Biometrika (2018), **105**, 2, *pp.* 389–402
Printed in Great Britain

doi: 10.1093/biomet/asy010
Advance Access publication 19 March 2018

On the number of principal components in high dimensions

By SUNGKYU JUNG

Department of Statistics, University of Pittsburgh, 1806 Wesley W. Posvar Hall,
230 Bouquet Street, Pittsburgh, Pennsylvania 15260, U.S.A.
sungkyu@pitt.edu

MYUNG HEE LEE

Center for Global Health, Department of Medicine, Weill Cornell Medicine,
1300 York Avenue, New York, New York 10065, U.S.A.
myl2003@med.cornell.edu

and JEONGYOUN AHN

Department of Statistics, University of Georgia, 310 Herty Drive, Athens,
Georgia 30602, U.S.A.
jyahn@uga.edu

Summary

We consider how many components to retain in principal component analysis when the dimension is much higher than the number of observations. To estimate the number of components, we propose to sequentially test skewness of the squared lengths of residual scores that are obtained by removing leading principal components. The residual lengths are asymptotically left-skewed if all principal components with diverging variances are removed, and right-skewed otherwise. The proposed estimator is shown to be consistent, performs well in high-dimensional simulation studies, and provides reasonable estimates in examples.

Some key words: High-dimensional data; Principal component analysis; Skewness test.

1. Introduction

Principal component analysis is widely used and has proven to be effective in dimension reduction of high-dimensional data. Let $\mathcal{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ be an $n \times d$ data matrix, where each vector X_i has zero mean and covariance matrix $\Sigma_d = \sum_{i=1}^{d} \lambda_i u_i u_i^{\mathrm{T}}$, and (u_i, λ_i) denotes the ith principal component direction and variance. The classical estimates $(\hat{u}_i, \hat{\lambda}_i)$ are obtained by the eigendecomposition of the sample covariance matrix. Determining the number of components to retain is crucial in applications.

A number of strategies have been proposed to tackle this problem when the sample size is large and the dimension is relatively low, i.e., $d \ll n$. These include graphical methods based on the scree plot of eigenvalues, model-based tests, and computer-intensive tools (Jolliffe, 2002; Josse & Husson, 2012). However, modern data challenges often involve high-dimension, low-sample-size data with $d \gg n$, for which those methods may be infeasible, computationally prohibitive, or chosen subjectively. In this article, we propose a novel estimator of the number of components when $d \gg n$.

The true number of components is defined in terms of eigenvalues λ_i of Σ_d. A popular approach is to assume that the first m eigenvalues are larger than a threshold τ^2 and the rest equal τ^2. This spike model (Johnstone, 2001; Paul, 2007) has been used in many different contexts (Baik & Silverstein, 2006; Kritchman & Nadler, 2009; Leek, 2011). For a diverging dimension d with limited sample size, the spike size should be increasing at least as fast as d in order to have nontrivial eigenvector estimators (Lee, 2012), so we assume the eigenvalues of Σ_d to be

$$\lambda_i = \sigma_i^2 d \quad (i = 1, \ldots, m), \quad \sigma_1^2 > \cdots > \sigma_m^2 > 0, \tag{1}$$

and the rest of the eigenvalues, $\lambda_{m+1}, \ldots, \lambda_d$, to be equal to τ^2 or to form a slowly decreasing sequence.

Hellton & Thoresen (2017) have shown that under the m-component model (1), even though the classical estimates of (λ_i, u_i) are inconsistent as $d \to \infty$, the first m estimated principal component scores contain useful information on the true scores. We further show in § 5 that the remaining estimated scores are mostly accumulated noise, which implies that the number of spikes m in (1) can be considered to be the number of asymptotically meaningful components.

To determine m from a sample $\mathcal{X}$, we propose to sequentially test the null hypothesis $H_k : m = k$ against the alternative hypothesis $H_{a,k} : m > k$, for increasing values of k, and to estimate m by the smallest k for which H_k is not rejected. To this end, we show that the squared lengths of residuals that are obtained by removing the first k leading principal components are asymptotically left-skewed under the null hypothesis, or right-skewed under the alternative hypothesis. This motivates us to consider test statistics based on the empirical distribution of the residual lengths. We adopt well-known tests for skewness (Randles et al., 1980; D'Agostino & Pearson, 1973). The resulting estimator is consistent under a mild condition.

2. Sequential tests to determine m

2·1. *Motivation*

We propose to sequentially test the null hypotheses $H_0, \ldots, H_M$ for some $M < n$, against one-sided alternatives:

$$H_k : m = k \qquad \text{versus} \qquad H_{a,k} : m > k, \tag{2}$$

where m is the number of components with fast-diverging variances in (1). These null hypotheses do not overlap; if H_k is true, then H_ℓ is not true for all $\ell \neq k$. However, H_k is nested within all lower-order alternatives; if H_k is true, then $H_{a,\ell}$ is true for all $\ell < k$. These observations suggest testing of H_k only if H_ℓ is rejected for all $\ell < k$. The number of effective components, m, is then determined by the smallest k for which H_k is not rejected at a given level.

To test these hypotheses, we first note that the squared lengths of the data vectors $\|X_j\|_2^2$ $(j = 1, \ldots, n)$ have different empirical distributions depending on which hypothesis is true. For example, let us assume that the global null hypothesis H_0 is true, $\Sigma_d = \tau^2 I_d$, and that the data are normal. Then, the squared length $\|X_j\|_2^2$ is normally distributed for large d: as $d \to \infty$,

$$d^{1/2}\left(d^{-1}\|X_j\|_2^2 - \tau^2\right) \to N(0, 2\tau^4) \tag{3}$$

in distribution. On the other hand, when $m \geqslant 1$ in (1), the squared length decomposes into a sum of two independent random variables: if $m = 1$, $d^{-1} \|X_j\|_2^2 = Z + Y$, where approximately

$$Z \sim N(\tau^2, 2\tau^4/d), \quad Y/\sigma_1^2 \sim \chi_1^2. \tag{4}$$

In the limit $d \to \infty$, Z degenerates to τ^2, so $d^{-1} \|X_j\|_2^2$ converges in distribution to a shifted-and-scaled chi-squared random variable, which is right-skewed.

