Chemometrics and Intelligent Laboratory Systems 94 (2008) 19–32

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Determining the number of components in a factor model from limited noisy data

Shira Kritchman, Boaz Nadler *

Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, P.O. Box 26, Rehovot 76100, Israel

ARTICLE INFO

Article history:
Received 28 January 2008
Received in revised form 3 June 2008
Accepted 9 June 2008
Available online 15 June 2008

Keywords:
Pseudorank estimation
Principal component analysis
Random matrix theory
Tracy–Widom distribution
Number of components in a mixture

ABSTRACT

Determining the number of components in a linear mixture model is a fundamental problem in many scientific fields, including chemometrics and signal processing. In this paper we present a new method to automatically determine the number of components from a limited number of (possibly) high dimensional noisy samples. The proposed method, based on the eigenvalues of the sample covariance matrix, combines a matrix perturbation approach for the interaction of signal and noise eigenvalues, with recent results from random matrix theory regarding the behavior of noise eigenvalues. We present the theoretical derivation of the algorithm and an analysis of its consistency and limit of detection. Results on simulated data show that under a wide range of conditions our method compares favorably with other common algorithms.

1. Introduction

Linear mixture models are one of the most common modelling approaches to multivariate data in many scientific fields. In spectroscopy, following Beer's law the (logarithm) of measured spectra is modelled as a linear mixture of the different chemical components, each multiplied by its characteristic spectral response [28]. Following basic laws of physics, and specifically acoustic or electromagnetic wave propagation, a similar modelling approach is also common in signal processing. Here, in a typical setting the vector of observations measured by an array of antennas or a collection of microphones is modelled as a superposition of a finite number of signals embedded in additive white noise [34].

One of the most fundamental tasks in the analysis of multivariate data from a linear mixture model is the determination of the number of components or sources present in it. In chemometrics, determination of the chemical rank is typically the first step in self modelling curve resolution, where correct estimation of the number of chemical components is crucial for correct curve resolution [9]. Similarly, in calibration of multi-component systems with non-vanishing correlations between different components and with interfering spectral responses, the theoretically optimal number of factors in common algorithms such as partial least squares or principal component regression, is typically equal to the number of components in the mixture [26]. While many calibration methods estimate the number of components by cross-validation, an a-priori independent estimate of the number of factors is valuable information. While the emphasis in this paper is on chemometrics, we remark that the problem of rank determination is important also in signal processing [34,6,15,30] and in other fields such as genomics, ecology and psychology, to name only a few. In signal processing, for example, determining the number of sources is typically the first step in blind source separation, detection of arrival and source localization tasks.

In the absence of noise, the number of components in a set of measured high dimensional signals is simply the rank of the data matrix, or equivalently the number of non-zero eigenvalues of the corresponding sample covariance matrix, also denoted as its *pseudorank*. In the presence of small additive noise, we expect the sample covariance matrix to have a few large eigenvalues, corresponding to the chemical signals, and a large number of small eigenvalues, corresponding to the noise. As such, the majority of algorithms for pseudorank determination are based on analysis of the eigenvalues of the sample covariance matrix.

Methods for determining the number of components date back at least to the works of Bartlett and Lawley, who developed likelihood ratio tests to check for sphericity, e.g. for equality of the smallest eigenvalues [13,17]. Their methods are based on asymptotic expansions for large sample sizes, and may not perform well in the common setting in chemometrics where the number of samples, n, is of the same order and often significantly smaller than the number of variables (wavelengths) p. Consequently, in the past four decades more than twenty different methods for rank determination have been suggested in the chemometrics community [3,4,5,7,18–21,23]. Independently, methods to detect the number of sources have also been developed both in the statistics and signal processing communities, with emphasis on model selection criteria and information theoretic approaches, such as minimum description length (MDL), Bayesian model selection, Bayesian information criteria (BIC) and more [34,36,24].

* Corresponding author. Tel.: +972 8 9342856; fax: +972 8 9346023.
E-mail address: boaz.nadler@weizmann.ac.il (B. Nadler).

0169-7439/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2008.06.002

The common thread to all algorithms for rank determination is the attempt to distinguish between small yet significant eigenvalues due to a signal, and large yet insignificant eigenvalues due to noise. In this paper, based on recent results in random matrix theory for the behavior of noise eigenvalues, and on a matrix perturbation approach for the interactions between noise and signal eigenvalues, we develop a novel algorithm for rank determination. Our proposed algorithm performs a sequence of hypothesis tests on the number of components, at each step testing the significance of the kth largest eigenvalue as arising from a signal rather than from noise. Our algorithm is thus intimately related to Roy's largest root test [31,16], with one additional key component—an accurate estimation of noise level. In our method we assume Gaussian homoscedastic noise (that is, equal noise variance in all directions). An interesting future research direction is to generalize our approach to the case of heteroscedastic noise.

As described in detail in Section 2 our method is based on firm statistical foundations and it is suitable for a wide range of values of p and n. Recent results in matrix random theory allow the analysis of its consistency and of its limit of detection. These issues are particularly important given that one of the main challenges in pseudorank estimation is the detection of minor components present in a mixture. The theoretical analysis of this limit of detection highlights the importance of incorporating prior knowledge for improved rank determination, and motivates the use of regularization methods for rank determination, for example using smooth PCA [35].

In Section 3 we present simulations comparing the performance of our algorithm to other common rank estimation methods, including Malinowski's F-test [21], its modification by Faber and Kowalski [4], and a recent algorithm suggested by Rao and Edelman [30]. For the last two algorithms we also present a theoretical analysis in the appendix. The simulation results show that under a wide range of conditions our algorithm is as good as, and often better, than these other methods.

We conclude the paper with a short discussion in Section 4. Proofs of consistency and other technical details appear in the appendices.

2. Problem formulation and main results

Notation. We denote random variables by lowercase letters, as in u, whereas specific realizations have an additional subscript, as in u_ν. Vectors and matrices are denoted by lowercase and uppercase boldface letters, e.g. $\mathbf{w}$ and $\mathbf{C}$, respectively. The identity matrix of order p is denoted $\mathbf{I}_p$, and a $p\times p$ matrix of all zeros is denoted $\mathbf{0}_p$.

Problem Formulation. Consider a dataset of n i.i.d. noisy samples $\{\mathbf{x}_\nu\}_{\nu=1}^n$ from the following p-dimensional linear mixture model with K components,

$$x = \sum_{j=1}^{K} u_j \mathbf{v}_j + \sigma\xi. \tag{1}$$

The random variables u_j are the K different components with corresponding response vectors $\mathbf{v}_j\in\mathbb{R}^p$, and σ is the level of noise. In this work we consider uncorrelated homoscedastic noise, i.e., we assume $\xi\in\mathbb{R}^p$ is a multivariate $\mathcal{N}(0,\mathbf{I}_p)$ Gaussian random noise vector with unit variance in all directions. Further, we assume that the $K\times K$ covariance matrix of the random variables u_j is of full rank, and that the response vectors $\mathbf{v}_j$ are linearly independent in $\mathbb{R}^p$. Under these assumptions the population covariance matrix $\mathbf{\Sigma}$ of the observations $\mathbf{x}$ can be diagonalized to have the form

$$\mathbf{W}'\mathbf{\Sigma}\mathbf{W} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_K \\ \hline & & \mathbf{0}_{(p-K)} \end{pmatrix} + \sigma^2\mathbf{I}_p. \tag{2}$$

The $p\times p$ matrix $\mathbf{W}$ is unitary and its columns $\mathbf{w}_j$ are the eigenvectors of $\mathbf{\Sigma}$, with eigenvalues $\lambda_j+\sigma^2$.

The problem considered in this paper is as follows: given a random sample of size n, $\{\mathbf{x}_\nu\}_{\nu=1}^n$ from Eq. (1), infer the value of K, i.e., the number of components in the model. The value K is also known as the *pseudorank* of the matrix, formally defined as the rank of the data matrix in the absence of noise. For future reference, we denote by $\mathbf{S}_n$ the (non-centered) $p\times p$ sample covariance matrix,

$$\mathbf{S}_n = \frac{1}{n}\sum_{\nu=1}^{n}\mathbf{x}_\nu\mathbf{x}_\nu'.$$

In this work we do not impose any further knowledge or prior assumptions on the shape of the response vectors $\mathbf{v}_j$ or on the (possibly non-Gaussian) distribution of the random variables u_j. The importance and benefits of incorporating prior knowledge are described in Section 2.1 below. Thus, we infer K only from the sample covariance matrix $\mathbf{S}_n$, and not from the n individual observations. Note that as we do not make any assumptions on the geometry of the response vectors, the space of vectors is isotropic, with all eigenvector directions having equal a-priori probability. In this setting, the eigenvalues of $\mathbf{S}_n$ are sufficient statistics that capture all useful information for our inference task [8]. Thus, following common practice of many other algorithms, we present a method to infer the number of components only from the eigenvalues of $\mathbf{S}_n$, denoted $\{\ell_j\}_{j=1}^p$, and sorted in decreasing order $\ell_1\geq\ell_2\geq\cdots\geq\ell_p$.

As the number of samples $n\to\infty$, inferring the number of components K from the eigenvalues of the sample covariance matrix is relatively easy. All noise eigenvalues converge with probability one to the same constant σ^2, whereas the signal eigenvalues converge with probability one to $\lambda_j+\sigma^2$. Therefore, σ^2 can be estimated and the number of components K is equal to the number of sample eigenvalues (significantly) larger than σ^2.

This approach, however, fails in the common setting of a noisy high dimensional dataset with a limited number of samples. The reason is that in this case the noise eigenvalues have a significant spread so they are far from being all equal to the same constant. We illustrate this phenomenon by the following example, borrowed from [12]. Consider, for example, $n=10$ samples in a $p=10$ dimensional space from a Gaussian distribution with variance $\sigma^2=1$. The population covariance matrix in this example is $\mathbf{I}_{10}$, yet a typical realization shows the extreme spread of the eigenvalues of the sample covariance matrix:

$$\ell^T = (3.33,\ 2.45,\ 1.78,\ 1.02,\ .564,\ .277,\ .237,\ .15,\ .04,\ .008).$$

Although as $n\to\infty$ all these eigenvalues converge to unity, at these specific values of p and n, $\lambda_{\max}/\lambda_{\min}=O(1000)$. We remark that this spread phenomenon is not due to the low number of samples n but rather depends primarily on the ratio p/n. This issue and its implications for pseudorank estimation are discussed in detail below. Moreover, when $p>n$, in addition to the significant spread of the non-zero eigenvalues, $p-n$ eigenvalues are strictly equal to zero. Thus, for finite values of p and n and particularly for large ratios $p/n>1$, the spectral gap between signal eigenvalues and noise eigenvalues may not be easily detected anymore.

As in other methods, in order to distinguish signal from noise we rely on the notion that large eigenvalues correspond to signal whereas small eigenvalues correspond to noise. The key novel ingredient here is a more precise statistical quantification of what is meant by "large" vs "small". This requires consideration of three theoretical issues. The first is the spread of eigenvalues of pure noise samples for any value of σ, p and n. The second issue is the interaction between noise and signal eigenvalues, and the third issue is an estimate of the a-priori unknown noise level σ. The first issue has been studied extensively in the random matrix theory literature. We present the relevant theoretical results in Section 2.1. For the second and third issues, we develop a novel approach based on matrix perturbation theory, presented in Section 2.2. The resulting algorithm is described in Section 2.3. Some of its theoretical properties are analyzed in Section 2.4.

2.1. Noise eigenvalues, detection limit and random matrix theory

Since our goal is to distinguish between noise and signal eigenvalues, we first describe some known results regarding the spread of pure noise eigenvalues. Consider thus a random sample $\{\mathbf{x}_\nu\}_{\nu=1}^n$ of pure noise observations where each $\mathbf{x}_\nu$ is multivariate Gaussian with zero mean and diagonal covariance matrix $\sigma^2\mathbf{I}_p$. Let $\mathbf{S}_n = \frac{1}{n}\Sigma_\nu \mathbf{x}_\nu \mathbf{x}_\nu'$, then $\frac{n\mathbf{S}_n}{\sigma^2}$ follows a Wishart distribution with parameters n, p. The distribution of its eigenvalues has been a subject of intensive research for many decades [1].

