Model Selection for Sinusoids in Noise: Statistical Analysis and a New Penalty Term

Boaz Nadler and Leonid (Aryeh) Kontorovich

***Abstract*—Detection of the number of sinusoids embedded in noise is a fundamental problem in statistical signal processing. Most parametric methods minimize the sum of a data fit (likelihood) term and a complexity penalty term. The latter is often derived via information theoretic criteria, such as minimum description length (MDL), or via Bayesian approaches including Bayesian information criterion (BIC) or maximum a-posteriori (MAP). While the resulting estimators are asymptotically consistent, empirically their finite sample performance is strongly dependent on the specific penalty term chosen. In this paper we elucidate the source of this behavior, by relating the detection performance to the extreme value distribution of the maximum of the periodogram and of related random fields. Based on this relation, we propose a combined detection-estimation algorithm with a new penalty term. Our proposed penalty term is sharp in the sense that the resulting estimator achieves a nearly constant false alarm rate. A series of simulations support our theoretical analysis and show the superior detection performance of the suggested estimator.**

***Index Terms*—sinusoids in noise, maxima of random fields, extreme value theory, periodogram, statistical hypothesis tests.**

I. INTRODUCTION

DETECTION of the number of sinusoids, along with their frequencies and amplitudes is a fundamental problem in signal processing and time series analysis. Classical treatments of this problem, suggesting (sequential) hypothesis tests and typically restricted to the Fourier frequencies date back to R.A. Fisher [15], see also [8], [29] and references therein. Since the introduction of Akaike information criteria (AIC), Schwarz' BIC criteria and the minimum description length (MDL) principle for model selection, many different detection algorithms based on these principles have been proposed, see [13], [16], [20], [22], [25], [37] and references therein.

In this paper we focus on parametric joint detection-estimation methods, which determine the number of sinusoids by minimizing

$$\hat{k} = \arg\min_k -\ln \mathcal{L}(\hat{\boldsymbol{\theta}}_k, \mathbf{x}) + k\, C_n$$

where $\mathbf{x}$ is the observed time series of length n, $\hat{\boldsymbol{\theta}}_k$ are parameter estimates in a model of order k, $\mathcal{L}(\hat{\boldsymbol{\theta}}_k, \mathbf{x})$ is the corresponding likelihood term and C_n is a model-complexity penalty term, typically of the form $C \ln n$ for some constant $C > 0$. Whereas the asymptotic consistency of such estimators has been thoroughly studied, in simulations, their finite sample performance has been observed to strongly depend on the specific penalty term. On the theoretical front, this finite sample detection performance, and in particular the probability of over-fitting when using particular penalty terms, remain unclear.

In this paper we present a detailed statistical study of the problem of sinusoid detection in white noise. We show that the detection performance is closely related to the extreme value distribution of the maximum of the periodogram and of related random fields. Based on this relation, we propose a combined detection-estimation algorithm with a new penalty term, that contains not only the familiar $\ln n$ term but also a $\ln\ln n$ term and an additive constant. We explain theoretically why these additional terms are *crucial* both to prevent over-fitting and to obtain sensible parameter estimates. We support our analysis via a series of simulations, which show the superior detection performance of the suggested estimator.

The paper is organized as follows. The problem formulation, along with a review of previous work appears in Section II. The motivation for our approach, and the resulting new penalty term is described in Section III, whereas its performance is analyzed in Section IV. Section V presents simulations supporting the theoretical analysis. We conclude with a discussion in Section VI.

II. DETECTION OF NUMBER OF SINUSOIDS IN NOISE AND INFORMATION THEORETIC CRITERIA

A. Problem Formulation

The problem of detecting the number of sinusoids embedded in noise is formulated as follows: Let $x(t)$ denote a one-dimensional real valued signal composed of an unknown number of sinusoids K, at unknown frequencies ω_j, amplitudes a_j, and phase shifts ϕ_j, and corrupted by additive noise

$$x(t) = \sum_{j=1}^{K} a_j \sin(\omega_j t + \phi_j) + \sigma\xi(t). \tag{1}$$

The signal is sampled at discrete times $\{t_j = j\}_{j=1}^{n}$.

Given the n observations $\mathbf{x} = \{x_t\}_{t=1}^{n}$, the problem is to estimate the unknown number of sinusoids K in Eq. (1). For simplicity, we perform a detailed analysis on the simple case whereby $\xi(t)$ is white and Gaussian, and moreover that the noise level σ is a-priori known. Later on we relax these assumptions, and show that our resulting model selection method performs well under a much broader class of non-Gaussian white noise with unknown strength or distribution.

B. Nadler is a faculty member at the Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. L. Kontorovich is a faculty member at the Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel. e-mails: boaz.nadler@weizmann.ac.il, karyeh@cs.bgu.ac.il.

In this paper we focus on *parametric* joint detection-estimation methods, which also require estimates of the unknown frequencies, amplitudes, and phase shifts. While in this paper we consider only real valued 1-D signals, complex exponentials embedded in complex valued noise, or 2-D sinusoids in noise [23] can be treated in a similar way.

Under the assumption of white Gaussian noise, given an estimate k for the number of sinusoids present, a common method to estimate the corresponding $3k$ parameters is by maximizing the likelihood $\mathcal{L}$ of the observed data,

$$\mathcal{L}(\boldsymbol{\theta}_k, \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{t=1}^{n} e^{-\left(x_t - \sum_{j=1}^{k} \ a_j \sin(\omega_j t + \phi_j)\right)^2/2\sigma^2} \tag{2}$$

where the vector $\boldsymbol{\theta}_k = \{a_j, \omega_j, \phi_j\}_{j=1}^k$ contains the $3k$ parameters of the k sinusoids. Given the number of sinusoids present, various clever and efficient methods for parameter estimation have been suggested in the literature [7], [27], [30], [33]. Note, however, that methods not directly maximizing the likelihood function may yield substantially less accurate parameter estimates. Since the maximum likelihood (ML) is a strictly increasing function of model complexity, unregularized ML cannot provide a consistent estimate for the true model order. Hence, some sort of penalty term is needed in order to estimate the unknown true number of terms in Eq. (1).

B. Previous and Related Work

Many different methods have been suggested in the literature for estimating the number of sinusoids. Some of the classical approaches considered the detection problem of a *single* sinusoid with an unknown frequency embedded in noise, and analyzed the maximum of the periodogram at the n Fourier frequencies [15], [8]. However, as discussed in [8], when the periodic components are not at Fourier frequencies, the detection performance of these tests is substantially reduced. A second family of estimators, which does not require estimation of the unknown frequencies and amplitudes, was proposed in [16], [32], as well as in the more recent works [6], [9], [28]. By exploiting the fact that (noiseless) data arising from a sum of trigonometric terms satisfy recursion relations with respect to a time shift, these works proposed computationally efficient (non-parametric) eigenvalue based estimators. Since these methods are non-parametric they may require a higher SNR to reliably detect the number of sinusoids present, as compared to parametric methods (see also Sec. V).

A third family of estimators, most relevant to our work, formulated the problem of estimating the number of sinusoids as a *model selection* problem. This naturally lead to estimators based on information theoretic criteria. For example, a straightforward application of the BIC/MDL principle yields

$$\hat{k}_{MDL} = \arg\min_k - \ln \mathcal{L}(\hat{\theta}_k, \mathbf{x}) + \frac{3k}{2} \ln n \tag{3}$$

where $\hat{\theta}_k$ denotes the vector of ML estimates of all parameters in a model of order k, and the penalty factor of $3k$ is the number of unknown parameters in such a model, not counting the (possibly unknown) noise variance σ^2, a parameter common to all models.

In [18], [22] Hannan and Kavalieris used MDL techniques to derive a model order selection method under a very general framework of sinusoids corrupted by colored and possibly non-Gaussian noise. In the case of white noise, their approach gives rise to the following penalty term

$$\hat{k}_{MAP} = \arg\min_k - \ln \mathcal{L}(\hat{\theta}_k, \mathbf{x}) + \frac{5k}{2} \ln n. \tag{4}$$

The same penalty term was also independently derived by Djuric [13]. By considering a Bayesian framework, Djuric showed that not all parameters in a model with k sinusoids should receive the same penalty (see also [30], section 3.6). Eq. (4) follows by selecting the model order with maximum a-posteriori probability (MAP), assuming a Bayesian prior.

Both (3) and (4) are specific examples of *efficient detection criteria* (EDC) type estimators, with a general form

$$\hat{k}_{EDC} = \arg\min_k - \ln \mathcal{L}(\hat{\theta}_k, \mathbf{x}) + kC_n \tag{5}$$

where C_n captures the dependency of the penalty on the number of samples n.

All estimators of the form (5) can also be interpreted as performing a generalized likelihood ratio test (GLRT), since a necessary condition to detect at least k sinusoids is that

$$\ln \left(\frac{\mathcal{L}(\text{model order } k)}{\mathcal{L}(\text{model order } k-1)} \right) > C_n.$$

The key question for model selection purposes is thus the following: what should be the penalty term C_n? One potentially desirable property of the resulting estimator is *asymptotic weak consistency*, that is, convergence in probability to the correct model order

$$\Pr[\hat{k} = K] \to 1 \quad \text{as} \quad n \to \infty.$$

From the work of An *et. al*. [4] on the asymptotic maximum of the periodogram (see also [29], [35], [36]), sufficient conditions on the penalty C_n to yield a consistent estimator are

$$\lim_{n\to\infty} \frac{C_n}{n} = 0, \qquad \lim_{n\to\infty} \frac{C_n}{\ln n} > 1. \tag{6}$$

A second desirable property is accurate detection performance for finite values of n. In particular, it is desirable to have control of the false alarm (over-detection rate),

$$\Pr[\hat{k} > K] \leq \alpha$$

say, for all $n > n_0$ and some $\alpha \ll 1$.

