Asymptotic MAP Criteria for Model Selection

Petar M. Djurić, *Member, IEEE*

***Abstract*—The two most popular model selection rules in the signal processing literature have been the Akaike's criterion AIC and the Rissanen's principle of minimum description length MDL. These rules are similar in form in that they both consist of data and penalty terms. Their data terms are identical, but the penalties are different, the MDL being more stringent toward overparameterization. The AIC penalizes for each additional model parameter with an equal incremental amount of penalty, regardless of the parameter's role in the model. In most of the literature on model selection, the MDL appears in a form that also suggests equal penalty for every unknown parameter. To this MDL criterion, we refer to as the *naive* MDL. In this paper, we show that identical penalization for every parameter is not appropriate and that the penalty has to depend on the model structure and type of model parameters. The approach to showing this is Bayesian, and it relies on large sample theory. We derive *maximum a posteriori* (MAP) rules for several different families of competing models and obtain forms that are similar to the AIC and the naive MDL. For some families, however, we find that the derived penalties are different. In those cases, our extensive simulations show that the MAP rule outperforms the AIC and the naive MDL.**

I. Introduction

A COMMON task in science and engineering is the selection of a model from a set of competing models. In signal processing, this problem is of great interest because the observed data are usually distorted and comprised of unknown number of signal components or even unknown types of signals. Then, one is faced with the problem of choosing a model for the data that describes them best in some predefined sense. Many examples can be found in a variety of areas such as underwater acoustics, vibration analysis, and medical imaging. The model selection is clearly a multiple hypothesis testing problem for which an optimal solution in the classical sense does not exist.

Researchers in signal and image processing often address this problem by utilizing two popular model selection rules: the AIC [1] and MDL [17]. More recent references where they are applied include [9], [13], [16], and [20]–[23]. The two criteria were derived under asymptotical assumptions using information and coding theoretic reasoning.

The AIC and MDL consist of two terms: a data term and a penalty term. As the model complexity increases, the data term usually decreases, whereas the penalty term always increases. The best model is the one that yields the minimum value of the criterion. A common fallacy of the AIC is that it penalizes for overmodeling the data independently of the "type" of parameters (amplitude, phase, frequency, damping factor, time delay, etc.) used in the models. For example, if we compare two models, with the same number of unknown parameters, their penalties will be the same, despite the difference in their structures. In most of the signal processing literature on model selection, the MDL shares the same feature, that is, every model parameter contributes to the model's overall penalty with identical amount. We refer to this MDL criterion as the *naive* MDL.

In this paper, we show that the penalization strongly depends on the types of models that are being used and that, in general, it cannot be simply obtained by counting the number of unknown parameters. This implies that one should *not* use the AIC or the naive MDL without careful examination of the models under investigation. It should be noted that many researchers have recognized the poor performance of these criteria in certain scenarios and have tried to improve them by modifying the penalties in a more or less *ad hoc* fashion [9], [13].

To obtain the new rules, we used Bayes' theory and large sample approximations [14], [19]. We followed the derivations in [10], [12], [15], and [19] and carefully investigated the results of five different families of models. They include models of sinusoidal signals with known frequencies, polynomials, autoregressions, models of sinusoidal signals with unknown frequencies, and models of chirp-type signals. In some of these cases, we obtained rules with different penalties from those of the AIC and the naive MDL. The computer simulations show that in those cases, the MAP rule has the best performance.

The paper is organized as follows. In Section II, we formulate the problem, and in Section III, we briefly outline the MAP criterion and present the general solution. Then, in Section IV, we exploit it on five different sets of models, and in Section V, we discuss some relevant issues. We present some simulation results that show the performance of the rules proposed here as well as the performance of the AIC and the naive MDL in Section VI. Finally, we conclude the paper with Section VII providing some final remarks.

II. Formulation of the Model Seletion Problem

A data vector $\mathbf{y}$ of length N is observed. There are q candidate models for $\mathbf{y}$ whose generic forms are given by

$$\mathcal{M}_k\colon \mathbf{y} = g_k(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta}), \quad k \in Z_q \tag{1}$$

where $g_k(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta})$ is a vector function that represents the kth model, $Z_q = \{0, 1, \ldots, q-1\}$, $\mathbf{e}$ is a noise vector, and $\boldsymbol{\theta}$ is a vector of model parameters taking values in the parameter

Manuscript received March 28, 1995; revised March 6, 1998. This work was supported by the National Science Foundation under Awards MIP-9110628 and MIP-9506743. The associate editor coordinating the review of this paper and approving it for publication was Dr. Geoffrey C. Orsak.

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(98)07073-1.

space $\boldsymbol{\Theta}_k \subseteq \mathbf{R}^{m_k}$, with m_k denoting the length of the vector $\boldsymbol{\theta}$. For example, if the kth model for $k \geq 1$ represents k sinusoids in additive noise, $g_k(\cdot)$ is given by

$$\mathcal{M}_k\text{: } g_k(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta}) = \mathbf{s}_k(\boldsymbol{\theta}) + \mathbf{e} \tag{2}$$

where $\mathbf{s}_k(\boldsymbol{\theta})$ is the signal vector that represents k superimposed sinusoids, $\boldsymbol{\theta}$ is the vector of signal parameters (the amplitudes, phases, and frequences of the sinusoids), $m_k = 3k$, and $\mathbf{e}$ is the additive noise. The probability density function of the noise is parameterized by $\boldsymbol{\phi}$, and it will be denoted by $f_k(\mathbf{e} \mid \boldsymbol{\phi})$, where $\boldsymbol{\phi} \in \boldsymbol{\Phi}_k$, $\boldsymbol{\Phi}_k \subseteq \mathbf{R}^{n_k}$, and n_k is the number of parameters necessary to describe the density $f_k(\mathbf{e} \mid \boldsymbol{\phi})$. The functional forms of $f_k(\mathbf{e} \mid \boldsymbol{\phi})$ and $g_k(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta})$ are assumed known, but the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are unknown.

The problem that we address here is the following: Given the observed vector $\mathbf{y}$ and the set of candidate models $\{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{q-1}\}$, choose the best model that describes the data $\mathbf{y}$, where the best model is the one that has the *maximum a posteriori* probability. Without loss of generality, it is supposed that the prior probability of each model is $p(\mathcal{M}_k) = \frac{1}{q}$, that is, there is no prior preference toward any of the models. In addition, it is tacitly assumed that one of the examined models may be the noise model whose function $g(\cdot)$ is given by

$$g(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta}) = \mathbf{e}. \tag{3}$$

It should be noted that the models in the set do not need to be related or nested.[1]

Before proceeding, we recall the forms of the AIC and naive MDL selection rules. They are given by

$$\text{AIC: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \{-2 \ln f(\mathbf{y} \mid \hat{\boldsymbol{\psi}}, \mathcal{M}_k) + 2d_k\}. \tag{4}$$

and

$$\text{MDL: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ -\ln f(\mathbf{y} \mid \hat{\boldsymbol{\psi}}, \mathcal{M}_k) + \frac{d_k}{2} \ln N \right\} \tag{5}$$

where $\mathcal{M}_s$ is the selected model, $s \in Z_q$, $\boldsymbol{\psi}$ is the vector of model parameters, $f(\mathbf{y} \mid \hat{\boldsymbol{\psi}}, \mathcal{M}_k)$ is the probability density function of the data given the model parameters and the model, $\hat{\boldsymbol{\psi}}$ is the maximum likelihood of $\boldsymbol{\psi}$, and d_k is the dimension of $\boldsymbol{\psi}$, or $d_k = m_k + n_k$.

