION STUDY

Method performance was investigated through simulation based on the three examples of Sect. 2.

For study of Type I error rates, data were repeatedly generated using Models (4), (5) and (6), for non-scaled data, data scaled by means and data scaled by standard deviations, respectively. In these models, $\boldsymbol{\Theta}_m$ was set equal to $\hat{\boldsymbol{\Theta}}_K$ as defined in Sect. 3.3, but with $\mathbf{X}$ defined as in either Eqs. (1), (2) or (3), depending on the model. Note that K equals m under the null hypothesis. The matrices $\mathbf{A}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\Sigma}$ were set equal to $\hat{\mathbf{A}}$, $\hat{\boldsymbol{\Delta}}$ and $\hat{\boldsymbol{\Sigma}}$, respectively, and the variance σ^2, which is inlcuded in $\mathbf{E}$, was set equal to $\hat{\sigma}^2_K$, all defined as in Sect. 3.3. With these settings, 100,000 datasets were randomly generated for each example, model, method and investigated value of K. The null hypotheses $H_0 : m = K$, $K = 0, 1, \ldots, M - 2$, were tested at significance level 0.05. For resampling methods, $B = 1000$ was used. Since 100,000 datasets were simulated, an approximate 0.95 tolerance interval for probability 0.05 can be computed as $0.05 \pm 1.96(0.05(1 - 0.05)/100{,}000)^{1/2} = 0.05 \pm 0.00135$.

For studies of power, data were repeatedly generated using the model

$$\mathbf{Y} = \mathbf{A} + (\psi \boldsymbol{\Theta}_1 + \mathbf{E})\boldsymbol{\Sigma}\,, \tag{14}$$

which apart from ψ is the same as Model (6) with $m = 1$. The additional parameter ψ was varied from 0.2 to 1.0 in steps of 0.1. Therefore, power was investigated at different strengths of the first principal component. The other parameters of Model (14) were set equal to values exactly as described for Type I error (Model 6, $m = 1$). For each example (peanut, sparrows and fish) and value of ψ, 10,000 datasets were generated. For each generated dataset, the null hypothesis $H_0 : m = 0$ was tested using the Tracy–Widom methods W-2001 and W-2007, and the full parametric bootstrap method for data scaled by standard deviations. Only $m = 0$ was tested, since the methods W-2001 and W-2007 were only defined for this hypothesis. The full parametric bootstrap method was conducted with $B = 1000$ bootstrap samples.

5.1. RESULTS OF THE SIMULATION STUDY

Table 2 presents observed Type I error rates for bootstrap and parallel methods. For non-scaled observations, the simple parametric bootstrap method mostly showed frequencies of Type I error close the nominal level 0.05. However, with the parameter settings of the peanut dataset, the observed frequency of Type I error was clearly smaller than 0.05 for $K = 3, 4, \ldots, 7$. The simple parametric bootstrap method is based on an approximation of the distribution of the first K squared singular values (Muirhead 1978). This approximation requires that the null hypothesis is correct and the first K singular values are large. In practice, using the simple parametric bootstrap method, hypotheses should always be tested sequentially. As will be seen in Sect. 6, for the non-scaled peanut dataset, a non-significant result was obtained at $K = 2$. Since, due to a non-significant result, the hypothesis testing procedure is terminated at $K = 2$, the inferior performance with regard to Type I error for $K = 3, 4, \ldots, 7$ is of less concern. For the peanut and sparrows datasets, the simple parametric bootstrap method usually worked well with regard to Type I error rates also when observations were scaled by means. For the fish dataset, however, this was not the case. The simulation study importantly revealed that the simple parametric bootstrap method does not

Table 2. Results of the simulation study of Type I error rates, $\hat{\alpha}$, for the simple parametric bootstrap method, the full parametric bootstrap method and the parallel method.

