Model Order Selection of Damped Sinusoids in Noise by Predictive Densities

William B. Bishop, *Student Member, IEEE,* and Petar M. Djurić, *Member, IEEE*

***Abstract*— We develop a procedure for the order selection of damped sinusoidal models based on the maximum *a posteriori* (MAP) criterion. The proposed method merges the concept of predictive densities with Bayesian inference to arrive at a complex multidimensional integral whose solution is achieved by way of the efficient Monte Carlo importance sampling technique. The importance function, a multivariate Cauchy probability density, is employed to produce stratified samples over the hypersurfaces support region. Centrality location parameters for the Cauchy are resolved by exploiting the special structure of the compressed likelihood function (CLF) and applying the fast maximum likelihood (FML) procedure of Umesh and Tufts [38]. Simulation results allow for a comparison between our method and the singular value decomposition (SVD) based information theoretic criteria in [28].**

I. Introduction

OBSERVATIONS described in whole or in part by an additively error-corrupted weighted sum of functions of the same nonlinear parametric family occur in many fields of applied science. Among these weighted sum models, multiple-damped sinusoids occurring in speech analysis [10], biomedicine [25], radio astronomy [3], and a variety of other applications are frequently encountered. Parameter estimation methods based on forward-backward linear prediction [21], component by component iterative schemes such as expectation-maximization (EM) [11], or those based on system identification such as KiSS/IQML [4] are able to provide reasonably accurate estimates of the signal parameters, but all rely on *a priori* knowledge regarding the actual number of signal components (i.e., the model order). In most practical situations this information is unavailable, and therefore, a reliable technique for estimating this number is required.

As this problem is an old one, an exhaustive retrospect of the extensive literature is almost impossible. Recently, however, two interesting criteria were developed [28]. Both are singular value decomposition-based (SVD-based) offshoots of the popular Akaike information criterion (AIC) and minimum description length (MDL) rules originally proposed by Akaike [1], Schwartz [35], and Rissanen [29]. Because the original AIC and MDL were derived by way of asymptotic assumptions, however, their utility in any form for model selection of decaying sinusoids (or any other transient data model for that matter) must be carefully examined.

In this paper, following the contemporary theory of Bayesian statistical inference, we derive a maximum *a posteriori* estimator for the number of damped sinusoids in additive white noise. Predictive densities and estimation-validation techniques [6], [7], [22], [26] are used to construct selection criteria for the models. The predictive densities are formed by splitting the data into two mutually exclusive sets ("training" data and "validation" data), one of which (i.e., the training data) is used to obtain prior predictive densities for the parameters of each model. The remaining validation data are used to assess the likelihood of each model. It is shown that the best results are obtained when the training data comprise the minimum possible number of the last samples of the time series.

Our approach results in a pair of complicated integrals that are solved numerically by Monte Carlo importance sampling integration. A multivariate Cauchy probability density function (p.d.f.) is employed to produce stratified random variates over the integrands support region. By exploiting the special structure of the integrand (i.e., the CLF), the FML procedure yields the centrality location parameters for the Cauchy. The spread parameters are set by matching the support region of the Cauchy with that of the integrand, one dimension at a time. Computer simulations on two-component damped sinusoidal data demonstrate the relative efficacy of our procedure in comparison with the SVD-based information theoretic criteria of Reddy and Biradar [28]. In particular, our results expose the shortcomings of these rules when the data record length is not properly coupled with the information bearing portion of the signal.

The paper is organized as follows. In Section II, the general form of the MAP criterion is derived and the philosophy behind predictive densities is explained in detail. Section III considers the special case of model order selection of multiple damped sinusoids in white Gaussian noise. A brief exposition into Monte Carlo importance sampling integration is provided in Section IV and a justification of the Cauchy importance function is provided therein. Discussions and simulation results are provided in Sections V and VI, and finally in Section VII, conclusions are drawn.

Manuscript received November 3, 1994; revised September 8, 1995. This work was supported by the National Science Foundation under Award MIP-9110682. The associate editor coordinating the review of this paper and approving it for publication was Dr. Yingbo Hua.

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

Publisher Item Identifier S 1053-587X(96)02389-4.

II. Order Selection via the MAP Criterion and Predictive Densities

A. The MAP Criterion

The general problem of interest may be characterized by the following:

$$\begin{aligned}\mathcal{M}_0 &: x[n] = \epsilon(n;\boldsymbol{\psi}), \quad n \in Z_N \\ \mathcal{M}_q &: x[n] = \sum_{i=1}^{q} s_i(n;\boldsymbol{\theta}_i) + \epsilon(n;\boldsymbol{\psi}), \\ &\qquad n \in Z_N, \quad q \in \{1,2,\cdots,Q-1\}. \end{aligned} \tag{1}$$

Here, $\mathcal{M}_q$ represents a qth-order model, and $\mathcal{M}_0$ the "noise only" model. Q represents the number of models under consideration, and $Z_N \equiv \{0,1,\cdots,N-1\}$ denotes a finite set of nonnegative integers. Both N and Q are presumed known. The signal components $s_i(n;\boldsymbol{\theta}_i)$ are completely specified up to the unknown parameter vectors $\{\boldsymbol{\theta}_i\}_{1\le i\le q}$. The noise samples $\epsilon(n;\boldsymbol{\psi})$ are a sequence of random variables whose population distribution is known, but whose characteristic parameters $\boldsymbol{\psi}$ are not. The model order q is also unknown, and the objective is to estimate q according to

$$\hat{q}_{\text{MAP}} = \arg\max_{q\in Z_Q}\{p(q\,|\,\mathbf{x})\} \tag{2}$$

where $p(q\,|\,\mathbf{x})$ is the posterior probability mass function of q given the data $\mathbf{x}$, and $Z_Q \equiv \{0,1,\cdots,Q-1\}$.

From Bayes' theorem we can write

$$p(q\,|\,\mathbf{x}) = \frac{f(\mathbf{x}\,|\,q)p(q)}{f(\mathbf{x})} \propto f(\mathbf{x}\,|\,q)$$

where we have assumed that all the models are equiprobable *a priori*. Marginalizing over the nuisance parameters in the usual way we obtain

$$f(\mathbf{x}\,|\,q) = \int_{\boldsymbol{\Psi}}\int_{\boldsymbol{\Theta}_q} f(\mathbf{x}\,|\,\boldsymbol{\theta}_q,\boldsymbol{\psi},q) f(\boldsymbol{\theta}_q,\boldsymbol{\psi}\,|\,q)\,d\boldsymbol{\theta}_q d\boldsymbol{\psi}. \tag{3}$$

The term $f(\mathbf{x}\,|\,\boldsymbol{\theta}_q,\boldsymbol{\psi},q)$ in (3) represents the likelihood of the parameters given the observed data $\mathbf{x}$, while the second, $f(\boldsymbol{\theta}_q,\boldsymbol{\psi}\,|\,q)$, denotes the prior p.d.f. of the unknown parameters for a q-component model. $\boldsymbol{\Theta}_q$ and $\boldsymbol{\Psi}$ are the parameter spaces of $\boldsymbol{\theta}_q$ and $\boldsymbol{\psi}$, respectively.