For our purposes, the key quantity of interest is the distribution of the *largest* noise eigenvalue as a function of p and n. While a closed form analytical expression is not available, substantial progress has been made in recent years. It was proven in [10] (for complex-valued observations) and in [11] (real-valued) that in the limit $p,n\to\infty$, with $p/n=c$ fixed, the distribution of the largest eigenvalue converges to a Tracy–Widom distribution,

$$\Pr\left\{\ell_1 < \sigma^2\left(\mu_{n,p} + s\sigma_{n,p}\right)\right\} \to F_\beta(s) \tag{3}$$

where F_β denotes the Tracy–Widom distribution of order β, and $\beta=1,2$ corresponds to real or complex-valued observations, respectively. As described in [12], for real-valued observations the following expressions

$$\begin{aligned}\mu_{n,p} &= \frac{1}{n}\left(\sqrt{n-1/2}+\sqrt{p-1/2}\right)^2,\\ \sigma_{n,p} &= \frac{1}{n}\left(\sqrt{n-1/2}+\sqrt{p-1/2}\right)\left(\frac{1}{\sqrt{n-1/2}}+\frac{1}{\sqrt{p-1/2}}\right)^{1/3},\end{aligned} \tag{4}$$

give an $O_P(p^{-2/3})$ rate of convergence in Eq. (3). For complex-valued observations the definitions of $\mu_{n,p}$ and $\sigma_{n,p}$ giving the same convergence rate are more involved and appear in [14].

The Tracy–Widom distribution F_β can be explicitly computed from the solution of a second order Painlevé ordinary differential equation [11,12]. While Eq. (3) holds in the limit $p,n\to\infty$, it has been numerically and theoretically shown to be a very good approximation for finite but large p and n.

Therefore, if the noise level σ is explicitly known, a statistical procedure to distinguish a signal eigenvalue ℓ from noise at an asymptotic significance level α is to check whether

$$\ell > \sigma^2\left(\mu_{n,p} + s(\alpha)\sigma_{n,p}\right) \tag{5}$$

where the value of $s(\alpha)$ depends on the required significance level, and can be found by inverting the Tracy–Widom distribution.[1]

We remark that for known σ, the test (Eq.(5)) is essentially Roy's largest root test to check for sphericity of a covariance matrix, and follows from the union-intersection principle [31,16]. Eq. (3) provides the asymptotic thresholds for a given confidence level α. One key result of the present paper is an accurate estimation of the unknown value of σ, needed to perform this test.

2.1.1. Limit of detection and identifiability of small variance components

Eq. (3) shows that unless $n \gg p$, the largest eigenvalue due to noise can be considerably larger than σ^2. This raises the question of identifiability of small variance components from the eigenvalues of the sample covariance matrix. This issue has also received considerable attention in recent years [2,29,25]. The key result is the presence of a *phase transition phenomenon*. For example, consider the model (1) with a single component and one large population eigenvalue λ. Then, in the joint limit $p,n\to\infty$, $p/n=c$, the largest eigenvalue of the sample covariance matrix converges with probability one to

$$\lambda_{\max}(\mathbf{S}_n) = ||\mathbf{S}_n|| \to \begin{cases} \sigma^2\left(1+\sqrt{c}\right)^2 & \text{if } \lambda < \sigma^2\sqrt{c} \\ \left(\lambda+\sigma^2\right)\left(1+c\frac{\sigma^2}{\lambda}\right) & \text{if } \lambda \geqslant \sigma^2\sqrt{c} \end{cases} \tag{6}$$

Therefore, for a single component to be identified, its population eigenvalue must be larger than the critical value

$$\lambda_{\text{crit}} = \sigma^2\sqrt{\frac{p}{n}}. \tag{7}$$

For the case of K components, in the limit $p,n\to\infty$, each component behaves "independently" and to be detectable its eigenvalue must be larger than the critical value of Eq. (7).

This result also shows the importance of incorporating possible prior knowledge, such as smoothness of the response vectors $\mathbf{v}_j$ or their approximate sparse representation in some basis of $\mathbb{R}^p$. Consider for example an orthonormal projection (dimensionality reduction / compression transformation) $T:\mathbb{R}^p\to\mathbb{R}^k$ of the general form

$$T\mathbf{x} = (\mathbf{x}\cdot\mathbf{a}_1, \ldots, \mathbf{x}\cdot\mathbf{a}_k),$$

where $\{\mathbf{a}_j\}_{j=1}^k$ are orthonormal vectors in $\mathbb{R}^p$. Then, if the original signals $\mathbf{x}_\nu$ are i.i.d. samples from Eq. (1), the compressed signals $T\mathbf{x}_\nu$ also follow a linear mixture model with additive Gaussian noise. Let u denote a low variance principal component with variance $\text{Var}(u)$ and direction $\mathbf{v}$. In the original space $\mathbb{R}^p$, the identifiability condition is $\text{Var}(u)||\mathbf{v}||^2 > \sigma^2\sqrt{p/n}$. However, in the lower dimensional space $\mathbb{R}^k$, the asymptotic condition for its identifiability is now

$$\text{Var}(u)||T\mathbf{v}||^2 > \sigma^2\sqrt{\frac{k}{n}}.$$

Therefore, if the dimensionality reduction scheme is able to represent the signals with a few significant features, $||T\mathbf{v}|| \approx ||\mathbf{v}||$ and $k \ll p$, then a small variance component may be identified in the reduced space but not in the original high dimensional space. This analysis provides a theoretical justification for compression of the signals or use of regularization methods prior to rank determination, for example, using smooth PCA [35]. We remark that a similar analysis, highlighting the importance of feature selection, also applies to the performance of multivariate calibration methods such as partial least squares, see [33,27].

Unless otherwise noted, in what follows we thus assume that all signal eigenvalues $\lambda_1,\ldots,\lambda_K$ are significantly *above* the critical value, and so we should be able to correctly identify the true number of components. Based on Eq. (5), our approach is to perform a sequence of hypothesis tests, at each step testing the significance of the kth largest eigenvalue as arising from signal or from noise. To employ this approach, an estimate of noise level is required and this is described in the next section. The resulting pseudorank estimation algorithm is described in Section 2.3.

2.2. Estimation of noise variance σ^2

Consider a model of rank K with covariance matrix (Eq.(2)). In the unknown basis $\mathbf{W}$ which diagonalizes the population covariance matrix $\boldsymbol{\Sigma}$, the sample covariance matrix takes the form

$$\mathbf{W}'\mathbf{S}_n\mathbf{W} = \begin{pmatrix} z_1 & & & & \\ & \ddots & & & \\ & & z_K & & \\ \hline & & & z_{K+1} & \\ & & & & \ddots \\ & & & & z_p \end{pmatrix} + \begin{pmatrix} \text{off} \\ \text{diagonal} \\ \text{elements} \end{pmatrix} \tag{8}$$

[1] Code for computing the Tracy–Widom distributions is freely available online, for example at http://math.arizona.edu/~momar/research.htm.

where for $j=1,\ldots,p$

$$z_j = \frac{1}{n}\sum_{\nu=1}^{n}\left(\mathbf{x}_\nu \cdot \mathbf{w}_j\right)^2$$

are random variables that capture the sample variances in the directions $\mathbf{w}_j$. Assuming a model of rank K, all projections $\mathbf{w}_j$ for $j>K$ contain only noise contributions, and hence, averaging over all noise realizations, $\mathbb{E}\{z_j\}=\sigma^2$. Therefore, an unbiased estimator of σ^2 is the average of $z_{K+1},\ldots,z_p$,

$$\begin{aligned}\sigma^2_{\text{unbiased}} &= \frac{1}{p-K}\sum_{j=K+1}^{p} z_j = \frac{1}{p-K}\left[\mathrm{Tr}(\boldsymbol{S}_n)-\sum_{j=1}^{K} z_j\right]\\ &= \frac{1}{p-K}\left[\sum_{j=K+1}^{p}\ell_j+\sum_{j=1}^{K}\left(\ell_j-z_j\right)\right].\end{aligned}$$

Unfortunately, the diagonalizing basis $\mathbf{W}$ is unknown, and to estimate σ^2 we need an estimate for $\sum_{j=1}^{K}(\ell_j-z_j)$. This gives

$$\sigma^2_{\text{est}} = \frac{1}{p-K}\left[\sum_{j=K+1}^{p}\ell_j+\left\{\sum_{j=1}^{K}\left(\ell_j-z_j\right)\right\}_{\text{est}}\right]. \tag{9}$$

Assuming K is known, a simple solution is to replace z_j by ℓ_j for $j=1,\ldots,K$. This gives the well known *real error function* (REF) as the estimate for the noise variance [20],

$$\sigma^2_{\text{REF}} = \frac{1}{p-K}\sum_{K+1}^{p}\ell_j. \tag{10}$$

In the PCA decomposition of the sample covariance matrix $\boldsymbol{S}_n$, for any integer q, the subspace with the largest variance is the one spanned by the first q principle components [13]. This means that $\sum_{j=1}^{q}\ell_j \geqslant \sum_{j=1}^{q} z_j$, or $\sum_{j=q+1}^{p}\ell_j \leqslant \sum_{j=q+1}^{p} z_j$. Therefore, Eq. (10) yields a downward biased estimator for σ^2. For high dimensional noisy data with relatively few samples this bias can become significant and lead to overestimation of the number of components.

2.2.1. Quantifying the bias of the real error function

We analyze the effect of estimating z_j by ℓ_j in Eq. (9). For simplicity, we consider a model with a single factor and eigenvalue λ, where we estimate z_1 by ℓ_1. We view the off diagonal elements in Eq. (8) as a small perturbation and expand the largest eigenvalue in terms of this perturbation (see [25]). This gives

$$\ell_1 = z_1 + \sum_{j=2}^{p}\frac{\left(\mathbf{W}'\mathbf{S}_n\mathbf{W}\right)^2_{ij}}{z_1-z_j}+o_P\left(\frac{1}{n}\right).$$

Taking averages gives

$$\mathbb{E}\{\ell_1\} = (\lambda+\sigma^2)\left(1+\frac{p-1}{n}\frac{\sigma^2}{\lambda}\right)+o\left(\frac{1}{n}\right). \tag{11}$$

Notice that this coincides with the asymptotic result in the joint limit $p,n\to\infty$, Eq. (6). Plugging this expected value in Eq. (10) with $K=1$ gives

$$\mathbb{E}\left\{\sigma^2_{\text{REF}}\right\}\approx\sigma^2\left(1-\frac{1}{n}\frac{\lambda+\sigma^2}{\lambda}\right)\leqslant\sigma^2\left(1-\frac{1}{n}\right). \tag{12}$$

This shows that even for significantly large signals with high SNR ($\lambda\gg\sigma^2$), the real error function estimate of Eq. (10) gives a relative downward bias of $O_P(n^{-1})$. This bias is significant when we have a limited number of samples n. As shown in the simulations in Section 3, using this noise estimator for rank determination may give higher rank estimates than the true rank.

2.2.2. An improved self-consistent method to estimate noise variance

To derive a less biased noise estimation we consider the term $\sum_{j=1}^{K}(\ell_j-z_j)$ in Eq. (9), neglected in the REF estimator. Denote by $\tilde{\mathbf{W}}$ the basis which diagonalizes the upper left submatrix of $\mathbf{W}'\mathbf{S}_n\mathbf{W}$. In this basis the sample covariance matrix takes the form

$$\tilde{\mathbf{W}}'\mathbf{S}_n\tilde{\mathbf{W}} = \left(\begin{array}{cc|cc} \rho_1 & & 0 & \\ & \ddots & & B' \\ 0 & & \rho_K & \\ \hline & & z_{K+1} & * \\ & B & & \ddots \\ & & * & z_p \end{array}\right) \tag{13}$$

where * denotes unknown non-zero random variables.