K	Simple parametric bootstrap			Full parametric bootstrap		Parallel
	Non-scaled	Mean-scaled	SD-scaled	Mean-scaled	SD-scaled	SD-scaled
(a) The peanut dataset						
0	0.050	0.059	0.006	0.049	0.050	0.050
1	0.049	0.053	0.018	0.049	0.045	0.026
2	0.049	0.051	0.034	0.049	0.044	0.002
3	0.045	0.049	0.038	0.049	0.043	0.000
4	0.047	0.043	0.042	0.045	0.046	0.000
5	0.038	0.045	0.044	0.049	0.048	0.000
6	0.038	0.043	0.041	0.045	0.044	0.000
7	0.043	0.044	0.043	0.048	0.046	0.000
(b) The sparrows dataset						
0	0.048	0.049	0.009	0.050	0.050	0.051
1	0.050	0.049	0.058	0.051	0.049	0.000
2	0.050	0.050	0.044	0.050	0.044	0.000
3	0.049	0.049	0.047	0.050	0.045	0.000
(c) The fish dataset						
0	0.051	0.188	0.006	0.021	0.050	0.049
1	0.051	0.095	0.032	0.042	0.045	0.018
2	0.049	0.066	0.048	0.049	0.050	0.001
3	0.049	0.054	0.047	0.050	0.049	0.000
4	0.049	0.051	0.050	0.050	0.050	0.000
5	0.051	0.049	0.049	0.050	0.050	0.000

The null hypothesis $H_0 : m = K$, where m is the unknown true number of principal components, was tested at significance level $\alpha = 0.05$.

work well when variables are standardized to unit variance. In this case, Type I error rates often deviated much from 0.05, especially for $K = 0$.

As a remedy to the problem with the poor performance of the simple parametric bootstrap method when PCA is performed on standardized data, Sect. 3.3 proposed full parametric bootstrap methods. These full parametric bootstrap methods performed much better with regard to Type I error rate than the simple parametric bootstrap method (Table 2). The observed frequency of Type I error was close to 0.05 in most cases, but for data scaled by means, there were some exceptions.

The last column of Table 2 presents observed Type I error rates for the parallel method. This method performed very well with regard to Type I error rate when testing the first principal component ($K = 0$). However, when testing higher-order principal components, the parallel method did not give Type I error rates close to the nominal level 0.05. In these cases, the null hypotheses were never or very rarely rejected.

Table 3 reports Type I error rates for Tracy–Widom methods. The methods V-2001 and V-2007, which simply compare the scaled largest eigenvalue, Eq (8), with the Tracy–Widom distribution of order 1, did not maintain the nominal level 0.05. The methods W-2001 and W2007, using the Johnstone (2001) ad hoc method for standardized data, performed much better, although some deviations from 0.05 were observed, especially for the peanut dataset.

In most cases, the full parametric bootstrap method for scaled data was more powerful than the Tracy–Widom methods (Table 4). The only exceptions were observed for the peanut

Table 3. Results of the simulation study of Type I error rates, $\hat{\alpha}$, for Tracy–Widom methods.

Example dataset	Method			
	V-2001	V-2007	W-2001	W-2007
(a) The peanut dataset	0.000	0.000	0.085	0.081
(b) The sparrows dataset	0.001	0.001	0.038	0.049
(c) The fish dataset	0.000	0.000	0.061	0.063

The null hypothesis $H_0 : m = 0$ was tested at significance level $\alpha = 0.05$.

Table 4. Results of the simulation study of power of Tracy–Widom and full parametric (FP) bootstrap methods.

ψ	Tracy–Widom		FP bootstrap
	W-2001	W-2007	SD-scaled
(a) The peanut dataset			
0.2	0.085	0.082	0.052
0.3	0.094	0.090	0.059
0.4	0.106	0.102	0.086
0.5	0.140	0.134	0.148
0.6	0.198	0.191	0.260
0.7	0.312	0.305	0.431
0.8	0.436	0.426	0.642
0.9	0.576	0.566	0.820
1.0	0.714	0.705	0.931
(b) The sparrows dataset			
0.2	0.081	0.099	0.156
0.3	0.309	0.347	0.626
0.4	0.736	0.768	0.969
0.5	0.967	0.975	1.000
0.6	0.999	0.999	1.000
0.7	1.000	1.000	1.000
0.8	1.000	1.000	1.000
0.9	1.000	1.000	1.000
1.0	1.000	1.000	1.000
(c) The fish dataset			
0.2	0.060	0.063	0.053
0.3	0.071	0.074	0.057
0.4	0.075	0.078	0.086
0.5	0.098	0.101	0.136
0.6	0.130	0.134	0.230
0.7	0.187	0.193	0.361
0.8	0.267	0.274	0.543
0.9	0.344	0.351	0.708
1.0	0.440	0.447	0.841

