Research Article

JOURNAL OF CHEMOMETRICS

Received: 5 August 2011, Revised: 7 October 2011, Accepted: 13 October 2011, Published online in Wiley Online Library: 9 December 2011

(wileyonlinelibrary.com) DOI: 10.1002/cem.1411

Tracy–Widom statistic for the largest eigenvalue of autoscaled real matrices

Edoardo Saccenti[a,c]*, Age K. Smilde[a,c], Johan A. Westerhuis[a,c] and Margriet M. W. B. Hendriks[b,c]

Eigenanalysis is common practice in biostatistics, and the largest eigenvalue of a data set contains valuable information about the data. However, to make inferences about the size of the largest eigenvalue, its distribution must be known. Johnstone's theorem states that the largest eigenvalues l_1 of real random covariance matrices are distributed according to the Tracy–Widom distribution of order 1 when properly normalized to $L_1 = \frac{l_1 - \eta_{np}}{\xi_{np}}$, where η_{np} and ξ_{np} are functions of the data matrix dimensions n and p. Very often, data are expressed in terms of correlations (autoscaling) for which case Johnstone's theorem does not work because the normalizing parameters η_{np} and ξ_{np} are not theoretically known. In this paper we propose a semi-empirical method based on test-equating theory to numerically approximate the normalization parameters in the case of autoscaled matrices. This opens the way of making inferences regarding the largest eigenvalue of an autoscaled data set. The method is illustrated by means of application to two real-life data sets. Copyright © 2011 John Wiley & Sons, Ltd.

Supporting information may be found in the online version of this paper.

Keywords: largest eigenvalue; covariance matrix; Tracy–Widom distribution; eigenanalysis; autoscaling

1. INTRODUCTION

Eigenanalysis has always been a hot topic in science. The most striking example probably comes from Quantum Mechanics where the properties of a physical system are modeled by Hermitian operators, that is, complex matrices: the study of the system is reduced to the study of the eigenvalues of such operators.

Eigenanalysis has also a tremendous importance in life sciences where many questions can be phrased in terms of investigating the largest eigenvalue. This requires knowledge about the distributional properties of the largest eigenvalue, which has become available only recently when the results of Tracy and Widom [1–3] provided an analytical expression for the distribution of the largest eigenvalue of the covariance matrix of certain classes of random matrices.

In a fundamental paper titled "On the distribution of the largest eigenvalue in principal components analysis", Johnstone showed that when suitably normalized, the largest eigenvalue of real covariance matrices follows a Tracy–Widom (TW) distribution of order 1 [4]. Johnstone's theorem allows the use of the TW statistic for hypothesis testing.

Recently, a few papers have appeared making direct use of the TW statistic in life sciences applications, and all of them deal with the use of principal component analysis (PCA) applied to (very) large biallelic genetic data sets. The possibility of using the TW statistic to select significant components is clearly indicated in Johnstone's paper; Patterson and coworkers empirically showed [5] the potential of the pairing of PCA and TW statistic to discover population structure. They argued that in a genetic data set, the nonindependency of the entries can induce deviations of the distribution of the largest eigenvalue from the theoretical TW distribution and proposed a strategy to overcome this problem by re-estimating the actual size of the data set. This approach was used also in other subsequent papers [6–8], while other authors do not make use of it [9].

Apart from genetics, many fields of data analysis can benefit from the use of the TW statistic because many problems can be reduced to testing the significance of the largest or of the kth largest eigenvalues. For instance, spectroscopic methods are often used to monitor biochemical processes; if such measurements are taken over time, a matrix $\mathbf{X}$ ($n \times p$) results from p measurements on the process performed at n time points. One of the first questions to ask is whether this matrix contains information other than measurement noise. This can be answered by testing whether the largest eigenvalue of the cross-product matrix $\mathbf{X}^T\mathbf{X}$ is statistically different from the one that can be expected from a matrix containing only measurement error, that is, a random matrix [10].

* *Correspondence to: E. Saccenti, Biosystems Data Analysis Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098XH, Amsterdam, The Netherlands.*
E-mail: e.saccenti@uva.nl

a *E. Saccenti, A. K. Smilde, J. A. Westerhuis*
Biosystems Data Analysis Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands

b *M. M. W. B. Hendriks*
Department of Metabolic Diseases, University Medical Center Utrecht, Utrecht, The Netherlands

c *E. Saccenti, A. K. Smilde, J. A. Westerhuis, M. M. W. B. Hendriks*
Netherlands Metabolomics Centre, Leiden, The Netherlands

In PCA, the problem of the selection of the number of significant components to model a data set often arises and naturally leads to testing the significance of the k-th largest eigenvalues of the data covariance matrix; several approximated methods have been proposed and used in chemometrics [11–13].

The issue whether a data set contains correlations beyond those that can be expected from chance results can also be formulated in terms of the largest eigenvalue. This issue is particularly relevant in functional genomics data where chance results are abundant [14–16] and has repercussions for the subsequent statistical analysis. In a classification context, the question may be whether one should use a method using correlations such as, for instance, partial least squares discriminant analysis [17,18] or a method neglecting correlations such as direct linear discriminant analysis [19] or nearest shrunken centroids (NSC) [20]. Association networks rely on correlations [21] in the data, but are these correlations really informative? As will be shown in this paper, distinguishing *true* correlation from chance correlation can be phrased in terms of testing whether the largest eigenvalue is significant or not.

As Patterson and coworkers already noticed in their paper, the assumptions underlying the theoretical results concerning the distribution of the largest eigenvalue hinder the use of this statistic in many cases of real-life data analysis. This is also true when preprocessing of data is required to answer precise biological questions [22].

Centering and scaling are two of the most used preprocessing techniques and are almost a standard procedure in the exploratory analysis of functional genomics data [24]. Centering converts data to fluctuations around the mean, while scaling divides each variable by a factor, which is different for each variable. Autoscaling—also known as standardization or converting to z-scores—combines centering and scaling using the standard deviation as scaling factor. Stated otherwise, autoscaling permits analyzing the data in terms of correlations.