By construction, the upper left submatrix of $\tilde{\mathbf{W}}'\mathbf{S}_n\tilde{\mathbf{W}}$ captures the signal subspace, with $\tilde{\mathbf{w}}_1,\ldots,\tilde{\mathbf{w}}_K$ spanning the same subspace of $\mathbb{R}^p$ as $\mathbf{w}_1,\ldots,\mathbf{w}_K$. Hence,

$$\sum_{j=1}^{K} z_j = \sum_{j=1}^{K}\rho_j$$

and

$$\sum_{j=1}^{K}\left(\ell_j-z_j\right) = \sum_{j=1}^{K}\left(\ell_j-\rho_j\right).$$

The matrix $\mathbf{B}$ captures the signal-noise interactions, whereas the lower right submatrix is pure noise. The diagonal elements of $\tilde{\mathbf{W}}'\mathbf{S}_n\tilde{\mathbf{W}}$ are all $O_P(1)$, while the off-diagonal elements are $O_P(n^{-0.5})$. Thus, we can decompose the matrix $\tilde{\mathbf{W}}'\mathbf{S}_n\tilde{\mathbf{W}}$ as a primary matrix with entries $O_P(1)$ and a perturbation matrix with entries $O_P(n^{-1/2})$,

$$\tilde{\mathbf{W}}'\mathbf{S}_n\tilde{\mathbf{W}} = \left(\begin{array}{cc|cc} \rho_1 & & 0' & \\ & \ddots & & \\ & & \rho_K & \\ \hline & & z_{K+1} & \\ & 0 & & \ddots \\ & & & z_p \end{array}\right) + \left(\begin{array}{cc|cc} 0 & & B' & \\ & & 0 & * \\ \hline & B & & \ddots \\ & & * & 0 \end{array}\right) \tag{14}$$

For n sufficiently large, the first K eigenvalues $\ell_1,\ldots,\ell_K$ of $\tilde{\mathbf{W}}'\mathbf{S}_n\tilde{\mathbf{W}}$ can be viewed as perturbations of $\rho_1\ldots,\rho_K$, and we can expand $\ell_j-\rho_j$ as a series in the elements of the perturbation matrix. This gives

$$\ell_j = \rho_j + \sum_{i=K+1}^{p}\frac{b^2_{ij}}{\rho_j-z_i}+O_P\left(\frac{1}{n^{1.5}}\right), \qquad 1\leqslant j\leqslant K. \tag{15}$$

A key property of the representation (14) and Eq. (15) is the approximate de-coupling between the different signal eigenvalues, namely that to leading order in $n^{-0.5}$, $\ell_j-\rho_j$ is affected only by the signal-noise interaction in the jth row.

Recall that our goal is to approximate $\sum_{j=1}^{K}(\ell_j-\rho_j)$. With this goal in mind, Eq. (15) is still not directly useful as it contains the unknown random variables b^2_{ij}, z_i and ρ_j. Therefore, we first perform a "moment-method" whereby we approximate the random variables b^2_{ij} and z_i by their expected means. The random variable b_{ij} is the sample covariance over n samples of the two independent random variables $\mathbf{x}_\nu\cdot\tilde{\mathbf{w}}_i$ and $\mathbf{x}_\nu\cdot\tilde{\mathbf{w}}_j$. Since for $i>K$, $\tilde{\mathbf{w}}_i=\mathbf{w}_i$ is fixed, the first random variable is Gaussian with variance σ^2. The second random variable has sample variance ρ_j. Therefore, applying the moment-method we replace b^2_{ij} by $\mathbb{E}\left(b^2_{ij}\right)=\frac{1}{n}\rho_j\sigma^2$ where both ρ_j and σ^2 are still unknown. Neglecting terms which are $o_P(n^{-1})$, gives

$$\ell_j-\rho_j\approx\frac{\sigma^2}{n}\sum_{i=K+1}^{p}\frac{\rho_j}{\rho_j-\sigma^2}=\frac{\sigma^2}{n}(p-K)\frac{\rho_j}{\rho_j-\sigma^2}. \tag{16}$$

According to Eq. (9) an estimate for the noise level depends on the bias in the eigenvalues $\ell_j-\rho_j$. Eq. (16) shows, in turn, that this bias depends on the (unknown) noise level itself. Thus, combining these two equations gives an approximately *self-consistent* method of noise

estimation. In practical terms, this amounts to solving the following non-linear system of $K+1$ equations involving the $K+1$ unknowns $\hat{\rho}_1,\ldots,\hat{\rho}_K$ and σ^2_{KN}:

$$\sigma^2_{\mathrm{KN}} - \frac{1}{p-K}\left[\sum_{j=K+1}^{p} \ell_j + \sum_{j=1}^{K}\left(\ell_j - \hat{\rho}_j\right)\right] = 0, \tag{17}$$

$$\hat{\rho}_j^2 - \hat{\rho}_j\left(\ell_j + \sigma^2_{\mathrm{KN}} - \sigma^2_{\mathrm{KN}}\frac{p-K}{n}\right) + \ell_j\sigma^2_{\mathrm{KN}} = 0. \tag{18}$$

This system of equations can be solved iteratively. Starting with an initial guess $\hat{\sigma}_0^2$ for σ^2_{KN}, solving Eq. (18) gives an estimate of $\hat{\rho}_1,\ldots,\hat{\rho}_k$, which then leads to an improved approximation of σ^2_{KN} via Eq. (17). We repeat this process iteratively until the absolute relative difference is below a small threshold.

To start from a relatively accurate initial guess, we use Eq. (6) to conclude that in the case of K signals,

$$\mathbb{E}\{\sigma^2_{\mathrm{REF}}\} \approx \sigma^2\left(1 - \sum_{j=1}^{K}\frac{1}{n}\frac{\lambda_j+\sigma^2}{\lambda_j}\right) \leqslant \sigma^2\left(1-\frac{K}{n}\right).$$

Therefore, a good initial guess is the following modification of the real error function:

$$\hat{\sigma}_0^2 = \frac{\sigma^2_{\mathrm{REF}}}{1-\frac{K}{n}}. \tag{19}$$

We remark that Eq. (19) was considered in [5] as an accurate estimate of noise variance (better than the REF estimate). Numerical simulations show that when starting from this initial guess, the iterative process typically converges in less than 10 iterations.

2.2.3. Quantifying the bias of the improved estimator σ^2_{KN}

For simplicity, we consider a model with a single factor and eigenvalue λ. In this case,

$$\sigma^2_{\mathrm{KN}} = \sigma^2_{\mathrm{REF}} + \frac{1}{p-1}\left(\ell_1 - \hat{\rho}_1\left(\sigma^2_{\mathrm{KN}}, \ell_1\right)\right).$$

Since the bias in σ^2_{REF} is $O_P\left(n^{-1}\right)$, we write $\sigma^2_{\mathrm{KN}} = \sigma^2\cdot\left(1+\frac{x}{n}\right)$, where x is an $O_P(1)$ random variable. Plugging this expression into Eq. (18), and expanding in powers of $O(1/n)$ gives that $\ell_1 - \hat{\rho}_1 = \frac{1}{n}\frac{\sigma^2\ell_1(p-1)}{\ell_1-\sigma^2}\left(1+O_P(1/n)\right)$, which to leading order does not depend on x. Combining this with Eq. (12) yields,

$$\begin{aligned}\frac{\sigma^2 - \mathbb{E}\left[\sigma^2_{\mathrm{KN}}\right]}{\sigma^2} &= \frac{1}{n}\frac{\lambda+\sigma^2}{\lambda} - \frac{1}{\sigma^2}\frac{1}{p-1}\mathbb{E}\left(\ell_1 - \hat{\rho}\left(\sigma^2_{\mathrm{KN}}, \ell_1\right)\right) + O(1/n^2)\\ &= \frac{1}{n}\frac{\lambda+\sigma^2}{\lambda} - \frac{1}{n}\mathbb{E}\left[\frac{\ell_1}{\ell_1-\sigma^2}\right] + O(1/n^2)\\ &= \frac{1}{n}\frac{\sigma^2}{\lambda}\mathbb{E}\left[1 - \frac{1}{(\ell_1-\sigma^2)/\lambda}\right] + O(1/n^2)\end{aligned}$$

Using Eq. (11) for the expectation of ℓ_1 gives

$$\frac{\sigma^2 - \mathbb{E}\left(\sigma^2_{\mathrm{KN}}\right)}{\sigma^2} = O\left(\frac{1}{n^2}\right) \tag{20}$$

2.3. Pseudorank estimation algorithm

We are now ready to present our algorithm.[2] It is based on a sequence of nested hypothesis tests, for $k=1,2,\ldots,\min(p,n)-1$,

$\mathcal{H}_0$: at : at least k components *vs*.

$\mathcal{H}_1$: at most $k-1$ components.

For each value of k we estimate the noise level σ assuming $\ell_{k+1},\ldots,\ell_p$ correspond to noise, and test the likelihood of the kth eigenvalue ℓ_k as arising from a signal or from noise, as follows:

$$\ell_k > \sigma^2_{\mathrm{KN}}(k)\left(\mu_{n,p-k} + s(\alpha)\sigma_{n,p-k}\right) \tag{21}$$

where α is a user-chosen confidence level, and $s(\alpha)$ is the corresponding value computed by inversion of the Tracy–Widom distribution. If Eq. (21) is satisfied we accept $\mathcal{H}_0$ and increase k by one. Otherwise, we output $\hat{K} = k-1$. In other words,

$$\hat{K} = \arg\min_k\left\{\ell_k \leqslant \sigma^2_{\mathrm{KN}}(k)\left(\mu_{n,p-k} + s(\alpha)\sigma_{n,p-k}\right)\right\} - 1.$$

Note that to test the kth component, we compare in Eq. (21) its eigenvalue to that of a random matrix of n samples with $p-k$ dimensions. This is consistent with the decomposition (Eq. (13)) where the noise matrix has dimension $p-k$.

2.4. Consistency of pseudorank estimation algorithm

In this section we discuss some limiting properties of our noise and pseudorank estimators. A detailed theoretical performance analysis for finite p,n is beyond the scope of this paper [37]. Since our main interest is in high dimensional settings with $p \gg 1$ and with sample sizes n comparable to the dimension p, we focus on the asymptotic limit $p,n\to\infty$ with $p/n\to c$. Since the bias of the simpler REF noise estimator is $O(1/n)$ which also converges to zero as $p,n\to\infty$, the theorems below hold also for the simpler REF noise estimation. The proofs appear in Appendix A.

2.4.1. Consistency of noise estimation

The following lemma shows that regardless of the assumed number of signals k and the true number K, in the asymptotic limit the noise estimator converges to the correct unknown value σ^2.

Lemma 1. *In the joint limit $p, n\to\infty$, $p/n\to c>0$, for any finite k, the noise estimator $\sigma^2_{KN}(k)$ given by* Eqs. (17) and (18), *is consistent,*

$$\lim_{p,n\to\infty}\sigma^2_{\mathrm{KN}}(k) = \sigma^2.$$

2.4.2. Consistency of rank estimation

The following theorem shows that in the asymptotic limit $p,n\to\infty$, provided all signal eigenvalues are above the phase transition threshold, our rank estimator reports at least the correct number of components.

Theorem 1. *Consider n i.i.d. samples from the model* (1) *with K components, whose population eigenvalues $\lambda_j > \lambda_{\mathrm{crit}} = \sigma^2\sqrt{p/n}$ for $j=1,\ldots,K$. Then, in the asymptotic limit $p,n\to\infty$, $p/n\to c>0$,*

$$\lim_{p,n\to\infty}\Pr\left\{\hat{K}\geqslant K\right\} = 1.$$

We now consider the misidentification probability. For simplicity we consider the case of a single signal above the phase transition ($K=1$), and in the following Theorem show that in the joint limit our rank estimator reports the exact number of signals with probability $1-\alpha$ and overestimates the number of signals with probability α.

Theorem 2. *Consider n i.i.d. samples from the model* (1) *with a single component, with signal eigenvalue $\lambda > \lambda_{\mathrm{crit}} = \sigma^2\sqrt{p/n}$. Then, in the asymptotic limit $p,n\to\infty$, $p/n\to c>0$, the misidentification probability converges to the significance level α,*

$$\lim_{p,n\to\infty}\Pr\left\{\hat{K} > K\right\} = \alpha.$$

[2] A matlab implementation of our algorithm can be downloaded from http://www.wisdom.weizmann.ac.il/ ~nadler.