While both the MDL and MAP estimators satisfy the conditions (6) and hence are asymptotically consistent, from a theoretical perspective their finite sample over-detection probability is unclear. Moreover, as noted in the simulation study by Kundu [24], for finite observation lengths n, estimators with different penalty functions C_n have quite different detection performances and over-estimation probabilities. Based on simulations, both [24] and Hwang and Chen [20] proposed $C_n = 2\ln n$, though, as noted by Kundu, 'no theoretical justification can be given in favor of this'.

A similar behavior, namely the strong dependency of an estimator's performance on the specific form of C_n, was also observed by Djuric [13], who showed via simulations that the

MAP estimator (4) with a larger penalty term achieves better detection performance than the standard MDL estimator (3), for samples of length $n = 64$ or $n = 128$. The empirical findings of [13] are rather surprising, given that according to [29] and Eq. (6) above, any penalty function $C_n = c \ln n$ with $c > 1$ yields an asymptotically consistent estimator. Hence, one would expect that the MDL estimator (3) with a coefficient $c = 3/2$ which is closer to the minimal value of $c = 1$, would have better detection performance than the MAP estimator with a larger penalty coefficient $c = 5/2$. Our analysis thus highlights the fact that minimal requirements for asymptotic consistency of an estimator may be misleading for finite sample sizes. Similar behavior is exhibited, for example, in the case of model selection for multivariate autoregression.

III. Main Results

In this paper we present a detailed statistical analysis of this problem, namely detection of the number of sinusoids embedded in white noise. Our analysis provides a theoretical explanation for various empirical findings of previous simulation studies. In particular, it explains why the MAP estimator enjoys a better detection performance than the MDL estimator for short records lengths (say $n < 1000$). More importantly, our analysis suggests a novel penalty term, which contains not only the familiar $\ln n$ term, but also a $\ln(\ln n)$ term and an additive constant term that depends on a user chosen false alarm rate. We show, both theoretically and empirically via simulations, the importance of these two additional penalty terms. These are verified in Section V via simulations, which show that the resulting estimator gives state-of-the-art detection performance for a wide range of values of n.

Throughout the paper we use the following notation: For two vectors $u, v \in \mathbb{R}^n$ we denote the standard inner product by $\langle u, v\rangle = \sum_j u_j v_j$, and the induced L_2 norm by $\|u\|^2 = \langle u, u\rangle$. Let ω_j, ϕ_j denote the frequencies and phase shifts of the K sinusoids. We denote by $\mathbf{sin}_{\omega,\phi} = (\sin(\omega + \phi), \sin(2\omega + \phi), \ldots, \sin(n\omega + \phi))$ the vector of length n containing the values of a sinusoid with frequency ω and phase shift ϕ, sampled at $t = 1, \ldots, n$. We denote by $V_K \subset \mathbb{R}^n$ the subspace spanned by the K vectors $\{\mathbf{sin}_{\omega_j,\phi_j}\}_{j=1}^K$. Finally, we denote by $P_K : \mathbb{R}^n \to \mathbb{R}^n$ the projection operator onto V_K and by $P_K^\perp$ its orthogonal complementary projection ($P_K \oplus P_K^\perp = I$).

A. GLRT and the maxima of stochastic fields

The key observation underlying our analysis is that detection of the number of sinusoids can be viewed as a *singular* hypothesis testing problem. Namely, fitting observed data sampled from a model of unknown order K with a model of order $K+1$ or higher, involves estimating *non-existent* (nuisance) parameters, the parameters of the $K+1$ sinusoid in our case. Hence, the standard statistical theory regarding the χ^2 distribution of the likelihood ratio test does not hold. Instead, we obtain the following asymptotic result:

Theorem 1: *Let $\mathbf{x} = \{x_t\}_{t=1}^n$ be n noisy observations from Eq. (1) with K sinusoids and with $\sigma = 1$. Denote by $\hat\theta_K$ and by $\hat\theta_{K+1}$ the ML estimates of the relevant parameters, assuming K or $K+1$ sinusoids, respectively. Denote the GLRT by*

$$G_K = \ln\left(\frac{\mathcal{L}(\hat\theta_{K+1}, \mathbf{x})}{\mathcal{L}(\hat\theta_K, \mathbf{x})}\right).$$

Then this random variable is asymptotically distributed as the maximum of a χ^2 random field

$$G_K \sim \sup_{\omega,\phi\in(0,2\pi)} \frac{1}{2}\eta_K(\omega,\phi)^2 \tag{7}$$

where

$$\eta_K(\omega,\phi) = \frac{\langle \xi, P_K^\perp \mathbf{sin}_{\omega,\phi}\rangle}{\|P_K^\perp \mathbf{sin}_{\omega,\phi}\|}, \tag{8}$$

and ξ is an n-dimensional Gaussian noise vector with zero mean and identity covariance.

The proof of the theorem appears in the appendix. Here we provide some intuition behind Eqs. (7) and (8). As $n \to \infty$, when fitting observed noisy data with the correct model order, from asymptotic consistency, the maximum likelihood estimates of the frequencies and phase shifts $\{\hat\omega_j, \hat\phi_j\}$ converge to the true ones $\{\omega_j, \phi_j\}$. The log likelihood is then proportional to the component of the noise that is *orthogonal* to the actual sinusoids, namely $P_K^\perp \xi$, since the component of the noise in the span of the true sinusoids is indistinguishable from the signal and is fitted by the ML estimates of the K sinusoidal amplitudes $\{\hat a_j\}$. When fitting observed data of order K with a model of order $K+1$, it can be proven that as $n \to \infty$, K frequencies converge to the true ones, and the remaining sinusoid is chosen to fit the remaining noise as best as possible. Eq. (8) measures the fit of an additional sinusoid having frequency ω and phase shift ϕ to the remaining noise. The sinusoid chosen is the one that maximizes this (over-)fit to noise, hence, the maximum in Eq. (7).

Example: We consider in detail the case $K = 0$, which corresponds to pure noise with no sinusoids present. Given the n observations $\{x_t\}_{t=1}^n$, the model selection algorithm first decides between the following two hypotheses,

$$\mathcal{H}_0 : \text{ no sinusoids present} \quad vs. \quad \mathcal{H}_1 : \text{one sinusoid present}$$

We now compute the two likelihoods. Under $\mathcal{H}_0$, we have

$$\ln \mathcal{L}_0 = -\tfrac{1}{2}\sum_t x_t^2$$

whereas under $\mathcal{H}_1$,

$$\ln \mathcal{L}_1 = -\tfrac{1}{2}\sum_t (x_t - \hat a \sin(\hat\omega t + \hat\phi))^2. \tag{9}$$

Since $\hat a$ is the ML estimate, it is given by

$$\hat a = \frac{\sum_t x_t \sin(\hat\omega t + \hat\phi)}{\sum_t \sin(\hat\omega t + \hat\phi)^2} \tag{10}$$

Substituting (10) into (9) the GLRT statistic is

$$G_0 = \ln\left(\frac{\mathcal{L}_1}{\mathcal{L}_0}\right) = \frac{1}{2} S(\hat\omega, \hat\phi) \tag{11}$$

where $S(\omega,\phi)$ is the random field corresponding to $\mathbf{sin}_{\omega,\phi}$,

$$S(\omega,\phi)=\frac{\left[\sum_{t=1}^{n} x_t \sin(\omega t+\phi)\right]^2}{\sum_{t=1}^{n}\sin(\omega t+\phi)^2}. \tag{12}$$

As $\hat{\omega},\hat{\phi}$ are the ML estimates of frequency and phase, respectively, it follows that

$$G_0=\frac{1}{2}\sup_{\substack{\omega\in(0,\pi)\\ \phi\in[0,2\pi]}} S(\omega,\phi). \tag{13}$$

Note that under $\mathcal{H}_0$, for any fixed values of ω,ϕ, the random variable $S(\omega,\phi)$ is a periodogram-like random field that follows a χ^2 distribution with one degree of freedom. In the case of no sinusoids ($K=0$), the GLRT is distributed as the *maximum* of this periodogram-like field. The χ^2 random field of Eq. (7) is its generalization to higher order models.

Remark: The GLRT approach considered in this paper yields maximization over a 2-dimensional random field, Eq. (12). Our approach is thus somewhat different from [12], [21] who considered certain one dimensional random fields, as well as from [35], who analyzed the classical periodogram

$$S_0(\omega)=\frac{2}{n}\sum_{j=1}^{n}|x_t e^{-i\omega t}|^2. \tag{14}$$

B. A new model selection criterion

Equation (13), or more generally Eq. (7) of Theorem 1 are the key for our analysis of existing model selection procedures as well as for the development of a new penalty term. In particular, Theorem 1 underscores the intimate connection between model selection and the distribution of the maxima of random fields. In general, the maximum of random processes and fields is a well studied challenging mathematical problem with a rich history, see for example [2], [3] and references therein. For our purposes, we shall use the following result (see Theorem 2 and Lemma 1 below): For large n, as $x\to\infty$

$$\Pr\left[\sup_{(\omega,\phi)\in(0,\pi)\times[0,2\pi]} S(\omega,\phi)>\ x\right]\le n\sqrt{\frac{\pi x}{6}}e^{-x/2}(1+o(1)) \tag{15}$$

We propose to use Eq. (15) as our point of departure for deriving a new model selection criterion. Let $\alpha\ll 1$ be a user specified false detection probability. In light of Eqs. (13) and (15), our goal is to find a threshold $x=x(n,\alpha)$ for which

$$n\sqrt{\frac{\pi x}{6}}e^{-x/2}=\alpha.$$

Taking logarithms in the above equation the corresponding threshold $x(n,\alpha)$ is the solution of

$$-\frac{x}{2}+\ln n+\ \frac{1}{2}\ln\frac{\pi x}{6}+O\left(\frac{1}{n\sqrt{x}}\right)=\ln\alpha \tag{16}$$

As is common in extreme value theory, we look for an asymptotic solution of Eq. (16) in the following form

$$x=A\ln n\left[1+B\frac{\ln\ln n}{\ln n}+\frac{C}{\ln n}(1+o(1))\right]. \tag{17}$$

Inserting (17) into Eq. (16) and equating terms of equal order gives $A=2, B=\frac{1}{2}$ and $C=-\frac{1}{2}\ln(3\alpha^2/\pi)$. To conclude,

$$\Pr\left[\sup_{\omega,\phi} S(\omega,\phi)>2\ln n+\ln\ln n-\ln(3\alpha^2/\pi)\right]\lesssim\alpha. \tag{18}$$

Note that even though our random field is slightly different, this result is identical to Turkman and Walker ([35], Theorem 3.1), for the maximum of the classical periodogram.