This example suggests considering test statistics based on the normality or the skewness of the empirical distribution of the squared lengths. We will show in § 3 that general asymptotic null and alternative distributions of the squared lengths are similar to those in (3) and (4), even under non-Gaussian assumptions.

2·2. *Test statistics*

In testing the global null hypothesis, the asymptotic normality, shown in (3) under H_0, can be used. Let $p_0^N = p^N(\|X_1\|_2^2, \ldots, \|X_n\|_2^2)$ be a p-value for testing the normality of $\|X_j\|_2^2$. Intuitively, if a principal component with large variance is present, p_0^N tends to be small, since the empirical distribution becomes right-skewed, as in (4).

For testing higher-order hypotheses H_k for $k \geqslant 1$, we propose to remove the first k estimated principal components from the data. We use the classical estimates $(\hat{u}_i, \hat{\lambda}_i)$ obtained by the eigen-decomposition of the sample covariance matrix $S_d = n^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{X} = \hat{U}\hat{\Lambda}\hat{U}^{\mathrm{T}}$. Denote the scaled squared length of the kth residual for the jth observation by

$$R_j(k) = \frac{1}{d}\left\| X_j - \sum_{i=1}^{k} \hat{u}_i\hat{u}_i^{\mathrm{T}} X_j \right\|_2^2 \quad (j = 1, \ldots, n;\ k = 0, \ldots, M). \tag{5}$$

The normality test may be adopted in computing p-values for testing H_k. We will show in § 3·2 that if $\hat{u}_i$ is a consistent estimator of u_i in the d-limit for $i \leqslant k$, then the asymptotic null distribution of $R_j(k)$ is Gaussian under H_k, thus leading to a uniform null distribution of the p-value.

The situations under which $\hat{u}_i$ is consistent are rare. In fact, under the fast-diverging eigenvalue assumption (1) and in the high-dimension, low-sample-size asymptotic scenario, the sample principal component directions are inconsistent (Jung et al., 2012b; Lee, 2012). Moreover, the true principal component variance λ_i is often overestimated by $\hat{\lambda}_i$ for $i \leqslant m$. Since the sum of squared scores equals the variance, i.e., $n^{-1}\sum_{j=1}^{n}(\hat{u}_i^{\mathrm{T}}X_j)^2 = \hat{\lambda}_i$, this overestimation affects (5) in such a way that $R_j(k)$ becomes smaller than desired and asymptotically left-skewed. We will revisit this phenomenon in § 3·3.

To incorporate the left-skewed $R_j(k)$, our first choice of the test statistic is from a test for skewness. For observations $y_j = R_j(k)$ $(j = 1, \ldots, n)$, suppose that the distribution of y_j is continuous with an unknown median θ. Randles et al. (1980) proposed a nonparametric test for symmetry about θ based on a U-statistic with kernel

$$f^*(y_i, y_j, y_k) = \mathrm{sign}(y_i + y_j - 2y_k) + \mathrm{sign}(y_i + y_k - 2y_j) + \mathrm{sign}(y_j + y_k - 2y_i),$$

called the triples test for symmetry. The triples test is an asymptotic test for large n, and Randles et al. (1980) recommended the use of its asymptotic normality when $n > 20$. A one-sided test for left- or right-skewed alternatives is also possible (Hollander et al., 2013, § 3.9). For our purpose, the p-value is obtained by the asymptotic normality of the one-sided triples statistic with the alternative of right-skewed distributions, and is denoted by

$$p_k^R = p^R\{R_1(k), \ldots, R_n(k)\}. \tag{6}$$

Our second choice of test statistic is obtained from a test for normality that is sensitive to skewed alternatives, based on the sample skewness coefficient $b_1 = m_3/(m_2)^{3/2}$, where $m_r = n^{-1} \sum_{j=1}^n (y_j - \bar{y})^r$. D'Agostino (1970) suggested a transformation of b_1, defining $Z = \delta \log[b_1/\lambda + \{(b_1/\lambda)^2 + 1\}^{1/2}]$, where δ and λ are functions of the theoretical moments of b_1 in samples of size n from the normal distribution. Under normal assumptions, the distribution of the transformed Z is well-approximated by the standard normal, even for a small sample size $n \geqslant 8$ (D'Agostino, 1970; D'Agostino & Pearson, 1973). Positive b_1 and Z indicate right-skewness, while negative values indicate left-skewness. The p-value of the skewness test is defined by

$$p_k^D = p^D\{R_1(k), \ldots, R_n(k)\} = 1 - \Phi(Z), \tag{7}$$

where Φ is the standard normal distribution function.

Both p-values in (6) and (7) have an approximately uniform distribution if the null distribution of $R_j(k)$ is normal. They are sensitive to right-skewed alternatives, as they tend to be close to zero under such cases. On the other hand, if the null distribution of $R_j(k)$ is left-skewed, the p-values are close to 1. Other tests of symmetry (e.g., Farrell & Rogers-Stewart, 2006) can be used in place of (6) and (7).

2·3. *Example*

Before proceeding with theoretical results, we demonstrate the proposed procedures on a microarray study (Bhattacharjee et al., 2001), which contains $d = 2530$ genes from $n = 56$ patients in four different lung cancer subtypes. An inspection of the principal component scores plot (Jung & Marron, 2009, Fig. 1) suggests that the four subtypes are visually separated by using the first few sample principal components, and there are no outliers.

We applied the tests discussed in § 2·2 to obtain sequences of p-values for testing (2). As a visual tool to determine the number of components, we plot p_k^R and p_k^D against k, as shown in Fig. 1(a). Graphical methods based on the scree plot, shown in Fig. 1(b), lead to either $\hat{m} = 2$ when locating an elbow, or $\hat{m} = 17$ when using a heuristic cut-off based on the cumulative percentage of variance explained, say 80%. In contrast, our estimate, using either of the two test statistics, is $\hat{m} = 9$, based on

$$\hat{m} = \min\{k : p_k > \alpha\}, \tag{8}$$

where $\alpha = 0{\cdot}1$ in this example. While $\alpha = 0{\cdot}1$ may be used as a default value in practice, we recommend inspecting the p-value sequence such as in Fig. 1.