Table 1
Settings for synthetic simulations

Description	β	K	λ	Setting	$c=p/n$
High signal variances	1	2	(200,50)	A1	4
				A2	1
Range of signal variances	1	4	(200,50,10,5)	B1	4
				B2	1
Low signal variances	2	2	(9,2)	C1	2
				C2	1

To prove Theorem 2 we will first prove the following lemma, which claims that in the presence of a single signal, in the joint limit, the second largest eigenvalue ℓ_2 has a Tracy–Widom distribution.

Lemma 2. *Consider a setting with a single signal* $\lambda>\lambda_{\text{crit}}=\sigma^2\sqrt{p/n}$. *Then, in the asymptotic limit* $p,n\to\infty$, $p/n\to c>0$, *the second largest eigenvalue (which corresponds to noise) has asymptotically the same Tracy–Widom distribution as that of a pure noise Wishart matrix.*

We conjecture that Theorem 2 and Lemma 2 hold also in the case of multiple signals. We remark that in [5] the assumption that the secondary eigenvalues of a test data matrix can be approximated by the eigenvalues of a random matrix with the same size were explored. Our Lemma 2 above provides a theoretical justification for this approximation.

3. Simulation results

To illustrate the performance of our algorithm, we present simulation results under a wide range of conditions. We compare our algorithm to the following other common methods: *i*) a simplified version of our algorithm which uses the real-error-function to estimate the noise level, *ii*) an algorithm recently suggested by Rao and Edelman [30], specifically designed for the large p small n setting, *iii*) Malinowski's F-test (at 5% significance level), and *iv*) a modified Faber–Kowalski F-test (at 1% significance level). For comparison purposes we denote our algorithm as KN-algorithm, for which we use a value of $\alpha_{\text{KN}}=0.5\%$ as the significance level. The simulation results show that our algorithm has a consistent high performance under a wide range of conditions and is typically as good as or better than the above other algorithms.

For the paper to be relatively self contained, we first describe the algorithms we compare to. Later we present two sets of simulations. In the first set we use synthetic data generated with a few parameters, chosen to represent a variety of conditions and to emphasize interesting behavior. For the second set of simulations, we consider the simulated dataset recently analyzed by Levina et. al. [18], consisting of a chemical mixture model with four or six components.

3.1. Other algorithms

3.1.1. Real error function algorithm (REF)

As described in Section 2, in the KN algorithm we use an improved estimator of the noise variance σ^2. Instead of using this estimator we can use the simpler real error function described in Eq. (10). We refer to this algorithm, which is a simplified version of the KN algorithm, as the REF algorithm. It is of course interesting to compare the two in order to examine the benefit of our noise estimator. We run the REF algorithm with a significance level of $\alpha_{\text{REF}}=0.5\%$.

3.1.2. Rao and Edelman's estimator (RE)

This estimator [30] is based on an information theoretic approach. It chooses a model which minimizes (an approximation of) the Akaike Information Criterion (AIC), essentially optimizing a trade-off between the complexity of the estimated model and how well the model fits the data. The RE estimator is defined as follows:

$$t_k = p\left[(p-k)\frac{\sum_{i=k+1}^{p}\ell_i^2}{\left(\sum_{i=k+1}^{p}\ell_i\right)^2}-\left(1+\frac{p}{n}\right)\right]-\left(\frac{2}{\beta}-1\right)\frac{p}{n}$$
$$\hat{K}_{\text{RE}} = \arg\min_k\left\{\frac{\beta}{4}\left(\frac{n}{p}\right)^2 t_k^2+2(k+1)\right\}0\leqslant k<\min(p,n) \tag{22}$$

The parameter $\beta=1$ for real-valued Gaussian noise and $\beta=2$ for complex Gaussian noise. As is typical for information theory based estimators, the RE estimator is parameter-free, and thus has no means of adjusting its significance level. An analysis of its asymptotic significance level and its interesting non-trivial behavior as a function of p/n appears in the appendix.

3.1.3. Malinowski's F-test (F-test)

Malinowski's F-test is also based on a sequence of nested hypothesis tests. Letting $q=\min\{n,p\}$, and advancing from $k=q-1$ to

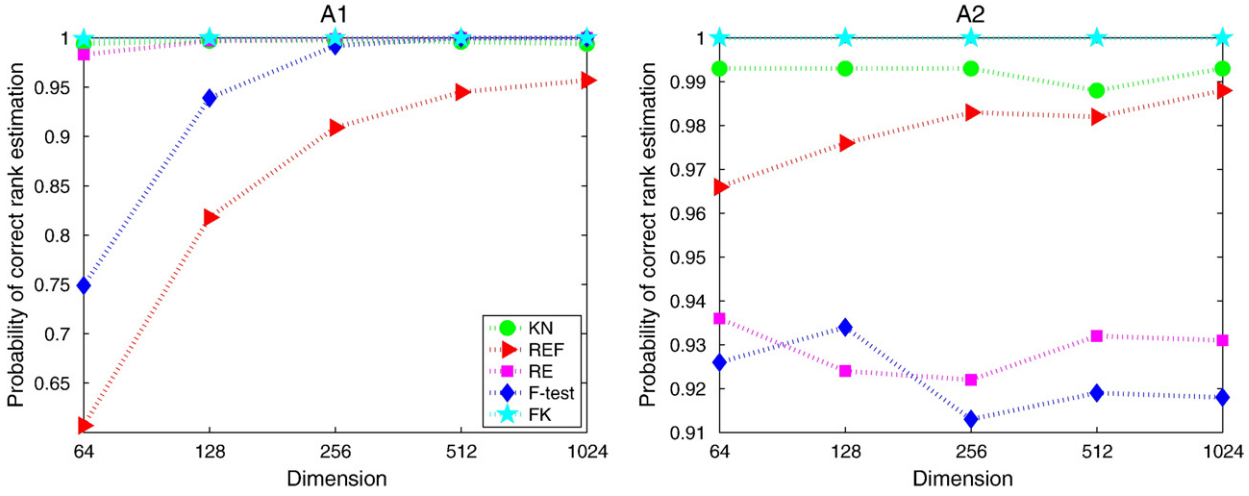

Fig. 1. Empirical probability of correct rank estimation for various p and various algorithms, for settings A1(left) and A2(right): real valued data, $K=2$, $\lambda=(200;\ 50)$, $c=4$ (left) or $c=1$ (right).

$k=1$, in each step it is tested whether the k'th population eigenvalue λ_k equals $\lambda_{k+1},\ldots,\lambda_q$ using the following F-ratio

$$F_k(1,q-k)=\frac{\ell_k}{\sum_{j=k+1}^{q}\ell_j}\frac{\sum_{j=k+1}^{p}(n-j+1)(p-j+1)}{(n-k+1)(p-k+1)}.$$

If the significance level of the test statistic is greater than α, then the null hypothesis is accepted and k is decreased by one. Otherwise, $\hat{K}=k$. In other words,

$$\hat{K}=\arg\max_k\{F_k(1,q-k)>f_{1,q-k}(1-\alpha/100)\}.$$

According to [21], a significance level of $\alpha=5\%$ tends to underestimate the rank, whereas the 10% level tends to overestimate it. Here we shall use $\alpha_{F\text{-test}}=5\%$.

3.1.4. Faber and Kowalski's modified F-test (FK)

In [4], Faber and Kowalski suggested a modification to Malinowski's F-test, by changing the degrees of freedom, based on Mandel's degrees of freedom analysis. They propose the following F-ratio

$$F_k(\nu_1,\nu_2)=\frac{\ell_k}{\sum_{j=k+1}^{p}\ell_j}\frac{\nu_2}{\nu_1}$$

where

$$\nu_1=n\cdot\mathbb{E}\{\ell_1(k)\},\qquad \nu_2=(n-k+1)(p-k+1)-\nu_1$$

and $\mathbb{E}\{\ell_1(k)\}$ is the expectation of the largest eigenvalue of a $(p-k)\times(p-k)$ pure noise sample covariance matrix with $n-k$ samples and $\sigma^2=1$. While Faber and Kowalski used simulations to approximate ν_1, for our simulations we simply approximate $\mathbb{E}\{\ell_1(k)\}$ by the explicit asymptotic formula (6). The FK algorithm starts from $k=1$ and thus the estimated pseudorank $\hat{K}$ is defined as

$$\hat{K}=\arg\min_k\{F_k(\nu_1,\nu_2)<f_{\nu_1,\nu_2}(1-\alpha/100)\}-1.$$

For the FK algorithm we use a significance level of $\alpha_{FK}=1\%$ as suggested by the authors.

In our simulations, the FK algorithm showed very good performance, except at low SNR's. In the appendix we present a theoretical analysis of this algorithm and its connection to our approach.

3.2. Synthetic simulations

We run simulations on both real and complex-valued data (where $\beta=1$ stands for real and $\beta=2$ stands for complex), with different values for the number of components K, their corresponding variances λ_j, and different ratios $c=p/n$ between the dimension and the number of samples. For each choice of K, c and λ_j, we consider a range of values for the dimension, $p=[64, 128, 256, 512, 1024]$. In all settings we use a noise variance of $\sigma^2=1$. We present results for three different choices of the signal variances, and for each such choice we use two different values for c, hence we have six different settings. For each setting we run 1000 MATLAB simulations. Table 1 reviews the parameters used in each of the six settings.

Figs. 1–4 compare the performance of the different algorithms for each of the six settings. Each plot shows the probability to correctly estimate K for different values of p and for the various algorithms. Tables 2–4 give more detailed results for each specific algorithm and setting, and show the probabilities to obtain various values of K. We present these tables only for the more insightful cases. Tables 6 and 7 summarize the results of the synthetic simulations.

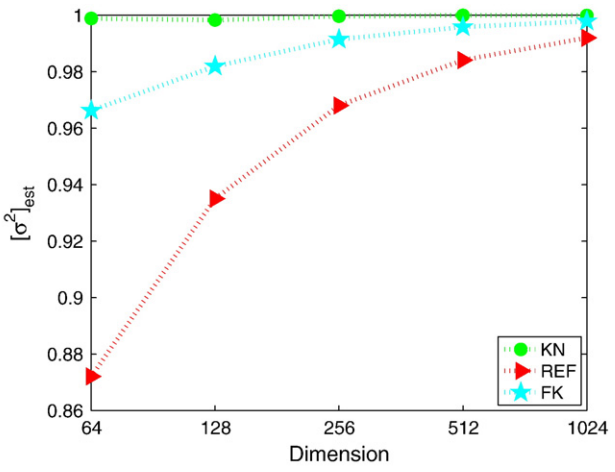

Fig. 2. Empirical mean of σ^2_{est} of the KN estimator, of the real error function, and of the FK algorithm, for setting A1: real data, $K=2$, $\lambda=(200; 50)$, $c=4$.

3.2.1. High signal variances: A1 and A2

In this case we have $K=2$ signals, with variances (200,50). Since $\sigma^2=1$, the population covariance matrix is

$$\Sigma=\left(\begin{array}{cc|c} 200 & & \\ & 50 & \\ \hline & & 0_{p-2}\end{array}\right)+\mathbf{I}_p=\left(\begin{array}{cc|c} 201 & & \\ & 51 & \\ \hline & & \mathbf{I}_{p-2}\end{array}\right).$$

We present results for $c=4$ (setting $A1$) and $c=1$ ($A2$).

According to Eq. (6), with the smallest noise free eigenvalue equal $\lambda=50$, asymptotically as $p,n\to\infty$, the phase transition occurs at $p/n>c_{\text{phase}}=50^2=2500$. Thus, in both settings A1 and A2, distinguishing between signal and noise eigenvalues is supposed to be quite an easy task, as $c\in\{1,4\}\ll c_{\text{phase}}$. Indeed, as shown in Fig. 1, most algorithms accurately estimate the rank for both settings A1 and A2 when p and hence n are sufficiently large. However, the REF algorithm shows rather poor results. This is due to the biased estimation of σ^2, as can be seen in Fig. 2, which compares our estimation of σ^2 with the real error function. The figure shows that our estimator is less biased than the real error function. The significant downward bias in the real error function leads to the wrong identification of some noise eigenvalues as signal, and hence leads to an overestimation of K, as seen in Table 2. As discussed in Section 2.2, our estimation of σ^2 is also downward biased, yet much less significantly. For this reason our method also slightly overestimates K, as seen in Table 3.