The null hypothesis $H_0 : m = 0$ was tested at significance level $\alpha = 0.05$ when $m = 1$. The strength of the first principal component was varied through the parameter ψ.

dataset when $\psi \leq 0.4$ and for the fish dataset when $\psi \leq 0.3$. These exceptions should be viewed in relation to the too high Type I error rate for these datasets (Table 3). For the fish dataset, $\psi = 0.9$, the full parametric bootstrap method was twice as powerful as the Tracy–Widom methods.

Table 5. Statistical analysis of the three example datasets.

K	Non-scaled			Mean-scaled			SD-scaled		
	$\hat{\tau}^2_{K+1}$	T	p-value	$\hat{\tau}^2_{K+1}$	T	p-value	$\hat{\tau}^2_{K+1}$	T	p-value
(a) The peanut dataset									
0	11.71	0.48	0.000	3.03	0.54	0.000	49.1	0.37	0.002
1	5.89	0.46	0.001	1.25	0.49	0.000	34.4	0.40	0.006
2	2.69	0.40	0.114	0.58	0.45	0.015	17.6	0.34	0.408
3				0.26	0.36	0.586			
(b) The sparrows dataset									
0	1695	0.86	0.000	0.169	0.73	0.000	174	0.72	0.000
1	222	0.82	0.000	0.037	0.59	0.000	26	0.38	0.185
2	30	0.62	0.000	0.015	0.46	0.283			
3	15	0.80	0.000						
(c) The fish dataset									
0	116,329,200	1.00	0.000	12.18	0.45	0.304	31.0	0.49	0.005
1	5775	0.73	0.000				17.0	0.53	0.049
2	1710	0.78	0.000				7.6	0.51	0.392
3	354	0.75	0.023						
4	106	0.90	0.007						
5	11	0.99	0.006						

Non-scaled observations were analysed using the simple parametric bootstrap method. Observations scaled by means and standard deviations (SD) were analysed using full parametric bootstrap methods. The null hypotheses $H_0 : m = K$, where m is the unknown true number of principal components, were tested using $B = 100{,}000$ bootstrap samples until $K = M - 2$ or a non-significant ($p > 0.05$) result was obtained.

6. ANALYSES OF THE EXAMPLES

Table 5 presents results of analyses using non-scaled data, Eq. (1), data scaled by column means Eq. (2) and data scaled by column standard deviations, Eq. (3). The non-scaled datasets were analysed using the simple parametric bootstrap method, whereas the datasets scaled by means and standard deviations were analysed using the proposed full parametric bootstrap methods, which are recommended based on the simulation study of Sect. 5. Tests were carried out sequentially, for $K = 0, 1, 2, \ldots$, up to $K = M - 2$, where $M = \min(n - 1, p)$, or a non-significant ($p > 0.05$) result was obtained.

For the peanut dataset, at most eight principal components could be tested ($M = 9$). When data were standardized using means, three principal components were significant. However, when observations were not scaled and when observations were scaled using standard deviations, non-significant results were obtained when testing the third component. In these two cases, PC2-by-PC1 biplots would illustrate the variation in all significant principal components. Specifically for environment E09, which was considered in Sect. 2, using observations standardized to zero means and unit variance, Model 6 with $m = 2$ principal components gives the estimate $\bar{y}_9 + \hat{\theta}_{39}\, s_9 = 2.90 + 0.82 \cdot 0.41 = 3.24$ for the yield of genotype G03. Here, $\bar{y}_9$ and s_9 are the mean and standard deviation, respectively, in environment E09, and $\hat{\theta}_{39}$ is the element in the third row and ninth column of $\hat{\mathbf{\Theta}}_2$, as defined in Sect. 3.3.

For the sparrows dataset, $M = 5$. Thus, a maximum of four principal components could be tested. When observations were not scaled, all four principal components were highly significant ($p = 0.000$) as a consequence of the large differences in standard deviation between the five variables. When data were made unitless through division by means, the first two principal components were highly significant ($p = 0.000$), but the third principal component was not ($p = 0.283$). A coefficient of variation biplot (Underhill 1990) would illustrate coefficients of variation and pairwise correlations in the space spanned by these two first principal components, which are significant. When data were scaled to unit variance, only the first principal component was significant ($p = 0.000$). In a biplot, patterns along the second principal component, which was not significant ($p = 0.185$), would not be larger than what could be expected by chance.