In this paper, we address the problem of the distribution of the k-largest eigenvalue of sample covariance matrices in the case in which the data matrix is autoscaled. To circumvent the theoretical limitations that hamper the use of the TW statistic for hypothesis testing in this special but often occurring case, we developed a numerical method by using an approach also common in test-equating theory. The performance of the method is illustrated with two real-life data analysis cases dealing respectively with the problem of chance correlations and with the selection of the optimal number of components in PCA.

2. BACKGROUND THEORY

2.1. The Tracy–Widom statistic

Much work has been carried out on describing and understanding the distribution of the largest eigenvalue l_1 of a matrix (l_1 will indicate the largest eigenvalue): we refer the reader to the excellent reviews by Bai [25] and El Karoui [26]. The three key papers by Tracy and Widom [1–3] marked a point towards the definition of the asymptotic distribution for the largest eigenvalue in sample covariance matrices. They found that the distribution

$$F_1(s) = exp\left(-\frac{1}{2}\int_s^{\infty} q(t) + (t-s)q^2(t)\mathrm{d}t\right) \tag{1}$$

is the limiting distribution of the largest eigenvalue of a certain class of $n \times n$ random matrices (the so called Gaussian Orthogonal Ensembles). The function $q(t)$ appearing in (1) is the unique Hastings–McLeod solution [27] of the nonlinear Painlevé differential equation $q''(t) = tq(t) + 2q^3(t)$ satisfying the boundary condition $q(t) \approx Ai(t)$ when $t \to \infty$; $Ai(t)$ is the Airy function [28].

In 2001, Johnstone [4], extending a previous result that Johansson derived for the complex case [29], showed that $F_1(s)$ is the limiting distribution of the largest eigenvalue of a real sample covariance matrix $\mathbf{C} = \mathbf{X}\mathbf{X}^T$ where $\mathbf{X}$ is a matrix whose entries are independent Gaussian variables. This result is summarized by the following fundamental theorem:

Theorem [4]
Let the entries x_{np} of the $n \times p$ matrix $\mathbf{X}$ be *i.i.d.* $N(0, 1)$, and let l_1 be the largest eigenvalue of $\mathbf{C} = \mathbf{X}\mathbf{X}^T$. Define

$$\eta_{np} = \left(\sqrt{n-1} + \sqrt{p}\right)^2 \tag{2}$$

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

If

$$\lim_{n\to\infty} \frac{p}{n} = \gamma \tag{4}$$

where $\gamma \in (0, 1]$, then

$$L_1 = \frac{l_1 - \eta_{np}}{\xi_{np}} \overset{\text{dist}}{\Rightarrow} TW_1 \tag{5}$$

where TW_1 is the TW distribution appearing in the study of the Gaussian Orthogonal Ensemble.

The probability density distribution TW_1 is shown in Figure 1. Johnstone's theorem is illustrated in Figure 2 by comparing the empirical distribution of the largest eigenvalue l_1 from a set of covariance matrices $\mathbf{C} = \mathbf{X}\mathbf{X}^T$, with the TW probability density distribution TW_1.

The effect of the normalization of l_1 by means of Johnstone's normalization parameters (5) η_{np} and ξ_{np} given by Equations (2) and (3) is shown in the right panel.

Johnstone's theorem has been followed by many universality results: relaxing of the normality constraint of the entries of $\mathbf{X}$ [30], different asymptotic behavior of the ratio $\frac{n}{p}$ (and $\frac{p}{n}$) when $n, p \to \infty$ [31,32], and nonidentity of the covariance matrix Σ [33].

2.2. Hypothesis testing with the Tracy–Widom statistic

The mathematical machinery developed in Johnstone's theorem allows for the use of the TW statistic in hypothesis testing: we will illustrate this by making use of an example inspired by Johnstone [34].

$$\mathbf{X} = \begin{pmatrix} 1.8638 & -0.0990 & 0.6065 & 0.8263 & 0.1963 & -1.5907 & -1.4364 & -0.7904 & 0.5238 & 1.3980 \\ 0.0829 & -1.6793 & 1.2225 & -0.6431 & -0.3942 & 1.6180 & -1.9627 & 1.1654 & 1.4753 & 1.4407 \\ 0.3674 & 1.0418 & -0.8928 & -1.1454 & -0.4583 & 1.0652 & -0.4120 & 0.4068 & 1.1628 & 0.1429 \\ -0.1312 & 0.0305 & 0.8269 & 0.6914 & -1.2540 & -0.0440 & -1.3624 & -0.0487 & 0.8703 & -0.4513 \\ -0.3735 & -0.6685 & 0.9230 & -0.0888 & -0.2264 & 0.0387 & -1.8868 & -0.3596 & 1.0525 & 1.1325 \\ 0.0080 & 1.3835 & -1.1887 & 2.0329 & -1.0426 & -0.0509 & -1.4490 & -0.4949 & 2.1055 & -1.0170 \\ 0.2980 & -0.4369 & 1.2984 & -0.9405 & -0.5406 & 0.3133 & 1.1276 & -0.5522 & -1.0429 & 0.3806 \\ -2.2005 & 1.1773 & -0.0524 & -1.3260 & 0.3835 & -0.8341 & 1.9981 & -0.0193 & -1.4198 & -1.3279 \\ 0.2157 & -1.0029 & 0.1521 & -0.5521 & 1.3878 & -1.0087 & 1.5635 & 0.5894 & 0.4116 & 0.6707 \\ 1.8281 & 0.9610 & -1.3584 & -1.1463 & 0.8328 & -0.6450 & 0.5721 & 0.9095 & -0.4085 & -0.7949 \end{pmatrix}$$

