

A Model Selection Rule for Sinusoids in White Gaussian Noise

Petar M. Djurić, *Member, IEEE*

Abstract—The model selection problem for sinusoidal signals has often been addressed by employing the Akaike information criterion (AIC) and the minimum description length principle (MDL). The popularity of these criteria partly stems from the intrinsically simple means by which they can be implemented. They can, however, produce misleading results if they are not carefully used. The AIC and MDL have a common form in that they comprise two terms, a data term and a penalty term. The data term quantifies the residuals of the model, and the penalty term reflects the desideratum of parsimony. While the data terms of the AIC and MDL are identical, the penalty terms are different. In most of the literature, the AIC and MDL penalties are, however, both obtained by apportioning an equal weight to each additional unknown parameter, be it phase, amplitude, or frequency. By contrast, in this paper, we demonstrate that the penalties associated with the amplitude and phase parameters should be weighted differently than the penalty attached to the frequencies. Following the Bayesian methodology, we derive a model selection criterion for sinusoidal signals in Gaussian noise which also contains the log-likelihood and the penalty terms. The simulation results disclose remarkable improvement in our selection rule over the commonly used MDL and AIC.

I. INTRODUCTION

MODEL SELECTION is an important area of research in signal processing, and its results are applied in many disciplines of science and engineering. Over the past two decades many of these problems have been addressed by utilizing two popular selection rules known as Akaike information criterion, or AIC [1], and minimum description length, or MDL [19]–[21]. For example, they have been applied in array processing to detect the number of sources that impinge on a passive sensor array [25], [26] in vibration analysis to decompose a nonstationary vibration record into stationary segments [7], and in image analysis to segment images [12], [13]. Many more examples can be cited from areas as diverse as econometrics, control theory, and psychometrics.

The widespread use of these rules is mainly due to their intrinsic simplicity. Neither requires the derivation of test statistics or selection of thresholds from statistical tables—tasks which are both rather difficult, particularly when there are several models to choose from, and/or the examined models are not nested. Instead, the AIC and MDL are simply applied by evaluating two terms, a data term and a penalty term.

Manuscript received September 27, 1994; revised January 12, 1996. This work was supported by the National Science Foundation under Award MIP-9110628. The associate editor coordinating the review of this paper and approving it for publication was Dr. Monique Fargues.

The author is with the Department of Electrical Engineering, State University of New York at Stony Brook, Stony Brook, NY 11794-2350 USA (e-mail: djuric@sbee.sunysb.edu).

Publisher Item Identifier S 1053-587X(96)04525-4.

These two terms are added together, and the model that yields the minimum sum is considered to be the best. In mathematical terms, if the set of competing models is denoted by $\mathcal{M}_k, k \in Z_Q$, where $Z_Q = \{0, 1, 2, \dots, Q-1\}$, and their associated parameters are θ_k , the model is selected according to [14]

$$\mathcal{M}_{\hat{k}} = \arg \min_{k \in Z_Q} \{-\mathcal{L}(\hat{\theta}_k) + p_c\} \quad (1)$$

where $(\hat{\theta}_k)$ is the log-likelihood function of the data, $\hat{\theta}_k$ is the maximum-likelihood (ML) estimate of the parameters θ_k , and p_c is the penalty of the criterion. The difference between the two criteria is in the penalty term p_c . The AIC imposes $p_c = p_{\text{AIC}} = d_k$, and the MDL, $p_c = p_{\text{MDL}} = d_k/2 \ln N/2$ [14]. Here d_k represents the number of parameters associated with the k th model, and N , the length of the observed data vector. These penalties obviously imply that each additional unknown parameter is equally weighted, regardless of its role in the model function.

In general, however, it is not appropriate to penalize for additional unknown parameters equally and without regard to their roles in the model. To establish the validity of this claim, we address a specific problem and show that the penalties are parameter dependent. In particular, we investigate a model selection problem where the competing models represent multiple sinusoids in white Gaussian noise. This problem has been of considerable interest since the beginning of the century [5], [23], and continues to attract the attention of many researchers [6], [8], [15], [16], [18]. When the selection is implemented by the AIC and MDL, the rules usually employed take the form of [10], [27], [28]

$$\hat{k}_{\text{AIC}} = \arg \min_{k \in Z_Q} \{-\mathcal{L}(\hat{\theta}_k) + 3k\} \quad (2)$$

$$\hat{k}_{\text{MDL}} = \arg \min_{k \in Z_Q} \left\{ -\mathcal{L}(\hat{\theta}_k) + \frac{3k}{2} \ln N \right\} \quad (3)$$

where \hat{k}_{AIC} and \hat{k}_{MDL} are the optimal numbers of signal components according to the AIC and MDL, respectively. Note that $d_k = 3k$ since each sinusoid is parameterized by its amplitude, phase, and frequency, and that the penalties due to each parameter are identical.

Recently, besides (2) and (3), similar selection rules have been proposed. In particular, in [24] a model selection rule for sinusoids in colored noise was proposed whose form is

$$\hat{k}_{\text{CON}} = \arg \min_{k \in Z_Q} \left\{ -\mathcal{L}(\hat{\theta}_k) + \frac{kc}{2} \ln N \right\} \quad (4)$$

where c is a constant that satisfies $c > \gamma$, with γ depending on the spectral density of the noise process. It was shown that (4) is strongly consistent. However, the use of this criterion is limited because the choice of c depends on the noise spectral density, which is usually unknown. It is important to point out that for white noise $\gamma = 2$, and $c > 2$. Noteworthy too are the facts that the results from [24] imply the strong consistency of (3), and when c is too small or too large, (4) tends to overestimate or underestimate the number of sinusoids, respectively [9].

For the colored noise problem another criterion was given in [9]. It is indirectly based on the MDL principle but is different from (3), and its evaluation is implemented in the frequency domain. There, the main idea is to model the continuous noise spectrum as constant over frequency bands of width $2\pi m/N$, where N is the number of observed data samples, and m is an integer. The obtained criterion has a penalty that has two terms, one depending on the number of sinusoids k , and the other being a function of m . The best signal model is obtained by optimizing the criterion simultaneously over k and m . For the same problem, one more MDL type criterion was suggested in [11], where the noise is modeled as an autoregression of unknown order.