For the fish dataset, using no scaling, all principal components were significant. When data were scaled using means, no principal components were significant. When standard deviations were used for scaling, the first principal component was clearly significant ($p = 0.005$), the second barely significant ($p = 0.049$) and the third not significant ($p = 0.392$).

7. DISCUSSION

In this article, we proposed parametric bootstrap methods for testing principal components in PCA when variables are standardized. Our interest in this problem derives from observing that researchers often present summaries of datasets using a few principal components without checking whether these are significant or not and whether the omitted components are indeed negligible. Previously proposed methods for hypothesis testing assumed non-standardized variables, although in practice standardization is usually needed.

An advantage of hypothesis tests, as compared to other approaches, is that p values are provided for each principal component. A significant result indicates incompatibility between the observed data and the model under the null hypothesis (Wasserstein and Lazar 2016); thus, suggesting a model with a more complex interaction would be preferable. By refraining from reporting insignificant components, researchers protect themselves from the risk of publishing random results, i.e. committing Type I errors. However, one should be aware that computation of p values does not account for data-dependent decisions that must be taken before the analysis (Gelman and Loken 2014), such as transformation and standardization of variables, and which variables to include in the analysis. Furthermore, the practice of publishing only significant results causes publication bias (Sterling 1959).

The simple parametric bootstrap method is exact, i.e. has the correct Type I error rate, at testing the first principal component, and almost exact at testing higher-order components. These are also the properties of the "exact" method proposed by Choi et al. (2017). Our research confirmed the good performance of the simple parametric bootstrap method, but using this method, variables must not be standardized to unit variance. This is a major drawback, because in practice such standardization is often needed. Thus, for analysis of standardized data, full parametric bootstrap methods were introduced that are just slightly more complicated than the simple parametric bootstrap method.

The results of the simulation study were similar between the three datasets. However, the full parametric bootstrap method for data scaled by means showed a small Type I error rate, 0.021, at testing the first principal component of the fish dataset. Being the smallest of the three datasets, the fish dataset is the most challenging. When observations were scaled by standard deviations, the full parametric bootstrap method clearly outperformed the simple parametric bootstrap method with regard to Type I error rate. In all datasets, the Tracy–Widom methods intended for non-scaled data, i.e. V-2001 and V-2007, scarcely yielded any significant results when the null hypothesis was true, i.e. the observed Type I error rate was considerably lower than 0.05. The Tracy–Widom methods intended for scaled data, i.e. W-2001 and W-2007, gave Type I error rates ranging from 0.038 to 0.085, whereas the full parametric bootstrap method for data scaled by standard deviations showed the correct Type I error rate, 0.050, in all three datasets. The full parametric bootstrap method was up to twice as powerful as the Tracy–Widom methods. Moreover, the full parametric bootstrap method can be used for testing all principal components, one at the time, whereas the Tracy–Widom methods are defined only for testing the first principal component.

In the simulation study, parallel analysis did not perform well. The reason for this is the following: In the three datasets studied, when the columns were standardized to unit variance, the first squared singular values $\hat{\tau}_1^2$ were quite large in comparison with the following squared singular values. Since for standardized data $\sum_{k=1}^{M} \hat{\tau}_k^2$ is fixed and equals $p(n-1)$, a large $\hat{\tau}_1^2$ implies small $\hat{\tau}_2^2, \hat{\tau}_3^2, \ldots, \hat{\tau}_M^2$. Specifically, these squared singular values $\hat{\tau}_2^2, \hat{\tau}_3^2, \ldots, \hat{\tau}_M^2$ become smaller than would be expected by chance if observations were standard normally distributed. In consequence, using parallel analysis, higher-order principal components typically do not become significant when the first principal component accounts for a large portion of the total variance, as in the three examples.