As we explain in the appendix, the FK algorithm implicitly estimates the noise variance σ^2. Hence, Fig. 2 shows the implicit FK noise estimator σ^2_{FK} as well. From the figure it is evident that the FK estimator is much less biased than the real error function, yet more biased than the KN estimator.

From a theoretical perspective, decreasing c from 4 to 1 makes the inference task easier. As seen in Fig. 1, both the KN, REF and FK algorithms indeed achieve better results for $c=1$. The RE and F-test algorithms, however, give worse results in this case. For the RE algorithm, this is in accordance with the theoretical analysis in Appendix B.

3.2.2. Wide range of signal variances, B1 and B2

In this case we have $K=4$ signals, with variances (200,50,10,5). We present results for $c=4$ (B1) and $c=1$ (B2). Here the phase transition occurs at $c_{\text{phase}}=5^2=25$, still sufficiently high to allow correct rank determination with high probability, even when $c=4$.

Fig. 3 shows that the KN algorithm performs very well, with the exception of the case $c=4$ and small p (hence also much smaller $n=p/c$), where all algorithms perform poorly. The F-test fails completely to estimate the true number of signals, as it wrongly identifies the small signal eigenvalues as noise, and hence estimates $\hat{K}=2$ or $\hat{K}=3$ instead of the true value $K=4$, see Table 4. The FK

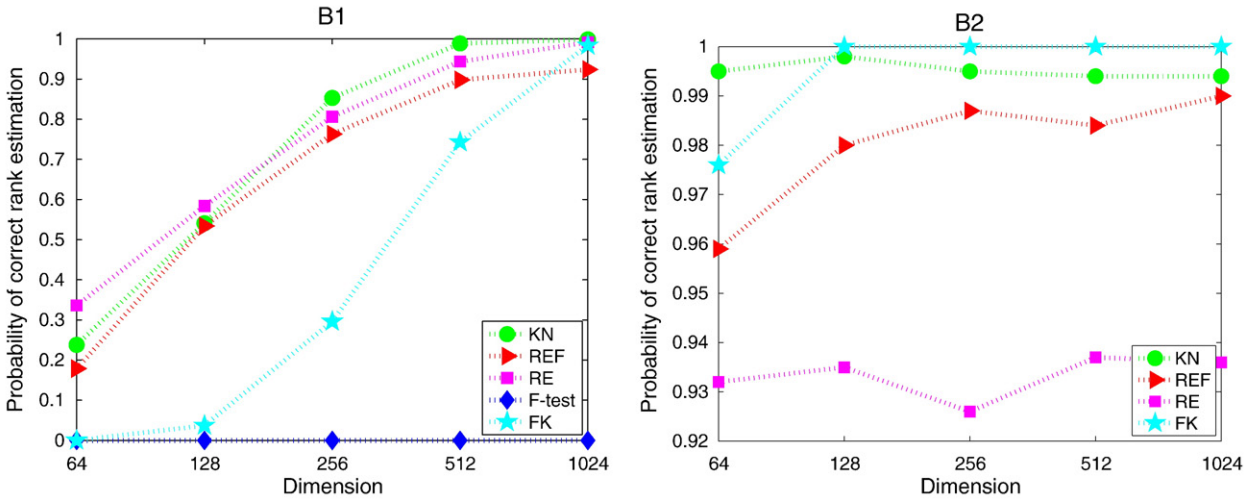

Fig. 3. Empirical probability of correct rank estimation for various p and various algorithms, for settings B1(left) and B2(right): real data, $K=4$, $\lambda=(200; 50; 10; 5)$, $c=4$ (left) or $c=1$ (right). In setting B2, the performance of the F-test is at $P=0$ and is not shown.

algorithm has better results than the F-test, yet not as good as the results of the KN, REF and RE algorithms. Only at $p=1024$ does the FK algorithm achieve a high success probability of 0.99. The RE algorithm behaves slightly worse than the KN algorithm for $c=4$, yet, for the case $c=1$, which is easier from a statistical point of view, its correct identification probability is always below 0.95, even for large p. Again, this is in accord with our theoretical analysis. In Appendix B we present a simulation with the same four signals, but over a wide range of $c=p/n$ values, which shows the complex dependence the RE algorithm has on the parameter c.

3.2.3. Complex data, small signal variances

Here we use the same parameters considered in [30]. Motivated by their focus on signal processing applications, Rao and Edelman tested their algorithm on data contaminated by complex-valued additive Gaussian noise. Hence we use $\beta=2$ and generate complex samples with $K=2$ signals and noise-free variances $(\lambda_1,\lambda_2)=(9,2)$. We present results for $c=2$ (C1) and $c=1$ (C2). Here, the phase transition occurs at $c_{\text{phase}}=2^2=4$. While the two values of c are below this threshold, the first value, $c=2$, is rather close to it. This makes the task of pseudorank estimation much more difficult, and indeed, Fig. 4 shows that for $c=2$ and for small values of p (and n), all five algorithms achieve poor results. Since c is smaller than the critical value, we do expect the algorithms to achieve a success probability close to one as $p,n\to\infty$. To examine this convergence we ran 100 simulations with $p=3000$. The results appear in Table 5, and show that the KN and REF algorithms indeed obtained very high success probabilities. The RE algorithm, however, achieved a success probability only slightly higher than 0.5.

In the setting C1 the REF algorithm seems to perform better than all other algorithms and deserves an explanation. At low SNR's it is likely to misclassify a signal eigenvalue as noise, and hence underestimate K. The REF estimate of σ^2 is downward biased, which, as we explained in setting A1, leads to an upward bias in K . Therefore, only at this specific setting, these two biases, one upward and one downward, roughly compensate each other and generate a better estimation of K.

When c is decreased to $c=1$, both the KN and the REF algorithm show excellent results. The F-test is still unable to detect the small signal eigenvalue. The FK algorithm performs better than before, and

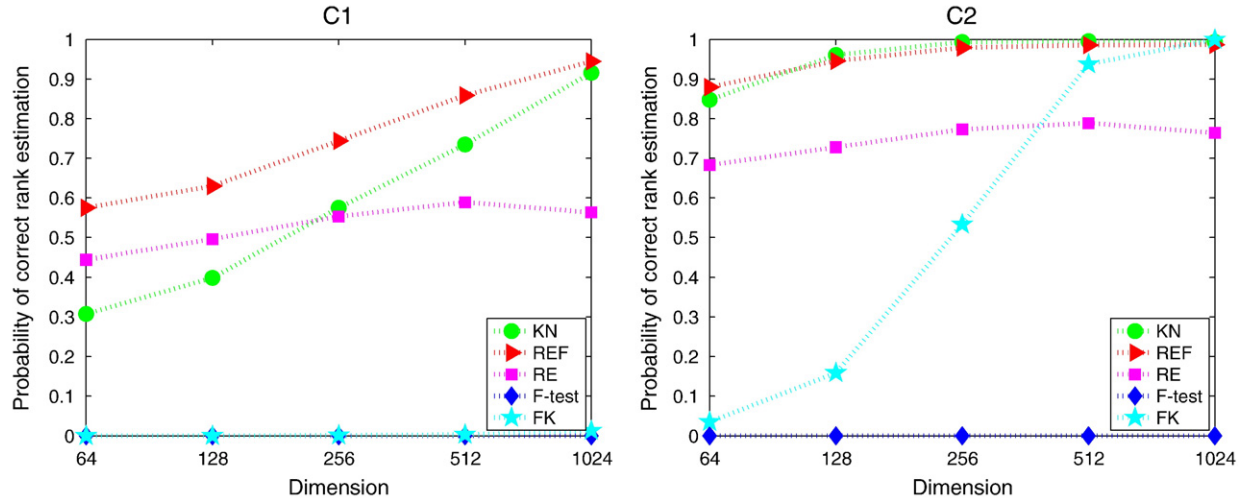

Fig. 4. Empirical probability of correct rank estimation for various p and various algorithms, for settings C1(left) and C2(right): complex data, $K=2$, $\lambda=(9; 2)$, $c=2$ (left) or $c=1$ (right).

Table 2
Empirical probabilities for various values of $\hat{K}$ for the REF algorithm for setting A1: real data, $K=2$, $\lambda=(200, 50)$, $c=4$

p	n	$P(\hat{K}=1)$	$P(\hat{K}=2)$	$P(\hat{K}=3)$	$P(\hat{K}=4)$
64	16	0	**0.607**	0.191	0.068
128	32	0	**0.818**	0.143	0.031
256	64	0	**0.909**	0.089	0.001
512	128	0	**0.945**	0.054	0.001
1024	256	0	**0.957**	0.043	0

The probability of detecting the true rank, $P(\hat{K}=K)$, appears in boldface letters.

Table 3
Empirical probabilities for various values of $\hat{K}$ for the KN algorithm for setting A1: real data, $K=2$, $\lambda=(200, 50)$, $c=4$

p	n	$P(\hat{K}=1)$	$P(\hat{K}=2)$	$P(\hat{K}=3)$	$P(\hat{K}=4)$
64	16	0	**0.992**	0.008	0.001
128	32	0	**0.997**	0.003	0
256	64	0	**0.997**	0.003	0
512	128	0	**0.996**	0.004	0
1024	256	0	**0.994**	0.006	0

The probability of detecting the true rank, $P(\hat{K}=K)$, appears in boldface letters.

shows good results when p is large enough. The RE algorithm achieves a relatively low success probability of only 0.78.

3.2.4. Summary of synthetic simulations

Tables 6 and 7 summarize the results of the synthetic simulations. Table 6 examines the case of low dimension, $p=64$, and shows the probability of correct rank estimation for each of the algorithms and for each of the six simulation settings, averaged over 1000 simulations. Table 7 examines the case of high dimension, $p=1024$, in a similar way.

Summarizing the results of the synthetic simulations, we examined the performance of the algorithms on various types of data: real valued data with two signals bearing high variance, real data with four signals, two of them with high variance and two with hard to detect low variance, and complex data with two very low signals. In all situations the signal eigenvalues are above the phase transition threshold. The KN algorithm shows good results in all situations but the most extreme (B1 and C1 with small p), and its success probability is close to one in all six settings. The REF algorithm shows good results when c is small enough, yet for the high values of c (4 and 2) its estimation of σ^2 is significantly downward biased and hence the estimation for K is upward biased. Even though the number of samples is increased, the RE algorithm typically exhibits worst performance when c is decreased from four to one. In all three cases its success probability does not converge to one as $p,n\to\infty$ when $c=1$. Malinowski's F-test is unable to detect low signals, even when $c=1$. The improved Faber–Kowalski F-test algorithm shows excellent performance when the signals are strong. However, in comparison to our proposed algorithm, it has a lower success probability at settings with low variance signals, especially when c is large and p is small.

Table 4
Empirical probabilities for various values of $\hat{K}$ for the F-test algorithm for setting B2: real data, $K=4$, $\lambda=(200, 50, 10, 5)$, $c=1$

P	n	$P(\hat{K}=1)$	$P(\hat{K}=2)$	$P(\hat{K}=3)$	$P(\hat{K}=4)$
64	16	0	0.605	0.319	**0**
128	32	0	0.477	0.457	**0**
256	64	0	0.359	0.554	**0**
512	128	0	0.251	0.668	**0**
1024	256	0	0.139	0.779	**0**

The probability of detecting the true rank, $P(\hat{K}=K)$, appears in boldface letters.

Table 5
Empirical probabilities for the various algorithms for $p=3000$ for setting C1: complex data, $K=2$, $\lambda=(9, 2)$, $c=2$

Algorithm	$P(\hat{K}=0)$	$P(\hat{K}=1)$	$P(\hat{K}=2)$	$P(\hat{K}=3)$	$P(\hat{K}=4)$
KN	0	0	**0.997**	0.003	0
REF	0	0	**0.989**	0.011	0
RE	0	0.253	**0.533**	0.142	0.06
F-test	1	0	**0**	0	0
FK	0	0.795	**0.205**	0	0

The probability of detecting the true rank, $P(\hat{K}=K)$, appears in boldface letters.