Eq. (18) and its relation to the GLRT via Eq. (13) naturally suggest the following penalty term

$$C_n=C_n(\alpha)=\ln n+\frac{1}{2}\ln\ln n-\frac{1}{2}\ln\left(\frac{3\alpha^2}{\pi}\right) \tag{19}$$

at least for testing $K=0$ vs. $K=1$. As discussed below, this penalty term is suitable for general model orders K, as long as $K\ll n$. Thus, our suggested novel estimator for the number of sinusoids is

$$\hat{k}_{EVT}=\arg\min_k -\ln\mathcal{L}(\hat{\theta}_k,\mathbf{x})+kC_n(\alpha) \tag{20}$$

where $\alpha\ll 1$ is a confidence level chosen by the user. Since this estimator is inspired by ideas from extreme value theory and the maxima of stochastic fields, we denote it by $\hat{k}_{EVT}$. A different interpretation of the estimator (20) is as performing a nested sequence of hypothesis tests, each time testing the statistical significance of fitting an additional sinusoid to the given noisy signal. In our approach, in fact, we indeed stop the first time the GLRT is smaller than the threshold (that is, we do not seek the optimal penalized likelihood over all model orders, which saves considerable un-necessary computations).

For a fixed $\alpha\ll 1$, the proposed estimator in Eq. (20), is technically speaking not asymptotically weakly consistent, as it has a positive false alarm probability $\alpha>0$ of detecting a single sinusoid when none is present ($K=0$). In order to obtain a consistent estimator it suffices to consider a monotonically decreasing sequence α_n of false alarm probabilities, with $\alpha_n\to 0$. For a similar approach in a different detection problem, see [31]. This analysis shows that the conditions of Eq. (6), which are sufficient for consistency, are in fact not necessary. In particular, choosing $\alpha_n=1/\ln n$ gives a weakly consistent estimator whose penalty term C_n satisfies

$$\lim_{n\to\infty}\frac{C_n}{\ln n}=1.$$

We note that a penalty term of the form (19) is implicit in [30] chapter 3.6, based on the asymptotics of the maximum of the periodogram as $n\to\infty$. As described below, this penalty term is in fact applicable also for finite n and $\alpha\ll 1$.

C. The Distribution of the GLRT for $K>0$

The penalty term in Eq. (19) follows from analyzing the model selection problem for the case $K=0$ vs. $K=1$. In this section we present the theoretical justification for using the same penalty term for higher order models, e.g. for comparing a model of order K with a model of order $K+1$.

First, recall that according to Eq. (7), we have that for sufficiently large n, (depending on the SNR of the sinusoids)

$$\begin{aligned}\Pr[G_K > x] &\approx \Pr[\tfrac{1}{2}\sup_{\omega,\phi}\eta_K^2(\omega,\phi) > x] \\ &\leq \Pr[\sup \eta_K(\omega,\phi) > \sqrt{2x}] \\ &\quad + \Pr[\inf \eta_K(\omega,\phi) < -\sqrt{2x}] \\ &\leq 2\Pr\left[\sup \eta_K(\omega,\phi) > \sqrt{2x}\right] \end{aligned} \quad (21)$$

Further, recall that by definition (8), η_K is a zero mean and unit variance Gaussian random field. Hence, model order overestimation is related to the distribution of the maxima of Gaussian random fields. The key theoretical result we shall use is the following general asymptotic result on the maxima of Gaussian random fields, see [3]:

Theorem 2: *Let $\eta(z)$ be a zero mean and unit variance Gaussian random field defined for $z = (z_1, z_2)$ inside a 2-D rectangle $T \subset \mathbb{R}^2$. Let $\rho(z,z') = \mathbb{E}[\eta(z)\eta(z')]$ be the covariance function of η. Then asymptotically in u*

$$\Pr[\sup_{z\in T}\eta(z) > u] \simeq \int_T |\det\Lambda(z)|^{1/2}dz \times \frac{ue^{-u^2/2}}{(2\pi)^{3/2}} \quad (22)$$

where $\Lambda(z)$ is given by

$$\Lambda(z) = \left.\begin{pmatrix} \frac{\partial^2\rho(z,z')}{\partial z_1^2} & \frac{\partial^2\rho(z,z')}{\partial z_1\partial z_2} \\ \frac{\partial^2\rho(z,z')}{\partial z_1\partial z_2} & \frac{\partial^2\rho(z,z')}{\partial z_2^2}\end{pmatrix}\right|_{z'=z} \quad (23)$$

We remark that the theorem above is a specific case of a more general theorem concerning the maximum of d-dimensional Gaussian processes defined over general domains $T \subset \mathbb{R}^d$.

In our case, $z = (\omega,\phi) \in (0,\pi)\times[0,2\pi] = T$, where due to periodicity of the random field in the phase variable ϕ, the domain T is a cylinder and not a rectangle. The theorem above continues to hold in this case as well. Next, according to Eq. (8), the Gaussian random field corresponding to our particular problem has the following specific form

$$\eta(z) = \frac{\langle f(z),\xi\rangle}{\|f(z)\|}$$

where $f(z) \in \mathbb{R}^n$ and ξ is an n-dimensional vector whose entries are i.i.d. $N(0,1)$ random variables. Therefore,

$$\rho(z,z') = \mathbb{E}[\eta(z)\eta(z')] = \frac{\langle f(z), f(z')\rangle}{\|f(z)\|\cdot\|f(z')\|}.$$

Thus, to analyze the distribution of the maximum of this random field, we study the behavior of $\det\Lambda(z)$. To this end, we denote $f_1 = \partial f(z)/\partial z_1$, $f_{11} = \partial^2 f(z)/\partial z_1^2$, and similar notations for derivatives w.r.t. the second variable z_2. Then

$$\frac{\partial\rho(z,z')}{\partial z_1} = \frac{\langle f_1(z), f(z')\rangle}{\|f(z)\|\cdot\|f(z')\|} - \frac{\langle f(z), f(z')\rangle}{\|f(z')\|}\cdot\frac{\langle f_1(z), f(z)\rangle}{\|f(z)\|^3}$$

whereas the expression for the second order derivative is

$$\begin{aligned}\frac{\partial^2\rho}{\partial z_1^2} &= \frac{\langle f_{11}(z), f(z')\rangle}{\|f(z')\|\cdot\|f(z)\|} - \frac{\langle f_1(z), f(z')\rangle}{\|f(z')\|}\frac{\langle f_1(z), f(z)\rangle}{\|f(z)\|^3} \\ &\quad -\frac{\langle f(z), f(z')\rangle}{\|f(z')\|\cdot\|f(z)\|^3}\left[\langle f_2(z), f(z)\rangle + \|f_1(z)\|^2\right] \\ &\quad +3\frac{\langle f(z), f(z')\rangle}{\|f(z')\|}\frac{\langle f_1, f\rangle^2}{\|f(z)\|^5}. \end{aligned} \quad (24)$$

At $z' = z$ we obtain

$$\left.\frac{\partial^2\rho(z,z')}{\partial z_1^2}\right|_{z'=z} = -\frac{\|f_1\|^2}{\|f\|^2} + \frac{\langle f_1, f\rangle^2}{\|f(z)\|^4}. \quad (25)$$

Similarly,

$$\left.\frac{\partial^2\rho(z,z')}{\partial z_1\partial z_2}\right|_{z'=z} = -\frac{\langle f_1, f_2\rangle}{\|f\|^2} + \frac{\langle f_1, f\rangle\langle f_2, f\rangle}{\|f\|^4}. \quad (26)$$

For the case $K = 0$, we have the following:

Lemma 1: *Let η_0 be the random field corresponding to the case of no sinusoids. Then, for all ω far away from $0,\pi$ (more precisely, $\min(\omega,\pi-\omega) \gg 1/n$),*

$$\det\Lambda(\omega,\phi) = \frac{1}{12}n^2(1+o(1)).$$

Proof: By definition, $f(\omega,\phi) = (\sin(\omega+\phi), \sin(2\omega+\phi),\ldots,\sin(n\omega+\phi))$. Hence, for ω bounded away from 0 or π, we have that

$$\begin{aligned}\|f\|^2 &= \sum_{t=1}^n f_t^2 = \frac{1}{n}\sum_{t=1}^n \sin(\omega t+\phi)^2 = \frac{n}{2}(1+o(1)) \\ \left\|\frac{\partial f}{\partial\omega}\right\|^2 &= \sum_{t=1}^n t^2[\cos(\omega t+\phi)]^2 = \frac{1}{6}n^3(1+o(1)) \\ \langle\frac{\partial f}{\partial\omega},\frac{\partial f}{\partial\phi}\rangle &= \sum_t t\cos(\omega t+\phi)^2 = \frac{1}{4}n^2(1+o(1))\end{aligned}$$

In addition, $\langle\partial f/\partial\omega, f\rangle$ and $\langle\partial f/\partial\phi, f\rangle$ are both $O(1)$. Hence, inserting these expressions into Eqs. (25) and (26) gives

$$\frac{\partial^2\rho}{\partial\omega^2} = -\frac{1}{3}n^2(1+o(1)), \quad \frac{\partial^2\rho}{\partial\phi^2} = -1(1+o(1)).$$

and

$$\frac{\partial^2\rho}{\partial\omega\partial\phi} = \frac{1}{2}n(1+o(1)).$$

To conclude, for $K = 0$ and most values of $(\omega,\phi) \in T$,

$$\det\Lambda(\omega,\phi) = \frac{\partial^2\rho}{\partial\omega^2}\frac{\partial^2\rho}{\partial\phi^2} - \left(\frac{\partial^2\rho}{\partial\omega\partial\phi}\right)^2 = \frac{1}{12}n^2(1+o(1)).$$

□

Similarly, it can be shown that $\det\Lambda(\omega,\phi)$ is well behaved also as $\omega \to 0$ or $\omega \to \pi$. Lemma 1 then implies that $\int_T \det\Lambda(\omega,\phi)^{1/2}d\omega d\phi = O(|T|n)$. Plugging this into Eq. (22) gives that for large values of u

$$\Pr\left[\sup_{(\omega,\phi)\in T}\eta(\omega,\phi) > u\right] = n\sqrt{\frac{\pi}{24}}ue^{-u^2/2}(1+o(1))$$

The change of variables $u = \sqrt{x}$, combined with Eqs. (11) and (21) yields Eq. (15).

