MODEL ORDER SELECTION OF DAMPED SINUSOIDS BY PREDICTIVE DENSITIES *

William B. Bishop, Petar M. Djurić

Department of Electrical Engineering
State University of New York at Stony Brook
Stony Brook, NY 11794, USA

ABSTRACT

In this paper, we investigate the problem of model order selection of damped sinusoids from a Bayesian perspective. We derive a maximum a posteriori (MAP) criterion through a combination of Bayesian inference and predictive densities. The MAP criterion is more appropriate for damped sinusoidal models (and transient data models in general) than are the SVD based information theoretic criteria in [1]. Simulation results are provided that display the breakdown of the AIC and MDL when the data record length is not properly coupled with the information bearing portion of the data model. This deterioration in performance is related to both, the underlying asymptotics upon which the AIC and MDL rules were originally based, and to their invalid penalty terms. Conversely, the MAP criterion is not based on asymptotics, and proves to be more reliable and consistent when the observation length is varied.

1. INTRODUCTION

The parameter estimation of multiple damped sinusoids superimposed in noise is of interest in many branches of applied science. Data in speech analysis, biomedicine, and other areas such as seismology and radio astronomy can be accurately represented by models of this type. Before an effective estimation procedure can be applied, however, the number of signal components (i.e., the model order) must be evaluated. This article specifically addresses this issue.

We have previously developed a MAP criterion [2], [3], for the determination of the number of damped signals in noise. In [2], the MAP procedure was implemented for damped sinusoids through the use of asymptotic approximations, while in [3], we considered the more difficult problem of damped exponentials.[1] This problem was approached through Monte Carlo adaptive importance sampling integration. A Gaussian importance function was appropriate for this case, since the integrands were relatively well behaved (i.e., they were not so sharply peaked).

*This work was supported by the National Science Foundation under Award No. MIP-9110628.

[1] When the signals have no frequency components, they are more highly correlated and are thus more difficult to discriminate between.

Following the approach in [3], we will once again invoke a numerical integration technique for realizing our criterion. The present procedure will differ from that in [3] in that a multivariate Cauchy probability density function (p.d.f.) will represent the importance function for the Monte Carlo procedure. The long tails of the Cauchy guard against instability while still capturing the essential features of the integrands kernel. The location parameters for the importance function are evaluated by applying a maximum likelihood estimation procedure for the parameters of each of the hypothesized model orders. The coverage region is set by matching the spread parameters of the Cauchy with the support region of the integrand.

A performance evaluation between the MAP criterion and its SVD-based counterparts is given by applying all three selection rules to the identical two component data model in [1]. The signal to noise ratio (SNR) will be fixed, and the data record length varied. This experiment clearly demonstrates the superior performance of the MAP criterion over both, the AIC and MDL.

This paper is organized as follows: In Section 2 we formulate the problem and state the objective. Section 3 will follow with the general derivation of the MAP model selection criterion, with the special case of damped sinusoids in independent identically distributed (i.i.d.) white Gaussian noise being considered. Simulation results are provided in Section 4, and detailed discussions are contained therein. Finally, in Section 5 we will conclude the paper.

2. PROBLEM FORMULATION

The general problem of interest can be characterized by the following model:

$$x[n] = \sum_{i=0}^{q} s_i[n;\underline{\theta}_i] + \epsilon[n;\underline{\psi}], \quad n \in Z_N, \quad q \in Z_Q \qquad (1)$$

where $Z_N = \{0, 1, \cdots, N-1\}$ is the finite set of non-negative integers, and Z_Q is similarly defined. The individual signal components $s_i[n;\underline{\theta}_i]$ are deterministic, and are completely specified up to the unknown parameter vectors $\underline{\theta}_i$, $i = 1, 2, \cdots, q$. The noise samples $\epsilon[n;\underline{\psi}]$ represent a sequence of random variables whose parametric distribution is known, but whose parameters, $\underline{\psi}$, are not. The model order q is also unknown. Given the observed data $x[n]$, the

0-7803-2431-5/95 $4.00

objective is to estimate q.

