Sensors **2015**, *15*, 20250-20266; doi:10.3390/s150820250

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

Statistical Analysis of the Performance of MDL Enumeration for Multiple-Missed Detection in Array Processing

Fei Du, Yibo Li * and Shijiu Jin

State Key Laboratory of Precision Measurement Technology and Instrument, Tianjin University, Tianjin 300072, China; E-Mails: dukemyy@tju.edu.cn (F.D.); shjjin@tju.edu.cn (S.J.)

* Author to whom correspondence should be addressed; E-Mail: slyb@tju.edu.cn; Tel.: +86-22-2740-2366; Fax: +86-22-2789-0026.

Academic Editors: Yunchuan Sun, Antonio Jara and Shengling Wang

Received: 8 June 2015 / Accepted: 13 August 2015 / Published: 18 August 2015

Abstract: An accurate performance analysis on the MDL criterion for source enumeration in array processing is presented in this paper. The enumeration results of MDL can be predicted precisely by the proposed procedure via the statistical analysis of the sample eigenvalues, whose distributive properties are investigated with the consideration of their interactions. A novel approach is also developed for the performance evaluation when the source number is underestimated by a number greater than one, which is denoted as "multiple-missed detection", and the probability of a specific underestimated source number can be estimated by ratio distribution analysis. Simulation results are included to demonstrate the superiority of the presented method over available results and confirm the ability of the proposed approach to perform multiple-missed detection analysis.

Keywords: performance analysis; minimum description length (MDL); array processing; multiple-missed detection; source enumeration

1. Introduction

Source enumeration is a critical step in array signal processing and widely used in many scenarios [1]. The accuracy and the tendency of enumeration will significantly affect the performance of succeeding algorithms, such as direction-of-arrival (DOA) estimation [2] or blind source separation [3]. Minimum description length (MDL) criterion derived by Rissanen [4], or its equivalent criterion derived by

Schwarz under the name of Bayesian information criterion (BIC) [5], is one of the most commonly used enumeration methods for its low complexity and asymptotic consistency which ensures a correct estimation as the sample size tends to infinity [6]. A lot of algorithms have been proposed to improve the MDL criterion for a performance promotion, low computation complexity or robustness in various environments [7–15]. Dayan and Rausley have presented a norm-based improved MDL (iMDL) algorithm in [7] by nonlinear rescaling of the sample eigenvalues and the corresponding normalized indexes. By using the training sequence of the desired signal, a minimum mean square error (MMSE) based MDL method has been developed by Huang in [8] to get a more accurate estimation of the source number. Huang and So have also employed the linear shrinkage estimation of noise subspace covariance matrix instead of sample covariance matrix in MDL criterion in [9] to achieve a more reliable detection in severe environments where the number of snapshots is comparable or even smaller than the number of sensors. For the cases at a small sample size, the probability density function of the sample eigenvalues has been taken into consideration in MDL as an essential supplement in [10,11]. To handle the coherent signals contaminated by colored noise, Zhen and Si [12] have whitened the sample eigenvalues to eliminate the inequality of eigenvalues caused by colored noise. Fishler and Poor [13] have proposed a robust-MDL (RMDL) method with proven consistency for source enumeration under non-uniform noise situations while Huang *et al.* [14,15] have improved MDL by introducing a multi-stage Wiener filter by using the filtered component variances or MMSE rather than the sample eigenvalues, which can offer computational simplicity and robustness to non-uniform noise.

Most of the above methods will give an accurate estimation of source number under the assumption of an infinite sample size. However, only a limited number of observations is available in practice particularly in the applications with latency requirements. Thus, the performance analysis of MDL at a finite sample size is of great practical value.

The statistical performance of MDL criterion for source enumeration has been extensively analyzed in [16–27]. Since the method is eigendecomposition-based, the statistical property of the sample eigenvalues has been investigated considerably. The distributions of the sample eigenvalues which were derived from the multivariate statistical theory [28], have been used by the authors in [16–18]. However, the performance estimations are found biased when the sample size is not sufficiently large. Recently, the random matrix theory [29] approach has been proposed to solve the enumeration problem in array processing in [30–33]. Asymptotic distributions of the sample eigenvalues have been given to rectify the bias by taking in the influence on signal eigenvalues from the noise subspace under the assumption of large dimension. For non-circular or non-Gaussian cases, statistical analysis has been made in [22,23] by taking the fourth-order statistics of the signals into consideration. In [34], the interactions between signal eigenvalues are considered by Lawley on the distribution analysis of the sample eigenvalues. The authors in [24–27] are able to predict the probability of underestimation in close accordance at a moderate sample size by a combination of different theories.

Although many analyses on the performance of MDL are available, few discuss the cases that the source number is underestimated by a number greater than one, which is denoted as “multiple-missed detection” here. Since the enumeration performance of MDL is signal-to-noise ratio (SNR) dependent, the source numbers estimated under varying noise levels may be different. This inconsistency in enumeration can be attributed to either the variation of actual source number or the disturbance of noise. By the statistical analysis on multiple-missed detection, the probability of a specific enumeration result can be estimated as an important reference for the attribution of the enumeration discrepancy.

In this paper, we propose a new procedure for the performance evaluation of MDL, which can predict the estimation results of MDL precisely at a finite sample size by considering the interactions between signal eigenvalues. A novel approach is also developed for the multiple-missed detection analysis. Thus the deterioration of enumeration performance with the degradation of SNR can be estimated.

