Non-Parametric Detection of the Number of Signals: Hypothesis Testing and Random Matrix Theory

Shira Kritchman and Boaz Nadler

***Abstract*—Detection of the number of signals embedded in noise is a fundamental problem in signal and array processing. This paper focuses on the non-parametric setting where no knowledge of the array manifold is assumed. First, we present a detailed statistical analysis of this problem, including an analysis of the signal strength required for detection with high probability, and the form of the optimal detection test under certain conditions where such a test exists. Second, combining this analysis with recent results from random matrix theory, we present a new algorithm for detection of the number of sources via a sequence of hypothesis tests. We theoretically analyze the consistency and detection performance of the proposed algorithm, showing its superiority compared to the standard minimum description length (MDL)-based estimator. A series of simulations confirm our theoretical analysis.**

***Index Terms*—Detection, number of signals, random matrix theory, statistical hypothesis tests, Tracy–Widom distribution.**

I. Introduction

DETECTION of the number of signals impinging on a collection of sensors is a fundamental problem in statistical signal and array processing and still a subject of ongoing research [3], [6], [34], [29]. It is typically a first step prior to computationally demanding parametric procedures that depend on this input, such as direction of arrival estimation, blind source deconvolution, etc.

Before proceeding, we mention that a similar, if not identical problem, also appears in other scientific fields, such as chemometrics, econometrics, population genetics, and classical statistics [23], [21], [31], [13], [27]. Examples include determining the number of different chemical components in a mixture, the number of multiplicative components in two way tables of interaction, and the order of an autoregression.

In the signal processing literature, the most common approach to solving this problem is by using information theoretic criteria, and in particular minimum description length (MDL), Bayesian information criterion (BIC) and Akaike information criterion (AIC); see [34] and their improvements such as [35], [9]. The paper [33] provides a review of these methods. While the MDL estimator is consistent as sample size $n \to \infty$ [38], as noted in various works [34], [37], [4], it fails to detect signals at low signal-to-noise ratio (SNR), hence underestimating the number of signals at small sample sizes. In contrast, while the AIC estimator is able to detect low SNR signals, regretfully it is not consistent as $n \to \infty$, having a non-negligible probability to overestimate the number of signals for $n \gg 1$. Furthermore, both estimators are based on large sample asymptotics, so none of them perform well when the number of sensors p is comparable to the number of observations. Finally, neither of these estimators is applicable to large aperture arrays with a number of sensors larger than the number of samples, $p > n$.

Various theoretical and practical questions arise with respect to these results: 1) Given the presence of noise, what is the signal strength required for reliable detection (i.e., with high probability)? 2) Is there an improved estimator, which enjoys both high detection performance at low SNR (similar to the AIC estimator) and (near) consistency at large sample sizes (similar to the MDL estimator)?

In this paper, we present a statistical analysis of the problem of detection of the number of signals, and present and analyze a new estimation algorithm, which together provide answers to the two questions raised above. The main tools used in our analysis are recent results in random matrix theory (RMT) regarding both the distribution of noise eigenvalues and of signal eigenvalues in the presence of noise. For the paper to be self-contained, these results are reviewed in Section II. One of these results shows that in the joint limit $p, n \to \infty$ there is a phase transition phenomenon, where only signals stronger than a certain threshold can be detected by the largest eigenvalue. We further show that the detection threshold of the MDL estimator is significantly larger than this asymptotic threshold, thus suggesting that the MDL procedure may not be an optimal one.

In Section III, we advocate a different approach for estimating the number of signals, via a sequence of hypothesis tests, at each step testing the significance of the kth eigenvalue as arising from a signal. To motivate this approach, we analyze a simple setting, in which the likelihood ratio test is an optimal procedure. We show that when deciding between the two hypotheses

$$\mathcal{H}_0 : \text{no signals} \quad \text{vs.} \quad \mathcal{H}_1 : \text{one signal of known strength}$$

with known noise variance and using only the sample covariance matrix, asymptotically as $n \to \infty$, only the *largest* sample eigenvalue should be used. Similarly, when testing $k-1$ signals versus k signals (with known noise variance and signal strengths), asymptotically only the kth sample eigenvalue should be used. Our proposed approach, of testing a single

Manuscript received August 31, 2008; accepted April 03, 2009. First published May 12, 2009; current version published September 16, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yonina C. Eldar. The work of B. Nadler was supported by a grant from the Ernst Nathan biomedical fund. Part of this work was done while B. Nadler participated in the SCH program at the Isaac-Newton Institute of Mathematical Sciences at Cambridge, U.K.

The authors are with the Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, 76100, Israel (e-mail: shira.kritchman@weizmann.ac.il; boaz.nadler@weizmann.ac.il).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2009.2022897

eigenvalue at a time, is similar in spirit to [4], [5], and [27], and is closely related to the classical largest root test proposed by Roy [22].

The asymptotic properties of our algorithm, as well as its performance for finite values of p and n, are analyzed in Section IV. We show that for finite p, n it has a detection threshold much smaller than that of the MDL estimator. Furthermore, our algorithm attains the asymptotic limit of detection when $p, n \to \infty$. Since the proposed method employs multiple tests with a fixed confidence level, it has a small probability of overestimation, that can be controlled by the user. This provides a positive answer to the second question raised above. We conclude our analysis by considering the case $p > n$ and the joint limit $p, n \to \infty$, and compare our algorithm to two recently suggested methods for estimating the number of signals under this setting [30], [32]. Section V presents simulations supporting the theoretical analysis. Section VI is summary and discussion.

The algorithm presented here was developed by us in a different context, that of rank determination in analytical chemistry, where the rank typically corresponds to the number of chemical components present in a mixture [23]. This paper extends our initial work with a detailed statistical analysis of the problem and of the performance of our proposed algorithm in comparison to the MDL and other estimators. Preliminary partial results were presented at the 2008 Asilomar Conference on Signals, Systems and Computers.

II. Non-Parametric Signal Detection, Random Matrix Theory and Information Theoretic Criteria

Notation: Vectors and matrices are denoted by lowercase and uppercase bold letters, as in $\mathbf{v}$ and $\mathbf{A}$, respectively. The conjugate transpose of $\mathbf{v}$ is denoted $\mathbf{v}^H, \mathbf{I}_p$ denotes the identity matrix of order p and $\mathrm{Tr}(\mathbf{A})$ denotes the trace of a matrix $\mathbf{A}$. The Gaussian distribution with mean μ and covariance $\boldsymbol{\Sigma}$ is denoted $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. An estimate of a parameter θ is denoted $\hat{\theta}$. Almost sure convergence of random variables, also known as convergence with probability one (w.p.1), is denoted $\xrightarrow{\text{a.s.}}$. Convergence in distribution is denoted $\overset{d}{\to}$.

A. Problem Formulation

We consider the following standard model for signals impinging on an array with p sensors. Let $\{\mathbf{x}_i = \mathbf{x}(t_i)\}_{i=1}^n$ denote p-dimensional i.i.d. observations of the form

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \sigma\mathbf{n}(t) \quad (1)$$

sampled at n distinct times t_i, where $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_K]$ is the $p \times K$ steering matrix of K linearly independent p-dimensional vectors. The $K \times 1$ vector $\mathbf{s}(t) = [s_1(t), \dots, s_K(t)]^T$ represents the random signals, assumed zero mean and stationary with full rank covariance matrix. σ is the unknown noise level, and $\mathbf{n}(t)$ is a $p \times 1$ additive Gaussian noise vector, distributed $\mathcal{N}(0, \mathbf{I}_p)$ and independent of $\mathbf{s}(t)$.

Under these assumptions, the population covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{x}(t)$ has a diagonal form,

$$\mathbf{W}^H\,\boldsymbol{\Sigma}\,\mathbf{W} = \sigma^2\mathbf{I}_p + \mathrm{diag}(\lambda_1, \dots, \lambda_K, 0, \dots, 0) \quad (2)$$

in an unknown basis $\mathbf{W}$ of $\mathbb{C}^p$ (or $\mathbb{R}^p$ for real-valued signals), where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_K > 0$. We denote by $\mathbf{S}_n$ the sample covariance matrix of the n observations $\mathbf{x}_i$ from the model (1),

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^H$$

and denote by $\ell_1 \geqslant \ell_2 \geqslant \ldots \geqslant \ell_p$ its eigenvalues.

In this paper we consider the problem of estimating the unknown number of sources K given the n observations $\{\mathbf{x}_i\}_{i=1}^n$ under the *nonparametric* setting, where no prior information is assumed about the matrix $\mathbf{A}$ beyond it being of rank K. Furthermore, we only consider methods to infer the number of signals that use the eigenvalues $\{\ell_j\}$ of the sample covariance matrix $\mathbf{S}_n$, rather than the original observations.

The problem at hand is thus a *model selection problem*, e.g., to determine which of the possible models of the form (2) is most likely given the sample eigenvalues $\{\ell_i\}_{i=1}^p$.

B. Mathematical Preliminaries: Random Matrix Theory and Asymptotic Limit of Detection

The key principle in nonparametric estimation of the number of sources is that for sufficiently large n, in the presence of K sources, the first K largest sample eigenvalues correspond to signals, whereas the remaining eigenvalues correspond to noise. Here we review some mathematical theory relevant to the problem at hand, in particular results from random matrix theory regarding the behavior of the largest eigenvalue of a pure noise matrix, and results about signal eigenvalues in the presence of noise.