Johnstone (2001) proposed W, Eq. (11), as a test statistic in PCA, using a scaled Tracy–Widom distribution as an approximate reference. However, since W is not computable from the data only, but is a function of the data and a random vector, W is not a statistic (Shao 2003). Considering the random component involved in the computation of W, the comparatively poor performance of this method with regard to power was not surprising.

The proposed simple and full parametric bootstrap methods are perhaps most useful for comparatively small datasets, such as those encountered in analysis of multi-environment crop variety trials. We tried the methods on several larger datasets, among them the genomic chicken dataset used by Husson et al. (2011). That dataset includes 7407 columns (genes) and 43 rows (chicken). Two problems were encountered when analysing these larger datasets: (i) the proposed methods took much time to complete, due to the many variables and singular value decompositions involved, and (ii) all or almost all principal components became significant, due to high pairwise correlations. Indeed, each null hypothesis requires B (e.g. 100,000) singular value decompositions. However, since nowadays huge computational resources exist, as well as fast algorithms for singular value decomposition, this problem could potentially be overcome.

The simple parametric bootstrap method uses an approximation (Muirhead 1978, p. 23) that improves as the values of the K positive singular values grow. Notably, n and p do not need to be large. This makes the simple parametric bootstrap method work for small datasets, as verified in Sect. 5, as long as hypotheses are tested sequentially. Many other results for

statistical inference on random matrices are asymptotically valid as n and p simultaneously approach infinity, while p/n approaches $\gamma \in (0, \infty)$ (Paul and Aue 2014), but in many applications either n or p or both of them are small.

Forkman and Piepho (2014) investigated power of the simple parametric bootstrap method. The procedure was powerful in datasets of similar sizes as the examples of the present article. However, the simple parametric bootstrap method is sensitive to the assumption of normality (Forkman and Piepho 2015). The full parametric bootstrap methods for standardized data have not been investigated with regard to robustness, but we would expect similar sensitivity to departures from assumptions. For non-standardized data, Malik et al. (2018) proposed nonparametric bootstrap and permutation methods with better performance when data are non-normally distributed. More research is needed on methods for significance testing of principal components in non-normally distributed datasets when variables are standardized.

The main contributions of the present article are: (i) specifications on how to apply the parametric bootstrap methodology to PCA when variables are standardized, (ii) the observation that the simple parametric bootstrap method (Forkman and Piepho 2014) does not work well when variables are standardized and (iii) the simulation study providing information about performance of parametric bootstrap methods, Tracy–Widom methods and parallel analysis.

As a practical conclusion, the full parametric bootstrap methods introduced in Sect. 3.3 are recommended when variables are standardized. The simple parametric bootstrap method (Forkman and Piepho 2014) is recommended when variables are not standardized.

8. SUPPLEMENTARY MATERIALS

R and SAS code for the recommended parametric bootstrap methods is available online. The supplementary materials also present a fourth example dataset, including analysis and simulation of Type I error and power. Results from that simulation study agree well with the results of Sect. 5.

ACKNOWLEDGEMENTS

H.P. Piepho was supported by the German Research Foundation (DFG grant no. PI 377/17-1).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

[Received November 2018. Accepted February 2019. Published Online February 2019.]

REFERENCES

Bai, J., and Ng, S. (2002), "Determining the number of factors in approximate factor models," *Econometrica*, 70, 191–221.

Bro, R., Kjeldahl, K., Smilde, A. K., and Kiers, H. A. L. (2008), "Cross-validation of component models: a critical look at current methods," *Analytical and Bioanalytical Chemistry*, 390, 1241–1251.

Bumpus, H. C. (1899), "The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*," *Biological Lectures*, Marine Biology Laboratory, Woods Hole, 11th lecture, 209–226.

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

Choi, B. Y., Taylor, J., and Tibshirani, R. (2017), "Selecting the number of principal components: estimation of the true rank of a noisy matrix," *The Annals of Statistics*, 45, 2590–2617.

Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D. et al. (2010), "Evaluation of parallel analysis methods for determining the number of factors," *Educational and Psychological Measurement*, 70, 885–901.

Crossa J., Yang, R. C., and Cornelius, P. L. (2004), "Studying crossover genotype x environment interaction using linear-bilinear models and mixed models," *Journal of Agricultural, Biological, and Environmental Statistics*, 9, 362–380.