The remainder of this paper is organized as follows: the problem formulations are given in Section 2, including the array signal model and the theoretical analysis of underestimation. Section 3 introduces the statistical analysis of underestimation by discussing the distributive property of the sample eigenvalues. Simulation results that illustrate the superior performance of the proposed method and the performance for multiple-missed detection are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Problem Formulation

2.1. Array Signal Model

Consider q narrowband far-field and incoherent sources impinging on a sensor array of p elements ($p > q$). The observed signals can be modeled as a superposition of source signals corrupted by additive circular Gaussian noise, which can be written as:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{A}$ is the $p \times q$ array steering matrix composed of q linearly independent column vectors, $\boldsymbol{s}$ is the q-dimensional source signal vector with nonsingular covariance matrix $\boldsymbol{R_S} = E\,[\boldsymbol{ss}^H]$ where $(.)^H$ stands for conjugate transpose, $\boldsymbol{n}$ is the source-independent noise vector with zero mean and covariance $\sigma^2\boldsymbol{I}$ where $\boldsymbol{I}$ is the $p \times p$ identity matrix. Signals and noises are assumed to be i.i.d. and complex circular Gaussian distributed. The $p \times p$ population covariance matrix $\boldsymbol{R}$ is calculated as:

$$\boldsymbol{R} = E[\boldsymbol{xx}^H] = \boldsymbol{AR_S A}^H + \sigma^2\boldsymbol{I} \tag{2}$$

whose population eigenvalues in descending order are given by:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q > \lambda_{q+1} = \ldots = \lambda_p = \sigma^2 \tag{3}$$

The first q eigenvalues of $\boldsymbol{R}$ are contributed by both the source signals and the noise, which are called the signal eigenvalues. The last $p - q$ eigenvalues are contributed by noise only, which are called the noise eigenvalues. The population covariance matrix can be estimated using the sample covariance matrix $\hat{\boldsymbol{R}}$:

$$\hat{\boldsymbol{R}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^H \tag{4}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are the independent and identically distributed snapshots of $\boldsymbol{x}$. The corresponding sample eigenvalues of $\hat{\boldsymbol{R}}$ in descending order are given by:

$$l_1 > l_2 > \ldots > l_q > l_{q+1} > \ldots > l_p \tag{5}$$

2.2. Source Enumeration of MDL

Assuming that all the observations are i.i.d. complex circular Gaussian random vectors with zero mean, the MDL estimator formulation is given by:

$$MDL(k) = -L(k) + y(k) = N(p-k)\ln\frac{A_k}{G_k} + \frac{1}{2}\left[k(2p-k)+1\right]\ln N \tag{6}$$

where $L(k)$ is the log-likelihood term, $y(k)$ is the penalty term:

$$A_k \triangleq \frac{1}{p-k}\sum_{i=k+1}^{p} l_i \tag{7}$$

and:

$$G_k \triangleq \prod_{i=k+1}^{p} l_i^{1/(p-k)} \tag{8}$$

represent the arithmetic and geometric means of the last $p - k$ eigenvalues respectively. The estimated source number is denoted by $\hat{q}$ which can be derived as follows:

$$\hat{q} = \arg\min_{k} MDL(k), k = 0,..., p-1 \tag{9}$$

Let H_q denote the hypothesis that the true number of sources is q. The probability of incorrect estimation P_e is defined as:

$$P_e = P\left(\hat{q} \neq q \mid H_q\right) = P_m + P_f \tag{10}$$

where the probability of missed detection P_m is defined as:

$$P_m = P\left(\hat{q} < q \mid H_q\right) \tag{11}$$

and the probability of false alarm P_f is defined as:

$$P_f = P\left(\hat{q} > q \mid H_q\right) \tag{12}$$

Ding and Kay [35] have proven that MDL is inconsistent at high SNR with a finite sample size. However, for the relatively large penalty term, MDL has a trend of underestimation of the source number at low SNR. For example, in the setting of p = 5, q = 2, N = 50 and SNR = 3 dB, the probability of false alarm P_f is 0.0013 while the probability of underestimation P_m is 0.5029 based on a 10,000 trial Monte Carlo simulation. Zhang *et al.*, showed in [18] that for a moderate number of snapshots, the probability of false alarm using the MDL criterion is approaching zero. So the probability of incorrect estimation P_e is dominated by the probability of underestimation P_m, which can be expressed as:

$$P_e \approx P_m = P(\hat{q} < q \mid H_q) \approx P\left(MDL(q-1) < MDL(q) \mid H_q\right) \tag{13}$$

Using Equation (6) in Equation (13) we obtain:

$$P_m \approx P\left(\frac{A_{q-1}^{p-q+1}}{l_q A_q^{p-q}} < \exp\left(\frac{(2p-2q+1)\ln N}{2N}\right)\right) \tag{14}$$

According to the definition of A_k in Equation (7), we can rewrite A_{q-1} as:

$$A_{q-1} = \frac{\sum_{i=q}^{p} l_i}{p-q+1} = \frac{l_q + \sum_{i=q+1}^{p} l_i}{p-q+1} = \frac{l_q + (p-q)A_q}{p-q+1} \tag{15}$$

In order to simplify Equation (14), we define:

$$\rho_i = l_i / A_i, \rho_i > 1 \tag{16}$$

when $i = q$, we will have:

$$\rho_q = l_q / A_q, \rho_q > 1 \tag{17}$$

Substituting Equations (15) and (17) into Equation (14), we can get:

$$P_m \approx P\left(f_1(\rho_q) < C_1\right) \tag{18}$$

where:

$$\mathrm{w}\, f_1(x) \triangleq \left(x + p - q\right) x^{-\frac{1}{p-q+1}} \tag{19}$$

and:

$$C_1 = (p-q+1)\exp\left(\frac{(2p-2q+1)\ln N}{2N(p-q+1)}\right) \tag{20}$$

The function $f_1(x)$ is a monotonically increasing function in the region of $x > 1$, therefore we can transform Equation (17) into a simpler form as:

$$P_m \approx P\left(\rho_q < \rho_{C1}\right) \tag{21}$$

in which:

$$f_1(\rho_{C1}) = C_1 \tag{22}$$

Since Equation (21) cannot be solved analytically, we can use the Newton-Raphson method to find a very accurate solution of ρ_{C1} numerically from the initial value derived by binomial expansion:

$$\rho_{C1}^0 = D_1 + \sqrt{D_1^2 - 1} \tag{23}$$

where:

$$D_1 = \left(\frac{C_1}{p-q+1}\right)^{p-q+1} \tag{24}$$

For a more complicated situation, we will discuss the cases of multiple-missed detection as follows. Consider that the true source number q is underestimated by a number greater than or equal to d, we define the corresponding probability as:

$$P_{md} = P\left(\hat{q} \le q - d \mid H_q\right), d < q \tag{25}$$

in order to distinguish the probabilities for different underestimated source numbers. So:

$$P_{md} \approx P\left(MDL(q-d) < MDL(q-d+1) \,|\, H_q \right) \tag{26}$$

By using Equation (6) in Equation (26), we can derive:

$$P_{md} \approx P\left(\frac{A_{q-d}^{p-q+d}}{l_{q-d+1} A_{q-d+1}^{p-q+d-1}} < \exp\left(\frac{(2p-2q+2d-1)\ln N}{2N} \right) \right) \tag{27}$$