In the absence of signals, the matrix $n \cdot \mathbf{S}_n$ follows a Wishart distribution with parameters n, p. While there is no simple explicit expression for the distribution of its largest eigenvalue, in recent years, its *exact* distribution under a certain asymptotic limit was derived [7], [14], [15] :

Theorem 1: Let $\mathbf{S}_n$ denote the sample covariance matrix of n pure noise vectors distributed $\mathcal{N}(0, \sigma^2\mathbf{I}_p)$. In the joint limit $p, n \to \infty$, with $p/n \to c \geqslant 0$, the distribution of the largest eigenvalue of $\mathbf{S}_n$ converges to a Tracy–Widom distribution

$$\Pr\left\{\frac{\ell_1/\sigma^2 - \mu_{n,p}}{\xi_{n,p}} < s\right\} \to F_\beta(s) \quad (3)$$

where $\beta = 1$ for real valued noise and $\beta = 2$ for complex-valued noise. The centering and scaling parameters, $\mu_{n,p}$ and $\xi_{n,p}$, respectively, are functions of n and p only.

For real valued noise, the following expressions provide $O(p^{-2/3})$ convergence rate in (3):

$$\mu_{n,p} = \frac{1}{n}(\sqrt{n - 1/2} + \sqrt{p - 1/2})^2, \quad (4)$$

$$\xi_{n,p} = \sqrt{\frac{\mu_{n,p}}{n}}\left(\frac{1}{\sqrt{n-1/2}} + \frac{1}{\sqrt{p-1/2}}\right)^{1/3}. \quad (5)$$

Explicit expressions for complex valued noise, having a similar yet more involved form, appear in [7]. These expressions provide good approximations for finite p, n as long as $\min(p, n) \gg 1$ and the ratio p/n or n/p is not too large [8].

Assuming the noise variance σ^2 is known, one possible test for the presence or absence of a signal, known as Roy's largest root test [22], is to check the significance of the largest eigenvalue, as follows:

$$\frac{\ell_1}{\sigma^2} > \mu_{n,p} + s(\alpha)\xi_{n,p}.$$

For this test to have a false alarm (type I error) with asymptotic probability α as $p, n \to \infty$, the threshold $s(\alpha)$ should satisfy

$$F_\beta(s(\alpha)) = 1 - \alpha. \tag{6}$$

The value of $s(\alpha)$ can be calculated by inverting the Tracy–Widom distribution. As there is no explicit closed form expression for the TW distribution, this inversion can be done numerically, for example using the software package[1] . However, for values $\alpha \ll 1$, from the asymptotics of the Airy function it follows that [15]

$$F_1(x) = 1 - \frac{e^{-2/3x^{3/2}}}{4\sqrt{\pi}x^{3/2}}(1 + O(x^{-3/2}))$$
$$F_2(x) = 1 - \frac{e^{-4/3x^{3/2}}}{16\pi x^{3/2}}(1 + O(x^{-3/2})).$$

Therefore, for sufficiently small α, approximate explicit thresholds are

$$s(\alpha) \approx \begin{cases} (-3/2\log 4\sqrt{\pi}\alpha)^{2/3} & \beta = 1 \\ (-3/4\log 16\pi\alpha)^{2/3} & \beta = 2. \end{cases} \tag{7}$$

We stress that (7) is valid only for $\alpha \ll 1$, with the approximation more accurate for real valued noise ($\beta = 1$).

In addition, we will use the following non-asymptotic bound on the largest eigenvalue of complex-valued noise observations (see [20, p. 187, (2.4)]).

Theorem 2: Let ℓ_1 be the largest eigenvalue as in Theorem 1, then

$$\Pr\left\{\frac{\ell_1}{\sigma^2} > \left(1 + \sqrt{\frac{p}{n}}\right)^2 + \varepsilon\right\} \leqslant \exp\{-nJ_{\text{LAG}}(\varepsilon)\} \tag{8}$$

where

$$J_{\text{LAG}} = \int_1^x (x-y)\frac{(1+c)y + 2\sqrt{c}}{(y+B)^2}\frac{dy}{\sqrt{y^2-1}} \tag{9}$$

with $c = (p/n), x = 1 + (\varepsilon)/(2\sqrt{c})$ and $B = (1+c)/(2\sqrt{c})$.

Next, we consider the behavior of signal eigenvalues in the presence of noise. As $n \to \infty, \mathbf{S}_n \xrightarrow{\text{a.s.}} \mathbf{\Sigma}$, and so the sample eigenvalues converge w.p.1 to the corresponding population eigenvalues. Hence, as $n \to \infty$ any (positive) signal strength can be eventually detected w.p.1 by inspection of the sample eigenvalues. The interesting question is which signal strengths can be identified with high probability, for given *finite* values of p, n and σ. For finite but large values of p, n, from (4) it follows that the largest eigenvalue due to noise is approximately $\sigma^2(1 + \sqrt{p/n})^2$. Thus, a weak signal cannot be detected by the largest sample eigenvalue, since the variance in its direction will be smaller than the largest variance in a random direction due to noise.

In the joint limit $p, n \to \infty$ with p/n fixed, there is a phase transition, where signals can be detected by the largest eigenvalues if and only if they are above a certain deterministic threshold [2], [28], [25]:

Theorem 3: Let $\mathbf{S}_n$ denote the sample covariance matrix of n observations from (1) with a single signal of strength λ. Then, in the joint limit $p, n \to \infty$, with $p/n \to c > 0$, the largest eigenvalue of $\mathbf{S}_n$ converges w.p.1 to

$$\lambda_{\max}(\mathbf{S}_n) \xrightarrow{\text{a.s.}} \begin{cases} \sigma^2(1 + \sqrt{p/n})^2 & \lambda \leqslant \sigma^2\sqrt{p/n} \\ (\lambda + \sigma^2)\left(1 + \frac{p}{n}\frac{\sigma^2}{\lambda}\right) & \lambda > \sigma^2\sqrt{p/n}. \end{cases} \tag{10}$$

Due to its importance, we denote this threshold as the non-parametric *asymptotic limit of detection*,

$$\lambda_{\text{DET}} = \sigma^2\sqrt{\frac{p}{n}}. \tag{11}$$

Finally, the following lemma considers the influence of a signal on the largest noise eigenvalue [23].

Lemma 1: Consider a setting with a single signal, $\lambda > \lambda_{\text{DET}}$. Then, in the asymptotic limit $p, n \to \infty, p/n \to c > 0$, the second largest sample eigenvalue (which corresponds to noise) has asymptotically the same Tracy–Widom distribution as the largest eigenvalue of a pure noise Wishart matrix, with parameters $n, p-1$.

By induction, the Proof of Lemma 1 can be generalized to any number of signals. Hence, in the case of K (sufficiently strong) signals, in the joint limit $p, n \to \infty$, the $(K+1)$th eigenvalue follows a Tracy–Widom distribution, with parameters $n, p-K$.

C. Previous Work: Information Theoretic Criteria

In the seminal paper [34], Wax and Kailath proposed the following minimum description length criteria to determine the number of sources,

$$\begin{aligned} \hat{K}_{\text{MDL}} &= \arg\min_k \text{MDL}(k) \\ &= \arg\min_k (p-k)\log\frac{\sum_{k+1}^p \ell_j}{p-k} - \log\prod_{k+1}^p \ell_j \\ &\quad + \frac{k(2p-k)}{2n}\log n. \end{aligned}$$

This estimator was proven to be strongly consistent [38], namely that

$$\lim_{n\to\infty} \Pr\{\hat{K}_{\text{MDL}} = K\} = 1. \tag{12}$$

These two properties, simplicity and consistency, made the MDL estimator the standard tool for detection of the number of signals.

D. Detection Performance of the MDL Estimator

The performance of the MDL estimator is analyzed in various works [36], [10], [26]. In the non-parametric case, the difference $\text{MDL}(K-1) - \text{MDL}(K)$ is asymptotically Gaussian

[1] Available [online]: http://math.arizona.edu/~momar/research.htm

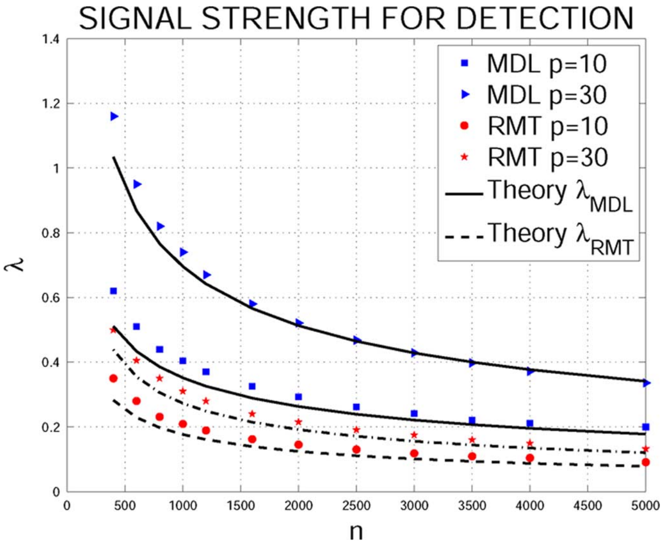

Fig. 1. Required signal strength λ for detection with probability 1/2 for the MDL and RMT algorithms (with $\alpha_{\mathrm{RMT}} = 0.5\%$) as a function of sample size n, for $p = 10$ or $p = 30$, and $\sigma = 1$. Comparison of simulation results with (14) for MDL and (25) for the RMT algorithm.