B. On the Choice of a Prior

To maintain the analytical tractability of the problem, we would certainly like to select a proper prior,[1] which is a member of the same natural conjugate family of distributions as is the likelihood function. This approach must be rejected, however, unless one can be found that is strongly justified by valid physical arguments. As was pointed out in [22], this is a common problem with the "fully" Bayesian approach to model selection, and is the main reason that modified versions are so common in practice.

[1] A "proper" prior is defined as one that retains the basic properties of a probability density function. That is, it is strictly nonnegative, and integrates or sums to unity over its admissible range of values.

To eliminate the bias that can result from a proper prior, one may opt for a noninformative one. Generally speaking, noninformative priors are improper, but are attractive because they reflect little information relative to that which is expected to be provided by the data. Unfortunately, the direct application of noninformative priors also encounters a serious setback—it results in arbitrary model selection rules [6], [26].

Formally, a noninformative prior for a model k is written as

$$f(\boldsymbol{\theta}_k,\boldsymbol{\psi}\,|\,k) = c_k \cdot g(\boldsymbol{\theta}_k,\boldsymbol{\psi}\,|\,k)$$

where $g(\cdot)$ is a function whose integral diverges over the parameter space and c_k is an unknown constant. When applying the purely noninformative Bayesian approach to the analysis of a single model k, the posterior of the model parameters is

$$\begin{aligned} f(\boldsymbol{\theta}_k,\boldsymbol{\psi}\,|\,\mathbf{x},k) &= \frac{c_k f(\mathbf{x}\,|\,\boldsymbol{\theta}_k,\boldsymbol{\psi},k) g(\boldsymbol{\theta}_k,\boldsymbol{\psi}\,|\,k)}{f(\mathbf{x}\,|\,k)} \\ &= \frac{c_k f(\mathbf{x}\,|\,\boldsymbol{\theta}_k,\boldsymbol{\psi},k) g(\boldsymbol{\theta}_k,\boldsymbol{\psi}\,|\,k)}{c_k \int_{\boldsymbol{\Theta}_k,\boldsymbol{\Psi}} f(\mathbf{x}\,|\,\boldsymbol{\theta}_k,\boldsymbol{\psi},k) g(\boldsymbol{\theta}_k,\boldsymbol{\psi}\,|\,k)\,d\boldsymbol{\theta}_k d\boldsymbol{\psi}}. \end{aligned} \tag{4}$$

As long as the integral in the denominator converges, the posterior is well defined despite the fact that c_k is unspecified (since a cancelation occurs between the numerator and denominator). This is not the case when evaluating the posterior odds of two models, however. For example, consider the following likelihood ratio $\mathcal{L}_{j,k}(\mathbf{x})$ of two models j and k with noninformative priors $f(\boldsymbol{\theta}_j,\boldsymbol{\psi}\,|\,j) = c_j \cdot g(\boldsymbol{\theta}_j,\boldsymbol{\psi}\,|\,j)$ and $f(\boldsymbol{\theta}_k,\boldsymbol{\psi}\,|\,k) = c_k \cdot g(\boldsymbol{\theta}_k,\boldsymbol{\psi}\,|\,k)$

$$\mathcal{L}_{j,k}(\mathbf{x}) = \frac{c_j}{c_k} \cdot \frac{\int_{\boldsymbol{\Theta}_j,\boldsymbol{\Psi}} f(\mathbf{x}\,|\,\boldsymbol{\theta}_j,\boldsymbol{\psi},j) g(\boldsymbol{\theta}_j,\boldsymbol{\psi}\,|\,j)\,d\boldsymbol{\theta}_j d\boldsymbol{\psi}}{\int_{\boldsymbol{\Theta}_k,\boldsymbol{\Psi}} f(\mathbf{x}\,|\,\boldsymbol{\theta}_k,\boldsymbol{\psi},k) g(\boldsymbol{\theta}_k,\psi\,|\,k)\,d\boldsymbol{\theta}_k d\boldsymbol{\psi}}.$$

From this example it is clear that the unspecified constants do not cancel, and they must now somehow be specified. This situation is similar to threshold setting in multiple hypothesis testing (a problem we certainly want to avoid). In order to overcome this problem, we will use an estimation-validation approach implemented by Bayesian predictive densities.

Proceeding with this method, we partition the data $\mathbf{x}$ into two mutually exclusive sets, $\mathbf{x}_R$ and $\mathbf{x}_{N-R}$. Here the subscripts R and $(N-R)$ denote that $\mathbf{x}_R$ and $\mathbf{x}_{N-R}$ are composed of R and $N-R$ samples, respectively. We then make the approximation

$$\begin{aligned} \max_{q\in Z_Q}\{p(q\,|\,\mathbf{x})\} &= \max_{q\in Z_Q}\{f(\mathbf{x}\,|\,q)\} \\ &\approx \max_{q\in Z_Q}\{f(\mathbf{x}_{N-R}\,|\,\mathbf{x}_R,q)\}. \end{aligned} \tag{5}$$

Now consider the marginalization in (5)

$$\begin{aligned} &f(\mathbf{x}_{N-R}\,|\,\mathbf{x}_R,q) \\ &\quad = \int_{\boldsymbol{\Psi}}\int_{\boldsymbol{\Theta}_q} \underbrace{f(\mathbf{x}_{N-R}\,|\,\mathbf{x}_R,\boldsymbol{\theta}_q,\boldsymbol{\psi},q)}_{\text{likelihood}} \underbrace{f(\boldsymbol{\theta}_q,\boldsymbol{\psi}\,|\,\mathbf{x}_R,q)}_{\text{prior}}\,d\boldsymbol{\theta}_q d\boldsymbol{\psi}. \end{aligned} \tag{6}$$

The first term $f(\mathbf{x}_{N-R}\,|\,\mathbf{x}_R,\boldsymbol{\theta}_q,\boldsymbol{\psi},q)$ in the integrand is the predictive density of $\mathbf{x}_{N-R}$ based on the data $\mathbf{x}_R$ and the model parameters. The second function $f(\boldsymbol{\theta}_q,\boldsymbol{\psi}\,|\,\mathbf{x}_R,q)$ is the

prior density of the unknown parameters, which can also be interpreted as the *posterior* density of the parameters given the data $\mathbf{x}_R$.

Note that (5) can be written as

$$\max_{q\in Z_Q}\{f(\mathbf{x}_{N-R}\,|\,\mathbf{x}_R,q)\}=\max_{q\in Z_Q}\left\{\frac{f(\mathbf{x}_N\,|\,q)}{f(\mathbf{x}_R\,|\,q)}\right\} \tag{7}$$

or

$$\hat{q}_{\mathrm{MAP}}=\arg\max_{q\in Z_Q}\left\{\frac{\int_{\boldsymbol{\Psi}}\int_{\boldsymbol{\Theta}_q} f(\mathbf{x}_N\,|\,q,\boldsymbol{\theta}_q,\boldsymbol{\psi})f(\boldsymbol{\theta}_q,\boldsymbol{\psi}\,|\,q)d\boldsymbol{\theta}_q d\boldsymbol{\psi}}{\int_{\boldsymbol{\Psi}}\int_{\boldsymbol{\Theta}_q} f(\mathbf{x}_R\,|\,q,\boldsymbol{\theta}_q,\boldsymbol{\psi})f(\boldsymbol{\theta}_q,\boldsymbol{\psi}\,|\,q)d\boldsymbol{\theta}_q d\boldsymbol{\psi}}\right\}. \tag{8}$$

Since the *identical* prior p.d.f.'s $f(\boldsymbol{\theta}_q,\boldsymbol{\psi}\,|\,q)$ are now present in both the numerator and denominator, when we specify them as noninformative and the arbitrary constants appear, a cancellation effect (similar to that in (4)) takes place between the numerator and denominator.