Let $i = q - d + 1$ in Equation (16), we will have:

$$\rho_{q-d+1} = l_{q-d+1} / A_{q-d+1}, \rho_{q-d+1} > 1 \tag{28}$$

and obtain a simpler form of Equation (26):

$$P_{md} \approx P\left(f_d(\rho_{q-d+1}) < C_d \right) \tag{29}$$

where:

$$f_d(x) \triangleq \left(x + p - q + d - 1 \right) x^{-\frac{1}{p-q+d}} \tag{30}$$

$$C_d = (p-q+d)\exp\left(\frac{(2p-2q+2d-1)\ln N}{2N(p-q+d)} \right) \tag{31}$$

Since $f_d(x)$ is also a monotonically increasing function in the region of $x > 1$, Equation (29) can be rewritten as:

$$P_{md} \approx P\left(\rho_{q-d+1} < \rho_{Cd} \right) \tag{32}$$

in which:

$$f_1(\rho_{Cd}) = C_d \tag{33}$$

The threshold ρ_{Cd} can be calculated numerically from the initial value:

$$\rho_{Cd}^0 = D_d + \sqrt{D_d^2 - 1} \tag{34}$$

where:

$$D_d = \left(\frac{C_d}{p-q+d} \right)^{p-q+d} \tag{35}$$

We can find that P_m is a particular form of P_{md} when $d = 1$. So the problem of underestimation probability turns into the statistical performance analysis of ρ_i which will be discussed in the next section. The expectation of the estimated source number can be calculated by:

$$E(\hat{q}) = q(1 - P_{m1}) + P_{m(q-1)} + \sum_{d=1}^{q-2} (q-d) P_{md} (1 - P_{m(d+1)}), \hat{q} \in [1, q] \tag{36}$$

which would be an effective indicator of the extent of underestimation.

3. Performance Analysis of Multiple-Missed Detection

According to Equation (16), the statistics of ρ_i are determined by the distributions of l_i and A_i. Many researches have been done on the statistical properties of the signal sample eigenvalue l_i and the arithmetic mean of the noise sample eigenvalues A_q, and can be mainly divided into multivariate statistical theory [28], random matrix theory [29] and Lawley's theory [34]. The multivariate statistical theory is derived based on large sample asymptotics and requires a large value of sample size N. Random matrix theory has been proposed to investigate the spectral properties of random matrices with the assumption of high-dimension and large sample asymptotic regime. Nadakuditi and Edelman have concluded in [32] that for a signal-free sample covariance matrix formed from a $p \times N$ matrix of observations with i.i.d. Gaussian samples of zero mean and identical variance σ^2 , the sample eigenvalues will follow the Marchenko-Pastur distribution and their arithmetic mean will converge to Gaussian distribution asymptotically as $p, N \rightarrow \infty$ with $p / N \rightarrow c \in (0, \infty)$, *i.e.*,

$$p\left(\frac{1}{p}\sum_{i=1}^{p} l_i - \sigma^2\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{2\sigma^4 c}{\beta}\right) \tag{37}$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution and β = 1 or 2 for real or complex values respectively. For a q-signal-bearing case defined in Section 2.1, as $p, N \rightarrow \infty$ with $p / N \rightarrow c \in (0, \infty)$, if all the signal eigenvalues of the population covariance matrix are larger than the critical value $\sigma^2(1+\sqrt{c})$ where σ^2 stands for the value of noise eigenvalue, which means no phase transition phenomenon, the distribution of signal sample eigenvalues can be described as following. If the signal eigenvalue $\lambda_i > \sigma^2(1+\sqrt{c})$ has multiplicity of one for $i \le q$ and $\sqrt{N}\left|c - p / N\right| \rightarrow 0$, the distribution of l_i converges to Gaussian distribution as:

$$\sqrt{N}\left(l_i - \lambda_i\left(1 + \frac{\sigma^2 c}{\lambda_i - \sigma^2}\right)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{2}{\beta}\lambda_i^2\left(1 - \frac{c\sigma^4}{\left(\lambda_i - \sigma^2\right)^2}\right)\right) \tag{38}$$

The distribution analysis of A_q can be performed as the arithmetic mean of the sample eigenvalues of a signal-free sample covariance matrix formed from a $(p - q) \times N$ matrix of observations. Although the random matrix theory is derived under the assumption of large dimension, some simulation results have shown that it may also work well in some low-dimension cases. Lawley has constructed a matrix with the same eigenvalues as the sample covariance matrix by using the sampling errors. By comparing the diagonal elements, the statistics of sample eigenvalues are derived under the assumption of Gaussian distribution. All the three theories assume that l_i and A_q follow Gaussian distribution asymptotically and the expectations and variances are listed in Table 1 for comparison.

Note that the expectations and variances of l_i in random matrix theory and Lawley's theory have an additional term compared with those in multivariate statistical theory. The random matrix theory has taken the disturbance from noise subspace into account while Lawley's theory has considered the interactions between signal eigenvalues in addition. When the sample size N is sufficiently large, the additional terms will diminish to zero and the distributions of all the three theories will equal to each other. Similarly, the expectation of A_q in Lawley's theory is different from the others for including the bias induced by the signal eigenvalues. Thus, we employ the expectation and variance of l_i in Lawley's theory for an accurate analysis, and additional terms with the order higher than $O(N^{-1})$ in the expectations or $O(N^{-2})$ for variances are omitted as they are relatively small and decay rapidly with increasing sample size. The phenomenon of phase transition is not considered for simplicity in this paper.

Table 1. Comparison of three theories on the expectations and variances of l_i and A_q.