distributed with explicitly known approximate formulas for the mean and variance. In case of K signals, where the Kth signal has strength $\lambda_K \ll \lambda_{K-1}$ and with $\sigma = 1$, the following holds, up to $o(1/n)$ terms [26]:

$$\begin{aligned}
&\mathbb{E}[\mathrm{MDL}(K-1) - \mathrm{MDL}(K)] \\
&\quad = -\log(1+\lambda_K) - \log\left(1 + \frac{p-K}{n}\frac{1}{\lambda_K} - \frac{K-1}{n}\right) \\
&\qquad - (p-K)\log\left(1 - \frac{1}{n}\frac{1+\lambda_K}{\lambda_K} - \frac{K-1}{n}\right) + \frac{2}{n} \\
&\qquad + (p-K+1)\log\left(1 + \frac{\left(1-\frac{K-1}{n}\right)\lambda_K}{p-K+1} - \frac{K-1}{n}\right) \\
&\qquad - \frac{1}{n(p-K+1)}\frac{(1+\lambda_K)^2 + p - K}{\left(1+\frac{\lambda_K}{p+1-K}\right)^2} \\
&\qquad - \frac{2p-2K+1}{2}\frac{\log n}{n}.
\end{aligned} \tag{13}$$

From (13), the approximate signal strength detectable by the MDL estimator with probability $> (1/2)$ can be derived as follows. For large n the detectable signal strength λ is small. A Taylor expansion of the logarithms in (13) for small λ gives that $\mathbb{E}[\mathrm{MDL}(K-1) - \mathrm{MDL}(K)] > 0$ if $\lambda > \lambda_{\mathrm{MDL}}$ where

$$\lambda_{\mathrm{MDL}} \approx \sqrt{\frac{p-K+1}{p-K}}\sqrt{\frac{2p-2K+1}{n}\log n - \frac{2(p-K+1)}{n}}. \tag{14}$$

This formula is in close agreement with simulation results, as shown in Fig. 1. The key point of (14) is that the detection threshold of the MDL estimator is significantly *larger* (by a factor of approximately $\sqrt{2\log n}$) than the asymptotic limit of detection, (11). Indeed, as noted in [10], while the goal of MDL is to minimize the description length, in detection problems the goal is to minimize the probability of misdetection of the true number of signals. Hence, (14) is not necessarily the lowest possible detection threshold.

In this paper we present and analyze a new algorithm for estimating the number of sources. To detect the presence of all K signals, the smallest signal eigenvalue should satisfy

$$\lambda_K \gtrsim \sqrt{\frac{p-K}{n}}\left(1 + \frac{C(1+\sqrt{(p-K)/n})^{1/2}}{(p-K)^{1/3}}\right)$$

where C is a constant independent of p, n. Thus, the proposed algorithm, described below, can detect much weaker signals than the MDL estimator. Furthermore, in the joint limit $p, n \to \infty$ it attains the asymptotic limit of detection of (11).

III. Non-Parametric Signal Detection by Statistical Hypothesis Tests

A. Motivation: Likelihood Ratio Asymptotics

In the previous section we saw that the detection threshold of the MDL estimator is larger by a factor of approximately $\sqrt{2\log n}$ compared to the asymptotic limit (11). An interesting question is what is the optimal procedure to detect the number of signals and what is its detection performance.

Regretfully, since determining the number of sources involves testing multiple composite hypotheses, there is no optimal statistical procedure for any prescribed false alarm rate α. However, one can gain insight into this problem by studying a specific case where an optimal procedure does exist, and that is the case of testing one simple hypothesis against a simple alternative.

We hence analyze the following scenario: consider testing the simple hypothesis $\mathcal{H}_0$ versus a simple alternative $\mathcal{H}_1$, defined as

$$\begin{aligned}
&\mathcal{H}_0 : \mathbf{\Sigma} = \mathbf{I}_p \quad \text{vs.} \\
&\mathcal{H}_1 : \mathbf{W^H \Sigma W} = \mathbf{I}_p + \mathrm{diag}(\lambda, 0, \ldots, 0)
\end{aligned}$$

where the noise variance is assumed known, $\sigma = 1$, and under the hypothesis $\mathcal{H}_1$ there is a single Gaussian distributed signal with an *a priori* known variance λ. In this setting, the eigenvalues $\{\ell_j\}_{j=1}^{p}$ of the sample covariance matrix $\mathbf{S}_n$ are sufficient statistics, and for any finite values of p and n, there are only two possible densities for them. By the Neyman–Pearson Lemma, for any false alarm probability α, the likelihood ratio test is the *optimal* procedure to distinguish between the two hypotheses, namely

$$\mathrm{LRT} = \frac{p(\ell_1, \ldots, \ell_p | \mathcal{H}_1)}{p(\ell_1, \ldots, \ell_p | \mathcal{H}_0)} \lessgtr C_\alpha \tag{15}$$

where the constant C_α is chosen such that $\Pr\{\text{reject } \mathcal{H}_0 | \mathcal{H}_0\} = \Pr\{\mathrm{LRT} > C_\alpha | \mathcal{H}_0\} = \alpha$.

For simplicity, we consider the case of real-valued samples. Recall that the joint density of all p eigenvalues of a sample covariance matrix of $n > p$ multivariate real-valued Gaussian observations with a population covariance $\mathbf{\Sigma}$ is given by

$$\begin{aligned}
&p(\ell_1, \ldots, \ell_p | \mathbf{\Sigma}) \\
&\quad = C_{n,p}\prod\nolimits_i \mu_i^{-n/2}\prod\nolimits_i \ell_i^{(n-p-1)/2} \\
&\qquad \times \prod\nolimits_{i<j}(\ell_i - \ell_j)\,{}_0F_0^p\left(-\frac{1}{2}nL, A\right)
\end{aligned} \tag{16}$$

where $C_{n,p}$ is a normalization constant, $L = \mathrm{diag}(\ell_1, \ldots, \ell_p)$, and $A = \mathrm{diag}(1/\mu_1, \ldots, 1/\mu_p)$ is the information matrix with

μ_j the eigenvalues of $\mathbf{\Sigma}$. ${}_0F_0^p$ is the hypergeometric function with matrix argument, given by

$$ {}_0F_0^p\left(-\frac{n}{2}L, A\right) = \int_{O(p)} \exp\left\{\operatorname{tr}\left(-\frac{1}{2}nLQ^HAQ\right)\right\} dQ $$

where dQ is the invariant measure on the group $O(p)$ of $p \times p$ orthogonal matrices.

Under $\mathcal{H}_0$, no signals are present ($\mathbf{\Sigma} = \mathbf{I}_p$), and

$$ {}_0F_0^p\left(-\frac{n}{2}L, \mathbf{I}_p\right) = \exp\left\{-\frac{n}{2}\sum\nolimits_{j=1}^{p}\ell_j\right\}. \tag{17} $$

Under $\mathcal{H}_1$, a single signal of strength λ is present. Here there is no simple explicit expression for the hypergeometric function. However, its asymptotic expansion for large n is [24]

$$ \begin{aligned} &{}_0F_0^p\left(-\frac{n}{2}L, A\right) \\ &= C_p \exp\left\{-\frac{n}{2}\left[\frac{\ell_1}{\lambda+1} + \sum\nolimits_{j=2}^{p}\ell_j\right]\right\} \\ &\times \prod\nolimits_{j>1}\left[\frac{2\pi}{n}\frac{1}{\ell_1-\ell_j}\frac{1}{1-1/(\lambda+1)}\right]^{1/2} \\ &\times \left[1+O\left(\frac{1}{n}\right)\right]. \end{aligned} \tag{18} $$

Combining (15)–(18) and taking logarithms gives

$$ \log\frac{p(\ell_1,\ldots,\ell_p|\mathcal{H}_1)}{p(\ell_1,\ldots,\ell_p|\mathcal{H}_0)} = \frac{n}{2}\left[\ell_1\frac{\lambda}{\lambda+1} - \log(1+\lambda)\right](1+o(1)) $$

Hence, asymptotically in sample size, we accept $\mathcal{H}_1$ if

$$ \ell_1 > \left(1+\frac{1}{\lambda}\right)\left[\log(1+\lambda) + \frac{2}{n}\log C_\alpha\right]. $$

The key point of this analysis is that as $n \to \infty$ with p fixed, distinguishing between $\mathcal{H}_0$ and $\mathcal{H}_1$ should be based *only* on the largest sample eigenvalue ℓ_1. The same conclusion holds also for complex-valued observations, where the density of sample eigenvalues has a slightly different formula. Hence, the optimal procedure to detect a single signal is closely related to Roy's largest root test for sphericity, originally derived by the union-intersection principle [22], [16].

A similar analysis, using formula (3.5) from [24], shows that to distinguish between the two hypotheses

$$ \mathcal{H}_0\colon\ k-1 \text{ signals} \quad \text{vs.} \quad \mathcal{H}_1\colon\ k \text{ signals} $$

with *a priori* known noise variance and signal strengths, asymptotically the LRT depends to leading order only on ℓ_k.

B. An RMT Based Estimation Algorithm

Motivated by the above analysis, we now present an estimator for the number of signals based on a sequence of hypothesis tests, each time testing the significance of the kth sample eigenvalue ℓ_k, as arising from a signal rather than from noise.[2] This algorithm was developed by us in a different context [23], of chemical rank determination in analytical chemistry, where typically $p \gg n$. For this paper to be self contained we present a brief description of the algorithm, and then focus on its analysis in the typical setting for signal processing applications, where $p \ll n$. Later on we also present some results where $p > n$. Further results in the latter setting can be found in [23].