3.3. Chemical mixtures

We now consider pseudorank estimation on the dataset recently considered by Levina et. al. [18]. This simulated data of chemical mixtures is based on two sets of six pure component spectra. The first set consists of five plastics and one bovine bone spectra, which are quite dissimilar from each other and are easy to discriminate. The second set contains five spectra of a fractured mouse tibia bone and one plastic. The differences between the five bone spectra are very subtle and hence are hard to detect. The case of dissimilar spectra is relatively easy and here we focus on the set of similar spectra.

Following Beer's law, or equivalently Eq. (1), each of the n mixture samples is generated in the following way: six weights are randomly drawn, each from a uniform distribution in an interval $[\alpha_j, \beta_j]$, and the weights are later normalized to sum up to $\sum w_j=1$. Each of the six component spectra $\mathbf{v}_j$ is multiplied by the corresponding weight w_j and the weighted spectra are summed up to give a noise-free mixture spectra. Finally, a Gaussian random noise vector $\boldsymbol{\xi}$ with diagonal covariance matrix $\sigma^2\mathbf{I}_p$ is added to the mixture,

$$\mathbf{x} = \sum_{j=1}^{6} w_j \mathbf{v}_j + \sigma\xi.$$

As in [18], we use three different sets for the six intervals and hence three different distributions for the weights, which are later referred to as simulation types 1–3. For simulation type 1, the first four intervals are equal, and the last two are zero, meaning that for this setting the sample weights mix only four spectra and hence $K=4$. For the other two simulation settings the first four intervals are again equal, and the last two intervals are much smaller than them, creating a distinction between four major components and two minor ones.

All in all we have three different distributions for the weights, resulting in three different settings for the noise-free population

Table 6
Summary for $p=64$, showing the empirical probability of correct rank estimation $P(\hat{K}=K)$ of each algorithm in each setting

Setting	KN	REF	RE	F-test	FK
A1	0.994	0.607	0.983	0.749	0.999
A2	0.993	0.966	0.936	0.926	1
B1	0.238	0.179	0.336	0	0
B2	0.995	0.959	0.932	0	0.976
C1	0.308	0.575	0.444	0	0
C2	0.848	0.880	0.683	0	0.035

Table 7
Summary for $p=1024$, showing the empirical probability of correct rank estimation $P(\hat{K}=K)$ of each algorithm in each setting

Setting	KN	REF	RE	F-test	FK
A1	0.994	0.957	1	1	1
A2	0.993	0.988	0.931	0.918	1
B1	0.999	0.924	0.992	0	0.948
B2	0.994	0.990	0.936	0	1
C1	0.916	0.945	0.563	0	0.013
C2	0.994	0.987	0.764	0	1

Table 8
Eigenvalues of the population covariance matrix for the various chemical mixtures

Type	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
1	79.2	$8.14\cdot10^{-3}$	$1.33\cdot10^{-4}$	$1.94\cdot10^{-5}$	0	0
2	83.2	$3.71\cdot10^{-3}$	$1.41\cdot10^{-4}$	$1.71\cdot10^{-5}$	$2.42\cdot10^{-6}$	$1.26\cdot10^{-6}$
3	79.7	$2.81\cdot10^{-3}$	$5.37\cdot10^{-5}$	$8.94\cdot10^{-6}$	$9.44\cdot10^{-7}$	$1.28\cdot10^{-7}$

covariance matrix, whose eigenvalues are given in Table 8. Each of these settings is examined with three different noise levels,

$$\sigma_1 = 5\cdot10^{-4}, \sigma_2 = 1\cdot10^{-3} \text{ and } \sigma_3 = 3\cdot10^{-3}.$$

The dimension of the spectra is $p=815$ and the number of samples is $n=3600$, hence $c=\frac{815}{3600}=0.226$. For each setting we run 1000 MATLAB simulations, and compare the performance of the KN, FK and RE algorithms. We do not consider here the REF and *F*-test algorithms, which generally had worse performance in the simulations of the previous section. Also, we do not compare to the maximum likelihood estimator of intrinsic dimension suggested in [18] for use on spectroscopic data, since from the results in [18] it appears to give worse results on this dataset. However, it should be noted that the MLE of intrinsic dimension algorithm is a very general algorithm which can be successfully used to infer intrinsic dimension of non-linear manifolds. More details regarding the pure spectra and the simulations can be found in [18].

According to Eq. (6), the eigenvalue detection limit for the three noise levels is

$$\lambda_{\text{crit}} = \sigma^2 c^{0.5} = \left(1.19\cdot10^{-7}, 4.76\cdot10^{-7}, 4.28\cdot10^{-6}\right).$$

A signal eigenvalue λ_i is identifiable if it is above this threshold. Table 8 shows that some of the population eigenvalues are *below* the thresholds, which means that at these noise levels we cannot detect some of the signals using the sample covariance eigenvalues. The threshold defines an effective number of identifiable signals, denoted K_{eff}. Table 9 shows the value of K_{eff} for the various settings, for the set of similar spectra. We can see that in three settings, $K_{\text{eff}}<K$.

As shown in Table 10, for simulation type 1 even with the strongest noise level $\sigma_3=3\cdot10^{-3}$ all algorithms achieve excellent results. As shown in Tables 11 and 12, for simulation types 2 and 3 the situation is more complicated. In each of these tables, and for the three possible noise levels, we emphasize the column with the effective number of signals, K_{eff}. These tables show that in most cases $\Pr\{\hat{K}_{\text{KN}}=K_{\text{eff}}\}$ is close to one. Yet, for simulation type 3, for the two borderline cases ($\sigma=\sigma_1$ or $\sigma=\sigma_2$), our algorithm does not detect the smallest signal which is still above the detection threshold. The reason is that although the ratio p/n is small enough, p and n themselves are not sufficiently large.

4. Summary and discussion

In this paper we derived a novel algorithm for pseudorank determination. It is based on a combination of results from random matrix theory and matrix perturbation. As we showed in simulations it achieves similar if not better results, compared to state of the art competing algorithms.

This work considered only the case of homoscedastic noise. However, the basic methodology can be extended to the case of heteroscedastic noise. We remark that modifications to the basic pseudorank estimation algorithm are crucial in this case. For example, if noise has different variance in different variables (such as different noise strengths in different wavelengths), the standard approach will fail. However, our algorithm can be easily modified to handle this case as well. If a specific model for the noise is assumed, then $\mu_{n,p}$ and $\sigma_{n,p}$ can be computed explicitly (at least numerically). Otherwise, at least the first few moments of the noise variance distribution can be estimated and from these approximate values for $\mu_{n,p}$ and $\sigma_{n,p}$ can be derived.

In the development of our algorithm we replaced various random variables by their expected values. Another possible improvement is to consider a Bayesian approach, where different realizations of these random variables yield possibly different rank estimates, and these are averaged according to various priors.

Finally, we remark that the approach presented in this paper can also be applied to other inference problems, such as derivation of confidence intervals for population eigenvalues, given a sample covariance matrix.

Table 9
K_{eff} for various settings

	σ_1	σ_2	σ_3
Sim. type 1 (real $K=4$)	4	4	4
Sim. type 2 (real $K=6$)	6	6	4
Sim. type 3 (real $K=6$)	6	5	4

Table 10
Empirical probabilities for simulation type 1 with noise variance $\sigma_2=3\cdot10^{-3}$

Algorithm	$P(\hat{K}=3)$	$P(\hat{K}=4)$	$P(\hat{K}=5)$
KN	0	**0.995**	0.005
RE	0	**0.998**	0.002
FK	0	**1**	0

The probability of detecting the effective rank, $P(\hat{K}=K_{\text{eff}})$, appears in boldface letters.

Acknowledgments

We thank Iain Johnstone, Peter Bickel and Ofer Zeitouni for interesting discussions and the latter also for his help in proving Lemma 2. We also thank Liza Levina and Amy Wagaman for kindly providing us with the pure spectra of [18]. Part of this work was performed while BN participated in the program "Statistical Theory and Methods for Complex High-Dimensional Data" at the Isaac Newton Institute at Cambridge, UK. BN would like to thank the organizers and the INI for their hospitality. The research of BN was supported by a grant from the Ernst Nathan Biomedical Fund.

Table 11
Empirical probabilities for simulation type 2 for different noise variances

σ_2	K_{eff}	Algorithm	$P(\hat{K}=3)$	$P(\hat{K}=4)$	$P(\hat{K}=5)$	$P(\hat{K}=6)$	$P(\hat{K}=7)$
$5\cdot10^{-4}$	6	KN	0	0	0	**0.995**	0.005
		RE	0	0	0	**1**	0
		FK	0	0	0	**1**	0
$1\cdot10^{-3}$	6	KN	0	0	0	**0.998**	0.002
		RE	0	0	0.717	**0.283**	0
		FK	0	0	0	**1**	0
$3\cdot10^{-3}$	4	KN	0	**0.992**	0.008	0	0
		RE	0.002	**0.994**	0.004	0	0
		FK	0	**1**	0	0	0

For each setting the probability of detecting the effective rank, $P(\hat{K}=K_{\text{eff}})$, appears in boldface letters.

Table 12
Empirical probabilities for simulation type 3 for different noise variances

σ_2	K_{eff}	Algorithm	$P(\hat{K}=3)$	$P(\hat{K}=4)$	$P(\hat{K}=5)$	$P(\hat{K}=6)$
$5\cdot10^{-4}$	6	KN	0	0	0.861	**0.139**
		RE	0	0	0.997	**0.003**
		FK	0	0	1	**0**
$1\cdot10^{-3}$	5	KN	0	0	**0.997**	0.003
		RE	0	0.848	**0.152**	0
		FK	0	0	**1**	0
$3\cdot10^{-3}$	4	KN	0	**0.994**	0.006	0
		RE	0.707	**0.291**	0.002	0
		FK	0	**1**	0	0

For each setting the probability of detecting the effective rank, $P(\hat{K}=K_{\text{eff}})$, appears in boldface letters.

Appendix A. Consistency of the KN algorithm

In this section we prove the theoretical statements of Section 2.4.

Proof of Lemma 1. (Consistency of noise estimation)

Consider Eq. (17). For any finite k, when $p,n\to\infty$, it is solved by $\sigma^2_{KN}=\sigma^2_{REF}$. Therefore,

$$\lim_{p,n\to\infty}\sigma^2_{KN}(k)=\lim_{p,n\to\infty}\sigma^2_{REF}(k)=\sigma^2.$$

□

Proof of Theorem 1. Consider a dataset sampled from Eq. (1) with K components whose eigenvalues are all above the phase transition threshold. As explained in Section 2.3, we estimate the pseudorank by a series of nested hypothesis tests. In the first K hypothesis tests we check, for $j=1,\ldots,K$, whether

$$\ell_j > \sigma^2_{KN}(j)\left(\mu_{n,p-j}+s\sigma_{n,p-j}\right) \tag{23}$$

where $\sigma^2_{KN}(j)$ denotes the estimate of noise variance assuming j components.

Under the assumption that $\lambda_j>\sigma^2\sqrt{p/n}$, it follows from Eq. (6) that

$$\lim_{p,n\to\infty}\ell_j=(\lambda_j+\sigma^2)\left(1+c\frac{\sigma^2}{\lambda_j}\right).$$

Further, according to Lemma 1, in the asymptotic limit $\sigma^2_{KN}(j)\to\sigma^2$. Finally, by definition Eq. (4),

$$\lim_{p,n\to\infty}\sigma_{n,p}=0,\quad \lim_{p,n\to\infty}\mu_{n,p}=\left(1+\sqrt{\frac{p}{n}}\right)^2.$$

Therefore, in the asymptotic limit the inequality in Eq. (23) holds with probability one and $\hat{K}\geqslant K$. □

Proof of Lemma 2. For simplicity, we consider the case $p=n$ ($c=1$) and w.l.g. $\sigma^2=1$. The proof can be easily generalized to arbitrary values of c. Let $\mathbf{U}=[\mathbf{w}_1,\mathbf{w}'_2,\ldots,\mathbf{w}'_p]$ be an (unknown) basis of $\mathbb{R}^p$ which diagonalizes the $(p-1)\times(p-1)$ lower right sub-matrix of $\mathbf{S}_n$, where $\mathbf{w}_1$ is the same as in Eq. (2). In this basis we have

$$\mathbf{U}'\mathbf{S}_n\mathbf{U}=\left(\begin{array}{c|ccc} z_1 & b_2 & \cdots & b_p \\ \hline b_2 & \mu_2 & & \\ \vdots & & \ddots & \\ b_p & & & \mu_p \end{array}\right)$$

where the lower right $(p-1)\times(p-1)$ sub-matrix is a pure noise Wishart matrix, and $\mu_2\geqslant\mu_3\geqslant\cdots\geqslant\mu_p$ are its eigenvalues. The random variables b_j capture the signal-noise interactions. Each b_j has mean zero, variance $z_1\mu_j/n$ and finite fourth moment. The matrix $\mathbf{U}'\mathbf{S}_n\mathbf{U}$ is in the form of an arrowhead matrix and so its eigenvalues ℓ_j satisfy the following secular equation,

$$\lambda-z_1=\sum_{j>1}\frac{b_j^2}{\lambda-\mu_j}. \tag{24}$$

We now show that $\mu_2-\ell_2=O_P\left(\frac{1}{n}\right)$. Since the fluctuations of μ_2 around its asymptotic value $\left(1+\sqrt{(p-1)/n}\right)^2$ are $O_P\left(p^{-2/3}\right)$ and $p^{-2/3}\gg n^{-1}$, the lemma follows.