C. On Partitioning of the Data into Estimation and Validation Subsets

An important issue concerning the application of predictive densities is the manner in which the data are partitioned. Many schemes have been devised over the years for accomplishing this task [8], [15], [24], [27], [30], [37], each with its relative advantages and disadvantages. The transient nature of damped sinusoidal data combined with our pursuit of a minimally informative proper prior suggests that the training data should comprise the samples that contain the least information. It therefore seems plausible that $\mathbf{x}_R$ should consist of the minimum possible number of the latter R samples of $\mathbf{x}$.[2] To clarify this notion, consider the following transient data model:

$$x[n]=ae^{-\alpha n}\cos(2\pi fn)+\epsilon[n],\quad n=0,1,\cdots,N-1 \tag{9}$$

where the error sequence $\epsilon[n]\overset{iid}{\sim}\mathcal{N}(0,\sigma^2)$. In matrix form, (9) can be written as

$$\mathbf{x}=\mathbf{h}a+\boldsymbol{\epsilon}$$

where $\mathbf{h}=[1\;\; e^{-\alpha}\cos(2\pi f)\;\; e^{-2\alpha}\cos(4\pi f)\;\cdots\; e^{-(N-1)\alpha}\cos(2(N-1)\pi f)]^T$. For clarity, take the noise variance σ^2 and the nonlinear parameters f and α to be known constants. Next we split the data $\mathbf{x}$ into $\mathbf{x}_R$ and $\mathbf{x}_{N-R}$. The prior density of the parameters in (6) can then be viewed as the *posterior* of the unknown amplitude a (given the training data $\mathbf{x}_R$). The quantity $(a\,|\,\mathbf{x}_R)\sim\mathcal{N}(\hat{a}_R,\sigma^2_{a_R})$, and thus

$$f(a\,|\,\mathbf{x}_R)\propto e^{-\frac{1}{2\sigma^2_{a_R}}(a-\hat{a}_R)^2}$$

where $\hat{a}_R=(\mathbf{h}_R^T\mathbf{h}_R)^{-1}\mathbf{h}_R^T\mathbf{x}_R$ and $\sigma^2_{a_R}=\sigma^2(\mathbf{h}_R^T\mathbf{h}_R)^{-1}$ [2]. Since the variance $\sigma^2_{a_R}$ is inversely proportional to $\mathbf{h}_R^T\mathbf{h}_R$, it is clear that increasing $\mathbf{h}_R^T\mathbf{h}_R$ causes $\sigma^2_{a_R}$ to decrease. Furthermore, $\mathbf{h}_R^T\mathbf{h}_R=\sum_{i\in K_R}h_i^2$ (here, h_i is the ith element of $\mathbf{h}_R$, and $K_R\subset Z_N$ consists of R elements of Z_N), so as R increases, $\mathbf{h}_R^T\mathbf{h}_R$ increases, causing $\sigma^2_{a_R}$ to decrease. The prior density of a is therefore increasingly more informative (smaller $\sigma^2_{a_R}$) for larger R. Also note that $\mathbf{h}_R^T\mathbf{h}_R$ is largest when the samples are taken from the beginning of the time series and smallest when they are taken from the end.

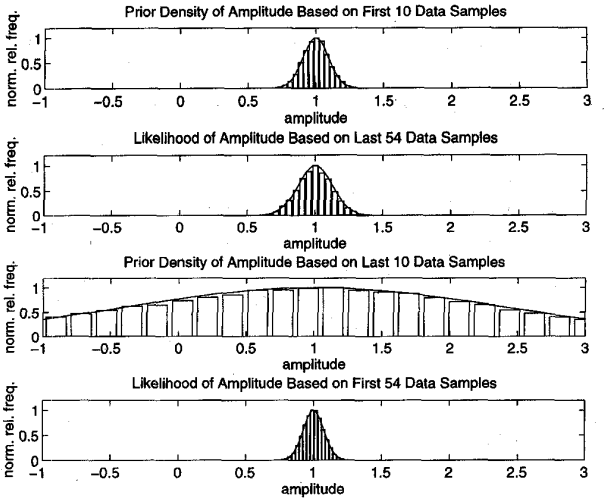


Fig. 1. Prior densities for a single unknown amplitude parameter and the corresponding likelihood functions. The top two distributions depict a prior based on the first ten samples and a likelihood based on the remaining 54 samples. The bottom two distributions show the prior based on the last ten samples and the likelihood based on the first 54 samples. Clearly, the topmost prior is informative for a, while the prior in the third diagram is relatively noninformative for a.

For demonstrative purposes, consider the model in (9) with $a=1,\ f=0.24,\ \alpha=0.05, N=64$, and σ^2 set to provide a signal-to-noise ratio (SNR)[3] of 15 dB. We generated 10 000 realizations according to (9) and plotted histograms of $\hat{a}_R$ (this is the empirical density $f(a\,|\,\mathbf{x}_R)$ based on 10 000 trials) for two cases. For the first we constructed $f(a\,|\,\mathbf{x}_R)$ with the *first* $R=10$ data samples (which entails that the corresponding likelihood of a was based on the *last* 54 observations), and in the second case we formed $f(a\,|\,\mathbf{x}_R)$ using the *last* $R=10$ samples (implying that the likelihood of a was based on the *first* 54 samples). Fig. 1 shows the normalized histograms of $\hat{a}_R$ superimposed on the theoretical density $f(a\,|\,\mathbf{x}_R)$, along with the normalized histograms of $\hat{a}_{N-R}$ superimposed on the theoretical likelihoods $f(\mathbf{x}_{N-R}\,|\,\mathbf{x},a)$. The first two distributions are nearly identical, implying that approximately the same information is contained in the prior as is contained in the likelihood function. That is, $f(a\,|\,\mathbf{x}_R)$ is highly informative (relative to its likelihood) when it is based on the first $R=10$ samples. Conversely, the second pair of histograms demonstrate that $f(a\,|\,\mathbf{x}_R)$ is locally uniform over the support region of the likelihood function, and thus, it is indeed noninformative relative to the likelihood function.

It can likewise be shown that decreasing the number of samples R decreases the information in the prior, and increasing R increases the information in the prior. Note that *this is consistent with the initial approximation in* (5). *That is, the*

[2] Since the signal is increasingly inundated by noise as the process evolves with time, the last samples of $\mathbf{x}$ will contain the least amount of information.