The algorithm works as follows: For $k = 1,\ldots,\min(p,n)-1$, we test

$$ \mathcal{H}_0\colon \text{at most } k-1 \text{ signals} \quad \text{vs.} \quad \mathcal{H}_1\colon \text{at least } k \text{ signals}. $$

Under the null hypothesis, ℓ_k arises from noise. Thus, we reject $\mathcal{H}_0$ if ℓ_k is too large,

$$ \ell_k > \hat{\sigma}^2(k)C_{n,p,k}(\alpha) $$

where $\hat{\sigma}^2(k)$ is an estimate for the unknown noise level σ^2 [see (22) below] and $C_{n,p,k}(\alpha)$ depends on the confidence level $\alpha \ll 1$ chosen by the user. Roughly speaking, $C_{n,p,k}(\alpha)$ is determined such that if only $k-1$ signals are present, and the kth eigenvalue is due to noise, then

$$ \Pr\{\text{reject } \mathcal{H}_0|\mathcal{H}_0\} = \Pr\{\ell_k > \sigma^2 C_{n,p,k}(\alpha)|\mathcal{H}_0\} \approx \alpha. $$

Hence, α controls the probability of model overestimation. We stop at the smallest index k where the above condition fails, i.e., the first time we accept $\mathcal{H}_0$. Our estimate of the number of signals is then $\hat{K} = k-1$.

Using Theorem 1 and Lemma 1 from Section II, for p, n large enough, under the null hypothesis of $k-1$ signals, ℓ_k approximately follows a TW distribution with negligible influence from the first $k-1$ signals. Hence, we set the threshold to

$$ C_{n,p,k}(\alpha) = \mu_{n,p-k} + s(\alpha)\xi_{n,p-k} $$

with $s(\alpha)$ given by (6).

To conclude, our estimator is defined by

$$ \hat{K}_{\mathrm{RMT}} = \arg\min_k\left\{\ell_k < \hat{\sigma}^2(k)(\mu_{n,p-k} + s(\alpha)\xi_{n,p-k})\right\} - 1. \tag{19} $$

We refer to this as the RMT estimator.

C. Noise Estimation

To apply (19), an accurate estimate of the unknown noise level is required. Under the assumption of K signals, a simple estimator of the noise level is via maximum likelihood [34]

$$ \hat{\sigma}^2 = \frac{1}{p-K}\sum_{K+1}^{p}\ell_j. \tag{20} $$

As discussed below, this estimator has a negative bias, which may lead to overestimation of the number of signals, specifically for small sample sizes [23].

To develop a more accurate noise estimator, an analysis of the interaction between signal and noise eigenvalues is needed. Specifically, let $\{\mathbf{w}_1,\ldots,\mathbf{w}_p\}$ be an orthogonal basis which diagonalizes the population covariance matrix $\mathbf{\Sigma}$ [see (2)] where $\operatorname{Span}\{\mathbf{w}_1,\ldots,\mathbf{w}_K\}$ is the signal subspace and $\operatorname{Span}\{\mathbf{w}_{K+1},\ldots,\mathbf{w}_p\}$ is the orthogonal noise subspace. Let

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

z_j denote the sample variance in the direction $\mathbf{w}_j$. Assuming K signals, 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 $z_{K+1}, \ldots, z_p$,

$$\sigma^2_{\text{unbiased}} = \frac{\sum_{j=K+1}^p z_j}{p-K} = \frac{\text{Tr}(\mathbf{S}_n) - \sum_{j=1}^K z_j}{p-K} = \frac{1}{p-K}\left[\sum_{j=K+1}^{p} \ell_j + \sum_{j=1}^{K}(\ell_j - z_j)\right]. \quad (21)$$

Unfortunately, the diagonalizing basis $\mathbf{W}$ is unknown, and to estimate σ^2, we need an estimate for $\sum_{j=1}^K(\ell_j - z_j)$ in (21). We now readily see that the simple estimator (20) follows by neglecting this positive term altogether, leading to a negative bias.

In [23], we developed an improved noise estimator using a matrix perturbation approach. The key idea is that an estimate of noise level depends on the bias of the signal eigenvalues, which, in turn, depends on the unknown noise level. Specifically, denote by $\widetilde{\mathbf{W}}$ the basis which diagonalizes the signal subspace of the sample covariance matrix $\mathbf{S}_n$. In this new basis,

$$\widetilde{\mathbf{W}}^H \mathbf{S}_n \widetilde{\mathbf{W}} = \left(\begin{array}{ccc|ccc} \rho_1 & & 0 & & & \\ & \ddots & & & \mathbf{B}^H & \\ 0 & & \rho_K & & & \\ \hline & & & z_{K+1} & & * \\ & \mathbf{B} & & & \ddots & \\ & & & * & & z_p \end{array}\right)$$

where the matrix $\mathbf{B}$ captures the interactions between the signal and the noise subspaces. We view the matrix $\mathbf{B}$ as a small perturbation, expand the eigenvalues and eigenvectors of $\mathbf{S}_n$ in terms of its entries, and take averages of the various random variables appearing in the resulting expressions. This principle yields an approximately self-consistent method to estimate the noise level via solution of the following non-linear system of equations:

$$\hat{\sigma}^2_{\text{RMT}} - \frac{1}{p-K}\left[\sum_{j=K+1}^{p} \ell_j + \sum_{j=1}^{K}(\ell_j - \hat{\rho}_j)\right] = 0,$$
$$\hat{\rho}_j^2 - \hat{\rho}_j\left(\ell_j + \hat{\sigma}^2_{\text{RMT}} - \hat{\sigma}^2_{\text{RMT}}\frac{p-K}{n}\right) + \ell_j\hat{\sigma}^2_{\text{RMT}} = 0. \quad (22)$$

We solve this system iteratively starting from an initial guess $\hat{\sigma}_0^2$ given by (20). The second equation defines each $\hat{\rho}_j$ as a function of $\hat{\sigma}^2$, where we take the larger root of the quadratic equation. Note that a real valued solution exists for any $\hat{\sigma}^2 \leq \sigma^2_{\max} = \ell_K/(1+\sqrt{(p-K)/n})^2$. Our next estimate is then

$$\hat{\sigma}^2_{t+1} = F(\hat{\sigma}^2_t) = \frac{1}{p-K}\left[\sum_{j=1}^{p} \ell_j - \sum_{j=1}^{K} \rho_j(\hat{\sigma}^2_t)\right].$$

It can be shown that $d\hat{\rho}_j/d\sigma < 0$ for any $1 \leqslant j \leqslant K < p$, and so $dF/d\sigma > 0$. If $\hat{\sigma}_0^2 \leqslant \sigma^2_{\max}$ then by definition $F(\hat{\sigma}_0^2) > \hat{\sigma}_0^2$ and so $\hat{\sigma}_1^2 > \hat{\sigma}_0^2$. Thus, if $F(\sigma^2_{\max}) < \sigma^2_{\max}$, then the sequence of refined guesses $\hat{\sigma}_t^2$ is guaranteed to monotonically converge to a solution $\sigma^2_{\text{RMT}} = F(\sigma^2_{\text{RMT}})$. Otherwise, for some t we get $\hat{\sigma}_t^2 \geqslant \sigma^2_{\max}$, no real-valued solution $\rho_K(\hat{\sigma}_t^2)$ exists, and our noise estimator is the current guess. In simulations the sequence $\hat{\sigma}_t^2$ typically converges within a few iterations, so that the computational load of the improved noise estimator is negligible compared to the complexity of eigenvalue calculations. As described in [23], whereas for fixed p the relative bias of the maximum likelihood estimator (20) is $O(1/n)$, the relative bias of the improved estimator is $O(1/n^2)$. This smaller bias can lead to a substantial improvement in the algorithm's detection performance.

Remark: We offer a different interpretation to our noise estimator. A related problem in the statistical literature is the determination of the number of interaction terms in a fixed two-way model [31]. To test the hypothesis of $k-1$ terms versus k terms, the likelihood ratio test statistic is

$$\frac{\ell_k}{\frac{1}{p-k}\sum_{j=k+1}^{p} \ell_j}.$$

However, for $k > 1$ the distribution of this statistic depends on unknown *nuisance* parameters, which are the strengths of the first $k-1$ signals. In this context, our noise estimator attempts to remove the mean effect of these nuisance parameters.

IV. Consistency and Performance Analysis of Our Algorithm

In this section we analyze the asymptotic behavior of our algorithm in certain asymptotic regimes. Proofs appear in the Appendix.

A. Consistency

We start by considering the asymptotic properties of our algorithm as $n \to \infty$ with p fixed. Recall that the MDL procedure is strongly consistent, see (12). Since our method is based on a sequence of hypothesis tests, we show that it is *weakly consistent*, in the sense that as $n \to \infty$ it detects all signals present (Theorem 4), but has a small probability of overestimation, that can be controlled by the user (Lemma 2). Later on we show that if the false alarm probability α decreases with sample size n at a certain rate, then the resulting algorithm is strongly consistent, just as the MDL procedure (Theorem 5).

Theorem 4: Let $\hat{K}_n$ be the RMT estimator of number of signals given n samples of the form (1) with K signals. Define the random variable $A_n = \{\hat{K}_n \geqslant K\}$. Then, as $n \to \infty$ with p fixed, $A_n \xrightarrow{\text{a.s.}} 1$. In particular, this implies that

$$\lim_{n\to\infty} \Pr\{\hat{K}_n \geq K\} = 1.$$

We now consider the probability of overestimation. For simplicity we assume that the noise level is known. The effects of its estimation are discussed in the Appendix and are argued to be small.