To prove the lemma it is instructive to see Fig. 5, where the two functions $\lambda-z_1$ and $\sum_{j>1}b_j^2/\left(\lambda-\mu_j\right)$ are plotted. Due to interlacing, $\mu_3<\ell_2<\mu_2$. Thus we introduce the following notation:

$$s_1=\sum_{j>3}\frac{b_j^2}{\ell_2-\mu_j},\qquad s_2=\ell_2-z_1,\qquad b_2^2=\frac{1}{n}\mu_2\xi_2^2,\qquad b_3^2=\frac{1}{n}\mu_3\xi_3^2$$

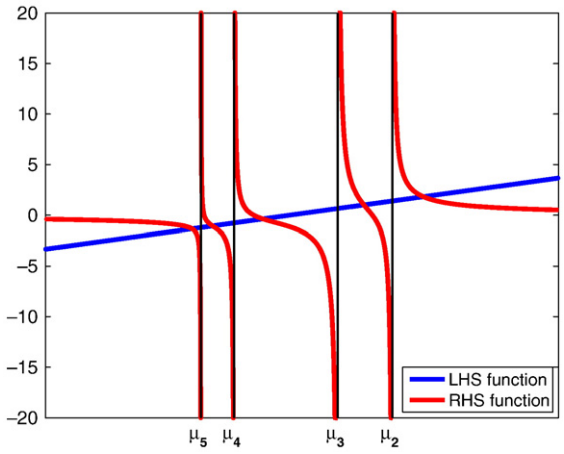

Fig. 5. This figure illustrates Eq. (24). The x axis corresponds to λ. The blue line plots the function on the LHS of the equation, while the red line plots the function on the RHS (which diverges whenever $\lambda=\mu_i$). Equality holds in Eq. (24) whenever the two curves meet. ℓ_2 corresponds to the second intersection from the right, between μ_3 and μ_2.

where $\xi_2,\xi_3\sim\mathcal{N}(0,1)$ and independent. Then at $\lambda=\ell_2$, we can rewrite Eq. (24) as

$$z_1\mu_2\frac{\xi_2^2}{n(\mu_2-\ell_2)}=s_1-s_2+z_1\mu_3\frac{\xi_3^2}{n(\ell_2-\mu_3)}. \tag{25}$$

We now consider the various quantities in Eq. (25) in the asymptotic limit $p,n\to\infty$. First of all, $z_1=(\lambda+1)(1+o_P(1))$ and $\ell_2=4+o_P(1)$. Furthermore, the empirical distribution of the pure noise eigenvalues converges to the Marchenko–Pastur distribution, whose density is given by [22]

$$f_{MP}(x)=\frac{1}{2\pi x}\sqrt{x(4-x)}\qquad x\in(0,4).$$

We have that

$$t_1=\lim_{p,n\to\infty}s_1=\int\frac{(\lambda+1)\mu}{\left(1+\sqrt{c}\right)^2-\mu}f_{MP}(\mu)d\mu=\lambda+1$$
$$t_2=\lim_{p,n\to\infty}s_2=\left(1+\sqrt{c}\right)^2-(\lambda+1)=3-\lambda$$

We define $d=t_1-t_2=2(\lambda-1)$. Note that for $\lambda>\lambda_{crit}=1$, we have that $d>0$. Thus, in the asymptotic limit,

$$(\mu_2-\ell_2)=\frac{1}{n}\frac{z_1\mu_2\xi_2^2}{s_1-s_2+z_1\mu_3\frac{\xi_3^2}{n(\ell_2-\mu_3)}}=\frac{1}{n}\frac{4(\lambda+1)\xi_2^2}{d+o_P(1)}\times(1+o_P(1))=O_P\left(\frac{1}{n}\right).$$

□

Proof of Theorem 2. We consider a model with a single factor and eigenvalue λ. By Lemma 2, $\ell_2-\mu_2=0_P\left(n^{-1}\right)$. By Eq. (20), $\sigma^2_{KN}(2)-\sigma^2=0_P\left(n^{-1}\right)$. According to Eq. (17), $\mu_{n,p-2}=O(1)$ and $\sigma_{n,p-2}=O(n^{-2/3})$. Therefore,

$$\frac{1}{\sigma^2_{n,p-2}}\left[-(\ell_2-\mu_2)+\left(\sigma^2_{KN}(2)-\sigma^2\right)\left(\mu_{n,p-2}+s(\alpha)\sigma_{n,p-2}\right)\right]=O_P\left(n^{-1/3}\right).$$

Thus,

$$\begin{aligned}\Pr\left\{\hat{K}>K\right\}&=\Pr\left\{\ell_2>\sigma^2_{KN}(2)\left(\mu_{n,p-2}+s(\alpha)\sigma_{n,p-2}\right)\right\}\\&=\Pr\left\{\mu_2+(\ell_2-\mu_2)>\left(\sigma^2+\left(\sigma^2_{KN}(2)-\sigma^2\right)\right)\left(\mu_{n,p-2}+s(\alpha)\sigma_{n,p-2}\right)\right\}\\&=\Pr\left\{\frac{\mu_2-\sigma^2\mu_{n,p-2}}{\sigma_{n,p-2}}>s(\alpha)+O_P\left(n^{-1/3}\right)\right\}\end{aligned}$$

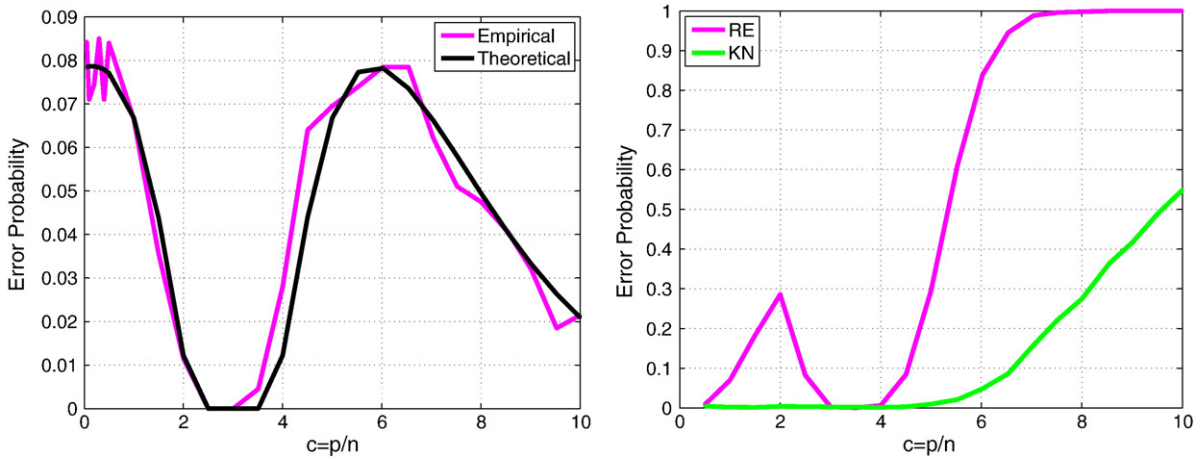

Fig. 6. (Left) Empirical and theoretical probabilities of inferring an incorrect number of signals, for signal-free real-valued samples of dimension $p=1000$, for various values of the number of samples n, as a function of $c=p/n$. (Right) Empirical probabilities of inferring an incorrect number of signals, for real-valued samples carrying four signals with variances $\lambda=(200;\ 100;\ 10;\ 5)$, as a function of $c=p/n$, where $p=1000$.

Therefore, by the Tracy–Widom law (3) applied on μ_2, and by the definition of $s(\alpha)$, we get

$$\lim_{p,n\to\infty}\Pr\left\{\hat{K}>K\right\}=\lim_{p,n\to\infty}\Pr\left\{\frac{\mu_2-\sigma^2\mu_{n,p-2}}{\sigma_{n,p-2}}>s(\alpha)+\frac{c_3(\omega)}{n^{1/3}}\right\}=\alpha.$$

□

Appendix B. Theoretical analysis of the RE and FK algorithms

RE algorithm

We here investigate the behavior of the RE algorithm in the limit $p,n\to\infty$ with $c=p/n$. Since in this limit signal eigenvalues above the critical threshold are identified with probability one, and since these have a negligible effect on individual noise eigenvalues, we study the probability of misclassifying the largest noise eigenvalue as a signal. For simplicity, we consider the signal-free setting ($K=0$), with population covariance matrix $\mathbf{\Sigma}=\mathbf{I}_p$. We consider here the real valued case, $\beta=1$, though a similar analysis can be carried for $\beta=2$. The following lemma gives an asymptotic lower bound for the probability of misidentifying a noise eigenvalue as a signal in this setting.

Lemma 2. *For a real-valued signal-free system ($K=0$), in the asymptotic limit $p,n\to\infty$, $p/n=c$,*

$$\lim_{p,n\to\infty}\Pr\left\{\hat{K}\neq 0\right\}\geqslant\Pr\{\eta>A(c)\}$$

where the random variable η follows a standard $\mathcal{N}(0,1)$ distribution and $A(c)=\frac{c^2-6c+17}{4|c-3|}$.

In Fig. 6 the expression $\Pr\{\eta>A(c)\}$ is shown as a function of c.

Proof. From Eq. (22) with $\beta=1$, a sufficient condition for the RE algorithm to report at least one signal is $t_0^2-t_1^2>8c^2$. Below we estimate the probability of this event in the joint limit $p,n\to\infty$. We introduce the following quantities,

$$T=\sum_{i=1}^{p}\ell_i,\qquad R=\sum_{i=1}^{p}\ell_i^2,\qquad f=p\frac{R}{T^2},\qquad h=1+c+\frac{c}{p}$$

and

$$q=\left(1+\sqrt{c}\right)^2\left[\left(1+\sqrt{c}\right)^2-2(1+c)\right]+(1+c)=c(3-c).$$

Using these notations, we have that

$$t_0=p^2\frac{R}{T^2}-ph=pf-ph$$

$$t_1=p(p-1)\frac{R-\ell_1^2}{(T-\ell_1)^2}-ph=f(p-1)\frac{1-\frac{\ell_1^2}{R}}{\left(1-\frac{\ell_1}{T}\right)^2}-ph$$

In [30,32] it is proved that in the asymptotic limit

$$pf-ph\to 2c\eta$$

where the random variable η follows a standard $\mathcal{N}(0,1)$ distribution. Further, from random matrix theory, the following random variables converge to deterministic quantities in the joint limit $p,n\to\infty$,

$$f\to 1+c,\quad \ell_1\to\left(1+\sqrt{c}\right)^2,\quad \frac{p}{T}\to 1$$

Replacing these expressions by their limiting values and taking into account that $\ell_1/T=O_P\left(p^{-1}\right)$, we get

$$\begin{aligned}t_0&\to 2c\eta\\ t_1&=f(p-1)\left(1-\frac{\ell_1^2}{R}\right)\left[1+\frac{2\ell_1}{T}+O_P\left(\frac{1}{p^2}\right)\right]-ph\\ &=t_0-\left[\frac{\ell_1}{S/p}\left(\frac{\ell_1}{S/p}-2f\right)+f\right]+O_P\left(\frac{1}{p}\right)\\ &\to 2c\eta-q\end{aligned}$$

and hence

$$t_0^2-t_1^2\to q(4c\eta-q).$$

Letting

$$A(c)=\frac{c^2-6c+17}{4|c-3|}$$

we obtain that

$$\Pr\left\{\hat{K}\neq 0\right\}\geqslant\Pr\left\{t_0^2-t_1^2>8c^2\right\}\to\Pr\{\eta>A(c)\}$$

□

In [30] (conjecture 6.3) the authors conjectured that the RE algorithm produces a consistent estimator of the effective number of components K_{eff} in the limit $p,n\to\infty$ with $p/n=c$. Note that Lemma 2 disproves this conjecture, as it shows that when $K=0$ the probability that $\hat{K}\neq 0$ does not converge to zero, but rather to a positive value which depends on c.