[3] In this context, SNR refers to the peak SNR; that is, the signal-to-noise ratio of the first sample defined as

$$\mathrm{SNR}=10\log_{10}\left(\frac{(\text{peak amplitude})^2}{2\sigma^2}\right)\text{ dB}.$$

approximation $f(\mathbf{x}\,|\,q) \approx f(\mathbf{x}_{N-R}\,|\,\mathbf{x}_R, q)$ *is better when* $\mathbf{x}_R$ *contains less information.*

III. Order Selection of Damped Sinusoids in White Gaussian Noise

Consider the data model in (1) wherein the signals $s_i(n;\boldsymbol{\theta}_i)$ represent real damped sinusoids and the noise samples $\epsilon[n] \sim \mathcal{N}(0,\sigma^2)$. For this case, the observed data $\mathbf{x}$ can be represented by

$$\begin{aligned}\mathcal{M}_0 &: x[n] = \epsilon(n,\sigma^2), \quad n \in Z_N \\ \mathcal{M}_q &: x[n] = \sum_{i=1}^{q} a_i e^{-\alpha_i n}\cos(2\pi f_i n + \phi_i) + \epsilon(n;\sigma^2), \\ &\qquad n \in Z_N, \quad q \in \{1,2,\cdots,Q-1\}. \end{aligned} \tag{10}$$

The unknown parameters associated with the ith signal are its amplitude (a_i), frequency (f_i), phase (ϕ_i), and damping factor (α_i). The noise power σ^2 is also assumed unknown. Given the data $\{x_n\}_{n\in Z_N}$, the objective is to estimate the model order q by applying the MAP criterion in (8).

Equation (10) can be written concisely in vector–matrix notation as

$$\mathbf{x} = \mathbf{A}_q\mathbf{a}_q + \boldsymbol{\epsilon}, \quad q \in Z_Q \tag{11}$$

where $\mathbf{x}$ and $\boldsymbol{\epsilon}$ are $N\times 1$ vectors, $\mathbf{a}_q$ is a $q\times 1$ vector of amplitude constants, and $\mathbf{A}_q$ is the $N\times q$ signal manifold matrix whose ith column is of the form

$$\begin{aligned}\mathbf{a}_i = [&\cos(\phi_i) \quad e^{-\alpha_i}\cos(2\pi f_i + \phi_i) \quad \cdots \quad e^{-\alpha_i(N-1)} \\ &\times \cos(2\pi f_i(N-1) + \phi_i)]^T.\end{aligned}$$

Since the noise process is white and Gaussian, the likelihood term can be expressed as

$$f(\mathbf{x}\,|\,q,\boldsymbol{\theta}_q,\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{A}_q\mathbf{a}_q)^T(\mathbf{x}-\mathbf{A}_q\mathbf{a}_q)}. \tag{12}$$

To depict a state of ignorance concerning the unknown parameters, we assign the noninformative Jeffreys' prior $f(\boldsymbol{\theta}_q,\sigma\,|\,q) \propto \sigma^{-1}$ (see [2]). Combining this and (12), the numerator of (8) can be expressed as

$$\begin{aligned}&\propto \Gamma\left(\frac{N-q}{2}\right)\int_{\boldsymbol{\alpha}_q}\int_{\mathbf{f}_q}\int_{\boldsymbol{\phi}_q} |\mathbf{A}_{q,N}^T\mathbf{A}_{q,N}|^{-\frac{1}{2}} \\ &\quad\times \left(\mathbf{x}_N^T\mathbf{P}_N^{\perp}\mathbf{x}_N\right)^{-\left(\frac{N-q}{2}\right)} d\boldsymbol{\phi}_q d\mathbf{f}_q d\boldsymbol{\alpha}_q, \quad N > q. \end{aligned} \tag{13}$$

The matrix $\mathbf{P}_{(\cdot)}^{\perp}$ is the projection operator for the left nullspace of $\mathbf{A}_{q,(\cdot)}$, and $\Gamma(\cdot)$ is the standard gamma function. The denominator in (8) can likewise be marginalized, the result of which is the following MAP model selection criterion for damped sinusoidal signals in white Gaussian noise (see (14) at the bottom of the page). Here the subscripts N and R in the numerator and denominator indicate that they are based on N, and the *last* R samples of the data vector $\mathbf{x}$, respectively. Note that the total dimensionality of the integrals is $3q$ for both numerator and denominator. To lower the total dimension to $2q$, we apply the following transformation to the data model in (10), as follows:

$$\begin{aligned}&\sum_{i=1}^{q} a_i e^{-\alpha_i n}\cos(2\pi f_i n + \phi_i) \\ &\quad = \sum_{i=1}^{q}[a_i\cos\phi_i e^{-\alpha_i n}\cos(2\pi f_i n) \\ &\qquad - a_i\sin\phi_i e^{-\alpha_i n}\sin(2\pi f_i n)]\end{aligned} \tag{15}$$

the right-hand side of which can be expressed in matrix form as

$$\begin{aligned}&\sum_{i=1}^{q}[a_i\cos\phi_i e^{-\alpha_i n}\cos(2\pi f_i n) - a_i\sin\phi_i e^{-\alpha_i n}\sin(2\pi f_i n)] \\ &\qquad = \mathbf{H}_q\mathbf{b}_q, \quad n \in Z_N\end{aligned}$$

where

$$\begin{aligned}\mathbf{H}_q &= [\mathbf{c}_1 \quad \mathbf{s}_1 \quad \mathbf{c}_2 \quad \mathbf{s}_2 \quad \cdots \quad \mathbf{c}_q \quad \mathbf{s}_q], \\ \mathbf{b}_q &= [\mathbf{b}_1^T \quad \mathbf{b}_2^T \quad \cdots \quad \mathbf{b}_q^T]^T\end{aligned}$$

with

$$\begin{aligned}\mathbf{c}_i^T = &\\ &[1 \quad e^{-\alpha_i}\cos(2\pi f_i) \quad \cdots \quad e^{-\alpha_i(N-1)}\cos(2\pi f_i(N-1))], \\ &\qquad\qquad i = 1,2,\cdots,q \\ \mathbf{s}_i^T = &\\ &[0 \quad e^{-\alpha_i}\sin(2\pi f_i) \quad \cdots \quad e^{-\alpha_i(N-1)}\sin(2\pi f_i(N-1))], \\ &\qquad\qquad i = 1,2,\cdots,q \\ \mathbf{b}_i^T = &[a_i\cos\phi_i \quad -a_i\sin\phi_i], \quad i = 1,2,\cdots,q.\end{aligned}$$

Applying the transformation in (15) to (10) and following all the steps that led to (14), we arrive at (16), shown at the bottom of the next page. Note that the dimensions of the integrals in (14) have been reduced by a factor of $\frac{1}{3}$ as a result of the transformation. Still, their evaluation is not trivial by any means. They will obviously not have a closed-form solution, and we therefore resort to numerical methods. Since the dimension of the parameter space is large, a classical technique such as numerical quadrature would be excessively computational. The Monte Carlo method, on the other hand, has long been recognized as a powerful alternative to performing calculations that are considered too complicated for classical techniques.

$$\hat{q}_{\mathrm{MAP}} = \arg\max_{q\in Z_Q}\left\{\frac{\Gamma\left(\frac{N-q}{2}\right)\int_{\boldsymbol{\alpha}_q}\int_{\mathbf{f}_q}\int_{\boldsymbol{\phi}_q}|\mathbf{A}_{q,N}^T\mathbf{A}_{q,N}|^{-\frac{1}{2}}\left(\mathbf{x}_N^T\mathbf{P}_N^{\perp}\mathbf{x}_N\right)^{-\left(\frac{N-q}{2}\right)}d\boldsymbol{\phi}_q d\mathbf{f}_q d\boldsymbol{\alpha}_q}{\Gamma\left(\frac{R-q}{2}\right)\int_{\boldsymbol{\alpha}_q}\int_{\mathbf{f}_q}\int_{\boldsymbol{\phi}_q}|\mathbf{A}_{q,R}^T\mathbf{A}_{q,R}|^{-\frac{1}{2}}\left(\mathbf{x}_R^T\mathbf{P}_R^{\perp}\mathbf{x}_R\right)^{-\left(\frac{R-q}{2}\right)}d\boldsymbol{\phi}_q d\mathbf{f}_q d\boldsymbol{\alpha}_q}\right\}, \qquad R > Q-1. \tag{14}$$