III. The MAP Criterion

The MAP criterion chooses the model with the largest posterior probability. Let the posterior probability of $\mathcal{M}_k$ be denoted by $p(\mathcal{M}_k \mid \mathbf{y})$. According to Bayes' theorem, $p(\mathcal{M}_k \mid \mathbf{y})$ is defined by

$$p(\mathcal{M}_k \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \mathcal{M}_k) p(\mathcal{M}_k)}{f(\mathbf{y})} \tag{6}$$

[1]The models will be called nested if the simpler models in the sets are identical to the more complex models when some parameters of the more complex models are set to zero.

where

- $f(\mathbf{y} \mid \mathcal{M}_k)$ marginal density of the data given they are generated by the model $\mathcal{M}_k$;
- $p(\mathcal{M}_k)$ prior probability of $\mathcal{M}_k$;
- $f(\mathbf{y})$ marginal density of the data, which is obtained by

$$f(\mathbf{y}) = \sum_{k=0}^{q-1} f(\mathbf{y} \mid \mathcal{M}_k) p(\mathcal{M}_k). \tag{7}$$

To find the MAP model, we evaluate $p(\mathcal{M}_k \mid \mathbf{y})$ for $k \in Z_q$ and select the model that has the maximum $p(\mathcal{M}_k \mid \mathbf{y})$. Formally, this is carried out according to

$$\begin{aligned} \mathcal{M}_s &= \arg \max_{(\mathcal{M}_k : k \in Z_q)} \{p(\mathcal{M}_k \mid \mathbf{y})\} \\ &= \arg \max_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{f(\mathbf{y} \mid \mathcal{M}_k) p(\mathcal{M}_k)}{f(\mathbf{y})} \right\} \end{aligned} \tag{8}$$

where $s \in Z_q$.

We already assumed that the models have equal prior probabilities, i.e.,

$$p(\mathcal{M}_k) = \frac{1}{q}, \quad k \in Z_q \tag{9}$$

and therefore, they do not affect the model selection in (8). This is also the case with the marginal density $f(\mathbf{y})$ since it is not a function of $\mathcal{M}_k$. Consequently, we may drop the factors $p(\mathcal{M}_k)$ and $f(\mathbf{y})$ from the model selection criterion, which then becomes

$$\mathcal{M}_s = \arg \max_{(\mathcal{M}_k : k \in Z_q)} \{f(\mathbf{y} \mid \mathcal{M}_k)\}. \tag{10}$$

Clearly, to find the MAP solution, we have to evaluate the marginal density of the data for each model. This density can be found from

$$f(\mathbf{y} \mid \mathcal{M}_k) = \int_{\boldsymbol{\Psi}_k} f(\mathbf{y} \mid \boldsymbol{\psi}, \mathcal{M}_k) f(\boldsymbol{\psi} \mid \mathcal{M}_k)\, d\boldsymbol{\psi} \tag{11}$$

where $f(\mathbf{y} \mid \boldsymbol{\psi}, \mathcal{M}_k)$ is the density of $\mathbf{y}$ obtained from (1) and $f(\mathbf{e} \mid \boldsymbol{\phi})$, $\boldsymbol{\Psi}_k$ is the model's parameter space, $\boldsymbol{\Psi}_k = \boldsymbol{\Theta}_k \cup \boldsymbol{\Phi}_k$, and $f(\boldsymbol{\psi} \mid \mathcal{M}_k)$ is the prior density of $\boldsymbol{\psi}$, where $\boldsymbol{\psi} = [\boldsymbol{\theta}^T \boldsymbol{\phi}^T]$. Note that $\boldsymbol{\Psi}_k \subseteq \mathbf{R}^{d_k}$, and $d_k = m_k + n_k$.

Obviously, the evaluation of the marginal density $f(\mathbf{y} \mid \mathcal{M}_k)$ requires, in general, multidimensional integration. Unfortunately, in most of the practical cases, the final result cannot be put in a closed analytical form. There are two ways to proceed. One is to employ a technique for numerical integration (see, for example [6]) or to resort to approximations that will allow a closed-form solution. The first approach is straightforward and usually more accurate but does not provide much insight into the model selection problem. On the other hand, the approximation may not lead to as accurate model selections, but the resulting closed-form solution may improve our understanding of the problem under study. We adopt the second approach and assume that we have long data vectors $\mathbf{y}$ so that standard asymptotical approximations can be applied.

Under certain regularity conditions and large data samples [2], we can use Laplace's method for integration and write

(see the Appendix)

$$\int_{\Psi_k} f(\mathbf{y} \mid \psi, \mathcal{M}_k) f(\psi \mid \mathcal{M}_k)\, d\psi \simeq (2\pi)^{\frac{d_k}{2}} |\hat{\mathcal{H}}_k|^{-\frac{1}{2}} f(\mathbf{y} \mid \hat{\psi}, \mathcal{M}_k) f(\hat{\psi} \mid \mathcal{M}_k) \quad (12)$$

where $\hat{\psi}$ is the maximum likelihood estimate of ψ, and $\hat{\mathcal{H}}_k$ is the Hessian of $-\ln f(\mathbf{y} \mid \psi, \mathcal{M}_k)$ evaluated at $\hat{\psi}$, which is also termed as the observed information matrix, or

$$\hat{\mathcal{H}}_k = -\frac{\partial^2 \ln f(\mathbf{y} \mid \psi, \mathcal{M}_k)}{\partial \psi \partial \psi^T}\bigg|_{\psi=\hat{\psi}}. \quad (13)$$

The approximation is particularly good when the likelihood function is highly peaked around $\hat{\psi}$. This is usually the case when the number of data samples N is large. Concerns for the accuracy of this approximation has led many researchers to try exact calculations of (11) by applying Monte Carlo methods [7], [11].

Now, from (10), (11), and (12), and neglecting the terms of order $O(1)$, we deduce that the asymptotical MAP criterion becomes

$$\text{MAP: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ -\ln f(\mathbf{y} \mid \hat{\psi}, \mathcal{M}_k) + \frac{1}{2} \ln |\hat{\mathcal{H}}_k| \right\}. \quad (14)$$

The first term of the criterion decreases when the complexity of the model increases, and at the same time, by contrast, the second term increases and acts as a penalty for using additional parameters to model the data. Parenthetically, it might be noted that the Hessian matrix in (14) can be replaced by the Fisher information matrix because in deriving (14), the error it introduces is of smaller order than the errors due to the neglected terms of order $O(1)$.