Lemma 2: Consider complex-valued observations with an *a priori* known noise strength. The false alarm (overestimation) probability of our algorithm is asymptotically bounded by

$$\lim_{n\to\infty} \Pr\left\{\hat{K}_{\text{RMT}} > K\right\} \leqslant C_{p-K}(\alpha)$$

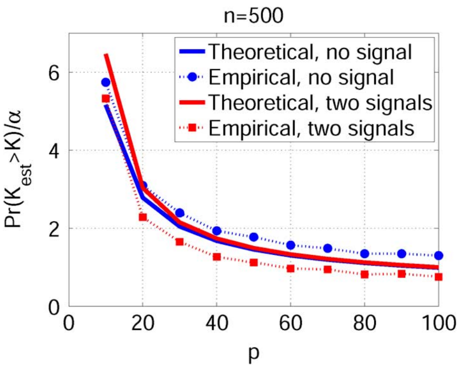

Fig. 2. Overestimation probability as a function of the number of sensors p and divided by α. The red curves are for a signal free system, whereas the blue curves are for a system carrying two signals with strengths $\lambda = (100, 50)$. Comparison of simulation results (dotted) with the theoretical results (solid) of Lemma 2, (23).

where

$$C_{p-K}(\alpha) = \exp\left\{-\frac{8\sqrt{2}}{3}\left(s(\alpha) - \frac{1}{(p-K)^{1/3}}\right)^{3/2}\right\}. \tag{23}$$

Fig. 2 compares $\Pr\{\hat{K} > K\}/\alpha$, for various values of α and K, as a function of p, with $n = 500$, to the theoretical result (23). Note that the empirical results are for an *a priori* unknown noise variance, estimated via (22). In the figure $s(\alpha)$ was approximated via (7) which introduces an additional source of error. Yet, the probability of overestimation is within a small multiplicative factor of the required confidence level α even for $p = 10$.

Finally, we show that our algorithm can be revised to be strongly consistent, e.g., satisfying (12). This can be done by replacing the fixed confidence level α, with a sample-size dependent one α_n, where $\alpha_n \to 0$ sufficiently slow as $n \to \infty$, so as to achieve both high detection performance at small sample size and consistency as $n \to \infty$.

Theorem 5: Let α_n be a sequence of significance levels and $s_n = s(\alpha_n)$ be their corresponding thresholds, computed by inverting the Tracy–Widom distribution. Assume $\alpha_n \to 0$ sufficiently slow such that $s_n \to \infty$ but $s_n/\sqrt{n} \to 0$. Let $\hat{K}_n$ denote the number of signals detected by our algorithm with n samples and with significance level α_n. Then,

$$\lim_{n\to\infty} \Pr\{\hat{K}_n = K\} = 1.$$

B. Performance Analysis

We now analyze the performance of our proposed algorithm for finite p and n. As in the analysis of the MDL estimator, we assume λ_K has multiplicity one and $\sqrt{p/n} < \lambda_K \ll \lambda_{K-1}$, so that the main source of error is misdetection of the kth sample eigenvalue. We first consider the case of known noise level $\sigma = 1$. The condition for our algorithm to report at least the correct number of signals K is

$$\ell_K > \mu_{n,p-K} + s(\alpha)\xi_{n,p-K}.$$

For $\lambda_K > \sqrt{p/n}$, according to [1] and [28], in the joint limit $p, n \to \infty$, the fluctuations of ℓ_K are Gaussian,

$$\sqrt{n}(\ell_K - \tau(\lambda_K)) \xrightarrow{d} \mathcal{N}(0, \delta^2(\lambda_K))$$

where

$$\tau(\lambda) = (\lambda+1)\left(1 + \frac{p-K}{n}\cdot\frac{1}{\lambda}\right)$$

is the asymptotic limit for the signal eigenvalue, see (10), and

$$\delta^2(\lambda) = \frac{2}{\beta}(\lambda+1)^2\left(1 - \frac{p-K}{n}\cdot\frac{1}{\lambda^2}\right).$$

While asymptotically there are no signal-signal interactions, for a more accurate performance analysis for finite p and n, we need to take into account the interaction between the signals. Consider the $K \times K$ submatrix of $\mathbf{S}_n$ which corresponds to the K-dimensional signal subspace, whose eigenvalues are $\rho_1 \geqslant \cdots \geqslant \rho_K$. A classical result by Lawley [17] shows that up to $o(1/n)$ terms

$$r_K = \mathbb{E}\{\rho_K\} = \lambda_K - \frac{1}{n}\sum_{j=1}^{K-1}\frac{\lambda_j}{\lambda_j - \lambda_K}.$$

Hence, an approximate expression for $\Pr\{\hat{K} \geqslant K\}$ is:

$$\Pr\{\hat{K} \geqslant K\} \approx \Pr\left\{\eta > \sqrt{n} \cdot \frac{\mu_{n,p-K} + s(\alpha)\xi_{n,p-K} - \tau(r_K)}{\delta(r_K)}\right\} \tag{24}$$

where $\eta \sim \mathcal{N}(0,1)$.

We now show that the effect of estimating σ^2 is small. Consider the threshold in the RHS of (24). For $\lambda > \sqrt{(p/n)}$, this threshold is negative and of order $O(\sqrt{n})$. Replacing $\sigma^2 = 1$ by any estimate $\hat{\sigma}^2 = \sigma^2(1 + o_P(1))$ does not change the leading order term of the threshold. Thus, asymptotically, (24) is a good approximation for the performance of our algorithm even when σ^2 is not known.

Furthermore, we can use (24) to answer the following question: what is the signal strength λ_K needed in order to detect it with probability at least $1/2$ by our method, i.e., for given finite values of n, p, and for fixed $\lambda_1, \ldots, \lambda_{K-1}$, what is λ_K s.t. $\Pr\{\hat{K} \geqslant K\} \approx (1/2)$? According to (24), this λ_K should approximately satisfy

$$\tau(r_K) = \mu_{n,p-K} + s(\alpha)\xi_{n,p-K}$$

where r_K is now a function of λ_K. Plugging (4) and (5) and the expression for $\tau(\lambda)$ above, and solving for λ gives

$$\lambda_{\mathrm{RMT}} \approx \sqrt{\frac{p-K}{n}}\left(1 + \frac{\sqrt{s(\alpha)(1 + \sqrt{(p-K)/n})}}{(p-K)^{1/3}}\right). \tag{25}$$

In Figs. 3 and 4 we present simulation results that support our theoretical analysis, and specifically the dependence of $\Pr\{\hat{K} \geqslant$

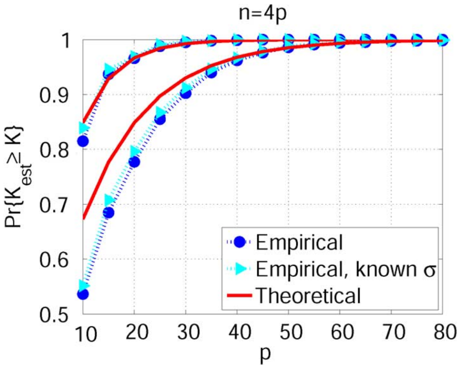

Fig. 3. Probability of estimating at least the correct number of signals as a function of the number of sensors p, with $n = 4 \cdot p$ samples. For the upper curve $K = 1$ and $\lambda = 1.5$. For the lower curve $K = 3$ and $\lambda = (20, 10, 1.2)$.

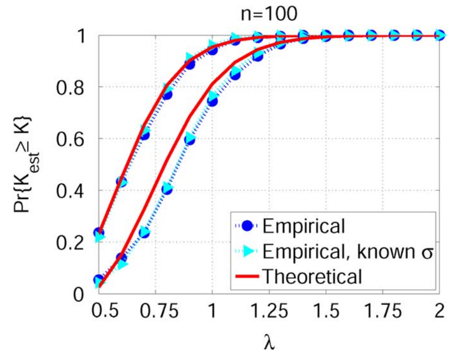

Fig. 4. Probability of estimating at least the correct number of signals as a function of signal strength λ. For the upper curve $K = 1, p = 10$ and $n = 100$. For the lower curve $K = 3, p = 20$ and $n = 100$. In this case, the first two signals are constant, $\lambda_1 = 20$ and $\lambda_2 = 15$, and the third signal is changing.

$K\}$ on the parameters p, n and λ, and that estimating σ^2 has very small effect on $\Pr\{\hat{K} \geqslant K\}$.

C. The Case $p > n$

In recent years, there is an increasing interest in statistical inference in the large p, small n case. When $p > n$ the standard MDL estimator is not applicable, since the smallest eigenvalues of $\mathbf{S}_n$ are identically zero. Hence, various alternative methods to estimate the number of signals have been developed. For example, [32] and [30] propose estimation algorithms based on the sequence of statistics

$$T_k = (p-k) \cdot \frac{\sum_{j=k+1}^{p} \ell_j^2}{\left(\sum_{j=k+1}^{p} \ell_j\right)^2} \quad k = 0, 1, \ldots, p-1.$$

One motivation for using these statistics is related to [18], where it was shown that the statistic T_0 can be used as a test of sphericity applicable for all values of p/n. Another convenient property is that under the null hypothesis of no signals, T_0 is scale free in the sense that its distribution is independent of the unknown noise level σ.

However, the main drawback of this statistic is that it attempts to distinguish the null hypothesis of no signals against *all* possible alternatives. Hence, as we show below, in the joint limit $p, n \to \infty$ estimators based on this statistic are not consistent for estimating the number of signals, in the sense that their limit of detection is strictly larger than the information limit of (11).

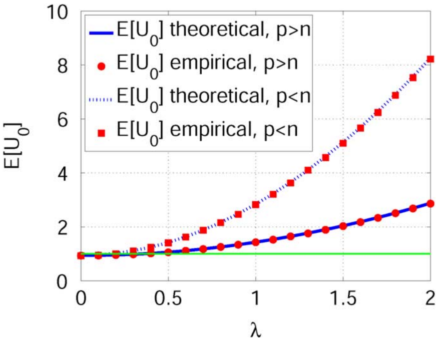

Fig. 5. Comparison of empirical mean of U_0 as a function of signal strength λ with the approximation (30).