	Multivariate Statistics	**Random Matrix**	**Lawley**
$E(l_i)$ *[1]	λ_i	$\lambda_i+\frac{c\lambda_i\sigma^2}{\lambda_i-\sigma^2}$ *[3]	$\lambda_i+\sum_{j\neq i}^{p}\frac{\lambda_i\lambda_j}{N(\lambda_i-\lambda_j)}$
$Var(l_i)$ *[1]	$\frac{\lambda_i^2}{N}$	$\frac{2\lambda_i^2}{\beta N}\left[1-\frac{c\sigma^4}{(\lambda_i-\sigma^2)^2}\right]$ *[2],*[3]	$\frac{2\lambda_i^2}{\beta N}\left[1-\frac{1}{N}\sum_{j\neq i}^{p}\left(\frac{\lambda_j}{\lambda_i-\lambda_j}\right)^2\right]$ *[2]
$E(A_q)$	σ^2	σ^2	$\sigma^2-\sum_{i=1}^{q}\frac{\lambda_i\sigma^2}{N(\lambda_i-\sigma^2)}$
$Var(A_q)$	$\frac{\sigma^4}{N(p-q)}$	$\frac{2\sigma^4}{\beta N(p-q)}$ *[2]	/

*[1] $i \le q$ for the expectations and variances of l_i. $E(.)$ and $Var(.)$ are the mathematical expectation and variance respectively; *[2] $\beta = 1$ for real-valued signals and $\beta = 2$ for complex-valued signals. The original formula of Lawley's $Var(l_i)$ without β is revised here; *[3] c is a positive finite value when $p, N \to \infty$, $p/N \to c$.

The expectation and variance of A_i are given only under the condition of $i = q$ in above theories. Now we will discuss the distributive property of A_i to get general expressions when $i < q$. According to the definition of A_i in Equation (7), we can rewrite A_i as:

$$A_i=\frac{1}{p-i}\left((p-q)A_q+\sum_{j=i+1}^{q}l_j\right) \tag{39}$$

Noting that:

$$E\left((p-i)A_i+\sum_{j=1}^{i}l_i\right)=E(tr(\hat{R}))=tr(R)=\sum_{j=1}^{p}l_j \tag{40}$$

we can obtain:

$$E(A_i)=\frac{1}{p-i}\left\{\sum_{j=i+1}^{q}\left[\lambda_j+\sum_{k=1}^{i}\frac{\lambda_j\lambda_k}{N(\lambda_j-\lambda_k)}\right]-\sum_{j=1}^{i}\frac{\lambda_j\sigma^2(p-q)}{N(\lambda_j-\sigma^2)}+(p-q)\sigma^2\right\},i\le q \tag{41}$$

which equals to the expectation of A_q in Lawley's theory as $i = q$. When $i < q$, A_i will contain signal sample eigenvalues. We assume that the covariance between signal and noise sample eigenvalues can be ignored. The covariance between signal sample eigenvalues is given in [34] as:

$$Cov(l_i,l_j)=\frac{2\lambda_i^2\lambda_j^2}{\beta N^2(\lambda_i-\lambda_j)^2},i\neq j,\quad i,j\le q \tag{42}$$

According to the properties of variance and complex circular Gaussian assumption, A_i is also asymptotically Gaussian distributed and the variance can be derived as:

$$Var(A_i) = Var\left(\frac{p-q}{p-i} A_q + \frac{\sum_{j=i+1}^{q} l_j}{p-i} \right) = \frac{(p-q)^2 Var(A_q)}{(p-i)^2} + \sum_{j=i+1}^{q} \frac{Var(l_j)}{(p-i)^2} + 2\sum_{j=i+1}^{q} \sum_{j<k\leq q} \frac{Cov(l_j, l_k)}{(p-i)^2} \tag{43}$$

Note that Equation (43) equals to the variance of A_q when $i = q$, so Equation (43) is the general expression of the variance of A_i for $i \leq q$. Since the distributions of l_i and A_i are asymptotically Gaussian, the distribution of ρ_i defined in Equation (16) would follow a ratio distribution of two Gaussian random variables. The probability density function of the ratio of two correlated Gaussian random variables has been derived by Hinkley in [36]. The correlation coefficient r_i between l_i and A_i can be derived as:

$$r_i = \frac{Cov(l_i, A_i)}{\sqrt{Var(l_i)Var(A_i)}} = \frac{1}{(p-i)\sqrt{Var(l_i)Var(A_i)}} \sum_{j=i+1}^{q} Cov(l_i, l_j), \quad i < q \tag{44}$$

So the probability density function $f(\rho_i)$ of ρ_i can be obtained by using Equation (1) of [36]. Furthermore, the probability of multiple-missed detection P_{md} can be calculated as:

$$P_{md} = F(\rho_{Cd}) = \int_{-\infty}^{\rho_{Cd}} f(\rho_i) d\rho_i \tag{45}$$

where:

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{46}$$

is the cumulative distribution function of the Gaussian ratio distribution.

4. Simulation Setup and Numerical Results

In the numerical simulations, a uniform linear array with an inter-sensor spacing of half-wavelength is employed. The observed signals are assumed to be uncorrelated complex circular Gaussian source signals contaminated by additive complex circular white Gaussian noise and some results may be invalid for arbitrary complex signals. The numbers of samples, sensors, true sources and underestimated sources are denoted by N, p, q and d respectively. The DOAs are denoted by the vector $\boldsymbol{\theta}$ and SNR is short for the signal-to-noise ratio. The probability of an underestimated source number greater than or equal to d is denoted by P_{md} and the expectation of the estimated source number is denoted by $E(\hat{q})$. All the simulation results are obtained based on 10,000 Monte Carlo trials.

4.1. Evaluation of the Proposed Method for Underestimation Analysis

The methods presented in [25–27] are used for comparison which based on the statistical analysis of sample eigenvalues as well. Haddadi *et al.* [25] use the expectations in Lawley's theory and the variance in multivariate statistical theory with the neglect of the variances of noise sample eigenvalues while Huang *et al.* [26] employ the variance in random matrix theory. Lu and Zoubir [27] have incorporated the expectations in Lawley's theory and the variances in random matrix theory to predict the estimation results of MDL precisely. All the four procedures use the same expectations of l_i

and A_q whereas different variances are selected. A comparison among different procedures is shown in Table 2.