For example, if i) the observed data $\mathbf{y}$ are real, ii) $f(\mathbf{e} \mid \phi)$ is a Gaussian density function whose form is

$$f(\mathbf{e} \mid \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{ \frac{1}{2\sigma^2} \mathbf{e}^T \mathbf{e} \right\} \quad (15)$$

and iii) $g_k(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta})$ is given by (2), the loglikelihood has the form

$$\ln f(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2, \mathcal{M}_k) = \text{const} - \frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{s}_k(\boldsymbol{\theta}))^T (\mathbf{y} - \mathbf{s}_k(\boldsymbol{\theta})) \quad (16)$$

and the Hessian becomes

$$\mathcal{H}_k = \begin{bmatrix} -\frac{\partial^2 \ln f(\mathbf{y}|\boldsymbol{\theta},\sigma^2,\mathcal{M}_k)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} & -\frac{\partial^2 \ln f(\mathbf{y}|\boldsymbol{\theta},\sigma^2,\mathcal{M}_k)}{\partial \boldsymbol{\theta} \partial \sigma^2} \\ -\frac{\partial^2 \ln f(\mathbf{y}|\boldsymbol{\theta},\sigma^2,\mathcal{M}_k)}{\partial \sigma^2 \partial \boldsymbol{\theta}^T} & -\frac{\partial^2 \ln f(\mathbf{y}|\boldsymbol{\theta},\sigma^2,\mathcal{M}_k)}{\partial^2 (\sigma^2)^2} \end{bmatrix}. \quad (17)$$

It is easy to show that the first term in (14) results in

$$-\ln f(\mathbf{y} \mid \hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \mathcal{M}_k) = \frac{N}{2} \ln 2\pi\hat{\sigma}^2 + \frac{N}{2}. \quad (18)$$

Obviously, when we examine nested models, $\hat{\sigma}^2$ decreases by including more parameters in the model, and so does $-\ln f(\mathbf{y} \mid \hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \mathcal{M}_k)$. The increase of $\frac{1}{2} \ln |\hat{\mathcal{H}}_k|$ with the complexity of the models, however, is not obvious since it depends on the type of models that are being tested.

When the observed data are independent and identically distributed, we can write

$$|\hat{\mathcal{H}}_k| = O\big(N^{\frac{d_k}{2}}\big). \quad (19)$$

This then reduces (14) to the naive MDL criterion, i.e.,

$$\text{MAP: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ -\ln f(\mathbf{y} \mid \hat{\psi}, \mathcal{M}_k) + \frac{d_k}{2} \ln N \right\}. \quad (20)$$

The expression (19), however, is not always valid. We will show in the sequel that there are several typical signal processing families of models for which (20) will not be appropriate selection rule.

IV. Examples

To put things in perspective, we briefly investigate several different sets of models. The first three are nested linear models, whereas the fourth and the fifth are nested nonlinear models. In all these examples, except for the last one, we assume that the noise vector $\mathbf{e}$ is real and zero mean Gaussian with a probability density function given by (15). In the last example, the observed data are complex, and the noise $\mathbf{e}$ is zero mean complex Gaussian whose probability density function is

$$f(\mathbf{e} \mid \sigma^2) = \frac{1}{(\pi\sigma^2)^N} \exp\left(-\frac{\mathbf{e}^H \mathbf{e}}{\sigma^2} \right) \quad (21)$$

where the real and imaginary components of the elements of $\mathbf{e}$ are independent and identically distributed with variance $\frac{\sigma^2}{2}$.

In the first three examples, the examined models are given by

$$\mathcal{M}_k\text{: } \mathbf{y} = \mathbf{H}_k \boldsymbol{\theta}_k + \mathbf{e}, \quad k \in Z_q \quad (22)$$

where $\mathbf{H}_k$ is a known $N \times m_k$ observation matrix with rank m_k, and $\boldsymbol{\theta}_k$ is the vector of unknown parameters.

The AIC and the naive MDL selection rules (4) and (5) are now simplified to

$$\text{AIC: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + m_k \right\} \quad (23)$$

and

$$\text{MDL: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + \frac{m_k}{2} \ln N \right\} \quad (24)$$

where $\hat{\sigma}_k^2$ is the estimated noise variance obtained by

$$\hat{\sigma}_k^2 = \frac{1}{N} \mathbf{y}^T \mathbf{P}_k^{\perp} \mathbf{y}. \quad (25)$$

Here, $\mathbf{P}_k^{\perp}$ is a projection matrix defined by

$$\mathbf{P}_k^{\perp} = \mathbf{I} - \mathbf{H}_k \big(\mathbf{H}_k^T \mathbf{H}_k \big)^{-1} \mathbf{H}_k^T \quad (26)$$

and $\mathbf{I}$ is an $N \times N$ identity matrix. It might be noted that the noise model $\mathbf{y} = \mathbf{e}$ is also included in the set of examined models, with $m_k = 0$, and $\hat{\sigma}_0^2 = \frac{1}{N} \mathbf{y}^T \mathbf{y}$.

On the other hand, from (14), (17), and (18), it is easy to show that the MAP criterion can be approximated by

$$\text{MAP: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + \frac{1}{2} \ln \left| \mathbf{H}_k^T \mathbf{H}_k \right| \right\}. \tag{27}$$

Note that we are still not able to make direct comparisons of the MAP with the other two rules. Therefore, we proceed by specifying the sets of models in more detail, and then, we derive the MAP selection rules for each of them.

A. Sinusoids with Known Frequencies

As a first example, let the models represent sets of sinusoids with known frequencies but unknown phases and amplitudes. Under $\mathcal{M}_k$, the data represent k sinusoids distorted by noise that can be expressed as

$$\mathcal{M}_k\text{: } \mathbf{y} = \mathbf{H}_k \boldsymbol{\theta}_k + \mathbf{e} \tag{28}$$

where

$$\mathbf{H}_k = [\mathbf{h}_{1s} \quad \mathbf{h}_{1c} \quad \mathbf{h}_{2s} \quad \mathbf{h}_{2c} \quad \cdots \quad \mathbf{h}_{ks} \quad \mathbf{h}_{kc}] \tag{29}$$

and

$$\begin{aligned} \mathbf{h}_{is} &= [0 \quad \sin(\omega_i) \quad \sin(2\omega_i) \quad \cdots \quad \sin((N-1)\omega_i)]^T \\ \mathbf{h}_{ic} &= [1 \quad \cos(\omega_i) \quad \cos(2\omega_i) \quad \cdots \quad \cos((N-1)\omega_i)]^T \end{aligned} \tag{30}$$

for $i = 1, 2, \ldots, k$, and $\boldsymbol{\theta}_k$ is the vector of amplitudes. Without loss of generality, it is assumed that $\omega_i \neq 0$ or π, $i = 1, 2, \ldots, k$. The set of frequencies ω_i for $i = 1, 2, \ldots, k$ is known, and $\omega_i \neq \omega_j$, for $i \neq j$. Note that the model of k sinusoids has $m_k = 2k$ unknown signal parameters. So, the AIC and the naive MDL become

$$\text{AIC: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + 2k \right\} \tag{31}$$

and

$$\text{MDL: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + k \ln N \right\}. \tag{32}$$

To obtain the form of the MAP criterion, we examine the elements of the matrix $\mathbf{H}_k^T \mathbf{H}_k$. It is not difficult to show that for the ijth element, we can write

$$\begin{aligned} \left[\mathbf{H}_k^T \mathbf{H}_k \right]_{ij} &= O(1), \quad i \neq j \\ \left[\mathbf{H}_k^T \mathbf{H}_k \right]_{ij} &= O(N), \quad i = j. \end{aligned}$$

This further implies that

$$\left| \mathbf{H}_k^T \mathbf{H}_k \right| = O(N^{2k}). \tag{33}$$

From (27) and (33), we deduce that the MAP criterion can be approximated by

$$\text{MAP: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + k \ln N \right\}. \tag{34}$$

We notice that for this set of models, the MAP and the naive MDL criteria are *identical*.