Note that the model in (1) can be written in a concise vector-matrix form as

$$\underline{x} = \underline{H}_q(\underline{\theta}_q) + \underline{\epsilon}, \quad q \in Z_Q \tag{2}$$

where $\underline{x}$ and $\underline{\epsilon}$ are $N \times 1$ vectors, and $\underline{H}_q(\underline{\theta}_q)$ is an $N \times q$ matrix whose i'th column is of the form

$$\underline{h}_i^T = [s_i(0;\underline{\theta}_i),\ s_i(1;\underline{\theta}_i),\ \cdots,\ s_i(N-1;\underline{\theta}_i)].$$

3. MAP MODEL SELECTION PROCEDURE

We define a model selection criterion which selects the "best" model as the one that maximizes the posterior probability mass of q given the observed data $\underline{x}$. That is,

$$\hat{q}_{\text{MAP}} = \arg \max_{q \in Z_Q} \{p(q|\underline{x})\}. \tag{3}$$

Applying Bayes' rule and marginalizing the posterior $p(q|\underline{x})$ over the nuisance parameters, we can write [3],

$$\hat{q}_{\text{MAP}} = \arg \max_{q \in Z_Q} \left\{ \frac{\int_{\tilde{\Theta}_q} f(\underline{x}|q,\tilde{\underline{\theta}}_q) f(\tilde{\underline{\theta}}_q|q) p(q) d\tilde{\underline{\theta}}_q}{\sum_{k=0}^{Q-1} \int_{\tilde{\Theta}_k} f(\underline{x}|k,\tilde{\underline{\theta}}_k) f(\tilde{\underline{\theta}}_k|k) p(k) d\tilde{\underline{\theta}}_k} \right\} \tag{4}$$

where $\tilde{\underline{\theta}}_k = [\underline{\theta}_1^T\ \underline{\theta}_2^T\ \cdots\ \underline{\theta}_k^T\ \underline{\psi}^T]^T$, and $\tilde{\Theta}_k$ is its parameter space.

Without loss of generality, we will assume that all models are equiprobable, thus $p(q) = \frac{1}{Q}$, $\forall q \in Z_Q$. Also note that the denominator in (4) is q-independent, and therefore

$$\hat{q}_{\text{MAP}} = \arg \max_{q \in Z_Q} \left\{ \int_{\tilde{\Theta}_q} f(\underline{x}|q,\tilde{\underline{\theta}}_q) f(\tilde{\underline{\theta}}_q|q) d\tilde{\underline{\theta}}_q \right\}. \tag{5}$$

Clearly, the employment of (5) requires the specification of a prior p.d.f. for $f(\tilde{\underline{\theta}}_q|q)$. To avoid the biasing which usually occurs with a proper prior, we would like to maintain objectivity in the criterion by directly applying the noninformative Jeffreys' prior.[2] Unfortunately, however, if we directly apply Jeffreys' prior to (5), the model selection rule will become arbitrary [4]. Still, we would like to use a noninformative prior because they are known to lead to the maximum expected information gained by the observed data.

In order to avoid the arbitrariness in the selection rule while still maintaining a high degree of objectivity, we will apply the concept of predictive densities and estimation-validation. This involves partitioning the observed data $\underline{x}$ into two mutually exclusive[3] subvectors, $\underline{x}_R$ and $\underline{x}_{N-R}$. The portion $\underline{x}_R$ is comprised of the latter R "training data" samples of $\underline{x}$, while $\underline{x}_{N-R}$ contains the remaining $N-R$ samples. The application of this approach to our initial criterion leads us to a slightly different selection rule

$$\hat{q}_{\text{MAP}} = \arg \max_{q \in Z_Q} \left\{ f(\underline{x}_{N-R}|q,\underline{x}_R) \right\} = \arg \max_{q \in Z_Q} \left\{ \frac{f(\underline{x}|q)}{f(\underline{x}_R|q)} \right\}. \tag{6}$$

Now upon marginalizing both the numerator and denominator in (6) we have the Bayesian MAP model selection criterion:

$$\hat{q}_{\text{MAP}} = \arg \max_{q \in Z_Q} \left\{ \frac{\int_{\Theta_q} \int_{\Psi} f(\underline{x}|q,\underline{\theta}_q,\underline{\psi}) f(\underline{\theta}_q,\underline{\psi}|q) d\underline{\psi} d\underline{\theta}_q}{\int_{\Theta_q} \int_{\Psi} f(\underline{x}_R|q,\underline{\theta}_q,\underline{\psi}) f(\underline{\theta}_q,\underline{\psi}|q) d\underline{\psi} d\underline{\theta}_q} \right\} \tag{7}$$

Note that now the identical prior p.d.f.'s appear in both the numerator and denominator, so that when we specify the noninformative prior, a cancellation of the arbitrary constants takes place, thus eliminating the overall arbitrariness in the criterion.