Crossa, J., Burgueño, J., Autran, D., Vielle-Calzada, J. P., Cornelius, P. L., Garcia, N., Salamanca, F., and Arenas, D. (2005), "Using linear-bilinear models for studying gene expression x teatment interaction in microarray experiments," *Journal of Agricultural, Biological, and Environmental Statistics*, 10, 337–353.

Forkman J. (2015), "A resampling test for principal component analysis of genotype-by-environment interaction," *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 19, 27–33.

Forkman, J., and Piepho H. P. (2014), "Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models," *Biometrics*, 70, 639–647.

Forkman, J., and Piepho H. P. (2015), "Robustness of the simple parametric bootstrap method for the additive main effects and multiplicative interaction (AMMI) model", *Biuletyn Oceny Odmian*, 34, 11–18.

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., and Fralish, J. S. (1995), "Parallel analysis: a method for determining significant principal components," *Journal of Vegetation Science*, 6, 99–106.

Galgani, E., Bocquene, G., Lucon, M., Grzebyk, D., Letrouit E., and Claisse D. (1991), "EROD measurements in fish from the northwest part of France," *Marine Pollution Bulletin*, 22, 494–500.

Gauch, H. G. (1992), *Statistical analysis of regional yield trials: AMMI analysis of factorial designs*, Amsterdam: Elsevier.

Gelman, A., and Loken, E. (2014), "The statistical crisis in science," *American Scientist*, 102, 460–465.

Glorfeld, L. W. (1995), "An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain," *Educational and Psychological Measurement*, 55, 377–393.

Green, S. B., Levy, R., Thompson, M. S., Lu, M., and Lo, W. J. (2012), "A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis," *Educational and Psychological Measurement*, 72, 357–374.

Hoyos-Villegas, V., Wright, E. M., and Kelly, J. D. (2016), "GGE biplot analysis of yield associations with root traits in a mesoamerican bean diversity panel," *Crop Science*, 56, 1081–1094.

Hoff, P. D. (2007), "Model avaraging and dimension selection for the singular value decomposition," *Journal of the American Statistical Association*, 102, 674–685.

Horn, J. L. (1965), "A rationale and test for the number of factors in factor analysis," *Psychometrika*, 30, 179–185.

Husson, F., Lê, S., and Pagès, J. (2011), *Exploratory multivariate analysis by examples using R*, Boca Raton, FL: CRC Press.

Johnson, R. A., and Wichern, D. W. (2007), *Applied multivariate statistical analysis*, 6th ed., Harlow: Pearson Education.

Johnstone, I. M. (2001), "On the distribution of the largest eigenvalue in principal components analysis," *The Annals of Statistics*, 29, 295–327.

— (2007), "High dimensional statistical inference and random matrices," In: M. Sanz-Sol, J. Soria, J. L. Varona, J. Verdera (eds.), *Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006*, Volume 1, p. 307–333, Zürich: The European Mathematical Society.

Jolliffe, I. T. (2002). *Principal component analysis*, 2nd ed., New York: Springer.

Jolliffe, I. T., and Cadima, J. (2016), "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A* 374, 20150202.

Josse, J., van Eeuwijk, F., Piepho H.P., and Denis, J. B. (2014), "Another look at Bayesian analysis of AMMI models for genotype-environment data," *Journal of Agricultural, Biological, and Environmental Statistics*, 19, 240–257.

Josse, J., and Husson, F. (2011), "Selecting the number of components in PCA using cross-validation approximations," *Computational Statistics and Data Analysis*, 56, 1869–1879.

Kang, M. S., Balzarini, M., and Guerra, J. L. L. (2004), "Genotype-by-environment interaction," In: A. M. Saxton (ed.). *Genetic analysis of complex traits using SAS*, p. 69–96, Cary, NC: SAS Institute.

Kaiser, H. F. (1960), "The application of electronic computers to factor analysis," *Educational and Psychological Measurement*, 20, 141–151.

Kollah, B., Ahirwar, U., Mohanty, S. R. (2017), "Elevated carbon dioxide and temperature alters aggregate specific methane consumption in a tropical vertisol", *Journal of Agricultural Science*, 155, 1191–1202.

Kritchman, S., and Nadler, B. (2008), "Determining the number of components in a factor model from limited noisy data," *Chemometrics and Intelligent Laboratory Systems*, 94, 19–32.