For simplicity, we analyze the case of a single signal with strength λ. The method described in [32] estimates the number of signals by a sequence of hypothesis tests which depend on a user chosen confidence level α. In particular, for large p and n, it reports at least one signal if

$$nT_0 - n - p > 1 + 2\phi^{-1}(1-\alpha) \tag{26}$$

where $\phi(z) = \Pr\{\mathcal{N}(0,1) < z\}$. The following Lemma characterizes the asymptotic limit of detection corresponding to this test.

Lemma 3: For real valued observations, to reliably detect a signal of strength λ by the test (26), the asymptotic requirement on the signal strength, as $p, n \to \infty$, with $p/n = c$ is

$$\lambda/\sigma^2 \gtrsim 2\sqrt{s}\sqrt{\frac{p}{n}} \tag{27}$$

where $\phi(s) = 1 - \alpha$.

Similar results can be derived for complex valued noise.

Equation (27) shows that the detection limit for the method of [32] is strictly larger than the limit (11). By a similar analysis, it is possible to show that the AIC type estimator proposed by Rao and Edelman [30] also has an asymptotic detection limit which, although smaller than (27), is still strictly larger than (11). Hence, conjecture 6.3 in [30] is not true. We note that the method [30] may also overestimate the number of signals; see the Appendix in [23].

In Fig. 5 we compare the approximation (30) used in the Proof of Lemma 3 with the empirical mean of $U_0 = nT_0 - n - p$ as a function of signal strength λ for two distinct pairs of values $(p,n) = (50,100)$ or $(p,n) = (100,50)$. Note the excellent agreement between the two curves in both cases. Fig. 6 presents histograms of U_0 and of the largest eigenvalue ℓ_1 for three different values of signal strength λ, showing that ℓ_1 has more statistical power than U_0 (or equivalently T_0).

Fig. 7 shows simulation results for the probability of misdetection as a function of signal strength for our algorithm with $\alpha_{\mathrm{RMT}} = 0.5\%$, the algorithm proposed by Rao and Edelman [30], and the algorithm by Schott [32], with a confidence level of $\alpha = 1\%$, for $p = 2000$ and $n = 500$ real valued observations. The left vertical line is λ_{RMT}, (25), whereas the right vertical line is (27) with $s = 2$. Fig. 8 shows similar results for

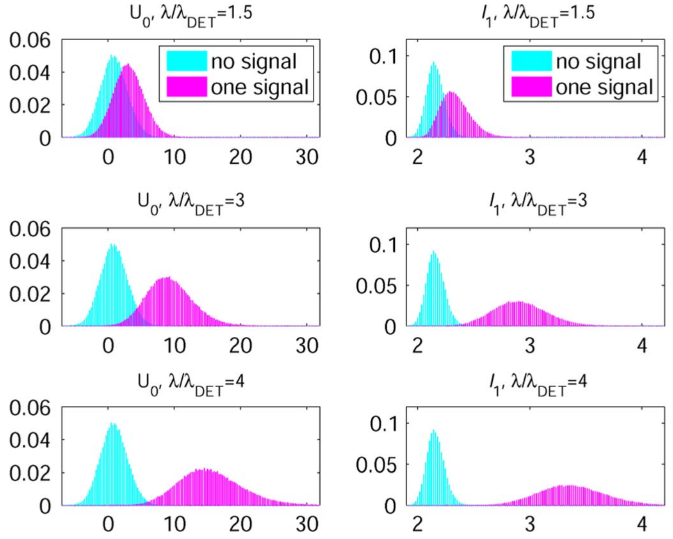

Fig. 6. The distribution of the statistic U_0 (left) and of the largest eigenvalue ℓ_1 (right) in the case of no signal versus the case of a single signal with strength λ for $p = 50$ and $n = 200$. In the top row, the signal is too weak to be detected by either of the two statistics. In the middle row the signal is sufficiently strong to be reliably detectable by the largest eigenvalue, but not by U_0. The bottom row shows the case of a strong signal, detectable by both methods.

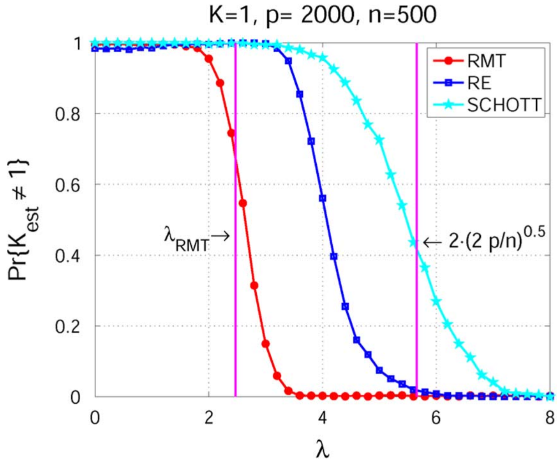

Fig. 7. Misdetection (error) probability as a function of signal strength, for $p = 2000, n = 500$ $(p > n)$.

$p = 200$ and $n = 800$. Note that in this case [30] overestimates the number of signals for $\lambda \gg 1$; see the Appendix in [23] for an explanation.

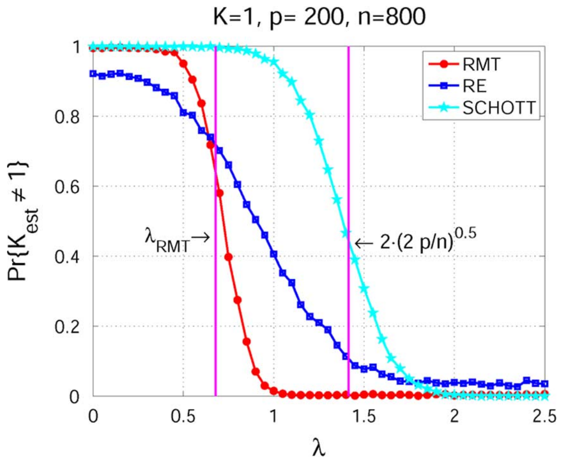

Fig. 8. Misdetection (error) probability as a function of signal strength, for $p = 200, n = 800$ $(p < n)$.

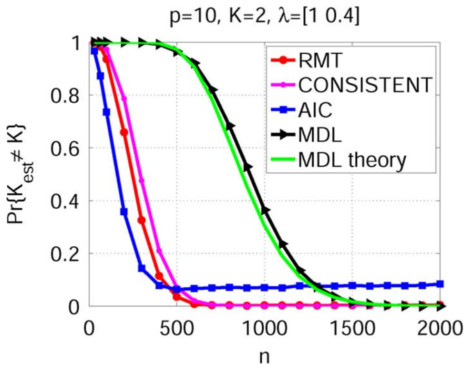

Fig. 9. Misdetection (error) probability as a function of sample size n.

V. Simulations

We compare the performance of our algorithm, namely (19) and (22) with a confidence level $\alpha_{\mathrm{RMT}} = 0.1\%$, to the standard MDL and AIC estimators [34], in a series of simulations all with complex valued signals and complex Gaussian noise with $\sigma = 1$. Our performance measure is the probability of misdetection,

$$\Pr\{\hat{K} \neq K\}.$$

In Fig. 9, we examine the performance of the different algorithms as a function of sample size n in the presence of two signals $(K = 2)$ with strengths $(\lambda_1, \lambda_2) = (1, 0.4)$, and with $p = 10$ sensors. We denote by CONSISTENT the RMT algorithm modified into a consistent algorithm by taking a decreasing confidence level $\alpha_n = 10^{-3}/\log n$; see Theorem 5. We also plot the theoretical misdetection probability of the MDL procedure, as derived in [10], [26]. This probability is given by $\phi(\eta)/(\sigma)$, where the quantities μ and σ^2 are approximations to the mean and variance of the random variable $\mathrm{MDL}(K-1) - \mathrm{MDL}(K)$ (see [10, (19) and (20)]). For μ we use the more accurate approximation of [26]; see (13). The simulation results are in very close agreement to the theoretical predictions.

In Fig. 10, results with $p = 25$ sensors are shown. Both figures show the superior detection performance of the RMT algorithm as compared to the MDL estimator. It is also evident that the AIC estimator is asymptotically inconsistent, having a non-negligible probability to overestimate the number of signals when n is large. This is also the reason for its seemingly better performance for small n. Even though the signal is too weak to be detected, the tendency for overestimation of the AIC estimator compensates for it. Finally, note that upon increasing the number of samples n, the RMT algorithm converges to (an almost) zero error probability. For example, with $p = 25$ sensors, $n = 3000$ and $K = 2$, we have $\Pr\{\hat{K}_{\mathrm{RMT}} = 2\} > 0.999$ whereas $\Pr\{\hat{K}_{\mathrm{MDL}} = 2\} < 0.8$. We also see that the modification of the RMT algorithm to a consistent one causes only a slight degradation in detection performance for small n.

In Fig. 11, we examine the performance of the different algorithms in the presence of $K = 3$ signals with strengths

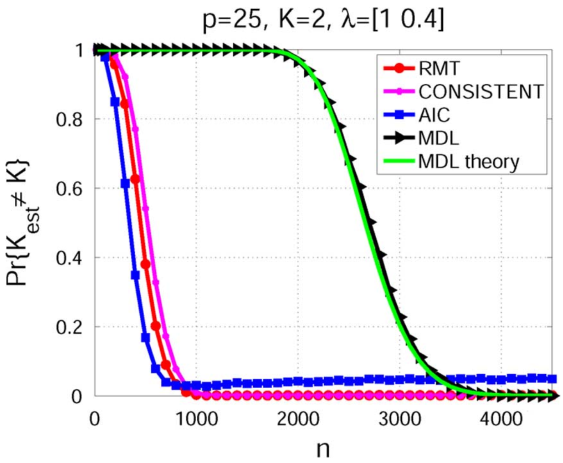

Fig. 10. Misdetection (error) probability as a function of sample size n.