B. Polynomial Models

Suppose next that the examined models are polynomials of various degrees. The models $\mathcal{M}_k$ are given again by (28), but the matrix $\mathbf{H}_k$ is defined according to

$$\mathbf{H}_k = [\mathbf{h}_0 \quad \mathbf{h}_1 \quad \mathbf{h}_2 \quad \cdots \quad \mathbf{h}_{k-2} \quad \mathbf{h}_{k-1}], \quad k \geq 1 \tag{35}$$

where

$$\mathbf{h}_l = [0^l \quad 1^l \quad 2^l \quad \cdots \quad (N-1)^l]^T \tag{36}$$

where $l \in Z_k$. Thus, for $k \geq 1$, the model $\mathcal{M}_k$ refers to a polynomial of degree $k-1$ embedded in noise. The number of signal parameters is $m_k = k$, and $\mathcal{M}_0$ refers to the noise model.

Again, we first determine the AIC and the naive MDL criteria. They are

$$\text{AIC: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + k \right\} \tag{37}$$

and

$$\text{MDL: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + \frac{k}{2} \ln N \right\}. \tag{38}$$

In deriving the MAP criterion, it is straightforward to show that for large N, $\mathbf{H}_k^T \mathbf{H}_k$ can be approximated by

$$\mathbf{H}_k^T \mathbf{H}_k \simeq \begin{bmatrix} N & \frac{N^2}{2} & \cdots & \frac{N^k}{k} \\ \frac{N^2}{2} & \frac{N^3}{3} & \cdots & \frac{N^{k+1}}{k+1} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{N^k}{k} & \frac{N^{k+1}}{k+1} & \cdots & \frac{N^{2k-1}}{2k-1} \end{bmatrix}.$$

This further implies that

$$|\hat{\mathcal{H}}_k| = O(N^{k^2+1})$$

and

$$\text{MAP: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + \frac{k^2}{2} \ln N \right\}. \tag{39}$$

This result is quite interesting since we see that the MAP criterion is different from the AIC and the naive MDL. The MAP penalty term is a quadratic function of the number of polynomial parameters, which implies that the MAP criterion penalizes for overparameterization much more rigorously than the AIC and the naive MDL. The MAP penalty for every additional parameter increases as the degree of the polynomial increases. For example, the penalty associated with the first parameter θ_0 is $\frac{1}{2} \ln N$, the penalty for the second parameter θ_1, $\frac{3}{2} \ln N$, and so on. Not surprisingly, the more accurately we can determine the model parameters, the higher the penalty to include them into the model.

C. Autoregressive Models

We assume now that the data are stationary and that they can be modeled by an autoregressive (AR) model. For instance, if the AR model is of kth order, we write

$$\mathcal{M}_k\colon y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_k y_{t-k} + e_t \quad t = k+1, \ldots, N \tag{40}$$

where the θ_i's, $i = 1, 2, \ldots, k$ are the AR parameters, and the e_t's are zero mean independent and identically distributed Gaussian random variables. In addition, we assume that

$$A(z) = 1 - \theta_1 z - \theta_2 z^2 - \cdots - \theta_k z^k \neq 0, \quad |z| \leq 1. \tag{41}$$

Equation (40) can be rewritten in a vector-matrix form as

$$\mathbf{y}_{k+1,N} = \mathbf{H}_k \boldsymbol{\theta}_k + \mathbf{e}_{k+1,N} \tag{42}$$

where

$$\mathbf{y}_{k+1,N} = [y_{k+1} \quad y_{k+2} \quad \cdots \quad y_N]^T$$
$$\mathbf{e}_{k+1,N} = [e_{k+1} \quad e_{k+2} \quad \cdots \quad e_N]^T$$
$$\boldsymbol{\theta}_k = [\theta_1 \quad \theta_2 \quad \cdots \quad \theta_k]^T$$

and

$$\mathbf{H}_k = \begin{bmatrix} y_k & y_{k-1} & \cdots & y_1 \\ y_{k+1} & y_k & \cdots & y_2 \\ \vdots & \vdots & \vdots & \vdots \\ y_N & y_{N-1} & \cdots & y_{N-k} \end{bmatrix}.$$

The AIC and the naive MDL are given by

$$\text{AIC: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + k \right\} \tag{43}$$

and

$$\text{MDL: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + \frac{k}{2} \ln N \right\}. \tag{44}$$

To derive the MAP criterion we proceed as follows. First, we claim that for large N, $|\mathbf{H}_k^T \mathbf{H}_k| = O(N^k)$. Indeed, we can write

$$\mathbf{H}_k^T \mathbf{H}_k = N\mathbf{R} \tag{45}$$

where

$$\mathbf{R} = \begin{bmatrix} \hat{r}_{11} & \hat{r}_{12} & \cdots & \hat{r}_{1,k} \\ \hat{r}_{21} & \hat{r}_{22} & \cdots & \hat{r}_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{r}_{k1} & \hat{r}_{k2} & \cdots & \hat{r}_{kk} \end{bmatrix} \tag{46}$$

and

$$\hat{r}_{ij} = \frac{1}{N} \sum_{t=k}^{N} y_{t-i+1} y_{t-j+1}. \tag{47}$$

For large N, the elements $\hat{r}_{ij}$, $i, j \in Z_{k+1}$ tend to the autocorrelation $r(i-j) = E(y_{t-i}\ y_{t-j})$ of the process, respectively, and the determinant of $\mathbf{R}$ remains of order $O(1)$. This entails that $|\mathbf{H}_k^T \mathbf{H}_k|$ is of $O(N^k)$. With this result, it is evident that the MAP rule can be approximated by

$$\text{MAP: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + \frac{k}{2} \ln N \right\}. \tag{48}$$

Thus, for AR models, the naive MDL and MAP criteria are identical.

D. Sinusoids with Unknown Frequencies

Our next example includes various sets of sinusoids as competing models again, but this time, all the parameters of the sinusoids are unknown. The models are described by

$$\mathcal{M}_k\colon \mathbf{y} = \mathbf{H}(\boldsymbol{\omega}_k)\mathbf{a}_k + \mathbf{e}, \quad k \geq 1 \tag{49}$$

where $\boldsymbol{\theta}_k = [\boldsymbol{\omega}_k^T \ \mathbf{a}_k^T]$, and the matrix $\mathbf{H}(\boldsymbol{\omega}_k)$ is defined by (29) and (30). The vector $\boldsymbol{\theta}_k$ now consists of $2k$ linear and k nonlinear parameters. The linear parameters are the amplitudes, and the nonlinear are the frequencies. The other assumptions are the same as in Example 1. The AIC and the naive MDL take the forms

$$\text{AIC: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + 3k \right\} \tag{50}$$

and

$$\text{MDL: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + \frac{3k}{2} \ln N \right\} \tag{51}$$

where

$$\hat{\sigma}_k^2 = \frac{1}{N} \mathbf{y}^T \mathbf{P}^{\perp}(\hat{\boldsymbol{\omega}}_k)\mathbf{y} \tag{52}$$

and

$$\mathbf{P}^{\perp}(\hat{\boldsymbol{\omega}}_k) = \mathbf{I} - \mathbf{H}(\hat{\boldsymbol{\omega}}_k)(\mathbf{H}^T(\hat{\boldsymbol{\omega}}_k) \quad \mathbf{H}(\hat{\boldsymbol{\omega}}_k))^{-1}\mathbf{H}^T(\hat{\boldsymbol{\omega}}_k). \tag{53}$$