Malik, W. A., Hadasch, S., Forkman, J., and Piepho H.P. (2018), "Non-parametric resampling methods for testing multiplicative terms in AMMI and GGE models for multi-environment trials," *Crop Science*, 58, 752–761.

Manly, B. F. J. (1986), *Multivariate statistical methods: a primer*, London: Chapman and Hall.

Marasinghe, M. G. (1985), "Asymptotic tests and Monte-Carlo studies associated with the multiplicative interaction-model," *Communications in Statistics – Theory and Methods*, 14, 2219–2231.

Muirhead, R. J. (1978), "Latent roots and matrix variates: A review of some asymptotic results," *Annals of Statistics*, 6, 5–33.

Muirhead, R. J. (1982), *Aspects of multivariate statistical theory,* New York: Wiley.

North Dakota State University (1997), Information Technology Services, https://www.ndsu.edu/pubweb/~doetkott/introsas/rawdata/bumpus.html (accessed Oct 28, 2018).

Onatski, A. (2009), "Testing hypotheses about the number of factors in large factor models," *Econometrica*, 77, 1447–1479.

Owen, A. B., and Wang, J. (2016), "Bi-cross-validation for factor analysis," *Statistical Science*, 31, 119–139.

Passimier, D., Li, Z., and Yao, J. (2017), "On estimation of the noise variance in high dimensional probabilistic principal component analysis." *Journal of the Royal Statistical Society B*, 79, 51–67.

Patterson, N., Price, A. L., Reich, D. (2006), "Population structure and eigenanalysis." *PLoS Genetics*, 2, 2074–2093.

Paul, D., and Aue, A. (2014), "Random matrix theory in statistics: A review," *Journal of Statistical Planning and Inference*, 150, 1–29.

Peres-Neto, P. R., Jackson, D. A., and Somers, K. M. (2005), "How many principal components? Stopping rules for determining the number of non-trivial axes revisited," *Computational Statistics & Data Analysis*, 49, 974–997.

Perez-Elizalde, S., Jarquin, D., and Crossa J. (2012), "A general Bayesian estimation method of linear-bilinear models applied to plant breeding trials with genotype x environment interaction," *Journal of Agricultural, Biological, and Environmental Statistics*, 17, 15–37.

Ruscio, J., and Roche, B. (2012), "Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure," *Psychological Assessment*, 24, 282–292.

Shao, J. (2003), *Mathematical statistics*, 2nd ed., New York: Springer.

Sobczyk, P., Bogdan, M., and Josse, J. (2017), "Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood," *Journal of Computational and Graphical Statistics*, 26, 826–839.

Sterling, T. D. (1959), "Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa," *Journal of the American Statistical Association*, 54, 30–34.

Underhill, L. G. (1990), "The coefficient of variation biplot," *Journal of Classification*, 7, 241–256.

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's statement on *p*-values: context, process, and purpose," *The American Statistician*, 70, 129–133.

Yan W., and Frgeau-Reid, J. (2018), "Genotype by yield*trait (GYT) biplot: a novel approach for genotype selection based on multiple traits," *Scientific Reports*, 8, 8242.

Yan, W., and Kang, M. S. (2003), *GGE biplot analysis: a graphical tool for breeders, geneticists, and agronomists*, Boca Raton: CRC Press.

Yan, W., and Tinker, N. A. (2006), "Biplot analysis of multi-environment trial data: principles and applications," *Canadian Journal of Plant Science*, 86, 623–645.

Yang, R. C., Crossa, J., Cornelius, P. L., and Burgueño, J. (2009), "Biplot analysis of genotype x environment interaction: proceed with caution," *Crop Science*, 49, 1564–1576.

Yeater, K. M., Duke, S. E., and Riedell, W. E. (2015), "Multivariate analysis: Greater insights into complex systems," *Agronomy Journal*, 107, 799–810.

Yochmowitz, M. G., and Cornell, R. G. (1978), "Stepwise tests for multiplicative components of interaction," *Technometrics*, 20, 79–84.

Zitko, V. (1994), "Principal component analysis in the evaluation of environmental data," *Marine Pollution Bulletin*, 28, 718–722.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.