Finally, in [4], a maximum *a posteriori* probability (MAP) criterion was proposed whose form is

$$\hat{k}_{\text{MAP}} = \arg \min_{k \in \mathbb{Z}_Q} \left\{ -\mathcal{L}(\hat{\theta}_k) + \frac{5k}{2} \ln N \right\}. \quad (5)$$

This criterion is identical to the MDL criteria from [9] and [11] when they are also applied to sinusoids in white noise. To make a distinction between the correct and the commonly used MDL criteria, we refer to (3) as the “MDL” criterion. A comparison of (5) with (2) and (3) clearly shows the differences in the penalties and that the penalty in (5) is the most stringent of the three.

In this paper, we derive (5) and show how the penalties are obtained for additional unknown parameters. In particular, we demonstrate that the penalization per additional unknown amplitude or phase is $\frac{1}{2} \ln N$ and per additional unknown frequency $\frac{3}{2} \ln N$, which makes a total of $\frac{5}{2} \ln N$ for every additional sinusoid. In using the Bayesian methodology, we apply asymptotic arguments and explain the derivation steps, which can readily be replicated in obtaining selection rules for other types of models. Aside from the derivation, we discuss important issues of the model selection problem, and we provide extensive simulation results that exhibit the performance of (5) and compare it to (3).

The article is organized as follows. In Section II the problem statement is given, followed by the definition and derivation

of our model selection criterion in Sections III and IV. Various points related to the selection problem are discussed in Section V, and a presentation of simulation results on the performance of the selection rules is given in Section VI. Finally, in Section VII some brief conclusions are drawn.

II. PROBLEM STATEMENT

Let \mathbf{y} be an observed vector of N real data samples. The elements of \mathbf{y} may represent samples of noise only,

$$y[n] = e[n], \quad n \in Z_N \quad (6)$$

or m superimposed sinusoids embedded in noise,

$$y[n] = \sum_{j=1}^m a_j \cos(\omega_j n + \phi_j) + e[n], \quad n \in Z_N. \quad (7)$$

Here $e[n]$ symbolizes a noise sample whereas a_j, ω_j and ϕ_j are the amplitude, radial frequency, and phase of the j th sinusoid, respectively. Z_N is the finite set of integers $\{0, 1, \dots, N-1\}$. Without loss of generality, we assume that

$$\begin{aligned} \omega_j &\neq \omega_l, & j &\neq l, & j, l &= 1, 2, \dots, m \\ \omega_j &\in (0, \pi), & j &= 1, 2, \dots, m. \end{aligned} \quad (8)$$

Alternatively, we may represent (7) as

$$y[n] = \sum_{j=1}^m (a_{jc} \cos(\omega_j n) + a_{js} \sin(\omega_j n)) + e[n], \quad n \in Z_N \quad (9)$$

where

$$a_{jc} = a_j \cos \phi_j, \quad a_{js} = -a_j \sin \phi_j, \quad j = 1, 2, \dots, m.$$

Finally, a concise depiction of (9) is provided in a vector-matrix form according to

$$\mathbf{y} = \mathbf{D}_{2m} \mathbf{a}_{2m} + \mathbf{e} \quad (10)$$

where \mathbf{D}_{2m} is an $N \times 2m$ matrix, shown at the bottom of the page, and the amplitude vector \mathbf{a}_{2m} is given by $\mathbf{a}_{2m}^T = [a_{1c} \ a_{1s} \ a_{2c} \ a_{2s} \ \dots \ a_{cm} \ a_{sm}]$, with T denoting transposition. The noise vector is assumed to be zero mean normal with density function

$$f(\mathbf{e}|\sigma) = (2\pi\sigma^2)^{-(N/2)} \exp \left\{ -\frac{1}{2\sigma^2} \mathbf{e}^T \mathbf{e} \right\}. \quad (11)$$

The number of sinusoids m and their parameters are unknown, as is the noise variance σ^2 . Given \mathbf{y} and the aforementioned assumptions, the objective is to determine the number of superimposed sinusoids in \mathbf{y} .

$$\mathbf{D}_{2m} = \begin{bmatrix} 1 & 0 & \dots & 1 & 0 \\ \cos(\omega_1) & \sin(\omega_1) & \dots & \cos(\omega_m) & \sin(\omega_m) \\ \cos(\omega_1 2) & \sin(\omega_1 2) & \dots & \cos(\omega_m 2) & \sin(\omega_m 2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cos(\omega_1(N-1)) & \sin(\omega_1(N-1)) & \dots & \cos(\omega_m(N-1)) & \sin(\omega_m(N-1)) \end{bmatrix}$$

III. MODEL SELECTION CRITERION

The formulated problem is clearly one of model selection (multiple hypotheses testing). Based on an accepted criterion, we want to choose the best model for the data from a set of predefined models. As an investigating strategy, the Bayesian methodology and MAP criterion are adopted as they allow for a consistent and optimal solution in a clearly defined way.

Let \mathcal{M}_k denote the hypothesized model, with k being the number of sinusoids in the data, and let $p(k)$ be the a priori probability of the k th model. Formally, when $k = 0$, the model \mathcal{M}_0 is represented by (6), and when $k > 0$, \mathcal{M}_k is given by

$$\mathcal{M}_k : \mathbf{y} = \mathbf{D}_{2k} \mathbf{a}_{2k} + \mathbf{e}. \quad (12)$$