The reason why the suggested penalty term C_n of Eq. (19) is suitable for testing higher order models is that $\det\Lambda(\omega,\phi) = O(n^2)$ also for $K > 1$ (as long as $K \ll n$). To illustrate this, consider a single sinusoid $K = 1$ with frequency and phase shift (ω_1,ϕ_1). As above, let $f(\omega,\phi) = (\sin(\omega+\phi), \sin(2\omega+\phi),\ldots,\sin(n\omega+\phi))$, and let $h = f(\omega_1,\phi_1)/\|f(\omega_1,\phi_1)\|$. According to Eq. (21), the random field of interest is $\eta_1(\omega,\phi) = \langle g(\omega,\phi),\xi\rangle/\|g(\omega,\phi)\|$, where

$$g(\omega,\phi) = f^{\perp}(\omega,\phi) = f(\omega,\phi) - \langle f(\omega,\phi), h\rangle h$$

First consider the case where ω, ϕ are far from ω_1, ϕ_1, (at distance significantly larger than $1/n$). Then,

$$\langle f, h\rangle = \frac{\sum_{t=1}^{n} \sin(\omega t+\phi)\sin(\omega_1 t+\phi_1)}{\sqrt{\sum_t \sin(\omega_1 t+\phi_1)^2}} = O\left(\frac{1}{\sqrt{n}}\right)$$

and

$$\langle \frac{\partial f}{\partial \omega}, h\rangle = \frac{\sum_{t=1}^{n} t\cos(\omega t+\phi)\sin(\omega_1 t+\phi_1)}{\sqrt{\sum_t \sin(\omega_1 t+\phi_1)^2}} = O\left(\sqrt{n}\right)$$

Therefore,

$$\left\|\frac{\partial g}{\partial \omega}\right\|^2 = \left\|\frac{\partial f}{\partial \omega}\right\|^2 - \left\langle \frac{\partial f}{\partial \omega}, h\right\rangle^2 = O(n^3).$$

Calculations analogous to those done above for $K=0$, give that $\det\Lambda(\omega,\phi) = O(n^2)$ as long as ω, ϕ is far from ω_1, ϕ_1 (and also $\min(\omega, \pi-\omega) \gg 1/n$). Since the behavior of the supremum of η_1 depends on the integral $\int_T \sqrt{\det\Lambda(\omega,\phi)}$, we need to show that $\det\Lambda(\omega,\phi)$ is also well behaved (does not explode) as $(\omega,\phi)\to(\omega_1,\phi_1)$. To this end, consider for example $(\omega,\phi) = (\omega_1,\phi_1)+(\delta\omega, 0)$. In this case

$$g = f^{\perp} = f(\omega_1+\delta\omega,\phi_1) - \langle f(\omega_1+\delta\omega,\phi_1), h\rangle h$$

A Taylor expansion gives

$$f(\omega_1+\delta\omega,\phi_1) = f(\omega_1,\phi_1)+\delta\omega\frac{\partial f}{\partial\omega}+\frac{1}{2}(\delta\omega)^2\frac{\partial^2 f}{\partial\omega^2}+O((\delta\omega)^3).$$

Inserting this expression into the previous equation gives

$$g = \delta\omega\left(F_1 + \frac{1}{2}\delta\omega F_2 + O(\delta\omega)^2\right)$$

where

$$\begin{aligned} F_1 &= \left(\frac{\partial f}{\partial\omega}\right)^{\perp} = \frac{\partial f}{\partial\omega} - \left\langle\frac{\partial f}{\partial\omega}, h\right\rangle h \\ F_2 &= \left(\frac{\partial^2 f}{\partial\omega^2}\right)^{\perp} = \frac{\partial^2 f}{\partial\omega^2} - \left\langle\frac{\partial^2 f}{\partial\omega^2}, h\right\rangle h \end{aligned}$$

The random field relevant to this case is $\eta_1 = \langle g,\xi\rangle/\|g\|$. Hence one can omit in the calculations the multiplicative factor of $\delta\omega$ in the definition of g above. Tedious but straightforward calculations similar to those done for $K=0$ above give that

$$\begin{aligned} \|F_1\|^2 &= \frac{1}{6}n^3(1+o(1)) \\ \|F_2\|^2 &= \frac{1}{10}n^5(1+o(1)) \end{aligned} \tag{27}$$

Further, note that

$$\|g\|^2 = \|F_1\|^2 + O(\delta\omega) \quad \text{and} \quad \frac{\partial g}{\partial\omega} = \frac{1}{2}F_2.$$

Hence,

$$\begin{aligned} \frac{\partial^2\rho}{\partial\omega^2} &= -\frac{\|\frac{\partial g}{\partial\omega}\|^2}{\|g\|^2} + \frac{\langle\frac{\partial g}{\partial\omega}, g\rangle}{\|g\|^4} \\ &= -\frac{1}{4}\frac{\|F_2\|^2}{\|F_1\|^2} + \frac{1}{4}\frac{\langle F_2, F_1\rangle^2}{\|F_1\|^4} \\ &= -\frac{3}{10}n^2(1+o(1)). \end{aligned} \tag{28}$$

Similar calculations give that $\partial^2\rho/\partial\phi^2 = -1(1+o(1))$ and $\partial^2\rho/\partial\omega\partial\phi = \frac{6}{16}n^2(1+o(1))$. Overall, this gives $\det\Lambda(\omega,\phi) = O(n^2)$ also when (ω,ϕ) is near (ω_1,ϕ_1). In a similar fashion it follows that $\det\Lambda(\omega,\phi) = O(n^2)$ also in the case of $K>1$ sinusoids, provided ω,ϕ are far from (ω_j,ϕ_j) of each of the K sinusoids.

The following proposition shows that if $\det\Lambda = O(n^2)$ then our proposed penalty term yields a nearly *constant false alarm rate* (CFAR) estimator for the number of sinusoids.

Proposition: *Let η_K be defined as in Eq. (8), and have a covariance function ρ_K, which may depend on the K unknown parameters $\boldsymbol{\theta}_K = \{\omega_j,\phi_j\}_{j=1}^K$. Assume that for the corresponding matrix $\Lambda_K(\omega,\phi)$ the following condition holds,*

$$\int_T \sqrt{\det\Lambda_K(\omega,\phi)}d\omega d\phi = n\,C(\boldsymbol{\theta}_K) \tag{29}$$

with $C(\boldsymbol{\theta}_K) = O(1)$ regardless of the unknown values $\boldsymbol{\theta}_K$. Then, for $n\gg 1$, the suggested penalty term

$$C_n = \ln n + \frac{1}{2}\ln\ln n + \frac{1}{2}\ln\frac{\pi}{3\alpha^2}$$

leads to a model order overestimation probability which is $O(\alpha)$ as $\alpha\to 0$.

Proof: Given a threshold $x = C_n$ (or a penalty term $k\,C_n$), model order overestimation occurs when $G_K > x$. According to Eq. (21), for sufficiently large n and x,

$$\Pr[G_K > x] \le 2\Pr\left[\sup_{(\omega,\phi)\in T}\eta_K(\omega,\phi) > \sqrt{2x}\right]$$

Next, by the definition (8), η_K is a zero mean unit variance Gaussian random field. Applying Eq. (22) of the theorem,

$$\Pr\left[\sup\eta_K > \sqrt{2x}\right] \approx \int_T\sqrt{\det\Lambda(\omega,\phi)}d\omega d\phi\,\frac{\sqrt{2x}e^{-x}}{(2\pi)^{3/2}}$$

Combining the last two equations with assumption (29), and the expression for $x = C_n$ gives

$$\begin{aligned} \Pr[G_K > C_n] &\le \tfrac{2nC(\theta_K)\sqrt{2\ln n}}{(2\pi)^{3/2}}\sqrt{1+\tfrac{1}{2}\tfrac{\ln\ln n}{\ln n}+\tfrac{1}{2}\tfrac{\ln\pi/3\alpha^2}{\ln n}} \\ &\quad \times e^{-\ln n - 1/2\ln\ln n - 1/2\ln\pi/3\alpha^2} \\ &\le A(\theta_K)\alpha(1+o(1)) = O(\alpha) \end{aligned} \tag{30}$$

since $A(\theta_K) = C(\theta_K)\sqrt{3}/\pi^2 = O(1)$. □

In simple words, the proposition above implies that for $K\ll n$ (the typical case of a few periodic components in a long time series), and for $\alpha\ll 1$

$$\Pr[G_K > C_n(\alpha)] \approx \Pr[G_0 > C_n(\alpha)] \tag{31}$$

Fig. 1 confirms this claim empirically for $n=128$. This analysis justifies the use of the same penalty term regardless of the number or parameters of the sinusoids present.

D. Unknown noise level σ

The analysis above assumed a known noise level. Next, we consider a more realistic case, where noise is still i.i.d. Gaussian but with an unknown standard deviation σ. For unknown σ, the log-likelihood function is given by

$$\mathcal{L}_k = -\frac{\|\mathbf{x}-\hat{\mu}_k\|^2}{2\hat{\sigma}_k^2} - \frac{n}{2}\ln\left(2\pi\hat{\sigma}_k^2\right) \tag{32}$$

where $\hat{\mu}_k$ denotes the sum of the k fitted sinusoids, and $\hat{\sigma}_k^2$ is the ML estimate of σ^2, assuming k sinusoids.