Four experimental settings are listed as following by varying the SNR or the sample size N:

Setting 1 (see Figure 1): $N = 50, p = 20, q = 3, \boldsymbol{\theta} = \{-5°, 0°, 5°\}$, SNR = [−10,−4] dB.
Setting 2 (see Figure 2): $N = 1000, p = 20, q = 3, \boldsymbol{\theta} = \{-5°, 0°, 5°\}$, SNR = [−15.5,−12.5] dB.
Setting 3 (see Figure 3): $N = [100,500], p = 30, q = 3, \boldsymbol{\theta} = \{-5°, 5°, 10°\}$, SNR = −12 dB.
Setting 4 (see Figure 4): $N = [1000,4000], p = 30, q = 3, \boldsymbol{\theta} = \{-5°, 5°, 10°\}$, SNR = −17 dB.

Table 2. Theoretical comparison of the four procedures.

	Haddadi *et al.*	**Huang *et al.***	**Lu &Zoubir**	**Ours**
$E(l_q)$	Lawley	Lawley	Lawley	Lawley
$Var(l_q)$	Multivariate Statistics	Random Matrix	Random Matrix	Lawley
$E(A_q)$	Lawley	Lawley	Lawley	Lawley
$Var(A_q)$	- *	-	Random Matrix	Random Matrix

* Means this term has been neglected.

As shown in Figure 1, the proposed method shows the best agreement with simulation results and outperforms the others. It is worth noting that the curves nearly intersect at the same point of $P_m = 0.5$ with different shapes which may correspond to the expectations and variances, respectively. The methods in [25,27] have very similar performance since the former uses a larger variance of signal sample eigenvalue while the latter considered the variance of noise sample eigenvalues as a counteract. An inconspicuous difference is found for the method in [26] due to the consideration of only the interactions between signal and noise subspaces.

The simulation results at a large sample size are presented in Figure 2. All the four methods match the simulation results pretty well and the superiority of the proposed method and the method in [26] can be confirmed in the details of Figure 2. The accurate prediction by the method in [26] may be attributed to the reason that the ignorance of the variance of noise sample eigenvalues may compensate the interactions between signal eigenvalues. Similar results are presented in Figures 3 and 4 by varying the sample size N instead of SNR and the proposed method outperforms the other methods.

The outstanding performance of the proposed method can be attributed to the fact that the interactions between signal eigenvalues have been taken into consideration in estimating the variance of signal sample eigenvalue. The performance of the methods in [25–27] is dependent of the sample size in the simulation settings, since they ignore the interactions between signal eigenvalues which are sensitive to the sample size. In the cases when the sample size is sufficiently large, all the methods are capable to yield satisfactory results. However, MDL is widely used in practical applications where only a limited number of samples is available. Thus, the proposed method is of more practical value for its accuracy in such cases.

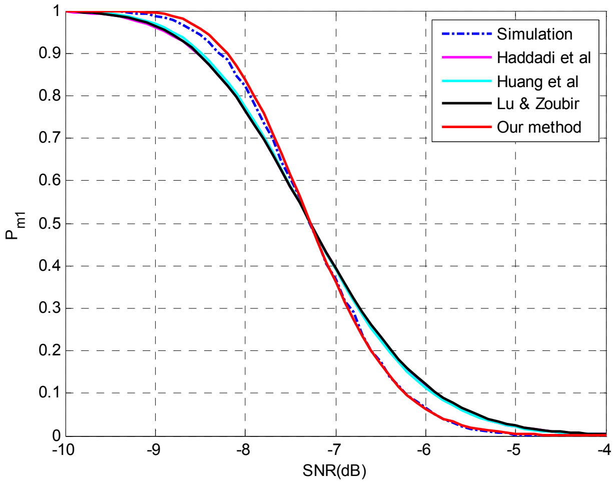

Figure 1. The probability of underestimation P_{m1} *versus* SNR with $N = 50$, $p = 20$, $q = 3$, $\mathbf{\theta} = \{-5°, 0°, 5°\}$.

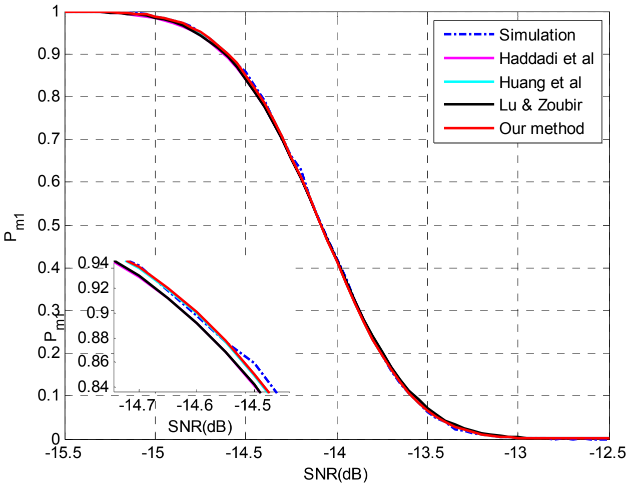

Figure 2. The probability of underestimation P_{m1} *versus* SNR and details with $N = 1000$, $p = 20$, $q = 3$, $\mathbf{\theta} = \{-5°, 0°, 5°\}$.

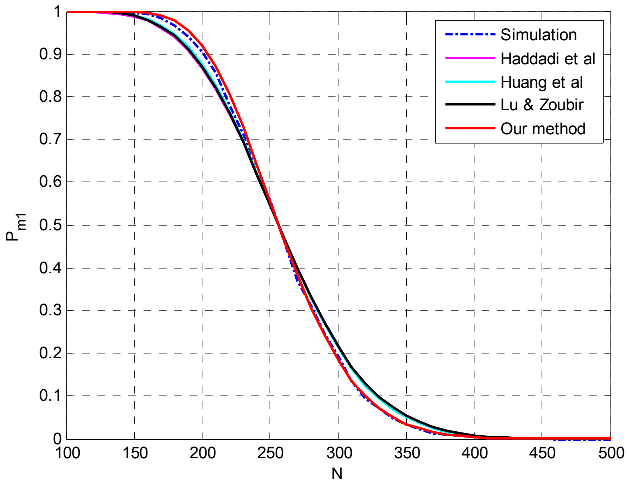

Figure 3. The probability of underestimation P_{m1} *versus* N with $p = 30$, $q = 3$, $\mathbf{\theta} = \{-5°, 5°, 10°\}$, SNR = −12 dB.