It is assumed that there are Q competing models, where $Q > m$, and that each model is equiprobable. That is

$$p(k) = \frac{1}{Q}, \quad k \in Z_Q. \quad (13)$$

The MAP estimate of m will be the value of k that maximizes the *a posteriori* probability $p(k|\mathbf{y})$, where $k \in Z_Q$, or

$$\begin{aligned} \hat{m}_{\text{MAP}} &= \arg \max_{k \in Z_Q} \{p(k|\mathbf{y})\} \\ &= \arg \max_{k \in Z_Q} \left\{ \frac{f(\mathbf{y}|k) p(k)}{f(\mathbf{y})} \right\} \\ &= \arg \max_{k \in Z_Q} \left\{ \frac{f(\mathbf{y}|k)}{Q f(\mathbf{y})} \right\} \\ &= \arg \max_{k \in Z_Q} \{f(\mathbf{y}|k)\} \end{aligned} \quad (14)$$

where $f(\mathbf{y}|k)$ is the marginalized density of \mathbf{y} given there are k sinusoids in the data,¹ and $f(\mathbf{y})$ is the marginal density of the data. Note that under the assumption of (13), the maximization of the *a posteriori* probability becomes simply a maximization of the marginalized densities of the models, $f(\mathbf{y}|k)$. Q and $f(\mathbf{y})$ were dropped from the criterion because they are independent of k . Note that Q is a constant, and

$$f(\mathbf{y}) = \sum_{l=0}^{Q-1} p(l) f(\mathbf{y}|l) = \frac{1}{Q} \sum_{l=0}^{Q-1} f(\mathbf{y}|l). \quad (15)$$

The marginalized density $f(\mathbf{y}|k)$ can be obtained from

$$f(\mathbf{y}|k) = \begin{cases} \int_{\sigma} f(\mathbf{y}|k, \sigma) f(\sigma|k) d\sigma, & k = 0 \\ \int_{\Omega_k} \int_{\sigma} \int_{\mathcal{A}_{2k}} f(\mathbf{y}|k, \omega_k, \sigma, \mathbf{a}_{2k}) f(\omega_k, \sigma, \mathbf{a}_{2k}|k) d\mathbf{a}_{2k} d\sigma d\omega_k, & k > 0 \end{cases} \quad (16)$$

where Ω_k , σ , and \mathcal{A}_{2k} in (16) represent the parameter spaces of ω_k , σ , and \mathbf{a}_{2k} , respectively, while $f(\sigma|k)$ and $f(\omega_k, \sigma, \mathbf{a}_{2k}|k)$ are the *a priori* densities of σ and ω_k , σ , and \mathbf{a}_{2k} , respectively, given the number of sinusoids in the data is k . In the next section we present the derivation of $f(\mathbf{y}|k)$ and provide the final form of the selection criterion.

¹For fixed \mathbf{y} , $f(\mathbf{y}|k)$ can be viewed as the likelihood of k superimposed sinusoids in the data.

IV. DERIVATION OF THE CRITERION

First we derive $f(\mathbf{y}|k)$ for $k = 0$ and then for $k > 0$. When $k = 0$, the form of $f(\mathbf{y}|k)$ is determined from

$$f(\mathbf{y}|k) = \int_{\sigma} f(\mathbf{y}|k, \sigma) f(\sigma|k) d\sigma, \quad k = 0. \quad (17)$$

From (6) and (11), the first factor of the integrand can be expressed as

$$f(\mathbf{y}|k, \sigma) = (2\pi\sigma^2)^{-(N/2)} \exp \left\{ -\frac{1}{2\sigma^2} \mathbf{y}^T \mathbf{y} \right\}, \quad k = 0. \quad (18)$$

The second factor, $f(\sigma|k)$, is the prior density of the standard deviation of the noise which quantifies our initial knowledge of σ . If this knowledge is vague, one typically adopts the Jeffreys' prior [2]

$$f(\sigma|k) \propto \sigma^{-1} \quad (19)$$

where \propto signifies proportionality. Since we want to derive a model selection criterion that will be based on as little prior knowledge as possible, we adopt (19) as a prior. When (18) and (19) are substituted into (17), and with the use of the integral [2]

$$\int_0^{\infty} x^{-(p+1)} \exp \left\{ -\frac{u}{x^2} \right\} dx = \frac{1}{2} u^{-(p/2)} \Gamma \left(\frac{p}{2} \right), \quad u > 0; p > 0 \quad (20)$$

where $\Gamma(\cdot)$ is the standard Gamma function, it is found that

$$f(\mathbf{y}|k) \propto \Gamma \left(\frac{N}{2} \right) (\mathbf{y}^T \mathbf{y})^{-(N/2)}, \quad k = 0. \quad (21)$$

Now we derive the expression for $f(\mathbf{y}|k)$ when $k > 0$. In order to do so, the following integrals need to be solved:

$$f(\mathbf{y}|k) = \int_{\Omega_k} \int_{\sigma} \int_{\mathcal{A}_k} f(\mathbf{y}|k, \omega_k, \sigma, \mathbf{a}_{2k}) f(\omega_k, \sigma, \mathbf{a}_{2k}|k) d\mathbf{a}_{2k} d\sigma d\omega_k, \quad k > 0 \quad (22)$$

where $f(\omega_k, \sigma, \mathbf{a}_{2k}|k)$ is the prior of the model parameters. Again, we adopt a vague prior whose form is

$$f(\omega_k, \sigma, \mathbf{a}_{2k}|k) \propto \sigma^{-1}. \quad (23)$$

First, the innermost integral of (22) is solved yielding $f(\mathbf{y}, \omega_k, \sigma|k)$. From (10) and (11), it is deduced that the first factor of the integrand in (22) is

$$f(\mathbf{y}|k, \omega_k, \sigma, \mathbf{a}_{2k}) = (2\pi\sigma^2)^{-(N/2)} \cdot \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{D}_{2k} \mathbf{a}_{2k})^T (\mathbf{y} - \mathbf{D}_{2k} \mathbf{a}_{2k}) \right\}. \quad (24)$$

Applying the prior in (23), one readily obtains

$$f(\mathbf{y}, \omega_k, \sigma|k) \propto |\mathbf{D}_{2k}^T \mathbf{D}_{2k}|^{-(1/2)} \sigma^{-(N-2k+1)} \cdot \exp \left\{ -\frac{1}{\sigma^2} \mathbf{y}^T \mathbf{P}_{2k}^{\perp} \mathbf{y} \right\} \quad (25)$$

where \mathbf{P}_{2k}^{\perp} is an $N \times N$ projection matrix defined by

$$\mathbf{P}_{2k}^{\perp} = \mathbf{I} - \mathbf{D}_{2k} (\mathbf{D}_{2k}^T \mathbf{D}_{2k})^{-1} \mathbf{D}_{2k}^T. \quad (26)$$