Application to Damped Sinusoids in Gaussian Noise

For this special case, the signal components $s_i[n;\underline{\theta}_i]$ in (1) are of the form

$$s_i[n;\underline{\theta}_i] = a_i e^{-\alpha_i n} \cos[2\pi f_i n + \phi_i] \tag{8}$$

and the noise samples are independent identically distributed as $\epsilon[n] \sim \mathcal{N}(0,\sigma^2)$. The unknown parameters of the i'th signal are its amplitude (a_i), frequency (f_i), phase (ϕ_i), and damping factor (α_i). The noise variance σ^2 is also assumed to be unknown.

To lower the dimensionality of the integrals that will ultimately result in the selection rule, we apply the following transformation to each of the signals

$$a_i e^{-\alpha_i n} \cos(2\pi f_i n + \phi_i)$$

$$= a_i \cos\phi_i e^{-\alpha_i n} \cos(2\pi f_i n) - a_i \sin\phi_i e^{-\alpha_i n} \sin(2\pi f_i n) \tag{9}$$

This allows the data model to be expressed as

$$x[n] = \sum_{i=1}^{q} [a_i \cos\phi_i e^{-\alpha_i n} \cos(2\pi f_i n)$$

$$- a_i \sin\phi_i e^{-\alpha_i n} \sin(2\pi f_i n)] + \epsilon[n],\ n \in Z_N,\ q \in Z_Q \tag{10}$$

or equivalently, in matrix form as

$$\underline{x} = \underline{H}_q \underline{b}_q + \underline{\epsilon}, \quad q \in Z_Q \tag{11}$$

where

$$\underline{H}_q = [\underline{c}_1 \underline{s}_1\ \underline{c}_2 \underline{s}_2\ \cdots\ \underline{c}_q \underline{s}_q], \quad \underline{b}_q = [\underline{b}_1^T\ \underline{b}_2^T\ \cdots\ \underline{b}_q^T]^T$$

with

$$\underline{c}_i^T = [1\ \ e^{-\alpha_i} \cos(2\pi f_i)\ \ \cdots\ \ e^{-\alpha_i [N-1]} \cos(2\pi f_i [N-1])],$$

[2] Jeffreys' prior is approximately noninformative if it is taken to be proportional to the square root of the Fisher information matrix [5].

[3] This type of partitioning is sometimes referred to as "honest" validation.

$$\underline{s}_i^T = [0 \;\; e^{-\alpha_i} \sin(2\pi f_i) \;\; \cdots \;\; e^{-\alpha_i[N-1]} \sin(2\pi f_i[N-1])],$$

and

$$\underline{b}_i = [a_i \cos\phi_i \;\; -a_i \sin\phi_i], \quad i = 1, 2, ..., q.$$

Applying (11) to the criterion in (7), and analytically integrating over the vector $\underline{b}$ and noise variance σ^2, we arrive at the MAP model selection criterion for damped sinusoidal signals in white Gaussian noise

$$\hat{q}_{\text{MAP}} = \arg \max_{q \in Z_Q} \left\{ \frac{\Gamma(\frac{N-2q}{2})}{\Gamma(\frac{R-2q}{2})} \right\}$$

$$\times \left\{ \frac{\int_{\underline{f}_q} \int_{\underline{\alpha}_q} |\underline{H}_{q,N}^T \underline{H}_{q,N}|^{-\frac{1}{2}} (\underline{x}_N^T \underline{P}_N^{\perp} \underline{x}_N)^{-(\frac{N-2q}{2})} d\underline{\alpha}_q d\underline{f}_q}{\int_{\underline{f}_q} \int_{\underline{\alpha}_q} |\underline{H}_{q,R}^T \underline{H}_{q,R}|^{-\frac{1}{2}} (\underline{x}_R^T \underline{P}_R^{\perp} \underline{x}_R)^{-(\frac{R-2q}{2})} d\underline{\alpha}_q d\underline{f}_q} \right\}. \tag{12}$$

Now all we need to do is evaluate the integrals in (12), but this is not exactly a trivial task. The dimensions can obviously be quite large, and our investigations into the integrands in (12) have shown them to be very sharply peaked, particularly for N large and high SNR's. For these reasons we will apply Monte Carlo importance sampling with a multivariate Cauchy p.d.f. as the importance function to implement our criterion.

The following section will present simulation results comparing the MAP criterion to the SVD-based AIC and MDL which are given by [1]

$$AIC(k) = (N-L) \cdot \mathcal{L}^{(k)}(\underline{\tilde{\theta}}_k, \underline{x}) \; + \; 2(2k+1)$$

$$MDL(k) = (N-L) \cdot \mathcal{L}^{(k)}(\underline{\tilde{\theta}}_k, \underline{x}) \; + \; (2k+1)\ln(N-L).$$

Here $\mathcal{L}^{(k)}(\underline{\tilde{\theta}}_k, \underline{x})$ is a likelihood term which is based on the singular value decomposition of the modified backward linear prediction data matrix, N is the length of the observed data $\underline{x}$, and L represents the prediction filter order.