Let us consider a 10×10 matrix **X**, describing 10 observations of 10 variables, which are *i.i.d.* $N(0, \Sigma)$;
The largest eigenvalue l_1 of $\mathbf{X}\mathbf{X}^T$ is $l_1 = 43.64$. Is this value consistent with an identity covariance matrix Σ=**I**? In other words, the null hypothesis

$$H_0 : \Sigma = \mathbf{I} \tag{6}$$

needs to be tested against the alternative

$$H_1 : \Sigma \neq \mathbf{I} \tag{7}$$

This can be accomplished by using the TW statistic and the results of Johnstone's theorem. The largest eigenvalue l_1 needs to be rescaled to L_1 using the moments η_{np} and ξ_{np} before being confronted with the reference TW distribution:

$$\eta_{10,10} = \left(\sqrt{10-1} + \sqrt{10}\right)^2 = 37.97 \tag{8}$$

$$\xi_{10,10} = \left(\sqrt{10-1} + \sqrt{10}\right)\left(\frac{1}{\sqrt{10-1}} + \frac{1}{\sqrt{10}}\right)^{\frac{1}{3}} = 5.34 \tag{9}$$

The TW statistic L_1 for l_1 is then

$$L_1 = \frac{l_1 - 37.97}{5.34} = 1.062 \tag{10}$$

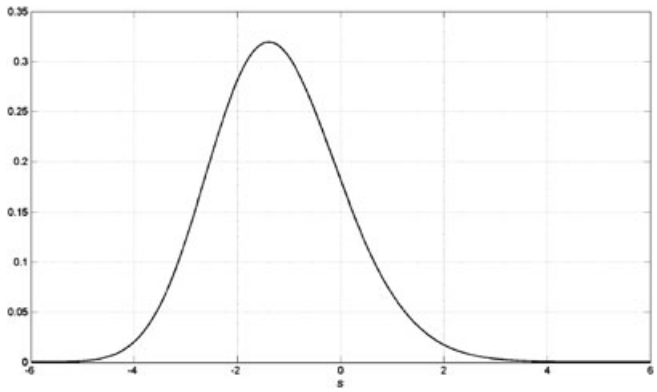

Figure 1. Tracy–Widom probability density function $TW_1 = \frac{d}{ds}F_1(s)$. Mean is $-\mathbf{1.206}\ldots$ and variance 1.607 . . .. Percentiles can be found in [23].

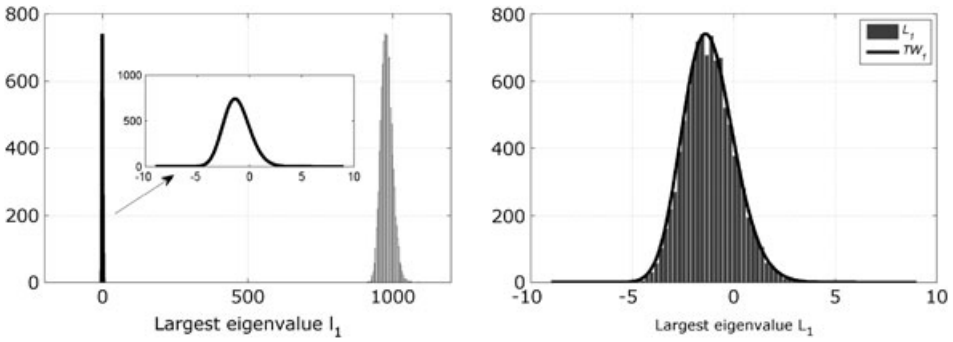

Figure 2. Graphical illustration of Johnstone's theorem. The left panel shows the theoretical Tracy–Widom distribution (solid curve) in comparison with the experimental distribution (histogram) of the largest eigenvalue l_1 derived from 10^4 300×300 random matrices whose entries are *i.i.d.* $N(0, 1)$. The right panel shows the distribution of the largest eigenvalue $L_1 = \frac{l_1 - \eta_{np}}{\xi_{np}}$, normalized according Johnstone's Theorem, together with the Tracy–Widom distribution.

By looking at *p*-values look-up tables as in any conventional test, it can be found that the null hypothesis cannot be rejected at the 0.01 level; *p*-values for the TW distribution can be found in [23]. Values for $\alpha = 0.05$ and 0.01 are given in Table I.

2.3. Autoscaling

In data analysis, manipulations of data matrices, such as centering or scaling, are common practice. These methods are commonly used to correct for aspects that can hinder the interpretation of data [24,22]. Autoscaling is known to have some undesirable properties such as inflating of measurement noise or introducing interdependencies among the entries of the data matrix: these aspects have been addressed elsewhere [24,35–38]. For what concerns the effects of autoscaling on the distributional properties of the eigenvalues, it is known since long time that scaling affects the distribution of the largest eigenvalue [39]. In this framework, Johnstone's theorem cannot be applied straightforwardly because the TW distribution does not hold anymore as shown in Figure 3. In every day biostatistics, it is often desired to have variables mean centered and standardized to unit variance. In the field of psychometrics, this is called *standardization*, while in chemometrics, it is commonly referred to as *autoscaling*. It is usually applied when the goal is to compare variables on the basis of correlation. After autoscaling, all variables become equally important: they are converted to fluctuations around zero instead of around the column mean, and their variance is unitary. Autoscaling the matrix **X** alters the distribution of the largest eigenvalues l_1 of the covariance matrix: this means that after rescaling l_1 to L_1 according to Johnstone's theorem, the largest eigenvalue is not distributed as TW_1 as shown in Figure 3. This fact practically impairs the use of the TW statistic for hypothesis testing where autoscaling of the matrix **X** is required before computing $\mathbf{C} = \mathbf{X}\mathbf{X}^T$.

The alternative hypothesis (7) deserves now special attention. For autoscaled data, diagonal elements of Σ are always one because autoscaling transforms covariances in correlations. Hence, finding an eigenvalue significantly larger than one (i.e. by rejecting H_0) necessarily points to an at least one nonzero off-diagonal element of Σ (see Appendix for a proof). The use of the TW statistic on testing for correlation was already suggested by Tracy and Widom in [40] but worked out in this paper.

To make use of the TW statistic in the case of autoscaled data, it would then be necessary to know the analytical expression of η_{np} and ξ_{np}. According to Péché [32], theoretically derived analytical expressions for the normalization parameters η_{np} and ξ_{np} for the case of mean-centered matrices are likely to appear soon; nevertheless, closed expressions for these parameters in the case of mean-centered or autoscaled matrices are not currently available. This problem can be tackled and bypassed by means of an empirical approach.