The empirical distribution of $R_j(0)$ in Fig. 1(c) is clearly right-skewed, strong evidence against $H_0 : m{=}0$. The components with large variances contained in $R_j(0) = d^{-1} \|X_j\|_2^2$ yield a heavy-tailed distribution skewed towards large values. On the other hand, noise-accumulated components are excessively subtracted in computing $R_j(15)$, leading to the left-skewed distribution of squared residual lengths, as shown in Fig. 1(d). In this example, p-values in the sequences are small for the first few tests, then rapidly increase. This pattern was also found in many real and simulated datasets.

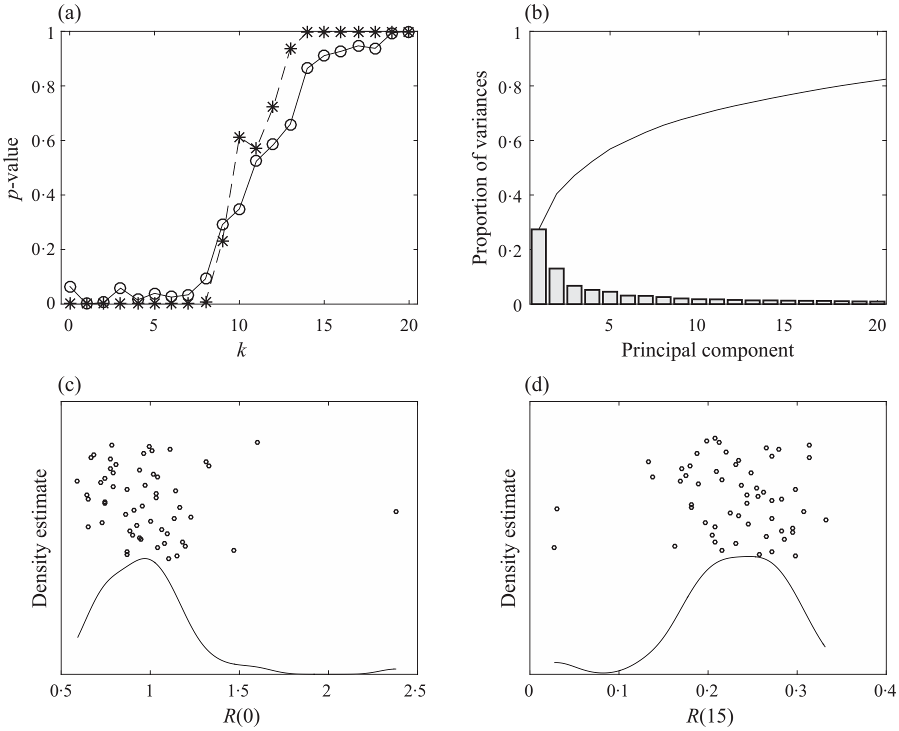

Fig. 1. Plots for determining the number of components for the data from Bhattacharjee et al. (2001). (a) sequences of p-values, p^R_k (circles, solid) and p^D_k (asterisks, dashed); (b) scree plot, where the bars indicate the proportion of the principal component variance, and the curve plots the cumulative proportion of variance; (c) and (d) show jitter plots of $R_j(0)$ and $R_j(15)$ and kernel density estimates.

3. Asymptotic null and alternative distributions

3·1. *Models*

We suppose that the variances of the first m components are diverging at rate d, while the rest are much smaller. For a fixed m, the m-component model is defined for increasing d in Conditions 1–4:

Condition 1. $\lambda_i = \sigma_i^2 d \ \ (i = 1, \ldots, m)$ with $\sigma_1^2 > \cdots > \sigma_m^2 > 0$;

Condition 2. $\sum_{i=m+1}^{d} \lambda_i/d \to \tau^2 \in (0,\infty)$, $\sum_{i=m+1}^{d} \lambda_i^2/d \to \upsilon_O^2 \in (0,\infty)$ as $d \to \infty$, and there exists $\delta \in (0,1]$ such that $\sum_{i=m+1}^{d} \lambda_i^{2+\delta} = o(d^{1+\delta/2})$.

By decomposing each observation into the first m principal components and the remaining term, we write $X_j = \sum_{i=1}^{m} \lambda_i^{1/2} u_i z_{ij} + \sum_{i=m+1}^{d} \lambda_i^{1/2} u_i z_{ij}\ (j = 1, \ldots, n)$, where z_{ij} is the normalized principal component score.

Condition 3. For each j, $Z_j = (z_{1j}, z_{2j}, \ldots)$ is a sequence of mutually independent random variables such that for any i, $E(z_{ij}) = 0$, $\mathrm{var}(z_{ij}) = 1$, and the second and third moments of z_{ij}^2 are uniformly bounded below and above. The sequences $Z_1, Z_2, \ldots$ are mutually independent.

Conditions 1 and 2 are quite general and encompass the spike models of Leek (2011) and Hellton & Thoresen (2017). In particular, they include equal, polynomially decreasing, and slowly diverging eigenvalues. Condition 2 is stronger than the conditions of Ahn et al. (2007), which are used in showing the high-dimension, low-sample-size geometric representation: modulo rotation, the data converge to a regular simplex (Hall et al., 2005). This stronger condition and the moment conditions in Condition 3 are needed for introducing a d-asymptotic normality in Theorem 1 and also in describing asymptotic behaviours of sample scores in Theorem 4. Conditions 1 and 2 imply a low effective rank assumption in the random matrix literature; see Koltchinskii & Lounici (2016, 2017).

A special case of our model is the high-dimensional approximate factor model with pervasive factors, defined below, which has recently gained popularity as it is believed to be more realistic than other models (Hellton & Thoresen, 2017). Let $X = \sum_{i=1}^{m} q_i z_i + \epsilon$ be an m-factor model, where z_i are standardized factor scores and $\epsilon = (\epsilon_1, \ldots, \epsilon_d)^{\mathrm{T}}$ is a zero-mean independent noise vector with uniformly bounded variances. The orthogonal factor loadings, $q_i \in \mathbb{R}^d$, are pervasive; that is, the proportion of nonzero entries of q_i is nonvanishing as d increases (Fan et al., 2013; Hellton & Thoresen, 2017). For example, q_i is pervasive if for $r \in (0, 1)$ the first $\lfloor rd \rfloor$ entries of q_i are 1, while the rest are zero for all d. The loading vector q_i is then expressed as $q_i = \lambda_i^{1/2} u_i$, where $\|u_i\|_2 = 1$ and $\lim_{d\to\infty} \lambda_i/d = \sigma_i^2$. Condition 1 makes the first m components pervasive. Intuitively, when more variables are added into the analysis, i.e., when the dimension d increases, these added variables are not simply noise terms but are correlated with the pervasive factors.