4. SIMULATION RESULTS

In order to demonstrate the performance of the MAP model order selection criterion, we considered the following two-component damped sinusoidal model:

$$x[n] = \sum_{i=1}^{2} a_i e^{-\alpha_i n} \cos(2\pi f_i n + \phi_i) + \epsilon[n], \;\; n = 0, \; 1, ..., N-1.$$

We conducted 100 independent trials for sequences whose lengths varied between $N = 64$ and $N = 256$ samples. The variance of the noise process $\epsilon[n]$ was adjusted so that the peak SNR was fixed at $15dB$. The true values of the signal parameters were set at $a_1 = a_2 = 1.0$, $\phi_1 = \phi_2 = 0$, $f_1 = 0.20$, $f_2 = 0.24$, $\alpha_1 = 0.10$, $\alpha_2 = 0.05$.

For each sequence length, N, the prediction filter order L of the SVD-based AIC and MDL was adjusted so that the backward linear prediction data matrix remained square. This allowed for these criteria to perform optimally, thus providing for a truly fair comparison between the MAP, AIC, and MDL selection rules. For each trial we fit the observed data to each of the models:

$$\begin{aligned} \mathcal{H}_0 : \quad x[n] &= \epsilon[n] \\ \mathcal{H}_k : \quad x[n] &= \textstyle\sum_{i=1}^{k} s_i[n; \underline{\theta}_i] + \epsilon[n], \quad k \in \{1, 2, 3\} \end{aligned} \tag{13}$$

where

$$s_i[n; \underline{\theta}_i] = a_i e^{-\alpha_i n} \cos(2\pi f_i n + \phi_i).$$

Note that $\mathcal{H}_0$ represents the "noise only" model. The results of this experiment are shown in Table 1, and are graphically depicted in Figure 1. They clearly indicate that the selection accuracies of the AIC and MDL both deteriorated as the length of the observation vector $\underline{x}$ superseded the actual information portion of the sequence. The MAP criterion on the other hand, seemed to improve somewhat as more data were obtained. For example, when $N = 64$ samples, the MAP, MDL, and AIC detection probabilities were 0.84, 0.89, and 0.68, respectively. At $N = 64$, the total sequence length was perfectly matched with the decay rate $\alpha_2 = 0.05$, and for this reason, the MDL may have been able to perform successfully. The MAP's detection performance was only slightly below that of the MDL at $N = 64$, while the AIC's performance was significantly below either of the other two criteria. Notice also that while the MAP and MDL's detection performance was nearly the same for $N = 64$, the MDL tended to overparameterize (choosing a 3'rd order model 9 times out of 11 incorrect decisions), and the MAP tended to adhere to the parsimony principle (i.e., it selected a lower order model for every incorrect decision). Now examine what happened when the number of data samples increased to $N = 128$. The relative performances of the three criteria already began to contrast considerably; for $N = 128$, the MDL's probability of correct decision decreased to 0.69, while the MAP's increased to 0.93. The AIC's detection performance also decreased to 0.54. This trend also seemed to persist as the value of N increased further. At $N = 256$ samples, the probability of correct selection was 0.97, 0.29, and 0.25 for the MAP, MDL, and AIC, respectively. This disparity in performance is certainly much larger than what it was when $N = 64$ - the data record length that was well matched to the signals effective decay rate of $\alpha_2 = 0.05$.

The interpretation of these results is fairly straightforward. The accuracy of the MAP strongly depends on how well one is able to approximate the integrand by the importance function, which in turn depends on the precision of its location parameters. As these location parameters were estimated by a maximum likelihood method, we can generally expect their accuracies to improve[4] with an increasing amount of data. The MDL criterion breaks down mainly because its penalty term is incorrect for transient data models. The MDL's penalty increases by the same amount for each additional data sample, but with transient data, each

[4] For transient data, the maximum likelihood estimates will only improve until the information in the data is exhausted. That is, we can expect improved estimation accuracy only until a particular value of N is reached.

additional sample carries a decreasing amount of information, so by logical deduction, a constant penalty should not be apportioned for each additional observation. The AIC's likelihood term is overly sensitive to N in comparison with its penalty. That is, the changes in the loglikelihood of the AIC due to different N are not properly compensated for by its penalty.