Fig. 6 (left) shows the results of a simulation which compares empirical values for Pr $\{t_0^2-t_1^2>8c^2\}$ with Pr $\{\eta>A(c)\}$ for various values of c. For each value of c we performed 2000 MATLAB simulations. We took $p=1000$ in all simulations and $n=p/c$. The figure shows a very good agreement between the theoretical approximation and the simulated results. It is interesting to note that the two curves are not monotonically increasing with c. This means that when using the RE algorithm with sufficiently large p, ignoring some of the samples and thus decreasing n and increasing c might lead to a better estimate of K.

Fig. 6 (right) shows empirical values for Pr {reporting incorrect number of signals} using both the RE and KN algorithms, when $K=4$ and $\lambda=(200, 100, 10, 5)$. These are the same eigenvalues considered in settings B1 and B2 in Section 3, only here we examine the two algorithms over a wide range of c values, keeping p constant. The KN algorithm achieves very low error probabilities, close to the chosen significance level α. The misidentification error naturally increases when sample size n becomes very small and hence c very large. In contrast, the error probability of the RE algorithm is not monotone in c, with a high peak of about 30% error at $c\approx 2$.

FK algorithm

In this section we discuss some limiting properties of the FK algorithm and its connection to our rank estimator. For simplicity, we consider the case $K=1$, although the analysis for general K is similar. The following lemma shows that in the asymptotic limit $p,n\to\infty$, provided the signal eigenvalue λ is above the phase transition threshold, the FK pseudorank estimator reports at least the correct number of signals.

Lemma 3. *For a single component system ($K=1$), with eigenvalue above the threshold $\lambda>\sigma^2\sqrt{p/n}$, in the asymptotic limit $p,n\to\infty$, $p/n=c$,*

$$\lim_{p,n\to\infty}\Pr\left\{\hat{K}_{\mathrm{FK}}\geq 1\right\}=1.$$

Proof. Recall that in the FK algorithm, the following statistic is computed:

$$F_k(\nu_1,\nu_2)=\frac{n\ell_k/\nu_1}{\left(n\sum_{j=k+1}^{p}\ell_j\right)/\nu_2}$$

where

$$\nu_1=n\cdot\mathbb{E}\{\ell_k\}=n\left(1+\sqrt{\frac{p-k}{n-k}}\right)^2,\qquad \nu_2=(n-k+1)(p-k+1)-\nu_1$$

The pseudorank estimation of the FK algorithm is

$$\hat{K}_{\mathrm{FK}}=\arg\min_k\left\{F_k(\nu_1,\nu_2)<f_{\nu_1,\nu_2}(1-\alpha/100)\right\}-1.$$

The algorithm will report at least one signal if $F_1(\nu_1,\nu_2)>f_{\nu_1,\nu_2}(1-\alpha/100)$. This condition can be written as

$$\ell_1>\frac{\sum_{j=2}^{p}\ell_j}{p-\left(1+\sqrt{\frac{p-1}{n-1}}\right)^2}\cdot\left(1+\sqrt{\frac{p-1}{n-1}}\right)^2\cdot f_{\nu_1,\nu_2}(1-\alpha/100).\qquad(26)$$

We examine the limiting properties of both sides of this equation. We start with the LHS. In the presence of a single signal above the phase transition threshold

$$\lim_{p,n\to\infty}\ell_1=(\lambda+\sigma^2)\left(1+\frac{p-1}{n}\frac{\sigma^2}{\lambda}\right).$$

On the RHS we have a product of three terms. For the first one,

$$\lim_{p,n\to\infty}\frac{\sum_{j=2}^{p}\ell_j}{p-\left(1+\sqrt{\frac{p-1}{n-1}}\right)^2}=\sigma^2.$$

For the second term,

$$\lim_{p,n\to\infty}\left(1+\sqrt{\frac{p-1}{n-1}}\right)^2=\left(1+\sqrt{c}\right)^2.$$

As for $f_{\nu_1,\nu_2}(1-\alpha/100)$, we first consider the distribution of an $F(\nu_1,\nu_2)$ random variable in the joint limit $p,n\to\infty$ (which implies $\nu_1\ \nu_2\to\infty$). From the closed form expressions for the expectation and variance of the F-distribution it follows that its mean tends to one and its variance to zero. Thus, an F-distributed random variable is increasingly highly concentrated near one approaching a delta function in the limit, and hence regardless of the value of $\alpha>0$,

$$\lim_{p,n\to\infty}f_{\nu_1,\nu_2}(1-\alpha/100)=1.$$

We conclude that the limiting value of the RHS is equal to $\sigma^2\left(1+\sqrt{c}\right)^2$, which is strictly smaller than the limiting value of the LHS, meaning that the algorithm will consistently recognize at least one signal. □

The proof of the lemma, and specifically the representation of the FK algorithm given in Eq. (26) reveals an interesting analogy to our KN algorithm. Recall that the KN algorithm identifies an eigenvalue as a signal if

$$\ell_{k+1}>\sigma_{\mathrm{KN}}^2(k)\left(\mu_{n,p-k}+s(\alpha)\sigma_{n,p-k}\right).\qquad(27)$$

Comparing Eqs. (27) and (26) shows that the two algorithms have a lot in common: Both estimate the noise variance, and detect an eigenvalue as a signal only if it is larger than some quantity which depends on α and on the noise estimate. The two main differences are that the FK algorithm does not account for the known true limiting Tracy–Widom distribution of the largest noise eigenvalue, and its noise estimator

$$\sigma_{\mathrm{FK}}^2=\frac{\sum_{j=2}^{p}\ell_j}{\frac{n-1}{n}(p-1)-\left(1+\sqrt{\frac{p-1}{n-1}}\right)^2},$$

does not take into account the interaction between noise and signal eigenvalues. As shown in simulations (see Fig. 2), this typically leads to a larger downward bias in comparison to the KN noise estimator σ_{KN}^2.

References

[1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1984.
[2] J. Baik, J.W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, Journal of Multivariate Analysis 97 (6) (2006) 1382–1408.
[3] A. Elbergali, J. Nygren, M. Kubista, An automated procedure to predict the number of components in spectroscopic data, Analytica Chimica Acta 379 (1999) 143–158.
[4] K. Faber, B.R. Kowalski, Modification of Malinowski's *F*-test for abstract factor analysis applied to the quail roost ii data sets, Journal of Chemometrics 11 (1997) 53–72.
[5] N.M. Faber, L.M.C. Buydens, G. Kateman, Aspects of pseudorank estimation methods based on the eigenvalues of principal component analysis of random matrices, Chemometrics and Intelligent Laboratory Systems 25 (1994) 203–226.
[6] E. Fishler, M. Grosmann, H. Messer, Detection of signals by information theoretic criteria: general asymptotic performance analysis, IEEE Transactions on Signal Processing 50 (2002) 1027–1036.
[7] R.C. Henry, E.S. Park, C.H. Spiegelman, Comparing a new algorithm with the classic methods for estimating the number of factors, Chemometrics and Intelligent Laboratory Systems 48 (1999) 91–97.
[8] A.T. James, Inference on latent roots by calculation of hypergeometric functions of matrix argument, Multivariate Analysis, 1966, pp. 209–235.

[9] J-H. Jiang, Y. Liang, Y. Ozaki, Principles and methodologies in self-modeling curve resolution, Chemometrics and Intelligent Laboratory Systems 71 (2004) 1–12.
[10] K. Johansson, Shape fluctuations and random matrices, Communications in Mathematical Physics 209 (2) (2000) 437–476.
[11] I.M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Annals of Statistics 29 (2) (2001) 295–327.
[12] I.M. Johnstone, High dimensional statistical inference and random matrices, Proc. International Congress of Mathematicians, 2006.
[13] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 2002.
[14] N. El Karoui, A rate of convergence result for the largest eigenvalue of complex white wishart matrices, Annals of Probability 34 (6) (2006) 2077–2117.
[15] K. Konstantinides, K. Yao, Statistical analysis of effective singular values in matrix rank determination, IEEE Transactions on Acoustics, Speech and Signal Processing 36 (5) (1988) 757–763.
[16] W.J. Krzanowski, Principles of Multivariate Analysis, Oxford University Press, 1988.
[17] D.N. Lawley, Tests of significance for the latent roots of covariance and correlation matrices, Biometrika 43 (1956) 128–136.
[18] E. Levina, A.S. Wagaman, A.F. Callender, G.S. Mandair, M.D. Morris, Estimating the number of pure chemical components in a mixture by maximum likelihood, Journal of Chemometrics 21 (1–2) (2007) 24–34.
[19] E.R. Malinowski, Determination of the number of factors and the experimental error in a data matrix, Analytical Chemistry 49 (1977) 612–617.
[20] E.R. Malinowski, Theory of error in factor analysis, Analytical Chemistry 49 (1977) 606–612.
[21] E.R. Malinowski, Statistical *F*-tests for abstract factor analysis and target testing, Journal of Chemometrics 3 (1989) 49–60.
[22] V.A. Marcenko, L.A. Pastur, Distribution for some sets of random matrices, Math USSR-Sb, 1967, pp. 457–483.
[23] M. Meloun, J. Capek, P. Miksik, R.G. Brereton, Critical comparison of methods predicting the number of components in spectroscopic data, Analytica Chimica Acta 423 (2000) 51–68.
[24] T.P. Minka, Automatic choice of dimensionality for PCA, in: T.K. Leen, T.G. Dietterich, K.-R. Muller (Eds.), Advanced in Neural Information Processing Systems, vol. 13, 2001, p. 598604.
[25] B. Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *to appear*, *Annals of Statistics*, 2008.
[26] B. Nadler, R.R. Coifman, Partial least squares, Beer law and the net analyte signal: statistical modeling and analysis, Journal of Chemometrics 19 (2005) 45–54.
[27] B. Nadler, R.R. Coifman, The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration, Journal of Chemometrics 19 (2005) 107–118.
[28] T. Naes, T. Isaksson, T. Fearn, T. Davies, User-friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, 2002.
[29] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Statistica Sinica 17 (2007) 1617–1642.
[30] D. Raj Rao, A. Edelman, Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples, IEEE Transactions of Signal Processing 56 (7) (2008) 2625–2638.
[31] S.N. Roy, On a heuristic method of test construction and its use in multivariate analysis, Annals of Mathematical Statistics 24 (2) (1953) 220–238.
[32] J.R. Schott, A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix, Journal of Multivariate Analysis 97 (4) (2006) 827–843.
[33] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Cote, Theoretical justification of wavelength selection in pls calibration: development of a new algorithm, Analytical Chemistry 70 (1) (1998) 35–44.
[34] M. Wax, T. Kailath, Detection of signals by information theoretic criteria, IEEE Transactions on Acoustics, Speech and Signal Processing 33 (2) (1985) 387–392.
[35] J.-H. Jiang, Y. Li, J.-Y. Qian, Z.-P. Chen, Y.-Z. Liang, R.-Q. Yu, Determination of the number of components in mixtures using a new approach incorporating chemical information, Journal of Chemometrics 13 (1999) 15–30.
[36] L.C. Zhao, P.R. Krishnaiah, Z.D. Bai, On detection of the number of signals in presence of white noise, Journal of Multivariate Analysis 20 (1986) 1–25.
[37] S. Kritchman, B. Nadler, Detection of the number of signals, statistical hypothesis testing and random matrix theory (in preparation).