The derivation of the MAP criterion is fairly technical, and its details are presented in [5]. Here, we only outline the main steps of the derivation. Recall that to obtain the MAP selection rule, we have to solve (11). The set of parameters $\boldsymbol{\psi}$ is defined by $\boldsymbol{\psi} = [\boldsymbol{\theta}^T \sigma]$. First, one integrates out the linear parameters $\mathbf{a}_k$. We adopt an improper prior for $\mathbf{a}_k$, and for the associated proportionality constant in the solution, we assume that is of order $O(1)$. Then, the standard deviation σ is integrated out. As a prior for σ, we use the improper Jeffreys' prior $f(\sigma) \propto \frac{1}{\sigma}$. Finally, we integrate out the frequencies $\boldsymbol{\omega}_k$ for which we assume a uniform prior. To carry out the integration, we exploit the Taylor expansion used for obtaining (12) and get

$$f(\mathbf{y} \mid \mathcal{M}_k) \propto \Gamma\left(\frac{N-2k}{2}\right) |\mathbf{H}^T(\hat{\boldsymbol{\omega}}_k)\mathbf{H}(\hat{\boldsymbol{\omega}}_k)|^{-\frac{1}{2}} \times (\mathbf{y}^T \mathbf{P}^{\perp}(\hat{\boldsymbol{\omega}}_k)\mathbf{y})^{-\frac{N-2k}{2}} N^{-\frac{3k}{2}} |\mathbf{R}|^{-1} \quad k \geq 1 \tag{54}$$

where $\Gamma(\cdot)$ is the standard Gamma function, and $\mathbf{R}$ is a matrix whose determinant is of order $O(1)$. The analysis of (54) shows that the MAP selection rule can be approximated by [5]

$$\text{MAP: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_q)} \left\{ \frac{N}{2} \ln \hat{\sigma}_k^2 + \frac{5k}{2} \ln N \right\}. \tag{55}$$

Therefore, the MAP is different from the AIC and the naive MDL, and once again, it has the most stringent penalty. As in the case of polynomials, the parameters do not contribute necessarily equal penalties to overparameterization. Each amplitude yields a penalty equal to $\frac{1}{2}\ln N$, whereas each frequency yields $\frac{3}{2}\ln N$. It is interesting to observe that the penalty in (55) can directly be obtained by using the Fisher information matrix $\mathcal{I}_k$. In other words, for large N, $\ln|\mathcal{I}_k|^{\frac{1}{2}} \simeq \frac{5k}{2}\ln N$. The same criterion has been proposed in [8], where the MDL criterion has been properly exploited.

E. Chirp Signals

As our final example, we consider the noise and the following three models:

$$\mathcal{M}_k\text{: } \mathbf{y} = \mathbf{h}_k a + \mathbf{e}, \quad k = 1,2,3 \tag{56}$$

where

$$\mathbf{h}_1 = \begin{bmatrix} \exp(j\omega t_0) \\ \exp(j\omega(t_0+1)) \\ \vdots \\ \exp(j\omega(N-1-t_0)) \end{bmatrix}$$

$$\mathbf{h}_2 = \begin{bmatrix} \exp\big(j\big(\alpha t_0^2 + \omega t_0\big)\big) \\ \exp(j(\alpha(t_0+1)^2 + \omega(t_0+1))) \\ \vdots \\ \exp(j(\alpha(N-1-t_0)^2 + \omega(N-1-t_0))) \end{bmatrix}$$

and (57), shown at the bottom of the page. In addition, we have $j = \sqrt{-1}$. Note that now, $\mathbf{y}$ is a vector of complex observations that represent noisy chirp-type signals, and the noise vector $\mathbf{e}$ has the density function (21). The unknown parameters are the complex amplitude a and the phase parameters ω, α, and β. The AIC and the naive MDL become

$$\text{AIC: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_4\})} \left\{ N \ln \hat{\sigma}_k^2 + (k+2)(1-\delta(k)) \right\} \tag{58}$$

and

$$\text{MDL: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_4\})} \left\{ N \ln \hat{\sigma}_k^2 + \frac{(k+2)(1-\delta(k))}{2} \ln N \right\} \tag{59}$$

where the estimate of the noise variance under $\mathcal{M}_k$ is found from

$$\hat{\sigma}_k^2 = \frac{1}{N} \left| \left(\mathbf{I} - \frac{1}{N} \mathbf{h}_k(\hat{\boldsymbol{\theta}}) \mathbf{h}_k^H(\hat{\boldsymbol{\theta}}) \right) \mathbf{y} \right|^2 \tag{60}$$

with $\hat{\boldsymbol{\theta}}$ being the estimated phase parameters of $\mathcal{M}_k$, and $\delta(k)$ is the Kronecker delta function.

The derivation of the MAP rule follows the same lines as the derivation of the MAP rule in the previous example. It shows that the final form of the rule is

$$\text{MAP: } \mathcal{M}_s = \arg \min_{(\mathcal{M}_k : k \in Z_4)} \left\{ N \ln \hat{\sigma}_k^2 + \frac{(k^2+2k+2)(1-\delta(k))}{2} \ln N \right\}. \tag{61}$$

This result can also be deduced from the determinant of the Fisher information matrix.

F. Discussion

An important issue for our approach of model selection is the accuracy of the made approximations. When the regularity conditions are satisfied, the relative error of the Laplace method for integration is of order $O(N^{-1})$, that is, if the integral in (12) is denoted by $\mathcal{J}$ and its approximation on the right-hand side by $\hat{\mathcal{J}}$, then [11]

$$\mathcal{J} = \hat{\mathcal{J}}(1 + O(N^{-1})). \tag{62}$$

If the inverse of the expected information matrix is used, the relative error becomes of order $O(N^{-\frac{1}{2}})$.

In this paper, we have assumed that the priors are noninformative, and that their omission in the evaluation of the integral $\mathcal{J}$ contributes to its overall relative error a value of order $O(1)$. We have also neglected another term of order $O(1)$: the factor $(2\pi)^{d_k/2}$ in (12). We could have retained all these terms, and by doing so, we could have obtained refined model selection rules. This, however, would have required and extensive analysis of the priors for the model parameters, which is a problem we wanted to alleviate in the paper. In Bayesian model selection, the assignment of priors is well recognized and considered to be nontrivial. When the number of observed data is large enough, these omissions are acceptable. The work on refined model selection rules where the priors cannot be ignored will be presented elsewhere.

The form of the penalty can be obtained with relative ease from the determinant of the Fisher information matrix, and its behavior can be found as a function of N when N becomes very large. Although the use of the Fisher information matrix increases the relative error of $\hat{\mathcal{J}}$ to the order of $O(N^{-\frac{1}{2}})$, it is acceptable to apply it because we have dropped terms that cause relative errors of order $O(1)$. It is also important to notice that the integral in (12) is approximated for every model $\mathcal{M}_k$ and that the most significant point about it is not so much the accuracy of the approximation but the preservation of ranking among the integral values after the approximations.