Next we solve the second integral

$$\begin{aligned} f(\mathbf{y}, \omega_k | k) &= \int_{\sigma} f(\mathbf{y}, \sigma, \omega_k | k) d\sigma \\ &\propto \int_{\sigma} \sigma^{-(N-2k+1)} |D_{2k}^T D_{2k}|^{-(1/2)} \\ &\quad \cdot \exp \left\{ -\frac{1}{\sigma^2} \mathbf{y}^T P_{2k}^{\perp} \mathbf{y} \right\} d\sigma. \end{aligned} \quad (27)$$

This integral is of the same form as (20). It results in

$$\begin{aligned} f(\mathbf{y}, \omega_k | k) &\propto \Gamma \left(\frac{N-2k}{2} \right) |D_{2k}^T D_{2k}|^{-(1/2)} \\ &\quad \cdot (\mathbf{y}^T P_{2k}^{\perp} \mathbf{y})^{-((N-2k)/2)}. \end{aligned} \quad (28)$$

Finally, we want to integrate out the frequencies ω_k , i.e., find

$$f(\mathbf{y} | k) = \int_{\Omega_k} f(\mathbf{y}, \omega_k | k) d\omega_k. \quad (29)$$

As the frequencies are nonlinear parameters, it is very difficult to obtain an analytical solution, and we thus resort to approximations. If $f(\mathbf{y}, \omega_k | k)$ is Taylor expanded around the ML estimate of ω_k , one can write

$$\begin{aligned} f(\mathbf{y}, \omega_k | k) &= \exp \{ \ln f(\mathbf{y}, \omega_k | k) \} \\ &\simeq \exp \{ \ln f(\mathbf{y} | \hat{\omega}_k, k) \\ &\quad - \frac{1}{2} (\omega_k - \hat{\omega}_k)^T \hat{\nabla}_k (\omega_k - \hat{\omega}_k) \} \end{aligned} \quad (30)$$

where $\hat{\omega}_k$ is the ML estimated of ω_k , and $\hat{\nabla}_k$ is the Hessian of $-\ln f(\mathbf{y}, \omega_k | k)$ evaluated at $\omega_k = \hat{\omega}_k$, or

$$\hat{\nabla}_k = \left[-\frac{\partial^2 \ln f(\mathbf{y}, \omega_k | k)}{\partial \omega_i \partial \omega_j} \right]_{\omega_k} = \hat{\omega}_k. \quad (31)$$

If (30) is substituted in (29), the integration becomes straightforward. The result can be expressed as

$$f(\mathbf{y} | k) \propto f(\mathbf{y} | \hat{\omega}_k, k) |\hat{\nabla}_k|^{-(1/2)}. \quad (32)$$

We would like to simplify (32) so that the evaluation of the Hessian may be avoided. To this end, we apply the following proposition.

Proposition: When N is large, the Hessian of $-\ln f(\mathbf{y}, \omega_k | k)$ can be written as

$$\hat{\nabla}_k = N^3 \hat{\mathbf{R}}_k \quad (33)$$

where $\hat{\mathbf{R}}_k$ is a $k \times k$ positive definite matrix whose determinant is of order $O(1)$.

The proof is not difficult but rather tedious. The following results will be used as we proceed [22]:

$$\begin{aligned} \frac{1}{N^{k+1}} \sum_{n=0}^{N-1} n^k \cos(\omega n) &\simeq 0 \\ \frac{1}{N^{k+1}} \sum_{n=0}^{N-1} n^k \sin(\omega n) &\simeq 0. \end{aligned} \quad (34)$$

In addition, if \mathbf{A} is a matrix whose entries are functions of ω_k , one can write [17]

$$\frac{d}{d\omega_i} \mathbf{A}^{-1} = -\mathbf{A}^{-1} \left(\frac{d}{d\omega_i} \mathbf{A} \right) \mathbf{A}^{-1}. \quad (35)$$

From (28), we have

$$\begin{aligned} -\ln f(\mathbf{y}, \omega_k | k) &\simeq C + \frac{1}{2} \ln |D_{2k}^T D_{2k}| \\ &\quad + \frac{N-2k}{2} \ln(\mathbf{y}^T P_{2k}^{\perp} \mathbf{y}) - \ln \Gamma \left(\frac{N-2k}{2} \right) \end{aligned} \quad (36)$$

where C is a constant. Upon taking the first derivative with respect to ω_i , one obtains

$$-\frac{\partial}{\partial \omega_i} \ln f(\mathbf{y}, \omega_k | k) \simeq \frac{N-2k}{2} (\mathbf{y}^T P_{2k}^{\perp} \mathbf{y})^{-1} \left(\frac{\partial}{\partial \omega_i} \mathbf{y}^T P_{2k}^{\perp} \mathbf{y} \right) \quad (37)$$

where the derivative with respect to the determinant $|D_{2k}^T D_{2k}|$ has been neglected since the determinant is almost constant over Ω_k when the conditions specified in (8) hold.

Now, after taking the partial derivative of (37) with respect to ω_j , we find

$$\begin{aligned} &-\frac{\partial^2}{\partial \omega_i \partial \omega_j} \ln f(\mathbf{y}, \omega_k | k) \\ &\simeq -\frac{N-2k}{2} (\mathbf{y}^T P_{2k}^{\perp} \mathbf{y})^{-2} \left(\frac{\partial}{\partial \omega_i} \mathbf{y}^T P_{2k}^{\perp} \mathbf{y} \right) \\ &\quad \cdot \left(\frac{\partial}{\partial \omega_j} \mathbf{y}^T P_{2k}^{\perp} \mathbf{y} \right) + \frac{N-2k}{2} (\mathbf{y}^T P_{2k}^{\perp} \mathbf{y})^{-1} \\ &\quad \cdot \left(\frac{\partial^2}{\partial \omega_i \partial \omega_j} \mathbf{y}^T P_{2k}^{\perp} \mathbf{y} \right). \end{aligned} \quad (38)$$

With (34) and (35) it can be shown that

$$\frac{\partial}{\partial \omega_i} \mathbf{y}^T P_{2k}^{\perp} \mathbf{y} = O(N) \quad (39)$$

$$\frac{\partial^2}{\partial \omega_i \partial \omega_j} \mathbf{y}^T P_{2k}^{\perp} \mathbf{y} = \begin{cases} O(N^3) & i = j \\ O(N^2) & i \neq j \end{cases} \quad (40)$$

$$\mathbf{y}^T P_{2k}^{\perp} \mathbf{y} = O(N). \quad (41)$$

If the results (39)–(41) are applied to (38), the claim in the proposition directly follows.