Table 1. Tracy–Widom statistic percentiles for the $\alpha = 0.05$ and $\alpha = 0.01$ confidence threshold for the *k*th largest eigenvalue of real covariance matrices

k	$\alpha = 0.05$	$\alpha = 0.01$
1	0.9793	2.0234
2	−1.5420	−0.7703
3	−3.3001	−2.6341
4	−4.7609	−4.1554

3. DISTRIBUTION MOMENTS FOR AUTOSCALED MATRICES

Given a data matrix **X** and its autoscaled version $\widetilde{\mathbf{X}}$, we aim to devise an empirical strategy to rescale the largest eigenvalue $\tilde{l}_1$ of the covariance matrix $\widetilde{\mathbf{C}} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ so that $\tilde{l}_1$ is approximately distributed like the TW distribution under the H_0 hypothesis (H_0: Σ=**I**). We indicate with $D = D\left(\tilde{l}_1\right)$ the experimental

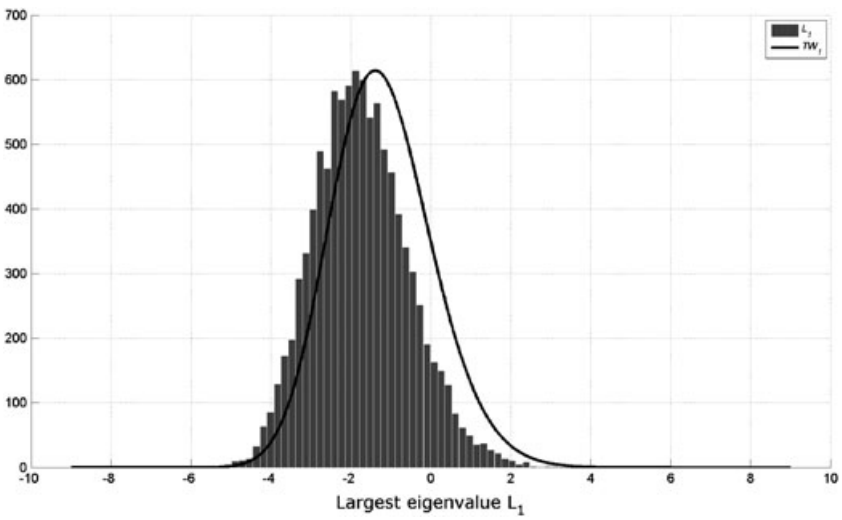

Figure 3. Comparison between the theoretical Tracy–Widom probability density $TW_1(s)$ with the experimental distribution $D(\tilde{L}_1)$ for the largest eigenvalue in the case of the covariance matrix of autoscaled matrices. Eigenvalues have been rescaled according to the parameters η_{np} and ξ_{np} as defined by Johnstone's theorem in Equations (2) and (3). The distribution has been generated from 10^4 300×300 matrices whose entries are *i.i.d.* $N(\mu_i, \sigma_i)$ before autoscaling.

probability distribution function for the largest eigenvalue $\tilde{l}_1$ of the autoscaled covariance matrix $\tilde{\mathbf{C}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ when $\mathbf{X}$ is autoscaled; it can be constructed by generating a set of random matrices *i.i.d.* $N(0,1)$, autoscaling them and collecting the largest eigenvalue $\tilde{l}_1$.

From the simulations shown in Figure 3, it is clear that autoscaling the matrix $\mathbf{X}$ introduces a shift and a shrinkage in the distribution of the largest eigenvalue of the autoscaled covariance matrix $\tilde{\mathbf{C}}$ after correction by means of Johnstone's η_{np} and ξ_{np} parameters; as shown in Figure 3, the difference between $D\left(\tilde{l}_1\right)$ and TW_1 is remarkably big.

From numerical evidence and from the results proven by the universality theorems [30–33,41,42], we speculated that $\tilde{l}_1$ still has a simple relation with the TW distribution and that the shift and the shrinking can be adjusted by assuming that normalization parameters $\tilde{\eta}_{np}$ and $\tilde{\xi}_{np}$ exist such that

$$\tilde{L}_1 = \frac{\tilde{l}_1 - \tilde{\eta}_{np}}{\tilde{\xi}_{np}} \tag{11}$$

follows the TW distribution, with $\tilde{l}_1$ the largest eigenvalue of the covariance matrix $\tilde{\mathbf{C}}$. By setting

$$a = \frac{1}{\tilde{\xi}_{np}} \qquad b = -\frac{\tilde{\eta}_{np}}{\tilde{\xi}_{np}} \tag{12}$$

Equation (11) is equivalent to assume that a linear function e_{TW} exist such as

$$e_{TW}(D) = TW_1 \tag{13}$$

where

$$e_{TW}(D) = aD + b. \tag{14}$$

With the writing $aD + b$, we indicate that each value x in the distribution D is transformed to $ax + b$.

To estimate a and b, we follow a procedure analogous to equipercentile equating, and the function e_{TW} can be regarded as an equipercentile equating function [43].