The examples in this section clearly show that model selection rules that penalize for model complexity have to be examined carefully before they are applied. One *cannot* use

$$\mathbf{h}_3 = \begin{bmatrix} \exp\big(j\big(\beta t_0^3 + \alpha t_0^2 + \omega t_0\big)\big) \\ \exp(j(\beta(t_0+1)^3 + \alpha(t_0+1)^2 + \omega(t_0+1))) \\ \vdots \\ \exp(j(\beta(N-1-t_0)^3 + \alpha(N-1-t_0)^2 + \omega(N-1-t_0))) \end{bmatrix}. \tag{57}$$

the AIC and the naive MDL rules in every scenario. In this context, it is important to point out that all the parameters of a model do not necessarily contribute equal penalties. For example, in the case of polynomial models, the coefficient that multiplies $\mathbf{h}_0$ adds a penalty of $\frac{1}{2}\ln N$, the coefficient that multiplies $\mathbf{h}_1$ a penalty $\frac{3}{2}\ln N$, and so on. Similarly, in the case of sinusoids with unknown frequencies, the penalty for each unknown amplitude and phase is $\frac{1}{2}\ln N$, and for each unknown frequency, it is $\frac{3}{2}\ln N$. Obviously, the penalty is larger if the achievable accuracy of the estimated parameter is larger.

Why does the MAP rule penalize discriminatively? An interesting insight might be gained by analyzing the simple example of a constant signal in Gaussian noise. Let $\mathbf{y}$ be generated according to

$$\mathcal{M}_0\text{: } \mathbf{y} = \mathbf{h}_0\theta_0 + \mathbf{e} \tag{63}$$

where $\mathbf{e}$ has the density from (15) with σ^2 known and equal to one, θ_0 is an unknown constant, and $\mathbf{h}_0$ is a vector of ones. Suppose that we have to choose from

$$\mathcal{M}_1\text{: } \mathbf{y} = \mathbf{h}_0\theta_0 + \mathbf{h}_1\theta_1 + \mathbf{e} \tag{64}$$

and

$$\mathcal{M}_2\text{: } \mathbf{y} = \mathbf{h}_0\theta_0 + \mathbf{h}_2\theta_2 + \mathbf{e} \tag{65}$$

where

$$\mathbf{h}_1^T = [0 \quad 1 \quad 2 \quad \cdots \quad N-1]$$

and

$$\mathbf{h}_2^T = [0 \quad 1 \quad 4 \quad \cdots \quad (N-1)^2].$$

Thus, the two models have equal number of parameters and they both contain the true model $\mathcal{M}_0$. Suppose next that the likelihood terms $f(\mathbf{y} \mid \hat{\theta}_0, \hat{\theta}_1, \mathcal{M}_1)$ and $f(\mathbf{y} \mid \hat{\theta}_0, \hat{\theta}_2, \mathcal{M}_2)$ have identical values, say, for $\hat{\theta}_1 = \hat{\theta}_2 = 0$. Then, the AIC and MDL rules would treat the two models as equally good, whereas our criterion would find that the penalty of $\mathcal{M}_2$ is greater than the penalty of $\mathcal{M}_1$ and, therefore, would consider $\mathcal{M}_1$ a better model for the data.

One interpretation for this choice is the following. If we perturb the parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ by the same amount $\Delta\theta$, it is interesting to examine the changes in the associated loglikelihoods $\mathcal{L}_1$ and $\mathcal{L}_2$ caused by these perturbations. If we denote the changes by $\Delta\mathcal{L}_1$ and $\Delta\mathcal{L}_2$, respectively, we readily find that the probability of the event $\Delta\mathcal{L}_1 > \Delta\mathcal{L}_2$ is given by

$$P(\Delta\mathcal{L}_1 > \Delta\mathcal{L}_2) = P\left(Z > \frac{(\mathbf{h}_2^T\mathbf{h}_2 - \mathbf{h}_1^T\mathbf{h}_1)|\Delta\theta|}{2\sqrt{\mathbf{h}_2^T\mathbf{h}_2 - 2\mathbf{h}_2^T\mathbf{h}_1 + \mathbf{h}_1^T\mathbf{h}_1}}\right) \tag{66}$$

where Z is the standard normal random variable. As N increases, this probability tends to zero quickly. Therefore, we conclude that a small perturbation of $\hat{\theta}_1$ would affect the loglikelihood $\mathcal{L}_1$ much less than the same perturbation of $\hat{\theta}_2$ the loglikelihood $\mathcal{L}_2$. Thus, the MAP rule simply chooses the more robust model of the two.

Finally, a few comments on the relationship between the MDL and the MAP model selection criteria are in order. Recall that the MDL rule chooses the model described by the shortest code length, and in general, it does not always choose the same model, as does the MAP criterion. The naive MDL has been used in many situations where $-\ln f(\mathbf{y} \mid \psi, \mathcal{M})$ does not grow proportionally with N and that is *inappropriate*. Normally, when the data come from i.i.d processes, the negative loglikelihood does grow proportionally with N, and the use of the naive MDL is then justified. Recently, Rissanen derived a formula for the ideal code length of models that are not necessarily models of i.i.d. processes, and he found that their code length $l(\mathbf{y} \mid \mathcal{M})$ is given by [18]

$$l(\mathbf{y} \mid \mathcal{M}) = -\ln f(\mathbf{y} \mid \hat{\psi}, \mathcal{M}) + \frac{d_k}{2}\ln\frac{N}{2\pi} + \ln\int_{\Psi}\sqrt{|\mathcal{I}(\psi)|}\,d\psi \tag{67}$$

where $\mathcal{I}$ is the Fisher information matrix. For (67) to hold, it is required that the maximum likelihood estimates satisfy the central limit theorem as well as some weak smoothness conditions. Since the square root of the determinant of the Fisher information matrix very often is not integrable, Rissanen developed modifications of (67). For the regression problem and the family of Gaussian distributions derived in [18], the obtained model selection rules are identical to our MAP selection criteria only after appropriate approximations.

V. Simulations

We validated the performance of the AIC, the naive MDL, and MAP rules by Monte Carlo simulations on observations generated according to Examples 2 and 5. Extensive examination of Example 4 is given in [5]. The results there show that the MAP criterion provides much more accurate model selections than the AIC and the naive MDL. The Examples 1 and 3 were not analyzed because the naive MDL and MAP have identical forms for the models cited there.

In the first set of experiments, we simulated polynomial-type signals embedded in Gaussian noise according to

$$\mathbf{y} = \mathbf{H}_3\boldsymbol{\theta}_3 + \mathbf{e} \tag{68}$$

where $\boldsymbol{\theta}_3 = [1 \;\; 4.5 \;\; -0.2 \;\; \theta_3]^T$. Throughout the experiment, θ_3 was changed to achieve a desired signal-to-noise ratio (SNR), which was defined by

$$\text{SNR} = 10\log_{10}\frac{\theta_3^2\mathbf{h}_3^T\mathbf{P}_2^{\perp}\mathbf{h}_3}{\sigma^2} \tag{69}$$

where $\mathbf{P}_2^{\perp}$ is obtained from (26) for $k = 2$. The SNR was varied in the range from 0–30 dB in steps of 1 dB. For each SNR, there were 1000 trials. The length of the data vectors was kept constant and equal to 50 samples, and q was set to 7. Thus, the most complex model was the polynomial of degree 5. A similar example with different parameters can be found in [4].