Note that the estimation of the amplitudes, frequencies and phase shifts of the k sinusoids is independent of that of the noise estimation. Hence, the maximum likelihood estimate of the noise variance is simply given by

$$\hat{\sigma}_k^2 = \frac{1}{n}\|\mathbf{x} - \hat{\mu}_k\|^2.$$

Therefore, we have that

$$\ln \mathcal{L}_k = -\frac{n}{2}(1 + \ln(2\pi)) - \frac{n}{2}\ln(\|\mathbf{x} - \hat{\mu}_k\|^2).$$

We claim that the same penalty is suitable in this setting as well, namely that model order selection via Eq. (20) would still perform well, even for unknown noise level.

To explain this point, consider the case of no sinusoids, where $\mathbf{x}$ is a vector of pure noise. Then, the GLRT for detecting a single signal is

$$G_0 = \ln\left(\frac{\mathcal{L}_1}{\mathcal{L}_0}\right) = \frac{n}{2}\ln\left(1 - \frac{|\langle \mathbf{x}, \mathbf{sin}_{\hat{\omega},\hat{\phi}}\rangle|^2}{\|\mathbf{x}\|^2 \cdot \|\mathbf{sin}_{\hat{\omega},\hat{\phi}}\|^2}\right)$$

As expected, this test statistic is invariant to scaling, since the noise level is assumed to be unknown. Note that under the null hypothesis of no signals, $\mathbf{x}$ is a Gaussian vector of length n, and so $\|\mathbf{x}\|^2 = n + O_P(\sqrt{n})$. The ratio inside the logarithm is $O(\ln n/n)$ and in particular it is significantly smaller than 1. Hence, a first order Taylor expansion of the logarithm gives that to leading order, the distribution of G_0 is approximately that of a normalized version of the random field $S(\omega, \phi)$ in Eq. (13) for the case of known noise. The normalization factor $\|\mathbf{x}\|^2/n$ is close to one for large n, and therefore the maximum has a similar distribution. In fact, the maximum of the normalized field is slightly more concentrated around its mean, and so its right tail probabilities are smaller in comparison to the un-normalized $S(\omega, \phi)$. A detailed study of this difference is beyond the scope of this paper. For a treatment of this normalized field using Hotelling's tube formula see [21].

E. Non-Gaussian white noise

Finally, we examine the robustness of our penalty term to a mismatch in the noise model. We thus consider the case where the noise ξ_t is white (i.i.d.) zero mean, with variance σ^2 and finite fourth moment, but is *not* necessarily Gaussian.

Despite this possible non-Gaussianity, in our approach we still assume the noise distribution to be Gaussian, and find the parameters of the k sinusoids by a least squares fit. As $n \to \infty$, when fitting data to the correct model order, the resulting parameter estimates are still asymptotically consistent.

The main claim is again that our proposed penalty term is suitable for this setting as well. To see this, note that for any fixed ω, ϕ, from a weighted version of the central limit theorem (a generalized Berry-Esseen theorem, see e.g. [14])

$$\frac{\sum_{t=1}^n \xi_t \sin(\omega t + \phi)}{\|\mathbf{sin}_{\omega,\phi}\|} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Hence, even with non-Gaussian noise, as $n \to \infty$, the periodogram-like random field $S(\omega, \phi)$ converges in distribution to a χ^2 random field with one degree of freedom, and thus its maxima follows a distribution similar to the case of Gaussian noise. We verify this claim empirically in Sec. V. A detailed study of this convergence, which depends on the higher order moments of the noise distribution, is an interesting research problem, beyond the scope of this paper.

Remark: We conclude this section with an important remark regarding the un-suitability of our simple penalty term for the case of strongly *correlated* noise. The reason is that in the presence of colored noise, the maxima of the periodogram can be significantly larger than in the white noise case, with the difference depending on the noise correlation structure. Deriving sharp model selection criteria under such settings is an interesting problem for future research.

IV. Theoretical Performance Analysis

A. Penalty Term Comparisons for Finite n

Let us compare the penalty term of our approach (Eq. (19)) to previously considered penalty terms. First, note that for a reasonable range of record lengths, $n \in [64, 2048]$, the term $\frac{1}{2}\ln\ln n$ is almost constant with a value bounded in the interval $[0.7, 1]$. Similarly, for a significance level of $\alpha = 0.5\%$, $\ln(\pi/3\alpha^2)/2 \approx 5.3$. Hence, in loose terms, for these values of n, one may view our suggested penalty term as having the approximate form

$$C_n = \ln n + const$$

where for a false alarm probability $\alpha = 0.5\%$ the constant is roughly 6.2.

We now explain why, for records of length $n = 64, 128$, Djuric's MAP estimator has better detection performance than the MDL estimator, despite its higher penalty constant. The reason is that for these specific values of n,

$$C_n(\text{MAP}) = \frac{5}{2}\ln n \approx \ln n + \ 6.5$$

whereas

$$C_n(\text{MDL}) = \frac{3}{2}\ln n \approx \ln n + \ 2.2$$

Hence, the MAP estimator has a sufficiently strong penalty leading to a negligible probability to overestimate the number of signals, whereas the MDL estimator, although asymptotically consistent, has too small a penalty term for these finite values of n, which leads to a non-negligible probability to overestimate the number of sinusoids.

Since the MDL estimator is asymptotically consistent, it is interesting to analyze at which minimal sample size n_α its overestimation probability drops to below a value of $\alpha \ll 1$ and becomes negligible. According to our analysis, this will occur roughly when

$$\tfrac{3}{2}\ln n = \ln n + \tfrac{1}{2}\ln\ln n + \ln\tfrac{\pi}{3\alpha^2}.$$

In other words,

$$n_\alpha \approx \frac{\pi^2}{(3\alpha^2)^2}.$$

For example, for the MDL estimator to have an overestimation probability of $\alpha = 1\%$ we need $n \gtrsim 10^8$ samples! Simply put, even though the MDL estimator is asymptotically consistent, this behavior begins to manifest itself only at extremely large sample sizes. We note that this finite sample size analysis is relevant also for many other model selection penalty terms that are asymptotically consistent.

B. Signal Detection Thresholds

Let us compare the signal detection performance of the different methods. In particular we analyze the signal strength required for detecting the presence of a single sinusoid with unknown amplitude a, frequency ω, and phase shift ϕ. For $k = 0$, we obtain

$$-\ln \mathcal{L}_0 = \frac{1}{2}\sum_{t=1}^{n}(a\sin(\omega t+\phi)+\xi_t)^2$$

whereas for $k = 1$

$$-\ln \mathcal{L}_1 = \frac{1}{2}\sum_{t=1}^{n}(a\sin(\omega t+\phi)-\hat{a}\sin(\hat{\omega} t+\hat{\phi})+\xi_t)^2$$

For $n \gg 1$, $\hat{\omega}, \hat{\phi}$ are close to ω, ϕ, and thus $\hat{a} \approx a + \langle \xi_t, \sin(\omega t+\phi)\rangle / \|\sin(\omega t+\phi)\|$. Therefore, upon opening the brackets above, the noise term $\sum \xi_t^2$ in the GLRT cancels out, and we obtain

$$\ln\left(\frac{\mathcal{L}_1}{\mathcal{L}_0}\right) \approx \frac{1}{2}\left[\hat{a}^2 \sum_t \sin(\omega t+\phi)^2\right]$$

For ω satisfying the condition that $\min(\omega, \pi-\omega) \gg 1/n$ we have that $\sum_t \sin(\omega t+\phi)^2 = \frac{n}{2}(1+o(1))$. Hence, the condition for detection by the MAP estimator is that

$$a^2_{\text{MAP}} > \frac{4}{n}\frac{5\ln n}{2} = 10\frac{\ln}{n}$$

In contrast, detection by the EVT estimator requires

$$a^2_{\text{EVT}} > \frac{4}{n}\left[\ln n + \frac{1}{2}\ln\ln n + \tfrac{1}{2}\ln\frac{\pi}{3\alpha^2}\right]$$

Assuming that a false alarm rate of $\alpha = 0.5\%$ is acceptable, the MAP and EVT have a similar detection performance for $n = 64$, where their penalty terms are comparable. However, for larger record lengths the penalty term of MAP becomes increasingly larger than that of the EVT estimator, thus affecting its detection performance. For example, for $n = 512$ the EVT estimator can detect sinusoids weaker by a factor of $10\log_{10}(a^2_{\text{MAP}}/a^2_{\text{EVT}}) = 1.5$ dB. This is also confirmed by simulations, see fig. 3.

V. Simulations

We compare the detection performance of our algorithm, denoted EVT, to the MDL and AIC estimators in a series of simulations, all with $\sigma = 1$. Our performance measure is the probability of correct model order estimation,

$$\Pr[\hat{k} = K].$$

We consider two settings similar to those of [13]. In the first setting $K = 2$, $n = 128$ or $n = 512$, and the sinusoid parameters are $\omega_1 = 2\pi \cdot 0.2, \omega_2 = \omega_1 + 2\pi/n$, $\phi_1 = 0$, $\phi_2 = \pi/4$. The amplitudes of the two sinusoids are equal $a_1 = a_2$. We use the same notation as in [13], where the SNR of a sinusoid with amplitude a is defined as $a^2/2\sigma^2$, or in dB units, SNR $= 10\log_{10}(a^2/2\sigma^2)$.

In our simulations we consider a wide range of amplitude strengths and corresponding SNR values. For each SNR value we performed 200 independent trials. In each trial, a random noise-corrupted signal is generated. For model orders $\hat{k} = 0, 1, 2, 3$ the relevant parameters are found by approximately maximizing the likelihood function. This is done iteratively, one frequency at a time, via a sequence of alternating projections, similar to [38]. Initial parameters for model order $k+1$ are those found previously for model order k, with an additional 1-d search for the remaining frequency.