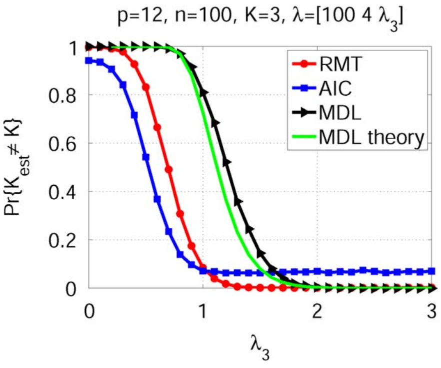

Fig. 11. Misdetection (error) probability as a function of signal strength λ.

$(100, 4, \lambda_3)$ as a function of λ_3, for $p = 12$ sensors and $n = 100$ samples. Again the RMT estimator has dominant performance over MDL, and the AIC estimator performs better for small λ_3 but does not converge to a zero misdetection error as λ_3 increases.

VI. Conclusion

In this paper, we presented a statistical analysis of the problem of non-parametric detection of the number of signals. We described the asymptotic limit of detection, showed that the highly popular MDL-based estimator is not optimal, presented a novel estimation method that asymptotically achieves this limit and analyzed its finite sample properties. The proposed algorithm is based on a sequence of hypothesis tests, at each step testing the significance of a single eigenvalue as arising from a signal. As shown both theoretically and by simulations, the proposed algorithm exhibits excellent performance under a wide range of parameter values.

In this paper, we focused on the (somewhat unrealistic) setting of homogeneous uncorrelated noise, with equal variance in all sensors. An interesting future research direction is to develop a similar estimator for more complicated settings such as heterogeneous sensors with unknown noise structure [11].

Appendix

Proof of Theorem 4: As $n \to \infty, \mathbf{S}_n \xrightarrow{\text{a.s.}} \mathbf{\Sigma}$, and hence the eigenvalues of $\mathbf{S}_n$ converge to those of $\mathbf{\Sigma}$ w.p.1. Similarly, as $n \to \infty$, for $1 \leqslant k \leqslant K, \hat{\rho}_j \xrightarrow{\text{a.s.}} \lambda_j + \sigma^2$ in (22), and hence

$$\hat{\sigma}^2_{\text{RMT}}(k) \xrightarrow{\text{a.s.}} \frac{1}{p-k} \sum_{j=k+1}^{p} (\lambda_j + \sigma^2)$$

with the convention that $\lambda_j = 0$ for $j > K$.

Further, as $n \to \infty$, $\mu_{n,p} \to 1$ and $\xi_{n,p} \to 0$, so the threshold in (19) converges to $\hat{\sigma}^2_{\text{RMT}}(k)$. Since for each $1 \leqslant k \leqslant K, \lambda_k + \sigma^2$ is strictly larger than the average of the remaining eigenvalues, it is asymptotically detected w.p.1. □

Proof of Lemma 2: Without loss of generality we assume $\sigma^2 = 1$. It is sufficient to consider the probability that the $(K+1)$th test passes. This occurs when $\ell_{K+1} > t_n$, where $t_n = \mu_{n,p-K-1} + s(\alpha)\xi_{n,p-K-1}$.

Let $\widetilde{\ell}_{K+1}$ denote the largest eigenvalue of the noise subspace of size $(p-K) \times (p-K)$ of the covariance matrix $\mathbf{S}_n$. Then, from Cauchy's Interlacing Theorem (see [12], Theorem 8.1.7), it follows that $\ell_{K+1} \leq \widetilde{\ell}_{K+1}$. Hence,

$$\Pr\{\ell_{K+1} > t_n\} \leqslant \Pr\{\widetilde{\ell}_{K+1} > t_n\}.$$

From (4) and (5) we have that

$$\begin{aligned} t_n &= \mu_{n,p-K-1} + s(\alpha)\xi_{n,p-K-1} \\ &\geqslant \left(1 + \sqrt{\frac{p-K}{n}}\right)^2 + \frac{s(\alpha)}{\sqrt{n}(p-K)^{1/6}} - \frac{1}{\sqrt{n(p-K)}} \\ &= \left(1 + \sqrt{\frac{p-K}{n}}\right)^2 + \frac{\delta}{\sqrt{n}} \end{aligned}$$

where

$$\delta = \frac{1}{(p-K)^{1/6}} \left(s(\alpha) - \frac{1}{(p-K)^{1/3}}\right).$$

Applying the bound of (8) with $\varepsilon = \delta/\sqrt{n}$ yields

$$\begin{aligned} \Pr\{\widetilde{\ell}_{K+1} > t_n\} &\leqslant \Pr\left\{\widetilde{\ell}_{K+1} > \left(1 + \sqrt{\frac{p-K}{n}}\right)^2 + \frac{\delta}{\sqrt{n}}\right\} \\ &\leqslant \exp\left\{-nJ_{\text{LAG}}\left(\frac{\delta}{\sqrt{n}}\right)\right\}. \end{aligned}$$

From the definition of the function J_{LAG}, (9), for $\delta > 0$

$$\begin{aligned} \lim_{n\to\infty} nJ_{\text{LAG}}\left(\frac{\delta}{\sqrt{n}}\right) &= \frac{8\sqrt{2}}{3}(p-K)^{1/4}\delta^{3/2} \\ &= \frac{8\sqrt{2}}{3}\left(s(\alpha) - \frac{1}{(p-K)^{1/3}}\right)^{3/2} \end{aligned}$$

which concludes the proof. □

Remark: The above analysis was conducted under the assumption that the noise variance σ^2 is known, but a similar conclusion holds also if we estimate the noise variance by any reasonable method [e.g., (20) or (22)], such that the noise estimator $\hat{\sigma}^2$ is a random variable of the form

$\hat{\sigma}^2 = \sigma^2\left(1-\eta/\sqrt{(p-K)n}\right)$ where $\eta = O_P(1)$. In this case we have that $\delta = 1/(p-K)^{1/6}\cdot\left(s(\alpha)-(1+\eta)/(p-K)^{1/3}\right)$, which slightly increases the overestimation probability.

Proof of Theorem 5: First we prove that $\lim_{n\to\infty}\Pr\{\hat{K}_n \geqslant K\} = 1$. As in the Proof of Theorem 4, we have that as $n \to \infty, \mathbf{S}_n \xrightarrow{\text{a.s.}} \mathbf{\Sigma}$ w.p.1, and thus for $1 \leqslant k \leqslant K$,

$$\hat{\sigma}^2_{\text{RMT}}(k) \xrightarrow{\text{a.s.}} \frac{1}{p-k}\sum_{j=k+1}^{p}(\lambda_j+\sigma^2).$$

Further, as $n\to\infty, \mu_{n,p}\to 1$, and the condition $s_n/\sqrt{n}\to 0$ guarantees $s_n\xi_{n,p}\to 0$, so the threshold in (19) converges to $\hat{\sigma}^2_{\text{RMT}}(k)$ as $n\to\infty$. Since for each $1\leqslant k\leqslant K, \lambda_k+\sigma^2$ is strictly larger than the average of the remaining eigenvalues, it is asymptotically detected w.p.1.

Next we prove $\lim_{n\to\infty}\Pr\{\hat{K}_n \neq K\} = 0$. It suffices to consider the $(K+1)$th test, which decides whether ℓ_{K+1} arises from a signal, and show that this test passes with probability that goes to zero as $n\to\infty$. Denote by t_n the threshold in (19). We show that $\lim_{n\to\infty}\Pr\{\ell_{K+1} > t_n\} = 0$.

Let $\widetilde{\ell}_{K+1}$ denote the largest eigenvalue of the noise subspace of size $(p-K)\times(p-K)$ of the covariance matrix $\mathbf{S}_n$. As in the Proof of Lemma 2, we have that

$$\Pr\{\ell_{K+1} > t_n\} \leqslant \Pr\{\widetilde{\ell}_{K+1} > t_n\}. \tag{28}$$

Thus, it suffices to show that the RHS goes to zero.

To this end, we consider the various components in the threshold $t_n = \hat{\sigma}^2(\mu_{n,p-K-1}+s_n\xi_{n,p-K-1})$ and how they scale when $n\to\infty$. From (4) and (5) we have

$$\mu_{n,p-K-1} = 1+2\sqrt{\frac{p-K-1}{n}}+O\left(\frac{1}{n}\right)$$

and $\xi_{n,p-K-1} \geqslant C_p/\sqrt{n}$ for some constant C_p that depends only on p. The noise estimator $\hat{\sigma}^2$ is a random variable of the form $\hat{\sigma}^2 = \sigma^2\left(1-\eta/\sqrt{pn}\right)$ where $\eta = O_P(1)$. Thus,

$$\begin{aligned} t_n &= \hat{\sigma}^2(\mu_{n,p-K-1}+s_n\xi_{n,p-K-1}) \\ &= \left(1-\frac{\eta}{\sqrt{pn}}\right)\left(1+2\sqrt{\frac{p-K-1}{n}}+C_p\frac{s_n}{\sqrt{n}}\right). \end{aligned}$$

Let b_n be some sequence such that $b_n\to\infty$ as $n\to\infty$, but $b_n = o(s_n)$, that is, $b_n/s_n\to 0$. For n sufficiently large, the later condition on the growth of b_n guarantees that with probability converging to 1 as $n\to\infty$,

$$t_n > \left(1+\sqrt{\frac{p}{n}}+\frac{b_n}{\sqrt{n}}\right)^2.$$

Random matrix theory provides us with the following non-asymptotic bound, that follows from Gordon's inequality [19]:

$$\Pr\left\{\widetilde{\ell}_{K+1} > \left(1+\sqrt{\frac{p}{n}}+\frac{b_n}{\sqrt{n}}\right)^2\right\} < e^{-b_n^2/2}.$$

Since $b_n\to\infty$, the RHS in (28) goes to zero when $n\to\infty$. □

Proof of Lemma 3: Without loss of generality we assume that $\sigma^2 = 1$, and denote $U_0 = nT_0-n-p$. As proven in [32], under the null hypothesis of no signals, in the joint limit $p,n\to\infty$ and regardless of the limiting value of $c = p/n, U_0 \xrightarrow{d} \mathcal{N}(1,4)$.