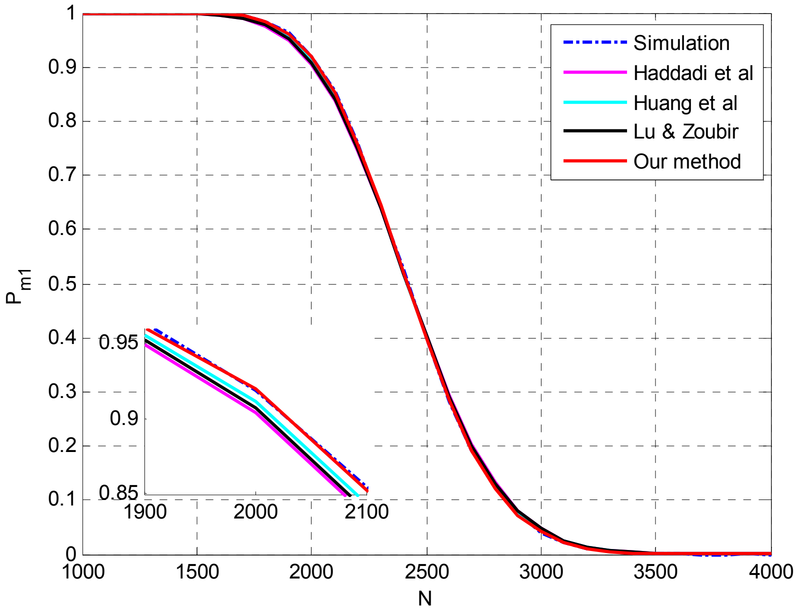

Figure 4. The probability of underestimation P_{m1} *versus* N and details with $p = 30$, $q = 3$, $\boldsymbol{\theta} = \{-5°, 5°, 10°\}$, SNR = −17 dB.

4.2. Evaluation of the Analysis on Multiple-Missed Detection

To evaluate the performance of the proposed method for multiple-missed detection, the algorithms in [25–27] have been adapted by the proposed approach as reference methods. Experimental settings are listed as following:

Setting 5 (see Figure 5): $N = 150$, $p = 30$, $q = 5$, $\boldsymbol{\theta} = \{-7°, -5°, 0°, 3°, 11°\}$, SNR = [−14,−10] dB, $d = 4$.

Setting 6 (see Figure 6): $N = 400$, $p = 30$, $q = 5$, $\boldsymbol{\theta} = \{-7°, -5°, 0°, 3°, 11°\}$, SNR = [−16.5,−13] dB, $d = 4$.

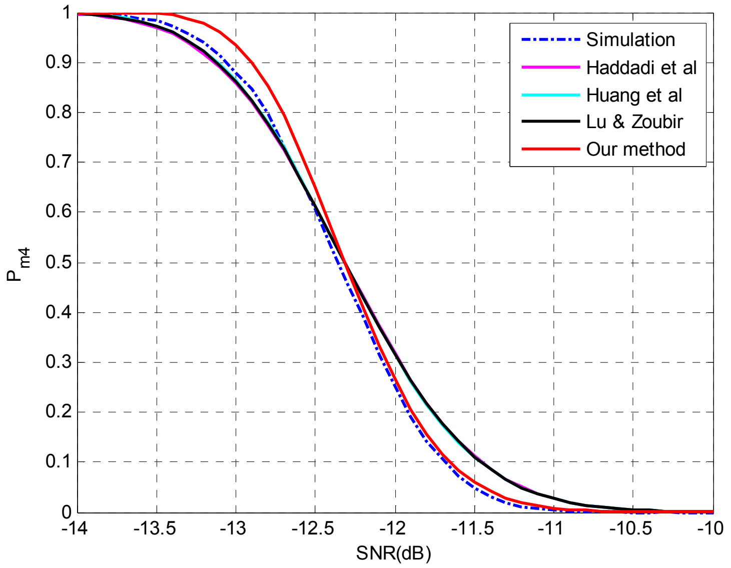

Figure 5. The probability of underestimation P_{m4} *versus* SNR with $N = 150$, $p = 30$, $q = 5$, $\boldsymbol{\theta} = \{-7°, -5°, 0°, 3°, 11°\}$, $d = 4$.

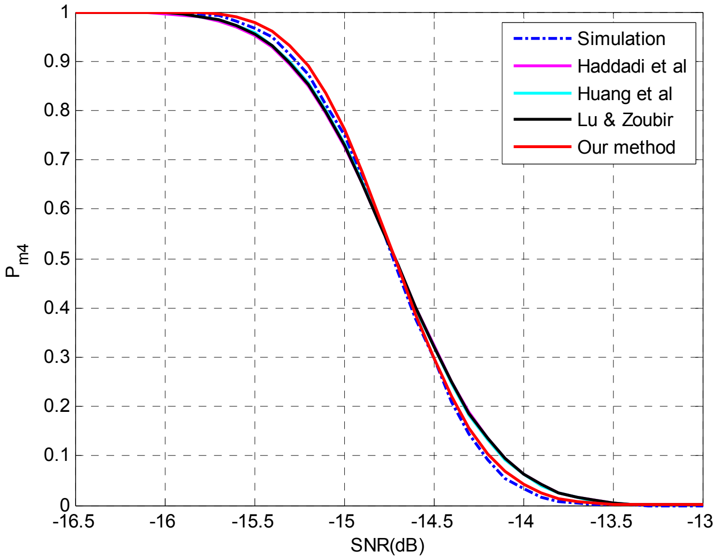

Figure 6. The probability of underestimation P_{m4} *versus* SNR with $N = 400$, $p = 30$, $q = 5$, $\boldsymbol{\theta} = \{-7°, -5°, 0°, 3°, 11°\}$, $d = 4$.

The simulation results are shown in Figures 5 and 6 at different sample sizes. As indicated in Figure 5, the proposed method is more accurate in the region of $P_m < 0.5$ while the other three methods perform better in the region of $P_m > 0.5$. When the sample size is increased to 400, the discrepancy among the methods become smaller and the superiority of the proposed method is reconfirmed in the entire region. We attribute this phenomenon to the omission of the higher-order terms, since a bias can be found between the simulation results and all the predictions in Figure 5, while the bias is perfectly rectified at a slightly larger sample size in Figure 6. The accumulated error by the neglect of higher-orders terms would affect the performance analysis for multiple-missed detection and a moderate sample size may be required for an accurate estimation.