IV. Monte Carlo Importance Sampling Integration

A. Overview of Importance Sampling

The fundamental concept behind simple Monte Carlo integration is to *uniformly* sample M points $\{\mathbf{y}_i\}_{1\le i\le M}$ from a multidimensional volume $\mathcal{V}$. Then the Monte Carlo estimate of the integral of a function ξ over $\mathcal{V}$ is [18], as follows:

$$\int \xi d\mathcal{V} \approx \underbrace{\mathcal{V}\cdot\langle\xi\rangle}_{\text{integral estimate}} \pm \underbrace{\mathcal{V}\cdot\sqrt{\frac{\langle\xi^2\rangle-\langle\xi\rangle^2}{M}}}_{\text{standard deviation of error}} \tag{17}$$

where

$$\langle\xi\rangle \equiv \frac{1}{M}\sum_{i=1}^{M}\xi(\mathbf{y}_i) \quad \text{and} \quad \langle\xi^2\rangle \equiv \frac{1}{M}\sum_{i=1}^{M}\xi^2(\mathbf{y}_i).$$

Clearly, the accuracy of the procedure depends on the error variance of the estimate. The variance can be reduced by applying importance sampling. With this technique, the M samples $\mathbf{y}_i$ are not taken uniformly, but rather, are stratified so that they are clustered in regions of $\mathcal{V}$ where the magnitude of $\xi(\cdot)$ is largest. This overpopulation is compensated for by reducing the effective weight of the function in this region. The reweighted function then becomes more nearly constant, thereby reducing the variance of the integral estimate.

To clarify this idea, let us suppose that M variates $\mathbf{y}_i$ are generated according to a general p.d.f. $h(\cdot)$, where

$$\int h(\cdot)d\mathcal{V} = 1.$$

Then the integral of any function $\xi(\cdot)$ can be estimated as

$$\int_{\mathcal{V}}\xi(\mathbf{y})d\mathcal{V} = \int_{\mathcal{V}}\frac{\xi(\mathbf{y})}{h(\mathbf{y})}\cdot h(\mathbf{y})d\mathcal{V} \equiv E\left(\frac{\xi(\mathbf{y})}{h(\mathbf{y})}\right) \approx \left\langle\frac{\xi(\mathbf{y}_i)}{h(\mathbf{y}_i)}\right\rangle = \frac{1}{M}\sum_{i=1}^{M}\frac{\xi(\mathbf{y}_i)}{h(\mathbf{y}_i)}$$

where $E(\cdot)$ is the well-known expectation operator. Note that if the function $\xi(\cdot)$ is reweighted to

$$\frac{\xi(\cdot)}{h(\cdot)} \approx c \equiv \text{constant}$$

then the standard deviation of the error in the integral estimate is

$$\mathcal{V}\cdot\sqrt{\frac{\langle\xi^2\rangle-\langle\xi\rangle^2}{M}} \Longrightarrow \mathcal{V}\cdot\sqrt{\frac{\left\langle\left(\frac{\xi}{h}\right)^2\right\rangle-\left\langle\frac{\xi}{h}\right\rangle^2}{M}} \longrightarrow 0, \quad \text{as } \frac{\xi}{h}\to c \tag{18}$$

and the error variance is reduced. It should also be noted that importance sampling is the only known means by which infinite singularities (or near singularities) can be "removed" from the integrand (this is accomplished by sampling from an importance function with a similar singularity in the same location).

The asymptotic error variance of an importance sampling estimate strongly depends on the density $h(\cdot)$ (also known as the importance function) [23]. The three most important properties of a good importance function are as follows.

- The simplicity by which the random variates $\mathbf{y}_i$ can be generated from $h(\cdot)$.
- $h(\cdot)$ should have longer tails than the integrand $\xi(\cdot)$.
- $h(\cdot)$ should be a close approximation to $\xi(\cdot)$.

The decision to use a Monte Carlo procedure was influenced mainly by the relationship between the dimensionality of the parameter space and the convergence rate of the integral estimate to the true value. From (17) it is clear that the uncertainty of a Monte Carlo integral estimate decreases as $M^{-\frac{1}{2}}$, independently of dimensionality. Classical multidimensional quadrature techniques, on the other hand, maintain a given integration accuracy only at the expense of exponentially increasing the number of functional evaluations. Therefore, a quadrature rule requiring M functional evaluations in one dimension will require M^p evaluations in p dimensions to maintain the same accuracy. This slows down the convergence rate in p dimensions by a factor of $\frac{1}{p}$. Since the convergence rate of Monte Carlo is independent of p, there is always some p for which Monte Carlo is more efficient[4] than the popular quadrature rules.

The importance sampling procedure decreases computation time even further, since sampling from $h(\cdot)$ allows for fewer samples to be taken. So, although the implementation of (16) by importance sampling integration is fairly intensive, it is certainly more efficient than simple Monte Carlo or the classical alternatives. For detailed discussions on importance sampling, cf. [5], [14], [20], [34] and [36].

B. On the Choice of an Importance Function

Investigations into the integrands in (16) have shown them to be very sharply peaked, particularly for high SNR and/or large N. Fig. 2 depicts a typical realization of the normalized 3-D surface and corresponding contour plot of the integrand $|\mathbf{H}_q^T\mathbf{H}_q|^{-\frac{1}{2}}(\mathbf{x}_N^T\mathbf{P}_N^{\perp}\mathbf{x}_N)^{-(\frac{N-2q}{2})}$ in the numerator of (16) over the (α, f) plane for $N = 64$ samples, and an SNR = 20 dB. The model order was set at $q = 1$. The true values of the signal parameters were set at $a_1 = 1.0$, $f_1 = 0.2$, $\phi_1 = 0.0$, $\alpha_1 = 0.15$. The characteristics of this integrand have been confirmed by observing many realizations of data from a variety of damped sinusoidal models. The surface in Fig. 2

[4] "More efficient" in the sense that it takes less computation time to achieve a given error in integration.

$$\hat{q}_{\text{MAP}} = \arg\max_{q\in Z_Q}\left\{\frac{\Gamma\left(\frac{N-2q}{2}\right)\int_{\boldsymbol{\alpha}_q}\int_{\mathbf{f}_q}|\mathbf{H}_{q,N}^T\mathbf{H}_{q,N}|^{-\frac{1}{2}}\left(\mathbf{x}_N^T\mathbf{P}_N^{\perp}\mathbf{x}_N\right)^{-(\frac{N-2q}{2})}d\mathbf{f}_q d\boldsymbol{\alpha}_q}{\Gamma\left(\frac{R-2q}{2}\right)\int_{\boldsymbol{\alpha}_q}\int_{\mathbf{f}_q}|\mathbf{H}_{q,R}^T\mathbf{H}_{q,R}|^{-\frac{1}{2}}\left(\mathbf{x}_R^T\mathbf{P}_R^{\perp}\mathbf{x}_R\right)^{-(\frac{R-2q}{2})}d\mathbf{f}_q d\boldsymbol{\alpha}_q}\right\}, \qquad R > 2(Q-1). \tag{16}$$