N	Criterion	$\mathcal{H}_0$	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$
64	AIC	0	0	68	32
	MDL	0	2	89	9
	MAP	0	16	84	0
100	AIC	0	1	61	38
	MDL	0	2	79	19
	MAP	0	12	88	0
128	AIC	0	5	54	41
	MDL	0	11	69	20
	MAP	0	7	93	0
150	AIC	0	7	38	55
	MDL	0	13	55	32
	MAP	0	5	94	1
200	AIC	0	21	34	45
	MDL	0	31	40	29
	MAP	0	2	98	0
256	AIC	0	24	25	51
	MDL	0	46	29	25
	MAP	0	3	97	0

Table 1: Performance comparison between the MAP and SVD-based AIC and MDL criteria for various data record lengths (N). 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 ($\mathcal{H}_2$), and for all trials the SNR was fixed at $15dB$.

These erratic effects are primarily due to the fact that the original AIC and MDL were derived by way of asymptotics, and as such, their validity strongly depends on large informative data records. For decaying sinusoids or any transient model, their utility as reliable model selection criteria is therefore questionable. The MAP criterion does not rely on asymptotical assumptions, and the decreasing information contained in each additional sample is automatically accounted for in the criterion. It is apparent that the AIC and MDL can perform reasonably well for transient models if there is some a priori knowledge regarding the length of the information bearing portion of the observation vector. Since this information is usually unavailable, the MAP criterion seems to be the more sound method for resolving the model order selection problem of transient signals in noise.

5. CONCLUSION

Following the Bayesian approach to statistical inference, we developed a MAP model order selection criterion by way of predictive densities and estimation-validation. We applied our criterion to the special case of multiple damped sinusoids in white Gaussian noise. The complicated integrals in our criterion were solved by Monte Carlo importance sampling. Simulation results were provided which displayed the improvement in performance of the MAP criterion over the SVD-based AIC and MDL for varying data record lengths.

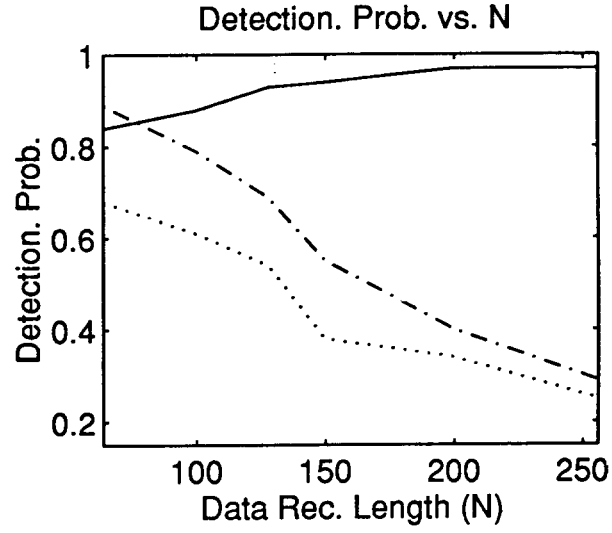

Figure 1: Probability of correct detection vs. data record length (N), for the MAP (solid), AIC (dotted), and MDL (dashed) model selection rules. These results are based on 100 independent trials, for a fixed signal-to-noise ratio (SNR) of $15dB$.

REFERENCES

[1] V. Umpathi Reddy and L. S. Biradar, "SVD-based information theoretic criteria for detection of the number of damped/undamped sinusoids and their performance analysis," *IEEE Transactions on Signal Processing*, vol. 41, No. 9, pp 2872-2881, Sept. 1993.

[2] P. M. Djurić, W. B. Bishop, D. E. Johnston "A Bayesian procedure the detection of damped signals," *International Symposium on Circuits and Systems*, Vol 2, pp 401-404, June 1994.

[3] W. B. Bishop, P. M. Djurić, D. E. Johnston, "Bayesian model selection of exponential time series through adaptive importance sampling," *IEEE Proc. of Seventh SP Workshop on Statistical Signal and Array Processing*, pp 51-54, June 1994.

[4] P.M. Djurić, *Selection of Signal and System Models by Bayesian Predictive Densities*, Ph.D. Dissertation, University of Rhode Island, Kingston: RI, 1990.

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

[6] D.W. Tufts and R. Kumaresan, "Estimation of frequencies of multiple sinusoids: making linear prediction perform like maximum likelihood," *Proc. of IEEE*, Vol. 70, No. 9, Sept. 1982.