The results are displayed in Fig. 1 (a)–(c). They show the probabilities of correct estimation, overparameterization, and underparameterization of the AIC, the naive MDL, and MAP,

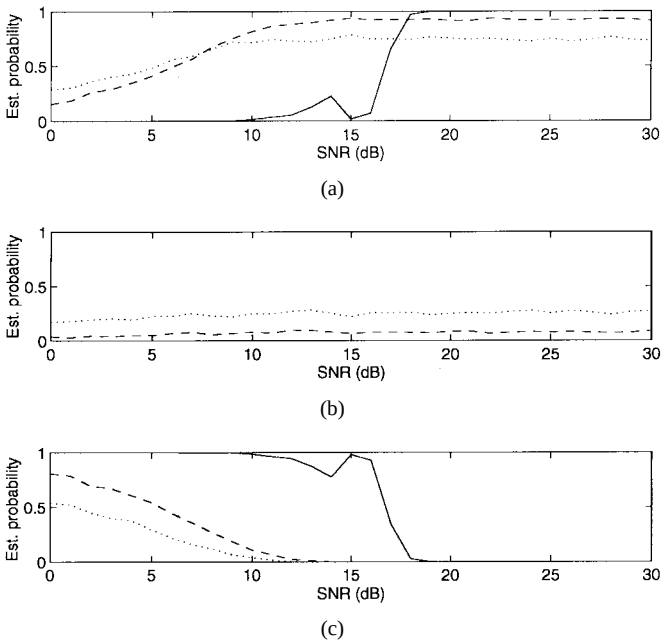

Fig. 1. (a) Probability of correct estimation versus SNR. (b) Probability of overestimation versus SNR. (c) Probability of underestimation versus SNR. The solid curve refers to the MAP rule, the dashed to MDL, and the dotted to AIC. The sequences had 50 samples, and for each SNR, there were 1000 trials.

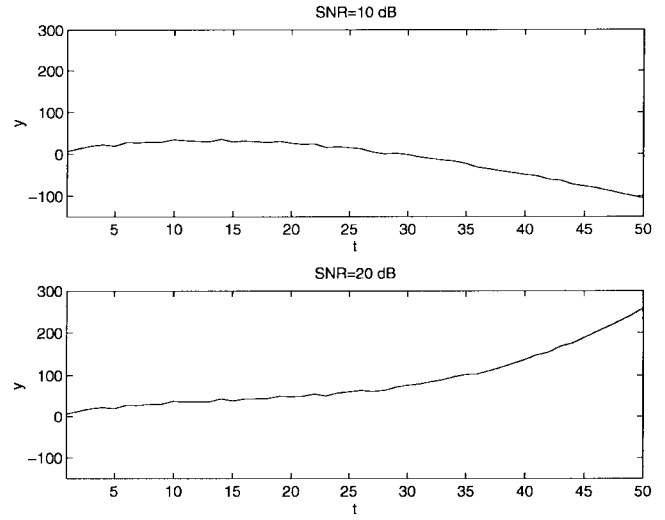

Fig. 2. Upper plot displays a third-degree polynomial embedded in noise for SNR $= 10$ dB. The lower figure displays a third-degree polynomial in noise for SNR $= 20$ dB.

respectively. We observe that for SNR's above 20 dB, the MAP criterion has perfect performance, whereas the AIC and the naive MDL tend to overestimate the polynomial degree. For lower SNR's than 20 dB, the MAP criterion chooses polynomials of smaller degree than the correct one.

To obtain a better perspective on the performance of the three rules, we tabulated the results obtained for SNR's equal to 10 and 20 dB, respectively. In Fig. 2, we show one realization of the observed data for each of these SNR's, and in Tables I and II, we list the results. Note that the correct model is $\mathcal{M}_4$ (a polynomial of third degree). The tables show the number of times the selection rules chose a particular model

TABLE I
PERFORMANCE COMPARISON OF THE AIC, MDL, AND MAP CRITERIA WHEN THE SNR $= 10$ dB. THE ENTRIES REPRESENT THE NUMBER OF TIMES A PARTICULAR MODEL WAS SELECTED OUT OF 1000 TRIALS. THE CORRECT MODEL IS $\mathcal{M}_4$

	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$
AIC	0	0	0	66	718	142	74
MDL	0	0	0	185	749	53	13
MAP	0	0	0	997	3	0	0

TABLE II
PERFORMANCE COMPARISON OF THE AIC, MDL, AND MAP CRITERIA WHEN THE SNR $= 20$ dB. THE ENTRIES REPRESENT THE NUMBER OF TIMES A PARTICULAR MODEL WAS SELECTED OUT OF 1000 TRIALS. THE CORRECT MODEL IS $\mathcal{M}_4$

	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$
AIC	0	0	0	0	764	135	101
MDL	0	0	0	0	929	48	23
MAP	0	0	0	2	998	0	0

out of 1000 trials. For SNR $= 10$ dB, the MAP criterion almost always selected the second degree polynomial and for SNR $= 20$ dB almost always the correct, third-degree polynomial. The performance of the AIC was slightly better for 20 dB than for 10 dB, and of the naive MDL much better for 20 dB but still with a significant number of overparameterizations. Note that 23 times out of 1000, the naive MDL selected polynomials of even fifth degree.

Clearly, the MAP rule did not perform well when the SNR $= 10$ dB. Since the MAP rule is derived under asymptotic assumptions, it is obvious that the approximation for this number of samples and SNR is not accurate enough. Thus, application of the MAP rule for low SNR's and/or small number of samples is not recommended.

In the next experiment, we simulated chirp-type signals according to (56) and (57). The SNR defined by

$$\text{SNR} = 10\log_{10}\frac{|a|^2}{\sigma^2} \tag{70}$$

was set to 12 dB, the data length to $N = 31$, and $t_0 = -15$. First, we generated a complex sinusoid whose frequency was $\omega = 2\pi\ 0.1$, and $a = \exp(j)$. There were four different models because we also included the noise only model. The correct model is $\mathcal{M}_1$. The number of trials was again 1000. We estimated the nonlinear parameters of all the models by the method described in [3]. The results of the simulations are shown in Table III. The MAP criterion performed perfectly. The AIC had correct selection in about 70% of the trials and the naive MDL around 90%. These two criteria sometimes selected even the most complex model in the set.

Next, we generated a chirp signal by keeping all the parameters as in the previous experiment except that $\alpha = 0.15$. Therefore, the correct model is $\mathcal{M}_2$. The results are shown in Table IV. The MAP criterion repeated the perfect performance

TABLE III
PERFORMANCE COMPARISON OF THE AIC, MDL, AND MAP CRITERIA FOR CHIRP-TYPE SIGNALS. THE ENTRIES REPRESENT THE NUMBER OF TIMES A PARTICULAR MODEL WAS SELECTED OUT OF 1000 TRIALS. THE CORRECT MODEL IS $\mathcal{M}_1$

	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
AIC	0	706	130	110
MDL	0	907	63	30
MAP	0	1000	0	0

TABLE IV
PERFORMANCE COMPARISON OF THE AIC, MDL, AND MAP CRITERIA FOR CHIRP-TYPE SIGNALS. THE ENTRIES REPRESENT THE NUMBER OF TIMES A PARTICULAR MODEL WAS SELECTED OUT OF 1000 TRIALS. THE CORRECT MODEL IS $\mathcal{M}_2$

	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
AIC	0	0	802	198
MDL	0	0	904	96
MAP	0	0	1000	0

by selecting the correct model in all 1000 trials. The naive MDL did so in around 90% and the AIC in about 80% of the total number of trials. The incorrect selections were always related to overparameterizations.