The following assumption on the pervasive factors plays a crucial role in our test procedures proposed in § 2·2.

Condition 4. The third central moment of z_{ij}^2 $(i = 1, \ldots, m; j = 1, \ldots, n)$ is positive.

Simply put, we require that z_{ij}^2 be right-skewed. Condition 4 holds for many distributions, including t_ν distributions with $\nu > 6$, the beta distributions with parameters (α, α) where $\alpha > 0{\cdot}5$, the gamma distributions, and a normal mixture $\xi X_1 + (1-\xi)X_2$, where X_1 and X_2 are independent normal random variables with a common variance and ξ follows a Bernoulli distribution.

3·2. *Known principal component directions*

We first investigate an ideal case where the principal component directions are known, to better understand the high-dimensional asymptotic behaviour of the residual lengths. Define the kth true residual length of the jth observation by

$$\tilde{R}_j(k) = d^{-1} \left\| X_j - \sum_{i=1}^{k} u_i u_i^{\mathrm{T}} X_j \right\|_2^2 = d^{-1} \sum_{i=k+1}^{d} w_{ij}^2, \tag{9}$$

where $w_{ij} = u_i^{\mathrm{T}} X_j = \lambda_i^{1/2} z_{ij}$ is the population principal component score.

The asymptotic behaviour of (9) can be understood by using a scaled Gram matrix $S_D = d^{-1} \mathcal{X}\mathcal{X}^{\mathrm{T}}$, whose (j, k)th element is $s_{jk} = d^{-1} X_j^{\mathrm{T}} X_k$ $(j, k = 1, \ldots, n)$. An immediate connection is that the jth diagonal element of S_D is $\tilde{R}_j(0)$. Under the assumption of m fast-diverging components, we denote the $n \times m$ matrix of the first m scaled components by $W_1^{\mathrm{T}} = d^{-1/2}\mathcal{X}(u_1, \ldots, u_m)$, where the (i,j)th element of W_1 is $d^{-1/2} w_{ij}$. It is known that S_D has a limiting expression (Jung et al., 2012b): as $d \to \infty$,

$$S_D \to W_1^{\mathrm{T}} W_1 + \tau^2 I_n \tag{10}$$

in probability, conditional on W_1. This result is now strengthened to provide a rate of convergence of S_D.

Theorem 1. *Assume the m-component model under Conditions 1–3. Let $m \geqslant 0$ be fixed. Conditioned on W_1, (10) holds. Moreover, each element of S_D has a d-asymptotic normal distribution: for $j \neq k$, as $d \to \infty$,*

$$d^{1/2}\left(s_{jj} - \sum_{i=1}^{m} \sigma_i^2 z_{ij}^2 - \tau^2\right) \to N(0, \upsilon_D^2),$$

$$d^{1/2}\left(s_{jk} - \sum_{i=1}^{m} \sigma_i^2 z_{ij} z_{ik}\right) \to N(0, \upsilon_O^2)$$

in distribution, where $\tau^2 = \lim_{d\to\infty} \sum_{i=m+1}^{d} \lambda_i/d$, $\upsilon_D^2 = \lim_{d\to\infty} \sum_{i=m+1}^{d} \lambda_i^2 \mathrm{var}(z_{ij}^2)/d$ and $\upsilon_O^2 = \lim_{d\to\infty} \sum_{i=m+1}^{d} \lambda_i^2/d$.

Theorem 1 provides the null and alternative distributions of $\tilde{R}_j(k)$.

Corollary 1. *Assume the m-component model under Conditions 1–4. Let $n > m \geqslant 0$ be fixed. Then for any $j = 1, \ldots, n$ and $k = 0, \ldots, n-1$, for large d:*

(i) *if $k \geqslant m$, $\tilde{R}_j(k)$ is asymptotically normal;*

(ii) *if $k < m$, $\tilde{R}_j(k)$ is asymptotically right-skewed.*

Intuitively, if all of the pervasive factors are removed from the residual, i.e., $k \geqslant m$, then the factors in the residual can be thought of as accumulated noise, and by Theorem 1, the residual length has a limiting normal distribution. On the other hand, if one or more pervasive factors remain in the residual, i.e., $k < m$, then the sum of squared factors appears in the residual length. Condition 4 ensures that the squared factors are right-skewed.

3·3. *Estimated principal component directions*

When the estimated principal component directions $\hat{u}_i$ are used, the residual lengths $R_j(k)$ have different limiting distributions from those of $\tilde{R}_j(k)$. We characterize the full family of asymptotic distributions of $R_j(k)$ under the null and alternative hypotheses. For this, we consider an asymptotic scenario where the limits $d \to \infty$ and $n \to \infty$ are taken progressively. This resembles the case where the dimension increases at a much faster rate than the sample size does, such as $d/n \to \infty$, but is not identical to it (Lee et al., 2014). Asymptotic null distributions of $R_j(k)$ for fixed n are discussed in the Supplementary Material.

Let $\hat{w}_{ij} = \hat{u}_i^{\mathrm{T}} X_j$ denote the sample projection score. The following decomposition is useful in explaining the limiting distribution of $R_j(k)$:

$$R_j(k) = \tilde{R}_j(k) + a_j(k), \quad a_j(k) = \frac{1}{d}\sum_{i=1}^{k}(w_{ij}^2 - \hat{w}_{ij}^2). \qquad (11)$$

First consider the asymptotic null distribution of $R_j(k)$ under $H_k : m = k$. The overestimation of λ_i $(i = 1, \ldots, m)$ by $\hat{\lambda}_i$ means that $\hat{w}_{ij}^2$ tends to be larger than w_{ij}^2, which is shown in the Supplementary Material. Thus one can expect that $a_j(m)$, the difference between the squared true score and squared sample score, is negative. It turns out that $a_j(m)$, and subsequently $R_j(m)$, are left-skewed in the limit, as shown in Theorem 2. Describing the alternative distribution of $R_j(k)$ for $k < m$ seems more challenging, because the two dependent variables in (11) exhibit different skewness: the first term $\tilde{R}_j(k)$ is asymptotically right-skewed, and the second term $a_j(k)$ is asymptotically left-skewed. Below, we also show that $a_j(k)$ is in fact asymptotically negligible.