The resulting performance curves are shown in figure 3. As predicted theoretically, for $n = 128$ the MAP and EVT estimators have a similar performance, whereas for $n = 512$ the EVT estimator outperforms the MAP estimator by approximately 1.5dB. For both record lengths, the MDL estimator has a non-negligible overestimation probability, of the order of 10% for $n = 128$ and 8% for $n = 512$. Finally, as shown in the figure, there is only a negligible performance degradation due to not knowing the noise level and having to estimate it.

Next, we investigate the accuracy of the frequency estimates in conjunction to the detection problem. First, suppose an a-priori knowledge that two sinusoids are present. Then at high SNR, their frequency estimates are very accurate, whereas at low SNR at least one of the estimated frequencies is far from the true one. This is the well known breakdown phenomenon of the Maximum-Likelihood estimator. In contrast, as shown in Fig. 4, limiting the frequency estimation step only to those realizations where two sinusoids are actually detected by our method, gives estimated frequencies close to the true ones, even at low SNR values. These results show the importance of detection prior to estimation in certain parametric problems. For a similar phenomena and a more detailed analysis in the setting of direction of arrival estimation in array processing applications, see [1], [5].

In the second setting, $K = 3$, $n = 128$, the values of $\omega_1, \omega_2, \phi_1, \phi_2$ are as above, while $\omega_3 = \omega_2 + 2\pi/n, \phi_3 = \pi/3$. Here the sinusoid amplitudes are of the form $(a_1, a_2, a_3) = a_0(1, \sqrt{6.3246/20}, 1)$, so the middle sinusoid is 5dB weaker than the others. A similar behavior to the case $K = 2$ is observed here as well (figure not shown), namely the MDL estimator overestimates the number of signals, whereas the EVT estimator has an improved detection performance, at the price of a slightly larger overdetection probability. Our analysis explains the simulation results of [13], which were performed at relatively high SNR (e.g. at the right edge of our figures), where the MAP estimator easily detects the present signals, whereas MDL overfits their number.

A. Non-Gaussian Noise

Next we study the performance of our model selection approach in the presence of noise mismatch. Fig. 5 shows the detection performance results for $n = 128$ and $k = 2$ (the

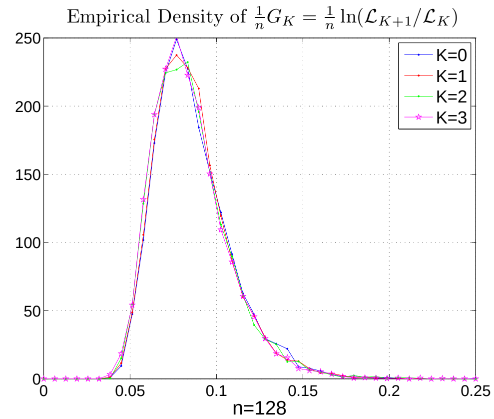

Fig. 1. Empirical density of the random variable G_K/n for different values of K. Note the very weak dependence on K, which justifies the use of our suggested penalty term.

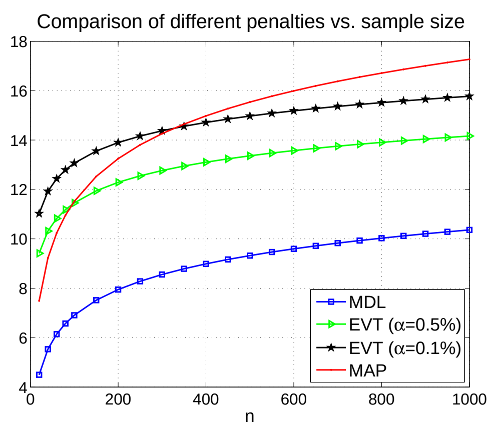

Fig. 2. Comparison of different penalty terms vs. sample size n. The MDL penalty (lower curve, blue squares) is clearly too small, and leads to a non-negligible probability of overestimation. The top curve (red dots) is the MAP penalty - which is well suited for $n \approx 100$ but is too large for $n > 500$.

same setting as in the previous section), when the noise is i.i.d. but with a Laplace distribution. As predicted theoretically, the detection performance is essentially unchanged.

B. Comparison to SAMOS

Finally, we compare our model selection approach to SAMOS [28], a recently suggested order selection method based on shift invariance principles.

Similar to the example considered in [28], we generate a signal composed of two sinusoids corrupted by Gaussian white noise, $x_t = \cos(2\pi\nu_1 t) + \cos(2\pi\nu_2 t) + \sigma\xi_t$, where $t = 0, \ldots, n-1$, $n = 65$, $\nu_1 = 0.2$, $\sigma = 0.4/\sqrt{2}$, and ν_2 is such that the difference $\delta\nu = (\nu_2 - \nu_1) \in [1/300, 1/50]$. Since we are dealing with real-valued signals and noise, the SAMOS procedure should output a model order of 4 (two real valued sinusoids can be described as the sum of four complex valued exponentials). Fig. 6 shows the success probability in model selection as a function of $1/(\nu_2 - \nu_1)$. Since here we have undamped exponentials, SAMOS performs slightly better than Fig. 2 in [28]. However, our fully parametric (and significantly

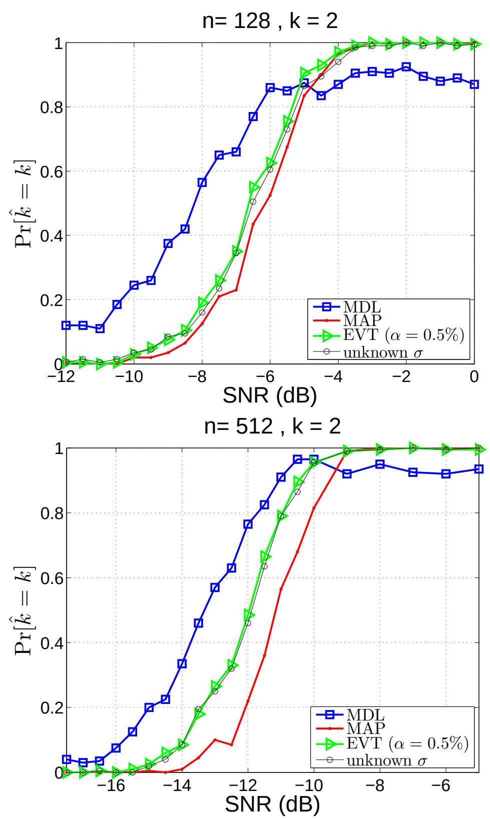

Fig. 3. Detection Performance of the MDL, MAP and EVT estimators as a function of SNR for record length $n = 128$ (left) and $n = 512$ (right) for the case of $K = 2$ sinusoids with closely spaced frequencies.

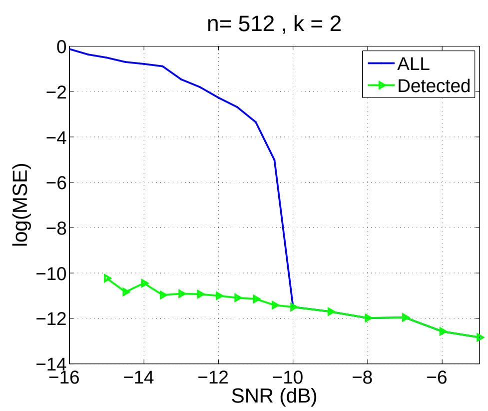

Fig. 4. Breakdown phenomenon in frequency estimation. The blue curve shows the log of the mean squared error as a function of SNR, averaged over many noise realizations. Note the breakdown phenomenon at roughly -10dB, where a sudden sharp increase in MSE is due to unreliable estimation of one of the frequencies. The green curve shows the log of the MSE, but averaged only over those realizations where two sinusoids were indeed detected by our algorithm. Note that at very low SNR, our estimator for the number of sinusoids can thus detect those realizations where reliable estimation is not possible.

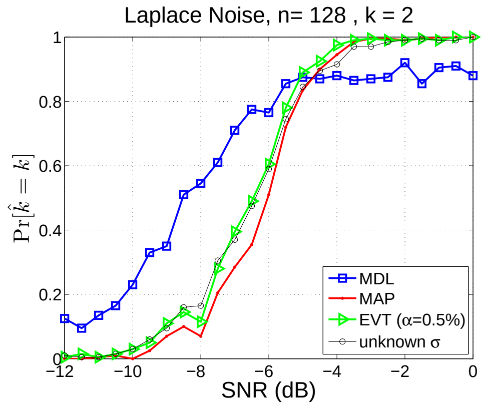

Fig. 5. Detection Performance with Laplace Noise.

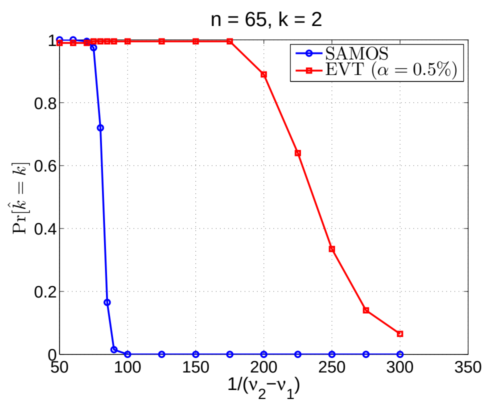

Fig. 6. Success probability vs. inverse frequency separation. Comparison of SAMOS with our EVT detector.

more computationally intensive) procedure is able to detect the two closely spaced sinusoids at much smaller frequency separation.

VI. Summary and Discussion

In this paper we presented a statistical analysis of the problem of detection of sinusoidal signals embedded in additive white noise. On the theoretical side, we showed that this model selection problem is closely related to problems of hypothesis testing when there are *nuisance parameters* not present under the null hypothesis. This observation implies, in turn, that parametric model selection is intimately related to the distribution of the maxima of random fields. We remark that these issues regarding model selection have been studied in other fields, for example in econometrics [19], and in mixture modes in statistics [11]. Our results are closely related to those of [11], in particular their theorem 3.2. However, our method of proof is different. We also note the potential relevance of our results to the penalty term in sparse representations (some of which also use a $C \ln n$ term) [17]. On the practical side, we showed that for reasonable record lengths, the MDL estimator has a non-negligible probability to overestimate the number of sinusoids. Furthermore, our analysis highlighted the crucial importance of including both a $\ln \ln n$ term as well as an *additive constant* in the penalty function for model selection.