In the presence of a signal of strength λ, the average $\mathbb{E}[U_0(\lambda)]$ is shifted upwards. Assuming that its variance is not changed significantly, the condition to reliably detect its presence via (26) is approximately

$$\frac{\mathbb{E}[U_0(\lambda)]-\mathbb{E}[U_0(0)]}{\sqrt{\text{Var}[U_0(0)]}} = \frac{\mathbb{E}[U_0(\lambda)]-1}{2} > 2s \tag{29}$$

where $\phi(s) = 1-\alpha$.

By definition, $\mathbb{E}[U_0(\lambda)] = n\mathbb{E}[T_0(\lambda)]-n-p$. Assuming both p,n are large, we approximate $\mathbb{E}[T_0(\lambda)]$ by $p\cdot(\mathbb{E}[\sum\ell_j^2])/(\mathbb{E}[(\sum\ell_j)^2])$, which gives

$$\mathbb{E}[T_0] \approx p\frac{A(\lambda)+(p-1)\left[1+c+\frac{1}{n}+\frac{2}{n}(\lambda+1)\right]}{A(\lambda)+(p-1)\left[2(\lambda+1)+(p-1)+\frac{2}{n}\right]} \tag{30}$$

where $A(\lambda) = (\lambda+1)^2(1+(2/n))$. Plugging (30) into (29), and solving for λ gives the asymptotic condition (27). □

Acknowledgment

The authors would like to thank P. Bickel, I. Johnstone, and O. Zeitouni for interesting discussions. They would also like to thank the anonymous referees for their constructive remarks which greatly improved the manuscript.

References

[1] J. Baik, G. Ben Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *Ann. Probab.*, vol. 33, no. 6, pp. 1643–1697, 2005.

[2] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *J. Multivariate Anal.*, vol. 97, no. 6, pp. 1382–1408, 2006.

[3] J. F. Böhme, , S. Haykin, Ed., "Array processing," in *Advances in Spectrum Analysis and Array Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1991, pp. 163–.

[4] W. Chen, K. M. Wong, and J. P. Reilly, "Detection of the number of signals: A predicted eigen-threshold approach," *IEEE Trans. Signal Process.*, vol. 39, no. 5, pp. 1088–1098, May 1991.

[5] P. Chen, M. C. Wicks, and R. S. Adve, "Development of a statistical procedure for detecting the number of signals in a radar measurement," *Proc. Inst. Electr. Eng.—Radar Sonar Navig.*, vol. 148, no. 4, pp. 219–226, 2001.

[6] P.-J. Chung, J. F. Böhme, C. F. Mecklenbraüker, and A. O. Hero, "Detection of the number of signals using the Benjamini–Hochberg procedure," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2497–2508, Jun. 2007.

[7] N. El Karoui, "A rate of convergence result for the largest eigenvalue of complex White Wishart matrices," *Ann. Prob.*, vol. 36, no. 6, pp. 2077–2117, 2006.

[8] N. El Karoui, "On the largest eigenvalue of Wishart matrices with identity covariance when n,p and n/p tend to infinity," *Bernoulli*, 2008, to be published.

[9] E. Fishler and H. Messer, "On the use of order statistics for improved detection of signals by the MDL criterion," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 2242–2247, Aug. 2000.

[10] E. Fishler, M. Grosmann, and H. Messer, "Detection of signals by information theoretic criteria: General asymptotic performance analysis," *IEEE Trans. Signal Process.*, vol. 50, no. 5, pp. 1027–1036, May 2002.

[11] E. Fishler and H. V. Poor, "Estimation of the number of sources in unbalanced arrays via information theoretic criteria," *IEEE Trans. Signal Process.*, vol. 53, no. 9, pp. 3543–3553, Sep. 2005.

[12] G. H. Golub and C. F. Van Loan*, Matrix Computations*, 3rd ed. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.

[13] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Stat. Soc., ser. E*, vol. 41, pp. 190–195, 1979.

[14] K. Johansson, "Shape fluctuations and random matrices," *Comm. Math. Phys.*, vol. 209, no. 2, pp. 437–476, 2000.

[15] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal component analysis," *Ann. Stat.*, vol. 29, pp. 295–327, 2001.

[16] W. J. Krzanowski, *Principles of Multivariate Analysis*. New York: Oxford Univ. Press, 1988.

[17] D. N. Lawley, "Tests of significance for the latent roots of covariance and correlation matrices," *Biometrika*, vol. 43, pp. 128–136, 1956.

[18] O. Ledoit and M. Wolf, "Some hypothesis tests for the covariance matrix when the dimension is large with respect to sample size," *Ann. Stat.*, vol. 30, no. 4, pp. 1081–1102, 2002.

[19] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer-Verlag, 1991.

[20] M. Ledoux, , V. D. Milman and G. Schechtman, Eds., *Deviation Inequalities on Largest Eigenvalues, in Geometric Aspects of Functional Analysis*, ser. Lecture Notes in Mathematics. New York: Springer, 2007, vol. 1910.

[21] P. Miksik, M. Meloun, J. Capek, and R. G. Brereton, "Critical comparison of methods predicting the number of components in spectroscopic data," *Anal. Chim. Acta*, vol. 423, pp. 51–68, 2000.

[22] S. N. Roy, "On a heuristic method of test construction and its use in multivariate analysis," *Ann. Math. Stat.*, vol. 24, no. 2, pp. 220–238, 1953.

[23] S. Kritchman and B. Nadler, "Determining the number of components in a factor model from limited noisy data," *Chem. Int. Lab. Syst.*, vol. 94, pp. 19–32, 2008.

[24] R. J. Muirhead, "Latent roots and matrix variates: A review of some asymptotic results," *Ann. Stat.*, vol. 6, no. 1, pp. 5–33, 1978.

[25] B. Nadler, "Finite sample approximation results for principal component analysis: A matrix perturbation approach," *Ann. Stat.*, vol. 36, no. 6, pp. 2791–2817, 2008.

[26] B. Nadler, "Detection of Signals by information theoretic criteria: Accurate performance analysis and a simple improved estimator," *IEEE Trans. Signal Process.*, submitted for publication.

[27] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS Genet*, vol. 2, no. 12, pp. 190–, 2006.

[28] D. Paul, "Asymptotics of sample eigenstruture for a large dimensional spiked covariance model," *Stat. Sinica*, vol. 17, no. 4, pp. 1617–1642, 2007.

[29] A. Quinlan, J. P. Barbot, P. Larzabal, and M. Haardt, "Model order selection for short data: An exponential fitting test," *EURASIP J. Adv. Signal Process.*, vol. 2007, Article ID 71953.

[30] N. R. Rao and A. Edelman, "Sample eigenvalue based detection of high dimensional signals in white noise using relatively few samples," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2625–2638, Jul. 2008.

[31] J. Schott, "A note on the critical values used in stepwise tests for multiplicative components of interaction," *Comm. Stat. Th. Meth.*, vol. 15, no. 5, pp. 1561–1570, 1986.

[32] J. Schott, "A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix," *J. Mult. Anal.*, vol. 97, no. 4, pp. 827–843, 2006.

[33] P. Stoica and Y. Selén, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, pp. 36–47, Jul. 2004.

[34] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 387–392, Apr. 1985.

[35] K. M. Wong, Q.-T. Zhang, J. P. Reilly, and P. C. Yip, "On information theoretic criteria for determining the number of signals in high resolution array processing," *IEEE Trans. Signal Process.*, vol. 38, no. 11, pp. 1959–1971, Nov. 1990.

[36] W. Xu and M. Kaveh, "Analysis of the performance and sensitivity of eigendecomposition-based detectors," *IEEE Trans. Signal Process.*, vol. 43, no. 6, pp. 1413–1426, Jun. 1995.

[37] Q. T. Zhang, K. M. Wong, P. C. Yip, and J. P. Reilly, "Statistical analysis of the performance of information theoretic criteria in the detection of the number of signals in array processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 10, pp. 1557–1567, 1989.

[38] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of the number of signals in presence of white noise," *J. Multivariate Anal.*, vol. 20, pp. 1–25, 1986.

Shira Kritchman received the B.Sc. degree in mathematics and physics (*summa cum laude*) from the Hebrew University of Jerusalem, Israel, in 2004 and the M.Sc. degree in applied mathematics from the Weizmann Institute of Science, Rehovot, Israel, in 2008.

She is currently a Research Assistant in the Department of Computer Science and Applied Mathematics at the Weizmann Institute of Science, exploring options for her Ph.D. studies.

Boaz Nadler was born in Israel in 1971. He received the B.Sc. degree in mathematics and physics (*cum laude*) in 1993, the M.Sc. degree in applied mathematics (*summa cum laude*), and the Ph.D. degree in applied mathematics, all from Tel Aviv University (TAU), Tel Aviv, Israel.

From 2002 to 2005, he was a Gibbs Instructor/Assistant Professor at the Department of Mathematics at Yale University, New Haven, CT. He is currently a Senior Research Scientist in the Department of Computer Science and Applied Mathematics at the Weizmann Institute of Science, Rehovot, Israel. His research interests are in stochastic processes, mathematical statistics, machine learning, and their applications in chemometrics and in signal processing.