Theorem 2. *Assume the m-component model under Conditions 1–4. Let $m \geqslant 0$ be fixed. Suppose that the limits $d \to \infty$ and $n \to \infty$ are taken successively.*

(i) *If $m = 0$, then in the limit, $R_j(0)$ is normally distributed.*
(ii) *If $m \geqslant 1$, then for each j, $n\{R_j(m) - \tau^2\} \to A_j(m)$ in probability, where $A_j(m) = -\tau^2 \sum_{i=1}^m z_{ij}^2$. Moreover, $A_j(m)$ $(j = 1, 2, \ldots)$ are identically distributed, left-skewed and mutually independent.*
(iii) *If $m > k \geqslant 0$, then for each j, $R_j(k) \to B_j(k, m)$ in probability, where $B_j(k, m) = \sum_{i=k+1}^m \sigma_i^2 z_{ij}^2 + \tau^2$. Moreover, $B_j(k, m)$ $(j = 1, 2, \ldots)$ are identically distributed, right-skewed and mutually independent.*

Theorem 2 provides a theoretical justification for the test procedures based on the skewness in § 2·2. The test statistics in (6) and (7) tend to be large under the nontrivial null hypothesis and small under the alternative. Theorem 2 also implies that the sharp transition of p-values from low to high, as shown in Fig. 1, is bound to happen for large enough d and n.

Our next result shows that our estimator (8) consistently estimates the true number of principal components. For this, we require that the test involved be consistent and the function p_k be continuous for each n. These hold if p_k^D is used. Although p_k^R does not satisfy the continuity condition, the estimator of m using the triples test appears to be consistent in our empirical results.

Theorem 3. *Assume the m-component model under Conditions 1–4. Let $\hat{m}(\alpha)$ be the estimator of m defined in (8), where p_k is computed using (7). Then for any $\alpha \in (0, 1)$,*

$$\lim_{n\to\infty} \lim_{d\to\infty} \mathrm{pr}\{\hat{m}(\alpha) = m\} = 1.$$

Theorem 3 not only shows consistency but also suggests that for large enough dimension and sample size, the estimator $\hat{m}$ should be nearly invariant with respect to the choice of α. This robustness against varying α is empirically confirmed in § 4·4.

4. Numerical studies

4·1. *Existing methods*

There are a number of existing methods for determining the number of components. For $d \ll n$, we refer to Jolliffe (2002) for discussion of heuristic and model-based methods.

Bai & Ng (2002) considered determining the number of principal components, m, when both the dimension and the sample size diverge. They proposed several information criteria-type estimators, but we found that using these estimators directly yields unsatisfactory results, so we use a modified estimator based on their information criteria, defined in the Supplementary Material. Simulation-based methods such as parallel analysis (Horn, 1965) have evolved into eigenvalue-based estimation of m, using an asymptotic random matrix theory for large d and n. Kritchman & Nadler (2008, 2009) and Passemier & Yao (2012, 2014) developed estimators of m using the Tracy–Widom distribution (Johnstone, 2001). Leek (2011) also proposed an eigenvalue-based estimator of m by choosing a stable threshold for the sample eigenvalues.

Our estimators obtained by (8) will be denoted by $\hat{m}_{\mathrm{R}}$ and $\hat{m}_{\mathrm{D}}$, when using the p-value sequences of (6) and (7), respectively. For simplicity, we used $\alpha = 0{\cdot}1$ for all numerical results. Our methods

Table 1. *Estimated number of principal components for example datasets*

Dataset	(d, n)	$\hat{m}_{\rm R}$	$\hat{m}_{\rm D}$	$\hat{m}_{\rm L}$	$\hat{m}_{\rm KN}$	$\hat{m}_{\rm PY}$	$\hat{m}_{\rm BN}$
Lymphoma	(7129, 77)	11	11	31	65	65	10
Prostate	(2135, 102)	22	22	14	52	25	14
NCI60 cDNA	(2267, 60)	5	5	2	31	9	2
NCI60 Affy	(2267, 60)	10	10	23	44	31	4
NCI60 combined	(2267, 120)	11	11	65	86	80	4
Leukemia	(3051, 38)	1	9	9	25	22	3
Lung	(2530, 56)	9	9	55	41	31	7
Lobular freeze	(16 615,817)	118	92	20	481	171	29
Hippocampi	(336, 51)	11	11	14	27	24	3
Liver	(12 813, 500)	71	151	171	416	290	137

$\hat{m}_{\rm R}$, our estimator using (6); $\hat{m}_{\rm D}$, our estimator using (7); $\hat{m}_{\rm L}$, Leek (2011)'s method; $\hat{m}_{\rm KN}$, Kritchman & Nadler (2008)'s method; $\hat{m}_{\rm PY}$, Passemier & Yao (2014)'s method; $\hat{m}_{\rm BN}$, Bai & Ng (2002)'s method.

are robust with respect to the choice of α, as discussed in § 4·4. In the numerical studies below, our estimators are compared with the methods of Kritchman & Nadler (2008), Passemier & Yao (2014), Leek (2011) and Bai & Ng (2002).

4·2. *Real data examples*

We report the estimated number of components for eight real datasets. The first six are from gene expression studies, which usually produce high-dimensional data with limited sample size. The latter two are different types of images. These datasets are described below; see Table 1.

The microarray datasets we tested include diffuse large B-cell lymphoma data (Shipp et al., 2002), prostate cancer data (Singh et al., 2002), and each of the two different platforms of the NCI60 cell line data (Shoemaker, 2006). We also tested the training set of leukemia data (Golub et al., 1999) and lung cancer data (Bhattacharjee et al., 2001). The lobular freeze dataset is a set of breast cancer gene expression data, measured by RNA sequencing (Ciriello et al., 2015).