The parameters a and b can be theoretically derived [44,45] and are given by

$$a = \frac{s_{TW}}{s_D} \tag{15}$$

$$b = m_{TW} - \frac{s_{TW}}{s_D} m_D \tag{16}$$

where m_{TW}, m_D, s^2_{TW}, and s^2_D are respectively the means and variances of TW_1 and D. Combining Equations (14), (15), and (16), we obtain the well-known formula for the linear equating function [44]:

$$TW_1 = e_{TW}(D) = m_{TW} + \frac{s_{TW}}{s_D}(D - m_D) \tag{17}$$

The mean and the variance of the TW distribution have been calculated [46] and are known:

$$m_{TW} = -1.206(5335745820)\ \ s^2_{TW} = 1.607(781034581) \tag{18}$$

The mean m_D is trivial to compute, while the variance s^2_D can be numerically calculated as the second centered moment of the distribution D:

$$s^2_D = \int \left(\tilde{l} - m_D\right)^2 D\left(\tilde{l}\right) \mathrm{d}\tilde{l} \tag{19}$$

Because D is a discrete distribution, the integral in (19) needs to be replaced by a finite sum running over all the elements in D. Comparing (12) with Equations (15) and (16), it is easy to express $\tilde{\eta}_{np}$ and $\tilde{\xi}_{np}$ as functions of m_{TW}, m_D, s^2_{TW}, and s^2_D:

$$\tilde{\eta}_{np} = m_D - \frac{s_D}{s_{TW}} m_{TW} \tag{20}$$

$$\tilde{\xi}_{np} = \frac{s_D}{s_{TW}} \tag{21}$$

In practice, the first step is to build the experimental distribution $D\left(\tilde{l}_1\right)$. Given n and p, $N = 10^4$ $n \times p$ random matrices *i.i.d.* $N(\mu_i, \sigma_i)$ (the index i indicating that μ and σ can differ for each column of $\mathbf{X}$) have been generated and autoscaled, and the largest eigenvalues l_1 have been calculated and collected, obtaining a distribution of $\tilde{l}_1$ largest eigenvalues. The largest eigenvalues $\tilde{l}_1$ are then rescaled to $\tilde{L}_1$ using the values of $\tilde{\eta}_{np}$ and $\tilde{\xi}_{np}$ calculated by means of Equation (11). The distribution of the $\tilde{L}_1$ values is actually $e_{TW}\left(D\left(\tilde{l}_1\right)\right) = TW_1$. The equivalence of the empirical $e_{TW}\left(D\left(\tilde{l}_1\right)\right)$ with TW distribution has been assessed by means of a Kolmogorov–Smirnov test [47], using a reference TW distribution generated by means of RMTstat (see Section 3.1). The p-values for all tests were $< 10^{-5}$. Figure 4 shows the agreement between the distribution of $\tilde{L}_1 = \frac{\tilde{l}_1 - \tilde{\eta}_{np}}{\tilde{\xi}_{np}}$ in the case of 300×300 matrices and the theoretical TW law of order 1.

For large values of n and p, this procedure can be computationally intensive, but it has the advantage that normalization parameters need to be calculated only once as results can be stored for subsequent use. For this reason, we limited ourselves to data matrix dimensions that are likely to occur in real-life practice ($10 \le n, p \le 200$). We have summarized our results in two reference tables (for values of $\tilde{\eta}_{np}$ and $\tilde{\xi}_{np}$), which are available as Supporting Information as well as for download at the author's website www.bdagroup.nl. Values for $n = 201$ and $n = 590$ have been also generated for the purpose of illustration of the method (see Sections 4.1 and 4.2). More comprehensive tables for larger values of n and p will be released in the future.

This method can be easily generalized to the 2nd, 3rd, ... kth largest eigenvalue for which an appropriate TW distribution of order 1 can be derived, thanks to a recurrence property as demonstrated in [48]. To extend the proposed method to the kth eigenvalue is only necessary to use the correct values for the mean and the variance of the TW distribution for the kth largest eigenvalue in Equation (18). Results for the first four largest eigenvalues are shown in Figure 5. It must be borne in mind that the TW limit is an asymptotic result that holds for $n, p \to \infty$: nevertheless, it has been shown that Johnstone's theorem holds also for moderate and even small finite values of n and p.

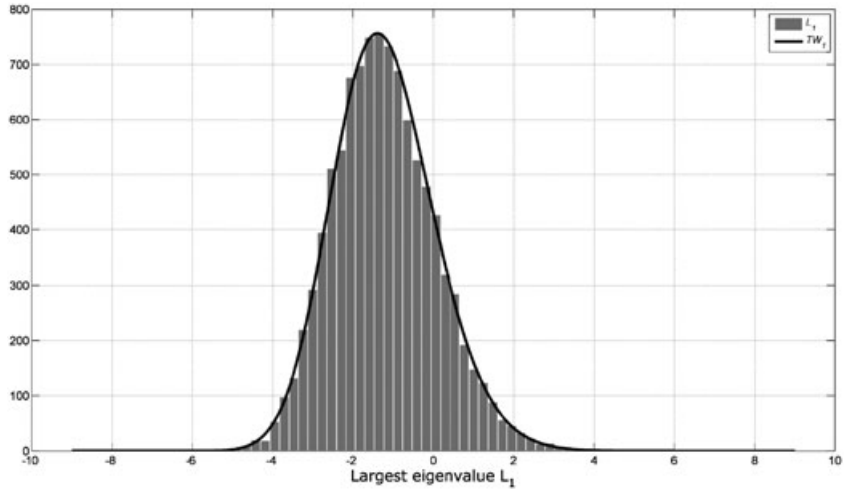

Figure 4. Comparison of the theoretical Tracy–Widom probability density TW_1 with the experimental distribution of $\tilde{L}_1$ for the largest eigenvalue in the case of the covariance matrix of autoscaled real matrices. Eigenvalues have been rescaled according the parameters $\tilde{\eta}_{np}$ and $\tilde{\xi}_{np}$ as numerically estimated according to Equation (20). The distribution $D(\tilde{L}_1)$ has been generated from 10^4 300×300 matrices whose entries are *i.i.d.* $N(\mu_i, \sigma_i)$ before autoscaling.