To assess the performance deterioration of enumeration with the degradation of SNR, we use the expectation of the estimated source number as the indicator which is calculated by Equation (36). Experimental settings are listed as follows:

Setting 7 (see Figure 7): $N = 150$, $p = 20$, $q = 5$, $\boldsymbol{\theta} = \{-10°, -6°, 0°, 6°, 8°\}$, SNR = [−15,0] dB.
Setting 8 (see Figure 8): $N = 300$, $p = 20$, $q = 5$, $\boldsymbol{\theta} = \{-11°, -7°, 0°, 2°, 10°\}$, SNR = [−16,−2] dB.

Figures 7 and 8 show that all the four methods are able to perform the analysis for multiple-missed detection precisely and the capability of the proposed approach for multiple-missed detection is verified. The mean absolute error (MAE) between the prediction and simulation results is selected to assess the performance of different methods quantitatively. The MAEs are listed in Table 3 and the superiority of our method can be confirmed.

Table 3. MAE of the methods in predicting the enumeration results.

MAE	Haddadi *et al.*	Huang *et al.*	Lu & Zoubir	Our Method
Setting 7	0.0143	0.0123	0.0140	0.0110
Setting 8	0.0054	0.0047	0.0055	0.0043

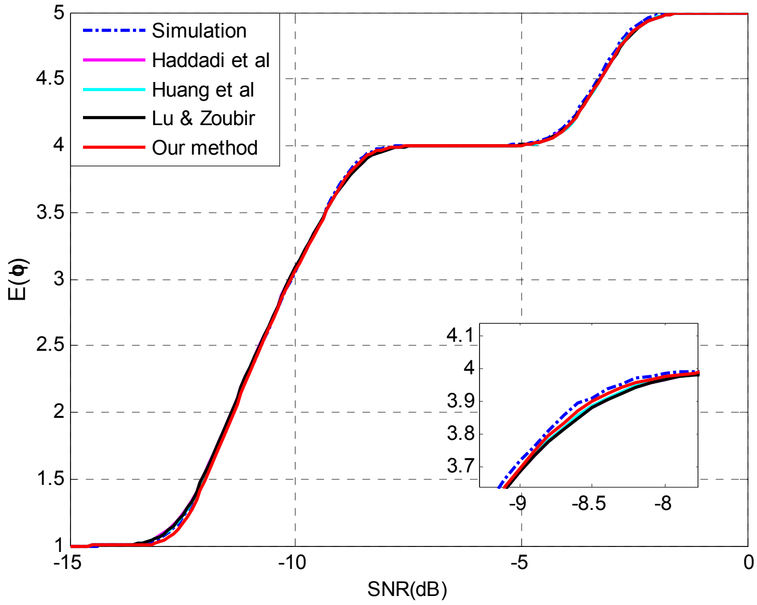

Figure 7. The expectation of the estimated source number $E(\hat{q})$ *versus* SNR and details with $N = 150$, $p = 20$, $q = 5$, $\boldsymbol{\theta} = \{-10°, -6°, 0°, 6°, 8°\}$.

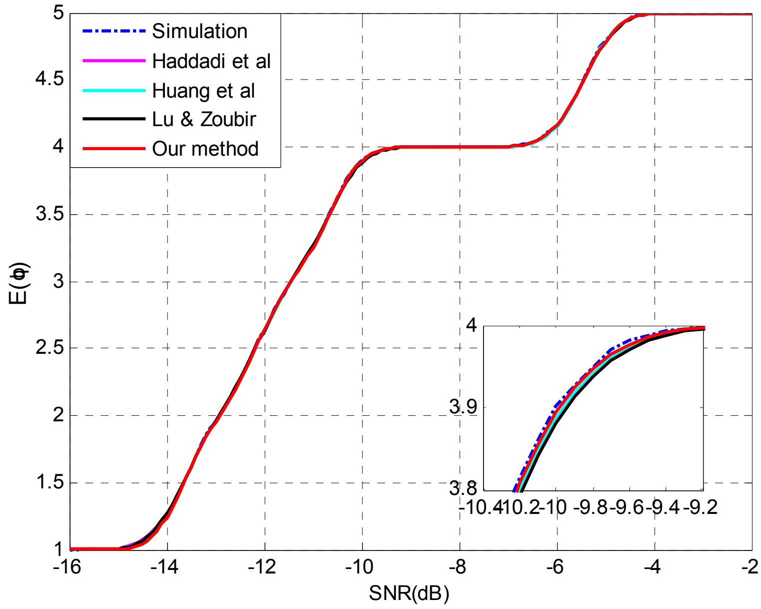

Figure 8. The expectation of the estimated source number $E(\hat{q})$ *versus* SNR and details with $N = 300$, $p = 20$, $q = 5$, $\boldsymbol{\theta} = \{-11°, -7°, 0°, 2°, 10°\}$.

5. Conclusions

This paper presents an accurate performance analysis for the underestimation performance of the MDL source enumeration method at a finite sample size in array processing. Theoretical derivations and statistical analyses have been performed with the consideration of the interactions between signal eigenvalues to obtain an improved estimation of the probability of underestimation. A new approach is also proposed to evaluate the performance of multiple-missed detection cases by ratio distribution analysis and can be employed by the eigenvalue-analyzing methods. Simulation results show the superiority of the proposed analysis, and verify the ability of the proposed approach in evaluating the deterioration of enumeration performance with the degradation of SNR, which may be a valuable reference for practical applications.

Acknowledgments

This paper is supported by the project supported by the National Natural Science Foundation of China—Research on Detection and Localization of Subsea Pipeline's Tiny Leakage (Grant No. 61473205) and State Key Laboratory Cultivation Plan—Research on Underwater Leak Detection Based on Bubble Acoustic.

Author Contributions

Fei Du completed the design and derivation of the algorithm and the numerical simulation setup. Yibo Li provided the guidance and recommendations for this research. Shijiu Jin contributed to the contents and writing of this manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Krim, H.; Viberg, M. Two decades of array signal processing research: The parametric approach. *IEEE Signal Process. Mag*. **1996**, *13*, 67–94.
2. Atashbar, M.; Kahaei, M. Direction-of-arrival estimation using AMLSS method. *IEEE Lat. Am. Trans*. **2012**, *10*, 2053–2058.
3. Naik, G.R.; Kumar, D.K. Dimensional reduction using blind source separation for identifying sources. *Int. J. Innov. Comput. Inf. Control* **2011**, *7*, 989–1000.
4. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471.
5. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
6. Wax, M.; Kailath, T. Detection of signals by information theoretic criteria. *IEEE Trans. Acoust. Speech Signal Process.* **1985**, *33*, 387–392.
7. Dayan, A.G.; Rausley, A.A.S. Simple and efficient algorithm for improving the MDL estimator of the number of sources. *Sensors* **2014**, *14*, 19477–19492.
8. Huang, L.; Long, T.; Mao, E.; So, H.C. MMSE-based MDL method for accurate source number estimation. *IEEE Signal Process. Lett*. **2009**, *16*, 98–801.
9. Huang, L.; So, H.C. Source enumeration via MDL criterion based on linear shrinkage estimation of noise subspace covariance matrix. *IEEE Trans. Signal Process*. **2013**, *61*, 4806–4821.
10. Lu, Z.; Zoubir, A.M. Source enumeration using the pdf of sample eigenvalues via information theoretic criteria. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 3361–3364.

11. Fishler, E.; Messer, H. On the use of order statistics for improved detection of signals by the MDL criterion. *IEEE Trans. Signal Process*. **2000**, *48*, 2242–2247.
12. Zhen, J.; Si, X. A method for determining number of coherent signals with arbitrary plane array. In Proceedings of the IEEE International Conference on Information and Automation, Harbin, China, 20–23 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 240–244.
13. Fishler, E.; Poor, H.V. Estimation of the number of sources in unbalanced arrays via information theoretic criteria. *IEEE Trans. Signal Process*. **2005**, *53*, 3543–3553.
14. Huang, L.; Wu, S.; Li, X. Reduced-rank MDL method for source enumeration in high-resolution array processing. *IEEE Trans. Signal Process*. **2007**, *55*, 5658–5667.
15. Huang, L.; Long, T.; Mao, E.; So, H.C. MMSE-based MDL method for robust estimation of number of sources without eigendecomposition. *IEEE Trans. Signal Process*. **2009**, *57*, 4135–4142.
16. Wang, H.; Kaveh, M. On the performance of signal-subspace processing—Part I: Narrow-band systems. *IEEE Trans. Acoust. Speech Signal Process*. **1986**, *34*, 1201–1209.
17. Kaveh, M.; Wang, H.; Hung, H. On the theoretical performance of a class of estimators of the number of narrow-band sources. *IEEE Trans. Acoust. Speech Signal Process*. **1987**, *35*, 1350–1352.
18. Zhang, Q.T.; Wong, K.M.; Yip, P.C.; Reilly, J.P. 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*. **1989**, *37*, 1557–1567.
19. Xu, W.; Kaveh, M. Analysis of the performance and sensitivity of eigendecomposition-based detectors. *IEEE Trans. Signal Process*. **1995**, *43*, 1413–1426.
20. Liavas, A.P.; Regalia, P.A. On the behavior of information theoretic criteria for model order selection. *IEEE Trans. Signal Process*. **2001**, *49*, 1689–1695.
21. Fishler, E.; Grosmann, M.; Messer, H. Detection of signals by information theoretic criteria: General asymptotic performance analysis. *IEEE Trans. Signal Process*. **2002**, *50*, 1027–1036.
22. Delmas, J.P.; Meurisse, Y. Performance analysis of MDL criterion for the detection of noncircular or/and non-Gaussian components. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 4188–4191.
23. Delmas, J.P.; Meurisse, Y. On the second-order statistics of the EVD of sample covariance matrices—Application to the detection of noncircular or/and nonGaussian components. *IEEE Trans. Signal Process*. **2011**, *59*, 4017–4023.
24. Nadler, B. Nonparametric detection of signals by information theoretic criteria: Performance analysis and an improved estimator. *IEEE Trans. Signal Process*. **2010**, *58*, 2746–2756.
25. Haddadi, F.; Malek-Mohammadi, M.; Nayebi, M.M.; Aref, M.R. Statistical performance analysis of MDL source enumeration in array processing. *IEEE Trans. Signal Process*. **2010**, *58*, 452–457.
26. Huang, L.; Xiao, Y.; Liu, K.; So, H.C.; Zhang, J. Bayesian information criterion for source enumeration in large-scale adaptive antenna array. *IEEE Trans. Veh. Technol.* **2015**, doi:10.1109/TVT.2015.2436060.
27. Lu, Z.; Zoubir, A.M. Flexible detection criterion for source enumeration in array processing. *IEEE Trans. Signal Process*. **2013**, *61*, 1303–1314.
28. Muirhead, R.J. *Aspects of Multivariate Statistical Theory*; Wiley: New York, NY, USA, 1982.

29. Bai, Z.D.; Silverstein, J.W. *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed.; Springer-Verlag: New York, NY, USA, 2010.
30. Yi, H. Local Signal Search Enhanced RMT Estimator for Estimating the Number of Signals based on Random Matrix Theory. Available online: http://arxiv.org/abs/1405.4713 (accessed on 14 October 2014).
31. Kritchman, S.; Nadler, B. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Signal Process*. **2009**, *57*, 3930–3941.
32. Nadakuditi, R.R.; Edelman, A. Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Trans. Signal Process*. **2008**, *56*, 2625–2638.
33. Nadakuditi, R.R.; Edelman, A. Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. *IEEE J. Sel. Top. Sign. Process.* **2010**, *4*, 468–480.
34. Lawley, D.N. Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika* **1956**, *43*, 128–136.
35. Ding, Q.; Kay, S. Inconsistency of the MDL: On the performance of model order selection criteria with increasing signal-to-noise ratio. *IEEE Trans. Signal Process*. **2011**, *59*, 1959–1969.
36. Hinkley, D.V. On the ratio of two correlated normal random variables. *Biometrika* **1969**, *56*, 635–639.