The hippocampi dataset (Pizer et al., 2013) consists of skeletal representations, three-dimensional models of human organs parameterized by spatial locations, lengths and directions of skeletal spokes. Pizer et al. (2013) proposed a nonclassical principal component analysis based on Jung et al. (2012a). This data example is chosen to show that our method can be applied to nonclassical principal component analysis through the scores matrix, since the residual lengths can be computed from the scores; see (11). The last dataset consists of cell nucleus greyscale images from human liver tissues (Wang et al., 2011). We chose $d = 12\,813$ variables with standard deviation greater than 0·01 from the original 36 864 pixels.

Table 1 shows that our estimates $\hat{m}_{\rm R}$ and $\hat{m}_{\rm D}$ are usually close to each other. Our estimates are generally larger than those of Bai & Ng (2002), but smaller than those of Kritchman & Nadler (2008) and Passemier & Yao (2014). Through a simulation study in § 4·3, we have come to believe that the method of Bai & Ng tends to underestimate, while those of Kritchman & Nadler and Passemier & Yao overestimate, for finite d and n. In particular, the estimates $\hat{m} = 25$ and 22 from the methods of Kritchman & Nadler and Passemier & Yao for the leukemia data seem unsuitably large considering the sample size $n = 38$. Our estimates, especially $\hat{m}_{\rm D}$, exhibit a balance between the two extremes. The seemingly biased other estimates are in part caused by the violation of distributional assumptions such as normality and equal tail-eigenvalues, which might be the case for real datasets. Our estimators do not need such assumptions.

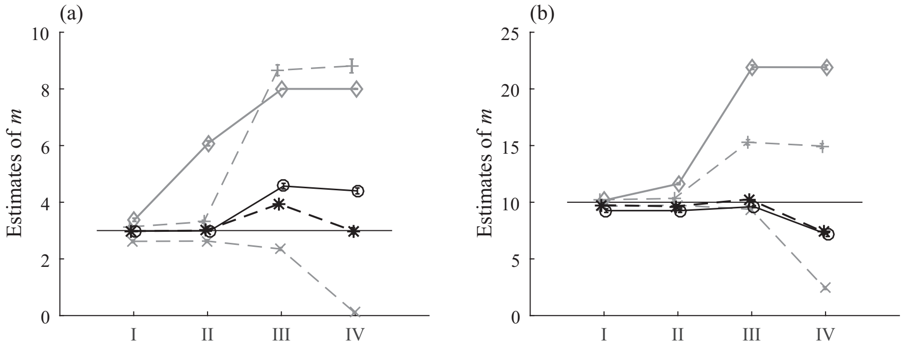

Fig. 2. Average estimates of the number of components for simulated data, computed by $\hat{m}_{\rm R}$ (circles, solid), $\hat{m}_{\rm D}$ (asterisks, dashed), Kritchman & Nadler (2008)'s method (grey +, dashed), Passemier & Yao (2014)'s method (grey diamonds, solid) and Bai & Ng (2002)'s method (grey ×, dashed): (a) $m = 3$ model; (b) $m = 10$ model. The estimates of Leek (2011) were similar to those of Bai & Ng (2002) and thus omitted. The largest standard error for all results is 0·25.

4·3. *Simulation*

To better understand the empirical performance of the estimators, we conducted a simulation study. The eigenvalues of Σ_d are modelled with $s > 0$ representing a signal strength, $0 \leqslant \beta < 1/2$ representing a decay rate of variances in noise components, and $g > 0$ controlling the gap between leading eigenvalues, by

$$\lambda_i = \begin{cases} \sigma_i^2 d, \quad \sigma_i^2 = s^2\{1 + g(m-i)\} & (i = 1, \ldots, m), \\ \tau_\beta i^{-\beta} & (i = m+1, \ldots, d), \end{cases} \tag{12}$$

where $\tau_\beta = \{\sum_{i=m+1}^{d} i^{-\beta}/(d-m)\}^{-1}$ is used to ensure that the average of λ_i $(i = m+1, \ldots, d)$ is 1. The eigenvectors of Σ_d are randomly chosen from the uniform distribution on the orthogonal group of dimension d.

We present simulation results for four different cases.

Case I. The standard normal distribution is used to sample the standardized scores z_{ij}. The eigenvalues of the population covariance matrix are defined by (12) with $(s, g, \beta) = (0{\cdot}2, 1, 0)$.

Case II. The standard normal model with $(s, g, \beta) = (0{\cdot}2, 1, 0{\cdot}3)$.

Case III. The t_3 model with $(s, g, \beta) = (0{\cdot}2, 1, 0{\cdot}3)$.

Case IV. The t_3 model with $(s, g, \beta) = (0{\cdot}1, 0{\cdot}5, 0{\cdot}3)$.

We set the true number of components $m = 3, 10$ for each of the cases and collected the estimated results for $(d, n) = (10\,000, 100)$ based on 100 simulation runs. Figure 2 shows that our estimators $\hat{m}_{\rm D}$ and $\hat{m}_{\rm R}$ perform as well as or better than the competing estimators.

Case I is an ideal situation for all methods considered. In particular, the variances of noise components are equal to each other, i.e., $\lambda_i = 1$ $(i = m+1, \ldots, d)$, and the normal assumption is satisfied. All methods perform similarly. In Case II, the methods of Kritchman & Nadler (2008) and Passemier & Yao (2014) tend to overestimate. This is because, for $\beta > 0$, the equal tail-eigenvalue assumption for the estimators of Kritchman & Nadler and Passemier & Yao to be consistent is not satisfied. The assumptions for consistency of our estimators are satisfied under Case II.