Instead of relying on information theoretic considerations, the penalty term we propose is based on testing the *statistical significance* of each additional estimated sinusoid. Beyond its good detection performance and its explicit control of the over-detection probability, perhaps the most important reason to detect signals only when the increase in the likelihood is statistically significant is the so-called *breakdown phenomenon* of the maximum likelihood estimator. Accepting the presence of sinusoidal signals when the observed increase in likelihood is smaller than our suggested penalty function (with an $\alpha \ll 1$) is rather meaningless from an estimation point of view, since the resulting estimates for the frequency are typically far from the true values. Finally, the observations and analysis made in this paper are not limited to the specific problem of detection of sinusoids in noise. Rather, they are applicable to many other parametric model selection problems in signal processing.

Appendix

A. Proof of Theorem 1

Before describing the technical details, let us first provide an overview of the proof. Let $\boldsymbol{\theta}_K$ be the true unknown vector of frequencies and phase shifts. For each record length n we consider maximum-likelihood estimates of the unknown parameters restricted to only a *finite* set Θ_m of size m, such that $(\omega_j, \phi_j) \in \Theta_m$. As $n \to \infty$ we consider increasingly finer sets Θ_m, with $m = m(n) \to \infty$ which become dense in $[0, \pi] \times [0, 2\pi]$. This formulation is not too restrictive as the size of the set is allowed to grow with sample size n, allowing arbitrary precision in the limit, and has the advantage of significantly simplifying the analysis.

Rather than proving Theorem 1 for the specific case of sinusoids, we consider a more general framework as follows: Let $\mathcal{F} = \{f_\theta : \mathbb{R} \to \mathbb{R} \,|\, \theta \in \Theta_m\}$ be a parametric class of continuous functions over a finite parameter space Θ_m, that satisfy the following conditions:

(i) Consider a discrete sampling of a function $f_\theta(t) \in \mathcal{F}$, at times $t_j = j$, for $j = 1, \dots, n$. Then, for all $\theta \in \Theta_m$,

$$c_1 \le \frac{1}{n} \sum_{j=1}^{n} (f_\theta(t_j))^2 \le c_2 \tag{33}$$

for some constants $0 < c_1 < c_2 < \infty$.

(ii) Let $g(t) = \sum_{j=1}^{k} a_j f_{\theta_j}(t)$, where $\theta_j \in \Theta_m$ are all distinct. Consider the $n-$dimensional vector $\mathbf{g} = (g(t_1), \dots, g(t_n))$. Then, for $k < n$

$$\|\mathbf{g}\|^2 \ge \epsilon_{m,k} \cdot n \cdot \sum_{j=1}^{k} a_j^2 \tag{34}$$

where $\epsilon_{m,k} > 0$. Note that Eq. (34) implies, in particular, that for any choice of k distinct values θ_j, the functions $f_{\theta_j}(t)$ evaluated at the n time points t_j are linearly independent vectors in $\mathbb{R}^n$.

Let $x(t)$ be a function of the following form

$$x(t) = \sum_{j=1}^{K} A_j f_{\theta_j}(t) + \xi(t) \tag{35}$$

where $\theta_j \in \Theta_m$ are all distinct and $\xi(t)$ is a Gaussian white noise process. The problem in this general framework is to detect the number of components K given observed data $x(t_j)$, $j = 1, \ldots, n$. Note that the collection of sinusoid functions $\mathcal{F} = \{\sin(\omega t + \phi) \,|\, (\omega, \phi) \in \Theta_m\}$ indeed satisfies conditions (i) and (ii) above.

The proof outline is as follows: First, in lemma 2 we prove that as $n \to \infty$, when fitting a model of order K to the data, all maximum likelihood estimates (MLE) of the various parameters converge to the true ones. While in principle this lemma follows trivially from the consistency of the MLE, we prove it in detail as it is informative in the case of finite parameter space, to understand theorem 1. Next, in lemma 3, we prove that when fitting data of order K with a model of order $K+1$, then as $n \to \infty$, K of the estimated parameters converge to their correct ones, and the last one is free to fit the remaining noise as best as possible.

As we now show, combining Lemmas 2 and 3 yield the theorem. Lemma 2 implies that

$$-\ln \mathcal{L}(\hat{\boldsymbol{\theta}}_K) \to \frac{1}{2}\|\xi^\perp\|^2 = \frac{1}{2}\sum_{t=1}^{n}(\xi_t^\perp)^2 \tag{36}$$

where $\xi^\perp$ is the part of the noise orthogonal to the span of $\{f_{\theta_j}\}_{j=1}^K$. From Lemma 3 it follows that

$$-\ln \mathcal{L}(\hat{\boldsymbol{\theta}}_{K+1}) \to \min_{a,\theta} \frac{1}{2}\|\xi^\perp - a f_\theta^\perp\|^2 \tag{37}$$

where $f_\theta^\perp$ is the part of f_θ orthogonal to span of $\{f_{\theta_j}\}_{j=1}^K$. For any given θ, the amplitude that minimizes the norm above is $a = \langle \xi^\perp, f_\theta^\perp\rangle / \|f_\theta^\perp\|$. Hence,

$$G_K = \ln\left(\frac{\mathcal{L}_{K+1}}{\mathcal{L}_K}\right) \to \frac{1}{2}\sup_\theta \frac{\langle \xi^\perp, f_\theta^\perp\rangle^2}{\|f_\theta^\perp\|^2}$$

and Theorem 1 follows.

Lemma 2: *Let $\{x(t_j)\}_j$ be a time series of length n of the form (35), with $K < n/2$ components. We further assume that conditions (i) and (ii) above hold. Denote by $\hat{\boldsymbol{\theta}}_K$ the MLE of $\boldsymbol{\theta}_K$ assuming a model of order K. Then, as $n \to \infty$,*

$$\Pr[\hat{\boldsymbol{\theta}}_K = \boldsymbol{\theta}_K] \to 1.$$

Proof: The proof consists of showing that the likelihood of the correct $\boldsymbol{\theta}_K$ is asymptotically larger that the likelihood of any other candidate. Indeed, the log likelihood at the correct parameters, $\hat{\boldsymbol{\theta}}_K = \boldsymbol{\theta}_K$ is given by

$$\ln \mathcal{L}_K(\boldsymbol{\theta}_K, x) = -\frac{1}{2}\|\xi^\perp\|^2 \tag{38}$$

where $\xi^\perp = P_K^\perp \xi$ is the component of the noise orthogonal to the subspace spanned by the K signal vectors $f_{\theta_i}(t_j)$.

Next, consider the log likelihood at some other parameter value $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_K) \neq \boldsymbol{\theta}_K$. Here,

$$\ln \mathcal{L}(\hat{\boldsymbol{\theta}}_K, x) = -\frac{1}{2}\|\sum_j \tilde{A}_j f_{\theta_j} - \sum_j \hat{A}_j f_{\hat{\theta}_j} + \xi^\perp\|^2 \tag{39}$$

where the coefficients $\tilde{A}_j$ incorporate also the component of the noise in the span of the signal vectors, such that $\sum_j \tilde{A}_j f_{\theta_j} = \sum_j A_j f_{\theta_j} + \xi - \xi^\perp$. For simplicity of notation, we write

$$g = \sum_{j=1}^K \tilde{A}_j f_{\theta_j} - \sum_{j=1}^K \hat{A}_j f_{\hat{\theta}_j} \tag{40}$$

We now examine the difference in the log-likelihoods,

$$\ln \mathcal{L}(\boldsymbol{\theta}_K, x) - \ln \mathcal{L}(\hat{\boldsymbol{\theta}}_K, x) = \frac{1}{2}\|g\|^2 - \|g\|\frac{\langle g, \xi^\perp\rangle}{\|g\|} \tag{41}$$

According to condition (ii), $\|g\|^2 \geq \epsilon_{m,2K}\min_j |A_j|^2 \cdot n$. On the other hand, for any function g, $\langle g, \xi^\perp\rangle/\|g\|$ is a $N(0,1)$ Gaussian random variable. For any choice of the K parameters $\hat{\theta}_1, \ldots, \hat{\theta}_K$, the dot product $|\langle g, \xi^\perp\rangle|/\|g\|$ is the norm of the projection of noise onto a specific vector in a linear subspace of dimension K. This quantity is thus maximized by the norm of the noise in this subspace, which is distributed as $\sqrt{\chi_K^2}$. Since the set Θ_m is finite, there are at most $|\Theta_m|^K$ possible choices for $\hat{\boldsymbol{\theta}}_K$. Applying the union bound yields that the maximum value of the second term is of the order of $\sqrt{n}K^{3/2}\log m$. As $n \to \infty$ the first term, which is $\Omega(n)$ clearly dominates, as long as $m = m(n)$ grows say polynomially with n. □

Lemma 2: Let $\{x(t_j)\}_j$ be a time series of length n of the form (35), with $K < n/2$ components. Let $\hat{\boldsymbol{\theta}}_{K+1}$ be the MLE assuming a model of order $K+1$. Then, as $n \to \infty$, K of the components in $\hat{\boldsymbol{\theta}}_{K+1}$ are equal to the true parameters $(\theta_1, \ldots, \theta_K)$.