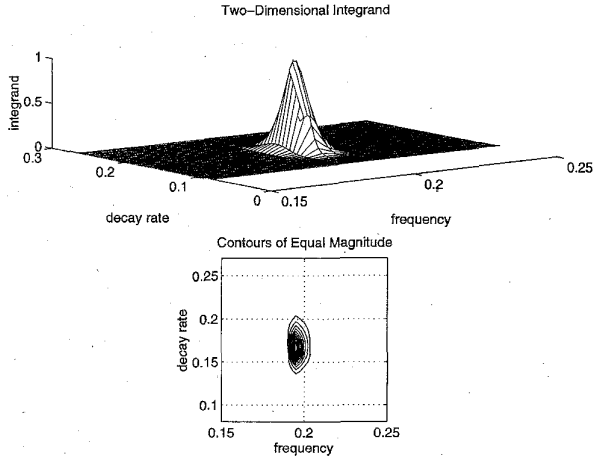


Fig. 2. Normalized 2-D integrand for a single damped sinusoid in noise. The SNR = 20 dB, and the number of data samples in $N = 64$. These diagrams should be compared with those in Fig. 3.

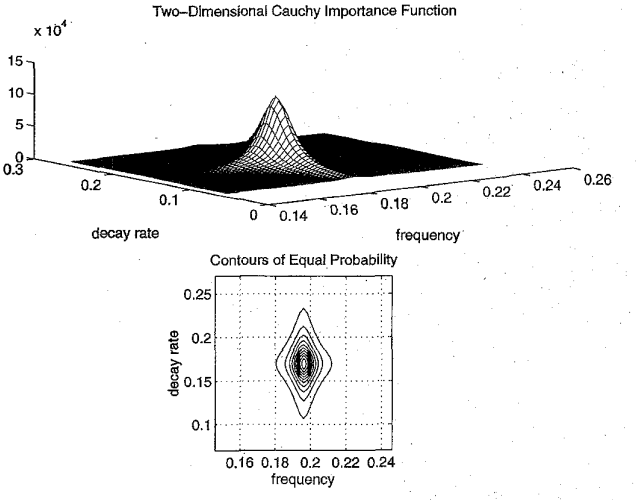


Fig. 3. Two-dimensional Cauchy importance function with corresponding contours of equal probability. Maximum likelihood estimates of frequency and decay location parameters are 0.1960 and 0.1700, respectively. The spread parameters are $\tilde{\sigma}_f = 1 \times 10^{-4}$ and $\tilde{\sigma}_\alpha = 1 \times 10,^{-2}$ respectively. These diagrams should be compared with those in Fig. 2.

represents a typical realization. This observation has also been verified in [38].

These properties prompted us to consider both multivariate normal and multivariate Cauchy distributions as importance functions. Based on stability considerations we decided upon the latter, for it is well known that Monte Carlo importance sampling is unstable for importance functions that pass through zero, or which approach zero quickly (such as Gaussian p.d.f.'s) [17]. The Cauchy has longer tails and a sharp peak, while Gaussians can be sharply peaked only at the expense of shorter tails. The short tails can cause stability problems since if the integrand should happen to approach zero slower than the Gaussian, then

$$\frac{\xi(\cdot)}{h(\cdot)} \longrightarrow \infty \tag{19}$$

and the resulting error variance in (18) will approach infinity. It is therefore dangerous to choose importance functions that approach zero quickly (such as Normal p.d.f.'s) [17], especially when one is attempting to match a highly concentrated integrand with a condensed importance sampling function.

Upon deciding on the Cauchy importance function, all that remains is the initialization of its location and spread parameters. The spread parameters are found by matching the support region of the Cauchy with that of the integrand. To see how this is accomplished, compare the contours of Fig. 2 with those of Fig. 3. Clearly, the importance function covers the same region as does the integrand. The location parameters are the maximum likelihood estimates of the peak of the CLF. They are found with the FML estimation procedure of Umesh and Tufts [38]. The importance sampling procedure is more sensitive to the initial frequency estimates $\hat{f}_i$ than to the those of the decay rates $\hat{\alpha}_i$. This is not unexpected since the compressed likelihood function is much smoother in the $\boldsymbol{\alpha}$ subspace than in the f subspace (cf. Fig. 2 for the 2-D case).

V. Discussion

The model selection rules in [28] are based on the information theoretic methods of Wax and Kailath [41], which are the SVD linear predictive versions of the AIC [1] and MDL [29], [35]. They are attractive in that they do not require the solution of nonlinear equations for the determination of the maximum likelihood estimates (MLE's) of the models parameters. This makes them computationally advantageous. There are, however, some key issues concerning the applicability of these rules to damped sinusoidal data (or any transient data) that need to be addressed, as follows.

- The $\mathrm{AIC}_{\mathrm{SVD}}$ criterion is simply a restatement of the original (1974) AIC, which was designed to be an *asymptotically* unbiased estimator of the Kullback–Leibler information. It is well known that the AIC is inconsistent and has a tendency to overparameterize.
- The $\mathrm{MDL}_{\mathrm{SVD}}$ criterion is a restatement of the original MDL in [29], which was based on the notion of compact encoding of data introduced by Wallace [39] in 1968 with the minimum message length (MML) procedure. The idea of combining optimal coding theory with inductive statistical inference was later expanded in [31], [32], and [40]. Since the MDL is based on large sample approximations, it is questionable in terms of its applicability to model order selection of transients (damped sinusoids being just one of the many examples).
- The penalty function of the $\mathrm{MDL}_{\mathrm{SVD}}$ monotonically increases with the data record length. Our claim is that this form of penalization is incorrect for transient signals. With transients, each additional sample contains less relevant information than the previous one, and, thus, the penalization should not continually increase with N. Instead, the penalty should be directly related to the determinant of the observed Fisher information matrix (i.e., the Hessian of the log likelihood of the model parameters). For example, consider Fig. 4 which shows the log determinant of the observed Fisher information

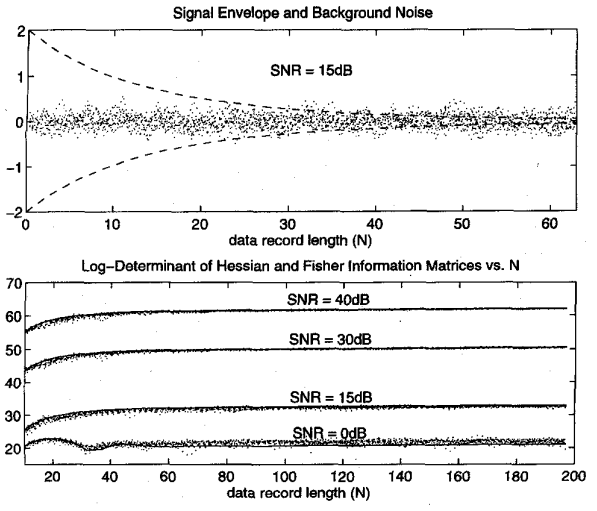


Fig. 4. Top figure displays the signal envelope on 15 realizations of background noise at SNR = 15 dB. The log determinant of the Hessian and Fisher information matrices as a function of data record length (N) are shown in the bottom figure. The curves depict 15 realizations of the Hessian for a model consisting of two damped sinusoids in white Gaussian noise for SNR's of 0 dB, 15 dB, 30 dB, and 40 dB. The log determinants of the Fisher information are shown by the solid curves.

matrix for the two-component damped sinusoidal model

$$x[n] = \sqrt{20}e^{-0.1n}\sin(2\pi(0.2)n) + \sqrt{20}e^{-0.05n}\sin\left(2\pi(0.24)n + \frac{\pi}{4}\right) + \epsilon(n), \quad n = 0, 1, \cdots, N-1.$$

The curves display the results of 15 realizations for SNR's of 0 dB, 15 dB, 30 dB, and 40 dB. Clearly, beyond 50–60 samples the determinant is nearly constant, and these are precisely the points at which the signal energy becomes negligible in comparison to the noise intensity (cf. Fig. 4 (top)). These results indicate that the MML or MDL will provide more accurate selection results than those obtained by the MDL in [28] if the penalties are derived from the Fisher information matrix [9], [19].