In Cases III and IV, a scaled t_3 distribution is used to sample the standardized scores z_{ij}. The coefficient of skewness in Condition 4 is not defined for this heavy-tailed distribution. Nevertheless, our estimators are less affected by the violation of the assumption than the more biased estimators of Kritchman & Nadler and Passemier & Yao, because the heavy-tailed scores

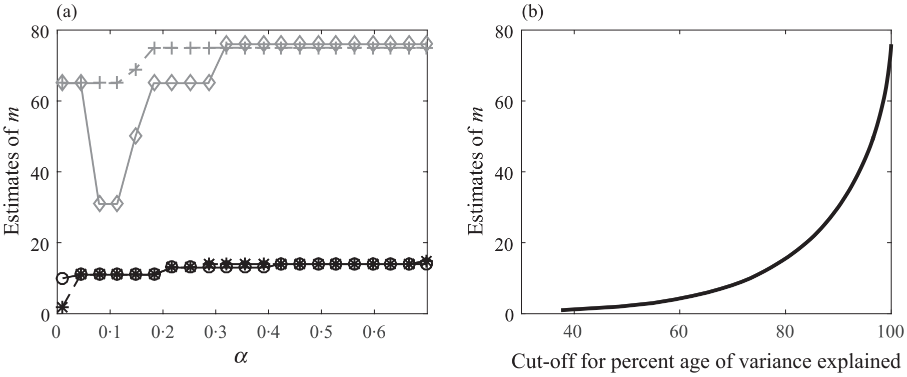

Fig. 3. Invariance of estimates for Shipp et al. (2002)'s data: (a) estimates as functions of α, computed by $\hat{m}_{\rm R}$ (circles, solid), $\hat{m}_{\rm D}$ (asterisks, dashed), Kritchman & Nadler (2008)'s method (grey plus signs, dashed) and Passemier & Yao (2014)'s method (grey diamonds, solid); (b) estimates as a function of the variance threshold.

exhibit more drastic distinctions of the left- and right-skewness than for the normal distribution. In Case IV, the leading m eigenvalues are smaller than in Case III. In this weak-signal setting, the estimators of Leek (2011) and Bai & Ng (2002) severely underestimate.

More simulation results are reported in the Supplementary Material.

4·4. *Empirical robustness against varying* α

The asymptotic invariance of $\hat{m}$ against varying $\alpha \in (0, 1)$, shown in Theorem 3, suggests some invariance for moderately large d and n. In fact, for most real and simulated data examples we considered, the values of $\hat{m}$ are stable against various values of α.

For the data of Shipp et al. (2002), introduced in § 4·2, our estimates $\hat{m}_{\rm R}(\alpha)$ and $\hat{m}_{\rm D}(\alpha)$ are stable for a wide range of α; see Fig. 3. As a comparison, we also experimented on the eigenvalue-based estimators of Kritchman & Nadler (2008) and Passemier & Yao (2014) by changing their threshold value, which is parameterized by the $1-\alpha$ quantiles of the Tracy–Widom distribution. These estimates change their values substantially. The methods of Leek (2011) and Bai & Ng (2002) are not subject to arbitrary choices of threshold, and they were excluded from this study. We further compare with a heuristic method using the cumulative percentage of variance explained. As shown in Fig. 3(b), changing the threshold, say from 80% to 90%, would drastically change the estimates.

The robustness of our estimators against varying α is also confirmed in simulated data. In Fig. 4, the estimates with varying α are plotted for data generated by Case II in § 4·3. Our estimates are stable, except for $\alpha < 0{\cdot}1$. The estimator of Kritchman & Nadler is also stable for larger values of α, but the estimates are clearly biased.

5. Principal component scores in high dimensions

We conclude with a formal statement on the usefulness of the sample principal component scores in high dimensions.

Recall that $W_1 = (\sigma_i z_{ij})_{i,j}$ is the $m \times n$ matrix of the scaled true scores, and $W_1 W_1^{\rm T}$ is proportional to the $m \times m$ sample covariance matrix of the first m scores. Similarly, we define $\hat{W}_1^{\rm T} = d^{-1/2}\mathcal{X}(\hat{u}_1, \ldots, \hat{u}_m)$. Let $\{\lambda_i(S), v_i(S)\}$ denote the ith largest eigenvalue-eigenvector pair of a nonnegative-definite matrix S and let $v_{ij}(S)$ denote the jth loading of the vector $v_i(S)$.

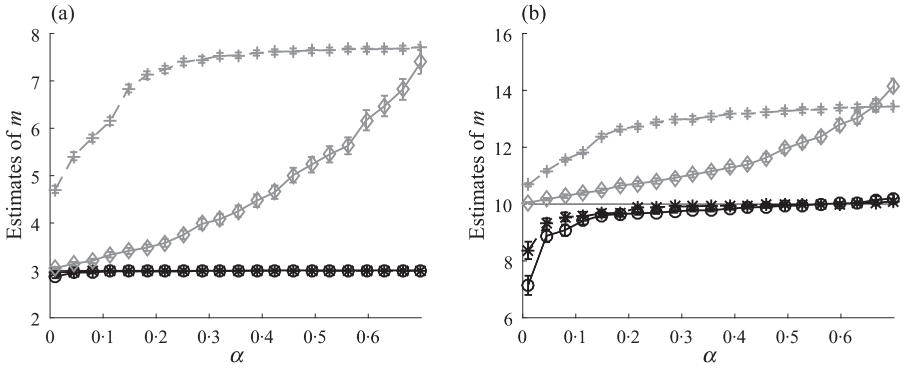

Fig. 4. Invariance of estimates for simulated data with the true values of $m = 3$ and 10, shown in (a) and (b), respectively. Average estimates, from 100 simulation runs, of $\hat{m}_{\mathrm{R}}$ (black circles, solid), $\hat{m}_{\mathrm{D}}$ (black asterisks, dashed), Kritchman & Nadler (2008)'s method (grey plus signs, dashed) and Passemier & Yao (2014)'s method (grey diamonds, solid) are shown with error bars representing the standard errors.

Theorem 4. *Assume the m-component model under Conditions 1–4 and let $n > m \geqslant 0$ be fixed. In addition, we assume that the scores w_{kj} are absolutely continuous.*

(i) *If $k \leqslant m$, then the ratio of the sample score to the true score of X_j for the kth component is asymptotically decomposed into a common factor, not depending on j, and an error specific to each data point. Specifically, for $j = 1, \ldots, n$,*

$$\frac{\hat{w}_{kj}}{w_{kj}} = \rho_k v_{kk}(W_1 W_1^{\mathrm{T}}) + \varepsilon_{kj} + O_{\mathrm{p}}(d^{-1/4}),$$

where $\rho_k = \{1 + \tau^2/\lambda_k(W_1 W_1^{\mathrm{T}})\}^{1/2}$ and $\varepsilon_{kj} = \rho_k \sum_{i=1,\ldots,m; i \neq k} \sigma_i z_{ij} (\sigma_k z_{kj})^{-1} v_{ki}(W_1 W_1^{\mathrm{T}})$. Moreover,

$$\hat{W}_1^{\mathrm{T}} = W_1^{\mathrm{T}} RS + O_{\mathrm{p}}(d^{-1/4}), \tag{13}$$

where $R = \{v_1(W_1 W_1^{\mathrm{T}}), \ldots, v_m(W_1 W_1^{\mathrm{T}})\}$ is an $m \times m$ orthogonal matrix and S is the $m \times m$ diagonal matrix whose kth diagonal element is ρ_k.