Returning to (32) and using the main result of the proposition as well as (28), we can express the marginal density $f(\mathbf{y} | k)$ by

$$\begin{aligned} f(\mathbf{y} | k) &\propto \Gamma \left(\frac{N-2k}{2} \right) |\hat{D}_{2k}^T \hat{D}_{2k}|^{-(1/2)} \\ &\quad \cdot (\mathbf{y}^T \hat{P}_{2k}^{\perp} \mathbf{y})^{-((N-2k)/2)} N^{-(3k/2)} |\hat{\mathbf{R}}_k|^{-(1/2)} \end{aligned} \quad (42)$$

where \hat{D}_{2k} and \hat{P}_{2k}^{\perp} are the matrices D_{2k} and P_{2k}^{\perp} , respectively, with ω_k substituted by $\hat{\omega}_k$. With this result, the final form of the model selection criterion can readily be established as

$$\begin{aligned} \hat{m}_{\text{MAP}} &= \arg \min_{k \in \mathcal{Z}_Q} \{ -\ln f(\mathbf{y} | k) \} \\ &= \arg \min_{k \in \mathcal{Z}_Q} \left\{ \frac{N-2k}{2} \ln(\mathbf{y}^T P_{2k}^{\perp} \mathbf{y}) \right\} \end{aligned} \quad (43)$$

$$+ \frac{1}{2} \ln |D_{2k}^T D_{2k}| - \ln \Gamma\left(\frac{N-2k}{2}\right) + \frac{3k}{2} \ln N \Big\} \quad (44)$$

where in the last expression we have dropped all the terms of order $O(1)$, and we have assumed that N is large. We further simplify (44) by observing that

$$|D_{2k}^T D_{2k}| = O(N^{2k}) \quad (45)$$

$$\mathbf{y}^T \mathbf{P}_{2k}^\perp \mathbf{y} = O(N) \quad (46)$$

and that the relative contribution of the Gamma function in (44) can asymptotically be replaced by $k \ln N$. To understand the latter approximation, divide $-\ln f(\mathbf{y}|k)$ in (44) by $\Gamma(N/2)$. This does not change the criterion because $\Gamma(N/2)$ is not a function of k , and the resulting penalty readily follows. With these approximations, one can obtain the final form of the criterion as

$$\hat{m}_{\text{MAP}} = \arg \min_{k \in Z_Q} \left\{ \frac{N}{2} \ln(\mathbf{y}^T \mathbf{P}_{2k}^\perp \mathbf{y}) + \frac{5k}{2} \ln N \right\}. \quad (47)$$

V. DISCUSSION

In this section we go on with discussion of several important issues related to (47).

- First we provide an interpretation of the result in (47). Note that the AIC and “MDL” rules in (2) and (3) can be written as

$$\hat{m}_{\text{AIC}} = \arg \min_{k \in Z_Q} \left\{ \frac{N}{2} \ln(\mathbf{y}^T \mathbf{P}_{2k}^\perp \mathbf{y}) + 3k \right\} \quad (48)$$

$$\hat{m}_{\text{MDL}} = \arg \min_{k \in Z_Q} \left\{ \frac{N}{2} \ln(\mathbf{y}^T \mathbf{P}_{2k}^\perp \mathbf{y}) + \frac{3k}{2} \ln N \right\} \quad (49)$$

which allow for a direct comparison with (47). Needless to say, the MAP criterion imposes a stricter penalty than either the AIC or MDL. Even more interesting is the fact that the MAP rule penalizes for the extra parameters with penalties that depend on the type of parameters used in the models, which is in contrast to the AIC and MDL. In particular, from our derivation, it is easy to deduce that the penalization for each amplitude or phase is $\frac{1}{2} \ln N$, and for each frequency $\frac{3}{2} \ln N$. The different penalties can be interpreted as follows. Suppose that the MAP criterion can be approximated as the minimizer of $-\ln(f(\mathbf{y}|\hat{\theta}_k, \hat{\sigma}, k)f(\theta_k))$ over k , where $\hat{\theta}_k$ and $\hat{\sigma}$ denote the ML estimates of the signal parameters and the noise variance, respectively, and $f(\theta_k)$ is the prior of the signal parameters which is chosen to be a constant that depends somehow on θ_k . Suppose also that the ML estimates of the signal parameters are obtained by a grid search. Once we decide about the grid size, the prior of the signal parameters is selected to be uniform over the grid. An open and sensitive question is how to choose the grid size. A reasonable choice is to adopt a grid which is a function of the data record length N , or more precisely, a function of the estimation accuracy with which $\hat{\theta}_k$ are obtained. Now, recall that the Cramer–Rao bound of the sinusoidal parameters is proportional to $1/N$ for the amplitudes and phases and $1/N^3$ for the frequencies. Then,

if the priors of the unknown signal parameters are chosen as proposed, i.e., for each amplitude and phase they are equal to $1/N$ and for each frequency $1/N^3$, the criterion function in (47) follows immediately from $-\ln(f(\mathbf{y}|\hat{\theta}_k, \hat{\sigma}, k)f(\theta_k))$.

- The result in (47) can also be obtained by a different derivation. If we expand the density $f(\mathbf{y}, \omega_k, \sigma, \mathbf{a}_{2k}|k)$ in a Taylor expansion around the ML estimates of the parameters, ignore the prior due to the asymptotic assumptions, and approximate the determinant of the Hessian by neglecting all the terms that are not a function of N , we obtain the same rule as (47).

- It can be shown that a similar rule holds for complex sinusoids. In [3] we have investigated this problem by exploiting the concept of predictive densities which is yet another way to find (47). To obtain the rule for complex data that is equivalent to (47), upon deriving the final results in [3], one has to make an additional step which includes asymptotical approximations.

- It may be tempting to use an algorithm which would avoid the evaluation of the criterion function for every model [24]. In particular, we might want to rely on the following phenomenon: As the model complexity increases, the criterion function decreases until it reaches the minimum at the correct model, and then increases as the models becomes more complex. In other words, if the criterion function is

$$J_k = \frac{N}{2} \ln(\mathbf{y}^T \mathbf{P}_{2k}^\perp \mathbf{y}) + \frac{5k}{2} \ln N \quad (50)$$

than we might propose to choose the best model according to

$$\tilde{m} = \min\{k: J_k \leq J_{k+1}\} \quad k \in Q. \quad (51)$$

The last expression is attractive because we start with evaluating the criterion function of the simplest (noise) model and then escalate the complexity of the models by sequentially adding one more sinusoid until the criterion functions of J_k and J_{k+1} meet the condition $J_k \leq J_{k+1}$. Obviously, by implementing the MAP rule with this strategy, we avoid the examination of all the models. This is, however, not recommended because we cannot guarantee that the first minimum of the criterion function is achieved at the correct model. To show this, consider the following experiment. Let \mathbf{y} be of fixed length N with a fixed noise sequence \mathbf{e} . If \mathbf{y} contains sinusoids, using (51), we will incorrectly choose the noise model \mathcal{M}_0 if

$$\ln \frac{\mathbf{y}^T \mathbf{y}}{\mathbf{y}^T \mathbf{P}_2^\perp \mathbf{y}} < \frac{5}{2} \ln N \triangleq \gamma. \quad (52)$$

Suppose that for some m (52) is not satisfied. Now, we start adding new sinusoids to the data, which will imply a steady decrease of the term on the left side of (52). After a sufficient number of sinusoids has been added, the value of the logarithm on the left side of the inequality will drop below γ , and we will choose the noise only model. This will happen when the data actually contain many sinusoids, thus incurring a gross error in

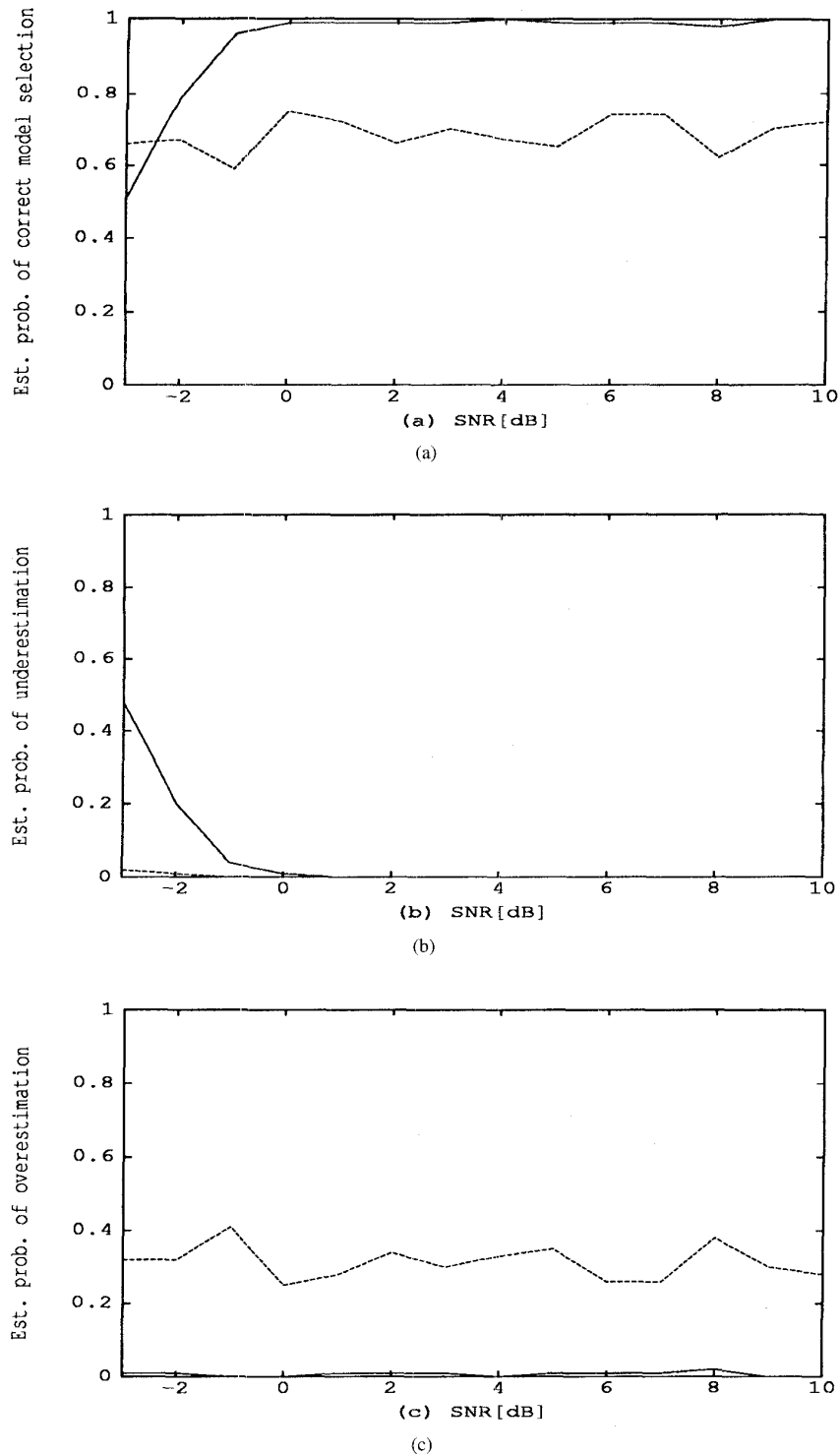


Fig. 1. (a) Estimated probabilities of correct selection. (b) Model underestimation. (c) Model overestimation. The solid and the dashed lines represent the MAP and the MDL performances, respectively.

the selection. In conclusion, one should not apply (51) when k is increasing. If we apply a modified rule by starting with the most complex model with k steadily decreasing, there still exists the possibility of a local minimum occurring before we reach the global minimum, and that may be a consequence of

noise modeling. Thus the use of (51) or an equivalent rule is not encouraged.

- Finally, the results of (47) depend critically on the quality of the ML estimates. We found that if an iterative algorithm is employed for parameter estimation, the final results can

TABLE I

PERFORMANCE COMPARISON OF MAP AND MDL CRITERIA FOR VARIOUS SNR'S WHEN THERE ARE TWO SINUSOIDS OF DIFFERENT POWERS. SNR IS DEFINED USING THE WEAKER SINUSOID (5 dB LARGER FOR THE STRONGER SINUSOID). FOR EACH SNR THERE WERE 100 TRIALS. ENTRIES REPRESENT NUMBER OF TIMES A PARTICULAR MODEL WAS SELECTED OUT OF 100 TRIALS. THE CORRECT MODEL IS COMPRISED OF TWO SINUSOIDS ($k = 2$)

		k=0	k=1	k=2	k=3	k=4	k=5
-3 dB	'MDL'	0	3	63	23	6	5
	MAP	0	28	70	2	0	0
-2 dB	'MDL'	0	0	64	20	9	7
	MAP	0	9	88	3	0	0
-1 dB	'MDL'	0	0	72	17	6	5
	MAP	0	3	96	1	0	0
0 dB	'MDL'	0	0	81	10	6	3
	MAP	0	0	99	1	0	0
1 dB	'MDL'	0	0	73	15	8	4
	MAP	0	0	98	2	0	0
2 dB	'MDL'	0	0	73	15	8	4
	MAP	0	0	99	1	0	0
3 dB	'MDL'	0	0	77	16	3	4
	MAP	0	0	100	0	0	0
4 dB	'MDL'	0	0	66	19	8	7
	MAP	0	0	100	0	0	0
5 dB	'MDL'	0	0	74	17	3	6
	MAP	0	0	99	1	0	0

strongly depend on the selected convergence criterion. For example, if convergence in the estimates is declared prematurely, despite some moderate differences in their values from one iteration to another, poor selections can be obtained.

VI. SIMULATION RESULTS

In this section we present experimental results that demonstrate the performance of the selection rules. Three experiments were considered. In the first, the data were generated according to

$$y[n] = a_1 \cos(\omega_1 n + \phi_1) + a_2 \cos(\omega_2 n + \phi_2) + e[n], \quad n \in Z_N \quad (53)$$

where

$$a_1 = a_2 = \sqrt{20}, \quad \phi_1 = 0 \text{ rad}, \quad \phi_2 = \pi/4 \text{ rad}, \\ \omega_1 = 2\pi \cdot 0.2, \quad \omega_2 = 2\pi \left(0.2 + \frac{1}{N}\right),$$

and $N = 64$. Throughout the experiment, the SNR defined by

$$\text{SNR} = 10 \log_{10} \frac{a_i^2}{2\sigma^2} \quad (54)$$

was varied from -3 to 10 dB in steps of 1 dB. The noise sequences were generated according to a Gaussian density function given by (11) with σ^2 appropriately chosen to yield the required SNR. For each SNR, there were 100 trials. The maximum number of sinusoids was assumed to be 5. Consequently, the selection rules had to choose the best model out of six nested models. The method for the frequency ML estimation was the one from [22]. The initial estimates were obtained by employing periodograms and notched periodograms [10].

The results of the simulations for the MAP and the MDL rules are shown in Fig. 1. In Fig. 1(a), we present the curves of correct model selection, and in Fig. 1(b) and 1(c), the curves of overestimation and underestimation, respectively. The AIC

TABLE II

PERFORMANCE COMPARISON OF MAP AND MDL CRITERIA FOR VARIOUS DATA RECORD LENGTHS. FOR EACH N THERE WERE 100 TRIALS. THE ENTRIES REPRESENT THE NUMBER OF TIMES A PARTICULAR MODEL WAS SELECTED OUT OF 100 TRIALS. THE CORRECT MODEL IS COMPRISED OF THREE SINUSOIDS ($k = 3$)

		k=0	k=1	k=2	k=3	k=4	k=5
N=64	'MDL'	0	0	0	63	21	16
	MAP	0	0	0	95	5	0
N=72	'MDL'	0	0	0	69	17	14
	MAP	0	0	0	100	0	0
N=80	'MDL'	0	0	0	70	16	14
	MAP	0	0	0	100	0	0
N=88	'MDL'	0	0	0	68	25	7
	MAP	0	0	0	100	0	0
N=96	'MDL'	0	0	0	74	19	7
	MAP	0	0	0	99	1	0
N=128	'MDL'	0	0	0	80	18	2
	MAP	0	0	0	100	0	0

always overestimated the model, most of the time choosing the most complex model which represented five sinusoids in noise, and, therefore, we did not include its results in the figures. The MAP rule had an excellent performance for SNR's above -1 dB, whereas the MDL was always yielding correct results ranging between 59 and 75 times out of 100. The deterioration in performance below -1 dB for the MAP estimator is not surprising because the ML estimator does not provide good estimates of the frequencies in that range. The performance of the MDL starts to deteriorate similarly for slightly lower SNR's. Again, the results in that range are questionable because the frequency estimates are not reliable.

In the second experiment we kept all the parameters from the first experiment the same except that we changed the amplitude of the second sinusoid to $a_2 = \sqrt{6.3246}$. The SNR was again varied, this time between 5 dB and -3 dB (for the second sinusoid, or equivalently between 10 and 2 dB for the first sinusoid). The results are given in Table I. Practically, identical performance was obtained as in the first experiment.

Finally, in the third experiment the data represented three closely spaced sinusoids. The data y were generated by

$$y[n] = a_1 \cos(\omega_1 n + \phi_1) + a_2 \cos(\omega_2 n + \phi_2) + a_3 \cos(\omega_3 n + \phi_3) + e[n], \quad n \in Z_N \quad (55)$$

where

$$a_1 = a_3 = \sqrt{20}, \quad a_2 = \sqrt{6.3246}, \quad \phi_1 = 0 \text{ rad}, \\ \phi_2 = \pi/4 \text{ rad}, \quad \phi_3 = \pi/3, \quad \omega_1 = 2\pi \cdot 0.2, \\ \omega_2 = 2\pi \left(0.2 + \frac{1}{N}\right), \quad \text{and} \quad \omega_3 = 2\pi \left(0.2 + \frac{2}{N}\right).$$

The SNR for the first and third sinusoids was 10 dB (and for the second 5 dB). The number of samples was varied and in the first 100 trials it was 64. Then it was increased to 72, 80, 88, 96, and 128, and for each data length, there were 100 trials. The results are shown in Table II.

Again, the MAP rule had excellent performance. The results of the MDL were similar as in the previous experiment showing a large percentage of overestimated models. However, as the number of samples was doubled from 64 to 128, its performance improved.

VII. CONCLUSION

In this paper, we proposed a model selection rule for sinusoids in Gaussian noise based on the MAP criterion. The rule has a log-likelihood and penalty terms, just like the AIC and MDL. The MAP criterion is different from the other two in the penalty term which is equal to $5k/2 \ln N$, where k is the number of sinusoids and N the length of the observed data. The derivation of the MAP criterion showed that the penalization due to additional unknown parameters depends on the parameters of the models. We penalize more for the unknown frequencies than for the unknown amplitudes or phases. The simulation results showed remarkable improvement in performance of our criterion over the MDL and AIC.

REFERENCES

- [1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.
- [2] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. New York: Wiley, 1992.
- [3] P. M. Djurić, "Simultaneous detection and frequency estimation of sinusoidal signals," in *Proc. ICASSP*, vol. IV, pp. 53–56, 1993.
- [4] ———, "Model selection based on asymptotic Bayes theory," in *Proc. 7th SP Workshop Stat. Signal Array Processing*, pp. 7–10, 1994.
- [5] R. A. Fisher, "Tests of significance in harmonic analysis," in *Proc. Royal Soc.*, London, U.K., Ser. A, vol. 125, pp. 54–59, 1929.
- [6] J.-J. Fuchs, "Estimating the number of sinusoids in additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1846–1853, 1988.
- [7] W. Gersch and T. Brotherton, "Estimation of stationary structural system parameters from nonstationary random vibration data: A locally stationary model method," *J. Sound Vibration*, vol. 81, pp. 215–227, 1982.
- [8] E. J. Hannan, "Testing for a jump in the spectral function," *J. Royal Stat. Soc.*, Ser. B, vol. 23, pp. 394–404, 1961.
- [9] ———, "Determining the number of jumps in a spectrum," in *Developments in Time Series Analysis*, T. Subba Rao, Ed. London, UK: Chapman and Hall, 1993, pp. 127–138.
- [10] J.-K. Hwang and Y.-C. Chen, "A combined detection-estimation algorithm for the harmonic retrieval problem," *Signal Processing*, vol. 30, pp. 177–197, 1993.
- [11] L. Kavalieris and E. J. Hannan, "Determining the number of terms in a trigonometric regression," *J. Time Series Anal.*, vol. 15, pp. 613–625, 1994.
- [12] Y. G. Leclerc, "Constructing simple stable descriptions for image partitioning," *Int. J. Comput. Vision*, vol. 3, pp. 73–102, 1989.
- [13] Z. Liang, "Tissue classification and segmentation of MR images," *IEEE Med. Biol. Mag.*, pp. 81–85, 1993.
- [14] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- [15] B. G. Quinn, "Testing for the presence of sinusoidal components," in *Time Series and Allied Processes, Papers in Honor of E. J. Hannan*. J. Gani and M. Priestley, Eds. Sheffield, U.K.: Applied Probability Trust, pp. 201–210, 1986.
- [16] ———, "Estimating the number of terms in a sinusoidal regression," *J. Time Series Anal.*, vol. 10, pp. 71–75, 1989.
- [17] L. Råde and B. Westergren, *Beta Mathematics Handbook*. Boca Raton, FL: CRC Press, 1992.
- [18] V. Umapathi 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. Signal Processing*, vol. 41, pp. 2872–2881, 1993.
- [19] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 468–478, 1978.
- [20] ———, *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1989.
- [21] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.
- [22] P. Stoica, R. L. Moses, B. Friedlander, and T. Soderstrom, "Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 378–392, 1989.
- [23] G. T. Walker, "Correlation in seasonal variation of weather. On the criterion for the reality of relationships of periodicities," *Memo. Indian Meteorol. Dept.*, pp. 13–15, 1914.
- [24] X. Wang, "An AIC type estimator for the number of cosinusoids," *J. Time Series Anal.*, vol. 14, pp. 433–440, 1993.
- [25] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 387–392, 1985.
- [26] M. Wax, "Detection and localization of multiple sources via the stochastic signal model," *IEEE Trans. Signal Processing*, vol. 39, pp. 2450–2456, 1991.
- [27] S. F. Yau and Y. Bresler, "Maximum likelihood parameter estimation of superimposed signals by dynamic programming," *IEEE Trans. Signal Processing*, vol. 41, pp. 804–820, 1993.
- [28] C. J. Ying, L. C. Potter, and R. L. Moses, "On model order determination for complex exponential signals: performance of an FFT-initialized ML algorithm," in *Proc. 7th SP Workshop Stat. Signal Array Processing*, pp. 43–46, 1994.
- [29] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of the number of sinusoids in presence of white noise," *J. Multivariate Anal.*, vol. 20, pp. 1–25, 1986.



Petar M. Djurić (S'86–M'91) received the B.S. and M.S. degrees from the University of Belgrade, Yugoslavia, in 1981 and 1986, respectively, and the Ph.D. degree from the University of Rhode Island, in 1990, all in electrical engineering.

From 1981 to 1986, he was with the Institute of Nuclear Sciences—Vinča, Computer Systems Design Department, where he conducted research in digital and statistical signal processing, communications, and pattern recognition. From 1986 to 1990, he was a Research and Teaching Assistant in the Department of Electrical Engineering at the University of Rhode Island. He joined the Department of Electrical Engineering at the State University of New York at Stony Brook, in 1990, where he is currently an Assistant Professor. His main research interests are in statistical signal processing and signal modeling.

Dr. Djurić is a member the American Statistical Association. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.