VI. Simulation Results

The relative accuracy of the MAP criterion was established by comparing it with the SVD-based AIC and MDL model selection rules in [28]. We considered two experiments. The first quantified the performances of the three criteria as a function of SNR with the data record length N held constant. In the second experiment, the data record length N was varied and the SNR fixed.

A. Experiment 1

The data model was given by

$$x[n] = 1.0e^{-0.1n}\cos(2\pi(0.2)n) + 1.0e^{-0.05n} \times \cos(2\pi(0.24)n) + \epsilon[n], \quad n = 0, 1, \cdots, 63.$$

The noise variance σ^2 was set to provide an SNR ranging between 10 dB and 21 dB, in 1-dB increments. The training data for the predictive densities were formed with the latter eight samples of $\mathbf{x}$. For each of the hypothesized model orders H_i, $i = 0, 1, 2, 3$, the importance sampling integration was based on $M = 2000$ random variates from a multivariate Cauchy density. The spread parameters for the frequencies were set to $\tilde{\sigma}_{f_i} = 1 \times 10^{-4}$, while those for the decay constants were $\tilde{\sigma}_{\alpha_i} = 1 \times 10^{-2}$, $i = 1, 2, 3$. The location parameters for the Cauchy were estimated via the FML algorithm. The SVD-based AIC and MDL were implemented with the modified backward linear prediction formulation, with a prediction filter order of length 32. This is the same order that was used in [28].

TABLE I
Performance Comparison Between the MAP and SVD-Based AIC and MDL Criteria for Various Signal-to-Noise Ratios. Entries Indicate the Number of Times Out of 100 Independent Trials That the Given Criterion Selected a Particular Model. The Number of Samples is $N = 64$ and the Correct Model Order is Two (H_2)

SNR	Criterion	H_0	H_1	H_2	H_3
10 dB	AIC	0	33	42	25
	MDL	0	52	37	11
	MAP	63	10	21	6
11 dB	AIC	0	22	50	28
	MDL	0	36	52	12
	MAP	31	11	58	0
12 dB	AIC	0	13	55	32
	MDL	0	29	64	7
	MAP	17	14	64	5
13 dB	AIC	0	6	58	36
	MDL	0	16	69	15
	MAP	1	23	76	0
14 dB	AIC	0	3	65	32
	MDL	0	5	84	11
	MAP	1	19	76	4
15 dB	AIC	0	0	68	32
	MDL	0	2	89	9
	MAP	0	16	84	0
16 dB	AIC	0	0	72	28
	MDL	0	0	95	5
	MAP	0	12	86	2
17 dB	AIC	0	0	73	27
	MDL	0	0	96	4
	MAP	0	6	94	0
18 dB	AIC	0	0	75	25
	MDL	0	0	96	4
	MAP	0	2	98	0
19 dB	AIC	0	0	75	25
	MDL	0	0	96	4
	MAP	0	1	98	1
20 dB	AIC	0	0	77	23
	MDL	0	0	96	4
	MAP	0	0	100	0
21 dB	AIC	0	0	77	23
	MDL	0	0	96	4
	MAP	0	0	100	0

Table I contains the selection results for 100 independent trials. The number of observations was set at $N = 64$. From Table I it is clear that the accuracies of the MAP and MDL

TABLE II
PERFORMANCE COMPARISON BETWEEN THE MAP AND SVD-BASED AIC AND MDL CRITERIA FOR VARIOUS DATA RECORD LENGTHS. ENTRIES INDICATE THE NUMBER OF TIMES OUT OF 100 INDEPENDENT TRIALS THAT THE GIVEN CRITERION SELECTED A PARTICULAR MODEL. THE CORRECT MODEL ORDER IS TWO (H_2) AND FOR ALL TRIALS THE SNR WAS FIXED AT 15 dB

Data Record Length N	Criterion	H_0	H_1	H_2	H_3
64 samples	AIC	0	0	68	32
	MDL	0	2	89	9
	MAP	0	16	84	0
100 samples	AIC	0	1	61	38
	MDL	0	2	79	19
	MAP	0	12	88	0
128 samples	AIC	0	5	54	41
	MDL	0	11	69	20
	MAP	0	7	93	0
150 samples	AIC	0	7	38	55
	MDL	0	13	55	32
	MAP	0	5	94	1
200 samples	AIC	0	21	34	45
	MDL	0	31	40	29
	MAP	0	2	98	0
256 samples	AIC	0	24	25	51
	MDL	0	46	29	25
	MAP	0	3	97	0

criteria were essentially the same, while the AIC was consistently below them. As is typically the case, the AIC tended to overparameterize. Although the MDL overparameterized less frequently than the AIC, it did so more often than the MAP criterion. The MAP rule rarely overestimated the model order, and for low SNR's it tended to select the noise-only model (H_0). (This is entirely reasonable since, for low SNR's, the information in the data is totally inundated by noise.)

B. Experiment 2

In the following, we will draw the distinction between the MAP and SVD-based AIC and MDL criteria as a function of N. The SNR was set to 15 dB, and we considered the original two-component data model

$$x[n] = 1.0e^{-0.1n}\cos(2\pi(0.2)n) + 1.0e^{-0.05n} \times \cos(2\pi(0.24)n) + \epsilon[n], \quad n = 0, 1, \cdots, N-1$$

for sequence lengths of $N = 64, 100, 128, 150, 200$, and 256 samples. Each case consisted of 100 independent trials. For each value of N, the prediction filter order was adjusted so that the backward linear prediction data matrix for the AIC and MDL remained square. This allowed the AIC and MDL to perform optimally [21]. The results of this experiment are shown in Table II.

Clearly, the performance of the AIC and MDL deteriorated as the number of data samples exceeded the information-bearing portion of the observation vector. Conversely, the performance of the MAP criterion improved with more data. For example, when $N = 64$ samples, the MDL and MAP performed almost identically, but as N increased to 128 samples, the disparity between the two criteria became apparent. For $N = 128$, the correct selection probability of the MDL *decreased* to 0.69, while that of the MAP *increased* to 0.93. This overall trend persisted as well. When $N = 256$ samples, the correct selection probabilities of the MDL and MAP decreased and increased to 0.29 and 0.97, respectively. This difference is certainly much larger than the 0.89 and 0.84 correct selection probabilities that resulted when $N = 64$ samples. It is obvious that the accuracy of both the AIC's and MDL's deteriorate with increasing N. The MAP criterion, on the other hand, showed an initial improvement as N increased. It then leveled off for $N > 128$ samples. This result certainly seems more logical. That is, the performance of any statistical criterion should improve to some extent when the phenomenon under study is observed over a longer time interval, and if it does not improve, it certainly should not deteriorate.