(ii) *If $k > m$, then the ratio diverges with the rate $d^{(1-\gamma_k)/2}$, for γ_k satisfying $\lambda_k \asymp d^{\gamma_k}$. Specifically, $\hat{w}_{kj}/w_{kj} = O_{\mathrm{p}}\{d^{(1-\gamma_k)/2}\}$ and $d^{-1} \sum_{j=1}^{n} \hat{w}_{kj}^2 \to \tau^2$ in probability as $d \to \infty$.*

The asymptotic relation (13) tells us that for large d, the first m sample scores in $\hat{W}_1$ converge to the true scores in W_1, uniformly rotated and scaled for all data points. It is thus valid to use the first m sample principal scores for exploration of important data structure, to reduce the dimension of the data space from d to m in the high-dimension, low-sample-size context.

Acknowledgement

The authors are grateful to the editor, associate editor and referees for their careful reading and thoughtful suggestions.

Supplementary material

Supplementary material available at *Biometrika* online includes technical details and additional data examples.

References

Ahn, J., Marron, J. S., Muller, K. M. & Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94**, 760–6.

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

Baik, J. & Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Mult. Anal.* **97**, 1382–408.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M. et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Nat. Acad. Sci.* **98**, 13790–5.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., Mclellan, M., Yau, C., Kandoth, C. et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–19.

D'Agostino, R. B. (1970). Transformation to normality of the null distribution of g_1. *Biometrika* **57**, 679–81.

D'Agostino, R. B. & Pearson, E. S. (1973). Tests for departure from normality: Empirical results for the distributions of b_2 and $\surd b_1$. *Biometrika* **60**, 613–22.

Fan, J., Liao, Y. & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Statist. Soc.* B **75**, 603–80.

Farrell, P. J. & Rogers-Stewart, K. (2006). Comprehensive study of tests for normality and symmetry: Extending the Spiegelhalter test. *J. Statist. Comp. Simul.* **76**, 803–16.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–7.

Hall, P., Marron, J. S. & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc.* B **67**, 427–44.

Hellton, K. H. & Thoresen, M. (2017). When and why are principal component scores a good tool for visualizing high-dimensional data? *Scand. J. Statist.* **44**, 581–816.

Hollander, M., Wolfe, D. A. & Chicken, E. (2013). *Nonparametric Statistical Methods*. Chichester: John Wiley & Sons.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179–85.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29**, 295–327.

Jolliffe, I. T. (2002). *Principal Component Analysis*. New York: Springer.

Josse, J. & Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Comp. Statist. Data Anal.* **56**, 1869–79.

Jung, S., Dryden, I. L. & Marron, J. S. (2012a). Analysis of principal nested spheres. *Biometrika* **99**, 551–68.

Jung, S. & Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37**, 4104–30.

Jung, S., Sen, A. & Marron, J. (2012b). Boundary behavior in high dimension, low sample size asymptotics of PCA. *J. Mult. Anal.* **109**, 190–203.

Koltchinskii, V. & Lounici, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Prob. Statist.* **52**, 1976–2013.

Koltchinskii, V. & Lounici, K. (2017). Normal approximation and concentration of spectral projectors of sample covariance. *Ann. Statist.* **45**, 121–57.

Kritchman, S. & Nadler, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemomet. Intel. Lab.* **94**, 19–32.

Kritchman, S. & Nadler, B. (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Sig. Proces.* **57**, 3930–41.

Lee, M. H. (2012). On the border of extreme and mild spiked models in the HDLSS framework. *J. Mult. Anal* **107**, 162–8.

Lee, S., Zou, F. & Wright, F. A. (2014). Convergence of sample eigenvalues, eigenvectors, and principal component scores for ultra-high dimensional data. *Biometrika* **101**, 484–90.

Leek, J. T. (2011). Asymptotic conditional singular value decomposition for high-dimensional genomic data. *Biometrics* **67**, 344–52.

Passemier, D. & Yao, J. (2014). Estimation of the number of spikes, possibly equal, in the high-dimensional case. *J. Mult. Anal.* **127**, 173–83.

Passemier, D. & Yao, J.-F. (2012). On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices: Theory Appl.* **1**, 1150002.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17**, 1617–42.

PIZER, S. M., JUNG, S., GOSWAMI, D., ZHAO, X., CHAUDHURI, R., DAMON, J. N., HUCKEMANN, S. & MARRON, J. S. (2013). Nested sphere statistics of skeletal models. In *Innovations for Shape Analysis: Models and Algorithms*, M. Breuß, A. Bruckstein & P. Maragos, eds. New York: Springer.

RANDLES, R. H., FLIGNER, M. A., POLICELLO, G. E. & WOLFE, D. A. (1980). An asymptotically distribution-free test for symmetry versus asymmetry. *J. Am. Statist. Assoc.* **75**, 168–72.

SHIPP, M. A., ROSS, K. N., TAMAYO, P., WENG, A. P., KUTOK, J. L., AGUIAR, R. C., GAASENBEEK, M., ANGELO, M., REICH, M., PINKUS, G. S. et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Med.* **8**, 68–74.

SHOEMAKER, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature Rev. Cancer* **6**, 813–23.

SINGH, D., FEBBO, P. G., ROSS, K., JACKSON, D. G., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A. A., D'AMICO, A. V., RICHIE, J. P. et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–9.

WANG, W., OZOLEK, J. A., SLEPČEV, D., LEE, A. B., CHEN, C. & ROHDE, G. K. (2011). An optimal transportation approach for nuclear structure-based pathology. *IEEE Trans. Med. Imag.* **30**, 621–31.

[*Received on* 8 *December* 2015. *Editorial decision on* 12 *December* 2017]