VI. CONCLUDING REMARKS

In this paper, we addressed the model selection problem from a Bayesian point of view and using large sample theory. We found that the asymptotical MAP rule, in general, has different penalties for model overparameterization than the AIC and the naive MDL. In the cases where it differs from the naive MDL and AIC, the computer simulations showed that the MAP had the best performance. When the number of samples is not large enough or the SNR is not sufficiently high, the MAP rule would usually choose a simpler model than the one that generated the data. The approximations used in this paper can be improved, which will lead eventually to even better selection performance. This, however, would require careful investigation of the employed priors for the parameters.

APPENDIX

Here, we quote a theorem from [2] that provides the conditions for the applicability of the Laplace integration method and the result used in (12).

Theorem: Let $g_1(\psi)$ and $g_2(\psi)$ be real-valued functions in $\psi \subseteq R^d$. If it is assumed that the following conditions hold:

a) $g_2(\psi)$ has an absolute maximum at an interior point $\hat{\psi}$ of $\mathbf{\Psi}$, and $g_2(\hat{\psi}) > 0$;
b) there exists a constant $c > 0$ such that $g_1(\psi)g_2^c(\psi)$ is absolutely integrable on $\mathbf{\Psi}$;
c) all the partial derivatives $\partial g_2(\psi)/\partial\psi_i$ and $\partial^2 g_2(\psi)/\partial\psi_i\partial\psi_j$ exist and are continuous in a neighborhood $\mathcal{N} - \hat{\psi}$ of $\hat{\psi}$;
d) there exists a constant $C < 1$ such that $g_2(\psi)/g_2(\hat{\psi}) < C$ for all $\psi \in \mathbf{\Psi} - \mathcal{N}$;
e) $g_1(\psi)$ is continuous in the neighborhood of $\hat{\psi}$, and $g_1(\hat{\psi}) \neq 0$;

then, when $N \to \infty$

$$\int_{\mathbf{\Psi}} u(\psi)(v(\psi))^N d\psi \sim u(\hat{\psi})\left(\frac{2\pi}{N}\right)^{\frac{d}{2}} |\mathcal{H}(\psi)|^{-\frac{1}{2}}(g_2(\hat{\psi}))^N \tag{71}$$

where $|H(\psi)|$ is the Hessian determinant of $-\log g_2(\psi)$ evaluated at $\psi = \hat{\psi}$.

Proof: See [2]. □

The sign $\sim$ in the above equations denotes asymptotic equivalence, which means that the ratio of the two sides in (71) tends to 1 as $N \to \infty$. Comparing (11) and (71), we easily identify the functions g_1 and g_2 as

$$g_1(\psi) = f(\psi \mid \mathcal{M}_k) \tag{72}$$

and

$$g_2(\psi) = \exp\left\{\frac{1}{N}\ln f(\mathbf{y} \mid \psi, \mathcal{M}_k)\right\}. \tag{73}$$

It is clear then that the prior $f(\psi \mid \mathcal{M}_k)$ has to be continuous around $\hat{\psi}$ and nonzero at $\hat{\psi}$. To check the remaining regularity conditions is also relatively straightforward.

REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.
[2] O. E. Barndorff-Nielsen and D. R. Cox, *Asymptotic Techniques for Use in Statistics*. New York: Chapman and Hall, 1989.
[3] P. M. Djurić and S. Kay, "Parameter estimation of chirp signals," *IEEE Trans. Signal Processing*, vol. 38, pp. 2118–2126, 1990.
[4] P. M. Djurić, "Model selection based on asymptotic Bayes theory," in *Proc. Seventh SP Workshop Statistical Signal Array Processing*, Quebec City, P.Q., Canada, 1994, pp. 7–10.
[5] ——, "A Model selection rule for sinusoids in white Gaussian noise," *IEEE Trans. Signal Processing*, vol. 44, pp. 1794–1751, 1996.
[6] N. Flournoy and R. K. Tsutakawa, Eds. Providence, RI: *Statistical Multiple Integration*, Amer. Math. Soc., 1989.
[7] A. E. Gelfand and D. K. Dey, "Bayesian model choice: Asymptotics and exact calculations," *J. R. Stat. Soc.*, vol. B, pp. 501–514, 1994.
[8] E. J. Hannan, "Determining the number of jumps in a spectrum," in *Developments in Time Series Analysis*, T. Subba Rao, Ed. London, U.K.: Chapman and Hall, 1993, pp. 127–138.
[9] J. Hwang, "A combined detection-estimation algorithm for the harmonic retrieval problem," *Signal Process.*, vol. 30, pp. 177–197, 1993.
[10] R. L. Kashyap, "Optimal choice of AR and MA parts in autoregressive moving average models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI–4, pp. 99–104, 1982.
[11] R. E. Kass and A. E. Raftery, "Bayes Factors," *J. Amer. Stat. Assoc.*, vol. 90, pp. 773–795, 1995.
[12] T. Leonard, "Comment on a paper by M. Lejeune and G. D. Faulkenberry," *J. Amer. Stat. Assoc.*, vol. 77, pp. 657–658, 1982.
[13] Z. Liang, R. J. Jaszczak, and R. E. Coleman, "Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing," *IEEE Trans. Nucl. Sci.*, vol. 39, pp. 1126–1133, 1992.
[14] A. O'Hagan, *Kendall's Advanced Theory of Statistics: Bayesian Inference*. New York: Wiley, 1994.
[15] D. S. Poskitt, "Precision, complexity and Bayesian model determination," *J. R. Stat. Soc. B*, vol. 49, pp. 199–208, 1987.
[16] 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. Signal Processing*, vol. 41, pp. 2872–2881, 1993.
[17] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–478, 1978.
[18] ———, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 42, pp. 40–47, 1996.
[19] G. Schwarz, "Estimation the dimension of the model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.
[20] D. B. Williams, "Counting the degrees of freedom when using AIC and MDL to detect signals," *IEEE Trans. Signal Processing*, vol. 42, pp. 3282–3284, 1994.
[21] G. Xu, R. H. Roy, and T. Kailath, "Detection of number of sources via exploitation of centro-symmetry property," *IEEE Trans. Signal Processing*, vol. 42, pp. 102–112, 1994.
[22] 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.
[23] 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. Seventh SP Workshop Stat. Signal Array Process.*, 1994, pp. 43–46.

Petar M. Djurić (S'86–M'90) received the B.S. and M.S. degrees from the University of Belgrade, Belgrade, Yugoslavia, in 1981 and 1986, respectively, and the Ph.D. degree from the University of Rhode Island, Kingston, 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, University of Rhode Island. He joined the Department of Electrical Engineering, State University of New York, Stony Brook, in 1990, where he is currently an Associate Professor. His main research interests are in statistical signal processing and signal modeling.

Dr. Djurić is a member of the American Statistical Association. He has served as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.