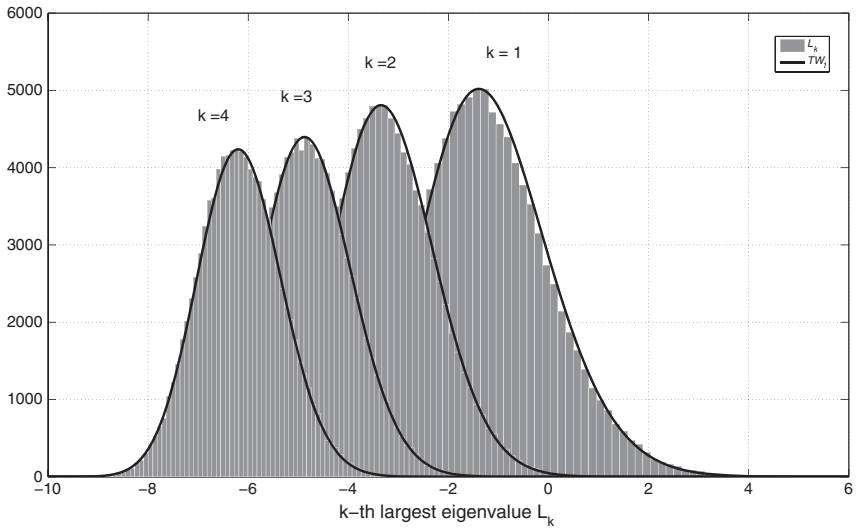

Figure 5. Comparison of the theoretical Tracy–Widom probability density TW_1 with the experimental distribution of $\tilde{L}_1$ for the kth largest eigenvalue in the case of the covariance matrix of autoscaled real matrices. Eigenvalues have been rescaled according the parameters $\tilde{\eta}_{np}$ and $\tilde{\xi}_{np}$ as numerically estimated according to Equation (20).

3.1. Software

Simulations have been carried out in the MATLAB environment (The MathWorks, Natick, MA, USA). Random matrices have been generated by means of the built-in function randn.

TW distributions have been generated by means of the Matlab Toolbox RMLab002 by M. Dieng [48], available via the web at the address http://math.arizona.edu /∼momar/research.htm.

Standard tail *p*-values of the TW law have been derived from the statistical look-up tables in [23].

Tracy–Widom distributed random variables have been generated by means of the R software package RMTstat by Johnstone, Ma, Perry, and Shahram available at http://cran.r-project.org/web/ packages/RMTstat.

Calculations have been performed on the LISA-SARA supercomputer facility: www.sara.nl.

4. EXAMPLES

4.1. Testing for correlations

Correlations play a crucial role in data analysis and a prior knowledge about the correlation structure of a data set can aid and guide towards a more efficient statistical analysis because the performance of a statistical method ultimately depends on the nature of the data set. It has been shown [49] that different methods applied on the same data set lead to different results, but deciding when and which method to apply is not a trivial task, and this choice mostly relies on the experience of the data analyst. If a significant correlation structure is expected to be present in the data, a method which uses the correlation information such as principal component discriminant analysis (PC-DA) should be preferred to methods such as NSC, which do not use

this kind of information. However, it must be taken into account that chance correlations are likely to appear, and its occurrence increases when the number of observations is small compared with the number of variables [14], as what almost usually happens in the case of functional genomics data sets. The question is then whether in a data set correlations are present other than those that can be expected by chance, that is, whether the correlation structure is significant. This is equivalent to test if the structure of the covariance matrix of the data is consistent with an identity covariance matrix, and it reduces to testing the null hypothesis as explained in Section 2. Autoscaling is usually applied when analyzing a data set on the basis of correlations, and the problem of assessing whether correlations in a data set are significant fits naturally in the framework we have developed for the use of TW statistic in the case of autoscaled matrices.

As a test example, we choose a data set that contains serum samples of 19 Gaucher patients and 20 controls. Serum samples were surveyed for basic proteins using surface-enhanced laser desorption/ionization–time of flight mass spectroscopy resulting in 590 variables: the final data matrix dimension is 39×590. Details about the data set generation can be found in [49,50].

The largest eigenvalue l_1 of the covariance matrix of the autoscaled Gaucher data set is $l_1 = 2345.9$, and the calculated normalization moments are $\tilde{\eta} = 924.2$ and $\tilde{\xi} = 17.7$. The TW statistic for this data set, calculated according to Equation (11), is $L_1 = 80.3$, which corresponds to a p-value $\ll 0.01$: the null hypothesis can be rejected, indicating that the correlation structure of the data is genuine and significant.

On the basis of this, it can be expected that PC-DA performs better than NSC on this data set when the two methods are applied with the purpose of class discrimination.

The optimal performance of the PC-DA for this data set was assessed in a previous work against five different methods in a joint effort of several research groups [49]. The analysis was carried out in a double cross-validation scheme. The cross-validation data split was kept the same for both methods so that the results could be directly compared. PC-DA lead to an error rate of 10% (four samples misclassified), while the error rate for the NSC was 15% (six samples misclassified). PC-DA clearly outperforms NSC. This trial and error approach cannot of course routinely be applied, so the use of the TW statistic is a valuable screening tool to aid in the choice of a statistical method among many others.

4.2. Application to principal component analysis

One of the more critical steps in PCA is the selection of the optimal number of principal components to build a model. In the PCA context, optimality means a model, which accounts for systematic variations in the data but (preferably) not the noise: noise can be defined as any variation, which is not correlated with any other variation in the data. In practice, the search for optimality reduces to finding the number of components for which adding more components does not provide a better description of the data not previously included in the model [51].

In PCA, the original variables are linearly combined to form orthogonal vectors (principal components) capturing maximum variance in the data. The coefficients of the kth principal component are the elements of the kth eigenvector of the covariance matrix of the data. The eigenvectors are ranked according to the rank of the eigenvalues, with the first eigenvector (and then the first principal component) corresponding to the largest eigenvalue and so on. It follows that the significance of the kth principal component is driven by the significance of the kth eigenvalue. The TW statistic can be applied to test the significance of the kth eigenvalue of the covariance matrix thus to test the significance of the kth component.

As principal components are sensitive to the variance of the original variables, data is usually autoscaled before PCA analysis to ensure that all variables have zero mean and the same variance: the problem is reduced to testing the significance of the kth largest eigenvalue of autoscaled real covariance matrices and can be addressed by means of the proposed TW statistic.

For the purpose of illustration, the method will be applied on a data set of UV–visible spectra of the two-step chemical reaction of 3-chlorophenylhydrazonopropane dinitrile with 2-mercaptoethanol to form 3-chlorophenyl-hydrazonocyanoacetamide and the byproduct ethylene sulfide [52]. The reduced data set consists of UV–visible spectra (wavelength $\times$ time) and has dimensions 34×201. The four largest eigenvalues are

$$l_1 = 4071.3, \quad l_2 = 2008.8, \quad l_3 = 408.2, \quad \text{and} \quad l_4 = 33.4$$

The calculated $\left(\tilde{\eta}, \tilde{\xi}\right)_{l_k}$ moments for the kth largest eigenvalue l_k are

$$(384.4, 10.3)_{l_1}, (380.0, 9.62)_{l_2}, (375.3, 9.17)_{l_3}, \text{ and } (370.1, 8.74)_{l_4}$$

The corresponding TW statistics L_k are

$$L_1 = 357.9, L_2 = 169.3, L_3 = 3.6, \text{ and } L_4 = -38.5$$

When confronted with the critical values for the 0.05 and 0.01 confidence threshold (see Table I), it appears that only the first three largest eigenvalue are significant so that only the first three principal components should be retained for model building.

The same data set was used in a previous study [51] to investigate the performance of different cross-validation [53] approaches to identify the number of principal components that best describes systematic variations in the data. Although the results of the different methods were not univocal, it was concluded that as many as three components were to be included in the model. This finding is strongly supported by the statistical significance of the first three components as assessed by the proposed TW statistic and is also confirmed by the inherent chemistry of the process in which three compounds are involved.

5. CONCLUSIONS

We developed a numerical method using an approach also common in test-equating theory to approximate the normalization parameters $\tilde{\eta}_{np}$ and $\tilde{\xi}_{np}$ in the case of autoscaled data. We are thereby introducing a formal test statistic for the largest eigenvalue of a covariance matrix using results from modern statistics. We showed the performance of the method on two real-life very important applications: testing on the existence of

significant correlation in a data set (which is very useful for functional genomics applications) and testing for the significance of principal components in PCA modeling.

Tables with the calculated values for $\tilde{\eta}_{np}$ and $\tilde{\xi}_{np}$ for a wide range of values of n and p, likely to appear in real life data analysis, are available as Supporting Information as well as for download at the author's website www.bdagroup.nl.

Acknowledgement

This project was financed by the Netherlands Metabolomics Centre (NMC) which is a part of the Netherlands Genomics Initiative/ Netherlands Organisation for Scientific Research.

REFERENCES

1. Tracy C, Widom H, Level-spacing distributions and the Airy kernel. *Phys. Lett. B* 1993; **305**: 115–118.
2. Tracy C, Widom H. Level-spacing distributions and the Airy kernel. *Commun. Math. Phys.* 1994; **159**: 151–174.
3. Tracy C, Widom H. On orthogonal and symplectic matrix ensembles. *Commun. Math. Phys.* 1996; **177**: 727–754.
4. Johnstone I. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* 2001; **29**: 295–327.
5. Patterson N, Price A, Reich D. Population structure and eigenanalysis. *PLoS Genet.* 2006; **2**: e190.
6. Bilal E, Rabadan R, Alexe G, Fuku N, Ueno H, Nishigaki Y, Fujita Y, Ito M, Arai Y, Hirose N, *et al*. Mitochondrial DNA haplogroup D4a is a marker for extreme longevity in Japan. *PLoS One* 2008; **3**: 2421.
7. Chih L, Ali A, Chun-Hsi H. PCA-based population structure inference with generic clustering algorithms. *BMC Bioinf.* 2009; **10** (Suppl 1): S73.
8. Kong X, Cho M, Anderson W, Coxson H, Muller N, Washko G, Hoffman E, Bakke P, Gulsvik A, Lomas D, *et al*. Genome-wide association study identifies BICD1 as a susceptibility gene for emphysema. *Am. J. Respir. Crit. Care Med.* 2011; **183**: 43–49.
9. Zhu C, Yu J. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics* 2009; **182**: 875.
10. Esteves da Silva J, Tavares M, Tauler R. Multivariate curve resolution of multidimensional excitation–emission quenching matrices of a Laurentian soil fulvic acid. *Chemosphere* 2006; **64**: 1939–1948.
11. Faber K, Kowalski B. Modification of Malinowski's f-test for abstract factor analysis applied to the Quail Roost II data sets. *J. Chemom.* 1997; **11**: 53–72.
12. Malinowski E. Adaptation of the Vogt–Mizaikoff f-test to determine the number of principal factors responsible for a data matrix and comparison with other popular methods. *J. Chemom.* 2004; **18**: 387–392.
13. Malinowski E. Determination of rank by median absolute deviation (DRMAD): a simple method for determining the number of principal factors responsible for a data matrix. *J. Chemom.* 2009; **23**: 1–6.
14. Topliss J, Costello R. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* 1972; **15**: 1066–1068.
15. Camacho D, de la Fuente A, Mendes P. The origin of correlations in metabolomics data. *Metabolomics* 2005; **1**: 53–63.
16. Steuer R. On the analysis and interpretation of correlations in metabolomic data. *Brief. Bioinform.* 2006; **7**: 151–158.
17. Wold H. *Estimation of Principal Components and Related Models by Iterative Least Squares*. Academic Press: NY, 1966; 391–420.
18. Barker M, Rayens W. Partial least squares for discrimination. *J. Chemom.* 2003; **17**: 166–173.
19. Guo Y, Hastie T, Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. *Bisostatistics* 2007; **8**: 86.
20. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.* 2003; **18**: 104–117.
21. Bumgarner R, Yeung K. Methods for the inference of biological pathways and networks. *Methods Mol Biol* 2009; **541**: 225–45.
22. van den Berg R, Hoefsloot H, Westerhuis J, Smilde A, van der Werf M. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 2006; **7**: 1471–2164.
23. Bejan A. Largest eigenvalues and sample covariance matrices. Tracy–Widom and Painleve II: computational aspects and realization in S-Plus with applications. Preprint: http://www.vitrum.md/andrew/MScWrwck/TWinSplus.pdf (2005).
24. Bro R, Smilde A. Centering and scaling in component analysis. *J. Chemom.* 2003; **17**: 16–33.
25. Bai Z. Methodologies in spectral analysis of large-dimensional random matrices, a review. *Stat. Sinica* 1999; **9**: 611–677.
26. El Karoui N. Recent results about the largest eigenvalue of random covariance matrices and statistical application. *Acta Phys. Pol., B* 2005; **36**: 2681.
27. Hastings S, McLeod J. A boundary value problem associated with the second Painleve transcendent and the Korteweg-de Vries equation. *Arch. Ration. Mech. An.* 1980; **73**: 31–51.
28. Airy G. On the intensity of light in the neighbourhood of a caustic. *Trans. Cambr. Phil. Soc.* 1838; 379–402.
29. Johansson K. Shape fluctuations and random matrices. *Commun. Math. Phys.* 2000; **209**: 437–476.
30. Soshnikov A. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *J. Stat. Phys.* 2002; **108**: 1033–1056.
31. El Karoui N. On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity. Arxiv preprint math/0309355 2003.
32. Péché S. Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probab Theor Relat Field* 2009; **143**: 481–516.
33. Baik J, Ben Arous G, Péché S. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* 2005; **33**: 1643–1697.
34. Johnstone I. High dimensional statistical inference and random matrices. Arxiv preprint math. ST/0611589 2006.
35. Paatero P, Hopke P. Discarding or downweighting high-noise variables in factor analytic models. *Anal. Chim. Acta* 2003; **490**: 277–289.
36. Khalheim O. Scaling of analytical data. *Anal. Chim. Acta* 1985; **177**: 71–79.
37. Deming S, Palasota J, Nocerino J. The geometry of multivariate object preprocessing. *J. Chemom.* 1993; **7**: 393–425.
38. Ivosev G, Burton L, Bonner R. Dimensionality reduction and visualization in principal component analysis. *Anal. Chem.* 2008; **80**: 4933–4944.
39. Anderson T. Asymptotic theory for principal component analysis. *Ann. Math. Stat.* 1963; **34**: 122–148.
40. Tracy C, Widom H. Distribution functions for largest eigenvalues and their applications. *Proceedings of the ICM* 2002; **1**: 587–596.
41. El Karoui N. A rate of convergence result for the largest eigenvalue of complex white Wishart matrices. *Ann. Probab.* 2006; **34**: 2077–2117.
42. El Karoui N. Tracy–Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Ann. Probab.* 2007; **35**: 663–714.
43. Dorans N. Equating methods and sampling designs. *Ethics & Behavior* 1990; **3**: 3–17.
44. Holland P, Rubin D. *Test Equating*. Academic Press: New York, 1982.
45. Kolen M, Brennan R. *Test Equating: Methods and Practices*. Springer: New York, 1995.
46. Bornemann F. On the numerical evaluation of distributions in random matrix theory: a review with an invitation to experimental mathematics. arXiv:0904.1581v4 [math.PR] 2009.
47. D'Agostino R, Stephens M. *Goodness-of-fit Techniques*. CRC: New York, 1986.
48. Dieng M. Distribution functions for edge eigenvalues in orthogonal and symplectic ensembles: Painlevé representations. *Int. Math. Res. Notices* 2005; **2005**: 2263.
49. Hendriks M, Smit S, Akkermans W, Reijmers T, Eilers P, Hoefsloot H, Rubingh C, de Koster C, Aerts J, Smilde A. How to distinguish healthy from diseased? Classification strategy for mass spectrometry-based clinical proteomics. *Proteomics* 2007; **7**: 3672–3680.
50. Smit S, van Breemen M, Hoefsloot H, Smilde A, Aerts J, de Koster C. Assessing the statistical validity of proteomics based biomarkers. *Anal. Chim. Acta* 2007; **592**: 210–217.
51. Bro R, Kjeldahl K, Smilde A, Kiers H. Cross-validation of component models: a critical look at current methods. *Anal. Bioanal. Chem.* 2008; **390**: 1241–1251.
52. Bijlsma S, Boelens H, Smilde A. Determination of rate constants in second-order kinetics using UV-visible spectroscopy. *Appl. Spectrosc.* 2001; **55**: 77–83.
53. Wold S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* 1978; **20**: 397–405.

Appendix A

In this appendix, we show how the alternative hypothesis ($H_1 : \Sigma \neq \mathbf{I}$) implies the existence of correlations (see Section 2.2).

Let Σ be a $p \times p$ correlation matrix. As data are autoscaled, it holds that $\mathrm{diag}(\Sigma) = 1$, where $\mathbf{1}$ is a $p \times 1$ with 1 as elements.

It holds that

$$\mathrm{tr}(\Sigma) = \sum_{i=1}^{p} \lambda_i = p \tag{22}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ are the p eigenvalues of Σ.

If there are no correlations, $\Sigma = \mathbf{I}_p$, where $\mathbf{I}_p$ is a $p \times p$ identity matrix. This implies that all the eigenvalues of Σ are equal to 1: $\lambda_1 = \lambda_2 = \ldots = \lambda_p = 1$.

Let us suppose that at least one of the eigenvalues of Σ is different from 1. Without loss of generality, we can assume $\lambda_1 > 1$. The fact that $\lambda_1 > 1$ implies that

$$\Sigma \neq \mathbf{I}_p \tag{23}$$

As autoscaling imposes that $\mathrm{diag}(\Sigma) = 1$, the inequality (22) can be satisfied if and only if at least two nonzero off-diagonal elements of Σ do exist. This fact implies the existence of correlations and can be illustrated by a small example.

Consider

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & 0 \\ \rho & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

with ρ a nonzero real number. The eigenvalues of Σ are $\lambda = 1 - \rho$ and $\lambda = 1 + \rho$ with multiplicity 1 and $\lambda = 1$ with multiplicity $p - 2$. If $\rho > 0$ then $\lambda = 1 + \rho > 1$; otherwise, if $\rho < 0$, then $\lambda = 1 - \rho > 1$. In any case

$$\sum_{i=1}^{p} \lambda_i = p \tag{24}$$

as it should be.