Proof: We first consider the case, where w.l.g., the first K parameters in $\hat{\boldsymbol{\theta}}_{K+1}$ are equal to the true values, $\hat{\boldsymbol{\theta}}_{K+1} = (\theta_1, \ldots, \theta_K, \hat{\theta}_{K+1})$. The estimated signal is then

$$\hat{x}(t) = \sum_{j=1}^K \hat{A}_j f_{\theta_j}(t) + \hat{A}_{K+1} f^\perp_{\hat{\theta}_{K+1}}(t)$$

where $f^\perp_{\hat{\theta}_{K+1}}$ is the part of the vector $f_{\hat{\theta}_{K+1}}$ orthogonal to the span of the first K vectors. The log-likelihood is then

$$\begin{aligned}-\ln \mathcal{L}_{K+1} &= \tfrac{1}{2}\Big\|\sum_{j=1}^K (\tilde{A}_j - \hat{A}_j) f_{\theta_j} + \xi^\perp - \hat{A}_{K+1} f^\perp_{\hat{\theta}_{K+1}}\Big\|^2 \\ &= \tfrac{1}{2}\Big\|\sum_{j=1}^K (\tilde{A}_j - \hat{A}_j) f_{\theta_j}\Big\|^2 \\ &\quad + \tfrac{1}{2}\Big\|\xi^\perp - \hat{A}_{K+1} f^\perp_{\hat{\theta}_{K+1}}\Big\|^2\end{aligned}$$

This sum is maximized by choosing $\hat{A}_j = \tilde{A}_j$ for $j = 1, \ldots, K$, and $\hat{\theta}_{K+1}$ such that it best fits the remaining noise component. Now consider an estimate $\hat{\theta}$ (of dimension $K+1$), which does not coincide with θ on any subset of K coordinates. The log-likelihood in this case is given by

$$\ln \mathcal{L}_{K+1} = -\frac{1}{2}\Big\|g + \xi^\perp\Big\|^2 \tag{42}$$

where

$$g = \sum_{j=1}^K \tilde{A}_j f_{\theta_j} - \sum_{j=1}^{K+1} \hat{A}_j f_{\hat{\theta}_j}$$

The first likelihood is of course bounded below by simply $-\frac{1}{2}\|\xi^\perp\|^2$. From this point on, the proof is analogous to that of the previous lemma. □.

ACKNOWLEDGMENTS

It is a pleasure to thank Peter Bickel, Bin Yu, Robert Adler and Mark Kliger for interesting discussions. We also thank the anonymous referees for valuable suggestions which greatly improved the manuscript. BN was supported by a grant from the Ernst Nathan biomedical fund.

REFERENCES

[1] Y. I. Abramovich, B.A. Johnson, Performance breakdown prediction for maximum likelihood DoA estimation, *ICASSP*, 2594–2597, 2010.

[2] R.J. Adler, On excursion sets, tube formulas and maxima of random fields, *Annals of Appl. Prob.* Vol. 10, No. 1, 174, 2000.

[3] R.J. Adler, and J.E. Taylor, *Random Fields and Geometry*, Springer, NY, 2007.

[4] H-Z. An, Z-G Chen and E.J. Hannan, The maximum of the periodogram, *J. Mult. Anal.*, vol. 13, pp. 383–400, 1983.

[5] Arkind, N., Nadler, B. Parametric joint detection-estimation of the number of sources in array processing, *6th IEEE SAM conference*, 2010.

[6] R. Badeau, B. David, G. Richard, A new perturbation analysis for signal enumeration in rotational invariance techniques, *IEEE Trans. Sig. Proc.*, vol. 54, no. 2., pp. 450–458, 2006.

[7] Bresler, Y. Macovski, A. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise *IEEE Trans. Sig. Proc.*, vol. 34, no. 5, pp. 1081–1089, 1986.

[8] S.T. Chiu, Detecting Periodic Components in a White Gaussian Time Series, *J. of the Royal Stat. Soc. Series B* , Vol. 51, No. 2, pp. 249-259, 1989.

[9] M.G. Christensen, A. Jakobsson and S.H. Jensen, Sinusoidal Order Estimation Using Angles between Subspaces, *EURASIP J. Adv. Sig. Pro.*, Article ID 948756, 2009.

[10] N.I. Cho, and S.U. Lee, On the adaptive lattice notch filter for the detection of sinusoids *IEEE Circuits and Systems*, vol. 40, no. 7, pp. 405-416, 1993.

[11] D. Dacunha-Castelle, and E. Gassiat, Testing the order of a model using locally conic parametrization: population mixtures and stationary arma processes, *Ann. Stat.*, vol. 27, no. 4, pp. 1178–1209, 1999.

[12] R.B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative *Biometrika*, vol. 74, no. 1, pp. 33–43, 1987.

[13] P.M. Djuric, A model selection rule for sinusoids in white Gaussian noise, *IEEE Sig. Proc.*, vol. 44, no. 7, pp. 1744–1751, 1996.

[14] W. Feller, On the Berry-Esseen theorem, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete,* vol. 10, pp. 261–268, 1968.

[15] R. A. Fisher, Tests of Significance in Harmonic Analysis, *Proceedings of the Royal Society of London. Series A*, Vol. 125, No. 796 (Aug. 1, 1929), pp. 54-59.

[16] J.-J. Fuchs, Estimating the number of sinusoids in additive white noise, *IEEE Trans. Sig. Proc.,* vol. 36, no. 12, pp. 1846–1853, 1988.

[17] J.-J. Fuchs, The generalized likelihood ratio test and the sparse representation approach, in *Proc. Intl. Conference on Image and Signal Processing, ICISP*, 2010.

[18] E.J. Hannan, Determining the number of jumps in a spectrum, In *Developments in Time Series Analysis*, Ed. T. Subba Rao, Chapman and Hall, pp. 127–138, 1993.

[19] B.E. Hansen, Inference when a nuisance parameter is not identified under the null hypothesis, *Econometrica*, vol. 64, no. 2, pp. 413–430, 1996.

[20] J-K Hwang and Y-C Chen, A combined detection-estimation algorithm for the harmonic retrieval problem, *Sig. Proc.*, vo. 30, pp. 177–197, 1993.

[21] I. Johnstone, D. Siegmund, On Hotelling's formula for the volume of tubes and Naiman's inequality, *Ann. Stat.*, vol 17, n. 1., pp. 184–194, 1989.

[22] L. Kavalieris and E.J. Hannan, Determining the number of terms in a trigonometric regression, *J. Time Ser. Anal.*, vol. 15, no. 6, pp. 613 – 625, 1994.

[23] M. Kliger, J.M. Francos, MAP model order selection rule for 2-D sinusoids in white noise, *IEEE Trans. Sig. Proc.*, vol. 53, no. 7, pp. 2563–2575, 2005.

[24] D. Kundu, Detecting the number of signals for undamped exponential models using information theoretic criterion, *J. Stat. Comp. Sim.* vol. 44, pp. 117–131, 1992.

[25] D. Kundu, Estimating the number of sinusoids and its performance analysis, *J. Stat. Comp. Sim.*, vol. 60, no. 4, pp. 347–362, 1998.

[26] Hongbin Li, P. Stoica and J. Li, Computationally efficient parameter estimation for harmonic sinusoidal signals *Sig. Proc.*, vol. 80, no. 9, pp. 1937–1944, 2000.

[27] M.D. Macleod, Fast nearly ML estimation of the parameters of real or complex tones or resolved multiple tones, *IEEE Trans. Sig. Proc.*, vol. 46, no. 1, pp. 141–148 ,1998.

[28] J-M. Papy, L. De Lathauwer, S. van Huffel, A shift invariance-based order-selection technique for exponential data modelling, *IEEE Sig. Proc. Lett.*, vol. 14, no. 7, pp. 473–476, 2007.

[29] B. G. Quinn, Estimating the number of terms in a sinusoidal regression, *J. Time Ser. Anal.*, Vol. 10, no. 1, pp. 71–75, 1989.

[30] B. G. Quinn, E.J. Hannan, *The estimation and tracking of frequency*, Cambridge University Press, 2001.

[31] S. Kritchman and B. Nadler, Non-parametric detection of the number of signals, hypothesis tests and random matrix theory, *IEEE Trans. Sig. Proc.*, vol. 57, no. 10, pp. 39303941, 2009.

[32] V.U. Reddy and L.S. Biradar, SVD-based information theoretic criteria for detection of the number of damped/undamped sinusoids and their performance analysis *IEEE Trans. Sig. Proc.*, 41(9), pp. 2872–2881, 1993.

[33] P. Stoica, R.L. Moses, B. Friedlander and T. Soderstrom, Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements, *IEEE Sig. Proc.*, vol. 37, no. 3, pp. 378–392, 1989.

[34] D.W. Tufts, and R. Kumaresan, Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood, *Proc. IEEE*, vol. 70, no. 9, pp. 975–989, 1982.

[35] K.F. Turkman and A.M. Walker, On the asymptotic distributions of maxima of trigonometric polynomials with random coefficients, *adv. appl. Prob.*, vol. 16, pp. 819–842, 1984.

[36] X. Wang, An AIC type estimator for the number of cosinusoids, *J. Time Series Anal.*, Vol. 14, no. 4, pp. 433–440, 1993.

[37] C.-H.J. Ying, A. Sabharwal and R.L. Moses, A combined order selection and parameter estimation algorithm for undamped exponentials, *IEEE Trans. Sig. Proc* vol. 48, no. 3, pp. 693–701, 2000.

[38] I. Ziskind and M. Wax, Maximum likelihood localization of multiple sources by alternating projection, *IEEE Trans. Acoustics, Speech and Signal Processing* vol. 36, no. 10, pp. 1553–1560, 1998.

Boaz Nadler received a B.Sc. degree in mathematics and physics (cum laude), an M.Sc. degree in applied mathematics (summa cum laude), and a 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. Since 2005 he is a senior research scientist in the department of computer science and applied mathematics at Weizmann Institute of Science, Rehovot, Israel. His research interests are in mathematical statistics, machine learning, stochastic processes, and their applications in chemometrics and in signal processing.

Aryeh (Leonid) Kontorovich received his undergraduate degree in mathematics with a certificate in applied mathematics from Princeton University in 2001. His M.Sc. and Ph.D. are from Carnegie Mellon University, where he graduated in 2007. After a postdoctoral fellowship at the Weizmann Institute of Science, he joined the Computer Science department at Ben-Gurion University of the Negev in 2009 as an assistant professor; this is his current position. His research interests are mainly in machine learning, with a focus on probability, statistics, and automata theory.