VII. CONCLUSION

In this paper, following the Bayesian approach to model selection, we investigated a MAP criterion for selecting the model order of superimposed signals in noise. The criterion was applied to the case of damped sinusoidal signals in i.i.d. white Gaussian noise, and its performance was compared to the SVD-based AIC and MDL. Our criterion proved to be more consistent than either of the others for damped sinusoidal data models. Computer simulations provided for a comparison between the MAP, AIC, and MDL criteria.

REFERENCES

[1] H. Akaike, "A new look at statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, Dec. 1974.

[2] G. E. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley, 1973.

[3] R. N. Bracewell, "Radio interferometry of discrete sources," *Proc. IRE*, vol. 46, pp. 97–105 1958.

[4] Y. Bresler and A. Macovski, "Exact maximum likelihood parameter estimation of superimposed exponential signals in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 5, 1986.

[5] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration.* New York: Academic, 1975.

[6] P. M. Djurić, "Selection of signal and system models by Bayesian predictive densities," Ph.D. dissertation, Univ. of Rhode Island, Kingston, 1990.

[7] P. M. Djurić and S. M. Kay, "Order selection of autoregressive models," *IEEE Trans. Signal Processing*, vol. 40, no. 11, pp. 2829–2833, 1992.

[8] ——, "Model selection based on Bayesian predictive densities and multiple data records," *IEEE Trans. Signal Processing*, vol. 42, no. 7, pp. 1685–1699, 1994.

[9] P. M. Djurić, "Model selection based on asymptotic Bayes theory," in *Proc. 7th SP Workshop Statist. Signal Array Processing*, P.Q., Canada, 1994, pp. 7–10.

[10] C. G. M. Fant, *Acoustic Theory of Speech Production.* The Hague, The Netherlands: Mouton, 1960.

[11] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 477–489, 1988.

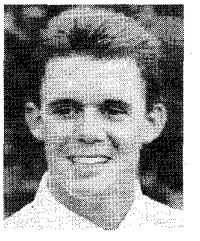
[12] N. Flourney and R. K. Tsutakawa, Eds., *Statistical Multiple Integration: American Mathematical Society Series in Contemporary Mathematics*, vol. 115, 1989.

[13] S. Geisser, "Aspects of the predictive and estimative approaches in the determination of probabilities," *Biometrics.* (supplement), vol. 38, pp. 75–85, 1982.

[14] J. Geweke, *Bayesian Inference in Econometric Models Using Monte Carlo Integration.* Durham, NC: Dept. of Econometrics, Duke University, 1986.

[15] P. Horst, *Prediction of Personal Adjustment.* New York: Social Science Research Council, Bulletin 48, 1941.

[16] C. M. Hurvich and C. L. Tsai, "A corrected Akaike information criterion for vector autoregressive model selection," *J. Time Series Anal.*, vol. 14, no. 3, pp. 271–279, 1993.
[17] F. James, "Monte Carlo theory and practice," *Reports on Progress in Physics*, vol. 43, pp. 1145–1175, 1980.
[18] M. H. Kalos and P. H. Whitlock, *Monte Carlo Methods.* New York: Wiley, 1986.
[19] L. Kavalieris and E. J. Hannan, "Determining the number of terms in a trigonometric regression," *J. Time Series Anal.*, vol. 15, no. 6, pp. 613–625, 1994.
[20] K. Kloek and H. K. van Dijk, "Bayesian estimates of equation system parameters: An application of integration by Monte Carlo," *Econometrica*, vol. 46, pp. 1–20, 1978.
[21] R. Kumaresan and D. W. Tufts, "Estimating the parameters of exponentially damped sinusoidal signals and pole-zero modeling in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, no. 6, pp. 833–840, 1982.
[22] P. W. Laud and J. G. Ibrahim,"Predictive model selection," *J. Royal Statist. Soc. B*, vol. 57, no. 1, pp. 247–262, 1995.
[23] M.-S. Oh, "Monte Carlo integration via importance sampling: dimensionality effect and an adaptive algorithm," in *Multiple Statistical Integration, vol. 115.* American Mathematical Society, 1989.
[24] F. Mosteller and D. L. Wallace, "Inference in an authorship problem," *J. Amer. Statist. Assoc.*, vol. 58, pp. 275–309, 1963.
[25] J. Myhill *et al.*, "Investigation of an operator method in the analysis of biological tracer data," *Biophysical J.*, vol. 5, pp. 89–107, 1965.
[26] A. O'Hagan, "Fractional Bayes factors for model comparison," *J. Royal Statist. Soc. B*, vol. 57, no. 1, pp. 99–138, 1995.
[27] M. H. Quenoille, "Notes on bias in estimation," *Biometrika*, vol. 42, pp. 353–360, 1956.
[28] 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, no. 9, pp. 2872–2881, 1993.
[29] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.
[30] ——, "A predictive least-squares principle," *IMA J. Math. Contr. Inform.*, vol. 3, pp. 211–222, 1986.
[31] ——, "Stochastic complexity," *J. Royal Statist. Soc., Series B, Methodol.*, vol. 49, no. 3, pp. 240–265, 1987.
[32] ——, "Stochastic complexity in scientific inquiry," in *World Scientific Series in Computer Science*, vol. 15, 1989.
[33] H. V. Roberts, "Probabilistic prediction," *J. Amer. Statist. Assoc.*, vol. 60, pp. 50–62, 1965.
[34] R. Y. Rubinstein, *Simulation and the Monte Carlo Method.* New York: Wiley, 1981.
[35] G. Schwartz, "Estimating the dimension of a model," *Annals Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[36] L. Stewart, "Bayesian analysis using Monte Carlo integration—A powerful methodology for handling some difficult problems," *Statistician*, vol. 32, pp. 195–200, 1983.
[37] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Royal Statist. Soc. B*, vol. 41, pp. 111–147, 1974.
[38] S. Umesh and D. W. Tufts, "Estimation of parameters of many exponentially damped sinusoids using fast maximum likelihood estimation with application to NMR spectroscopy data," submitted to *IEEE Trans. Signal Processing*.
[39] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Computer J.*, vol. 11, no. 2, pp. 185–194, 1968.
[40] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *J. Royal Statist. Soc., Series B, Methodol.*, vol. 49, no. 3, pp. 240–265, 1987.
[41] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 2, 1985.



William B. Bishop (S'94) was born in Long Island, New York. He received the B.E. and M.S. degrees from the State University of New York (SUNY) at Stony Brook, NY, USA, in 1990 and 1991, respectively.

From 1992 to present he has been with the Department of Electrical Engineering, SUNY, as both a research assistant and teaching assistant, and is currently a candidate for the Ph.D. degree. His research interests are in the areas of statistical signal modeling and estimation theory.

Mr. Bishop is a member of the American Statistical Association.

Petar M. Djurić (S'86-M'90) was born in Strumica, Yugoslavia. He 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, Providence, RI, USA, 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, NY, USA, 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. Currently, he serves as an associate editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING.