Model Order Selection for Summation Models

A. Sabharwal C. J. Ying L. Potter R. Moses

Department of Electrical Engineering
The Ohio State University
Columbus OH 43210
{sabhara,yingc,randy,potter}@er4.eng.ohio-state.edu

Abstract

In this paper, we propose two model order selection procedures for a class of summation models. We exploit the special structure in the class of candidate models to provide a data dependent upper bound on the model order. The proposed upper bound is also a consistent estimator of model order. Further, MDL, AIC and MAP when accompanied with the data dependent prior exhibit an improved rate of convergence to their asymptotic behaviour and an improved detection rate for finite SNR and finite data lengths. Asymptotic properties of the maximum likelihood parameters are used to derive the proposed methods. All simulations use the complex undamped exponential model.

1. Introduction

In parametric modeling problems, one often derives parameter estimation algorithms by first assuming that the model order is known. However model order is typically not known in practice. Further, the use of an incorrect model order can lead to serious loss in performance [9] of many frequently used estimators (for example MLE). Thus, effective parametric modeling requires accurate model order estimation from observed data.

A number of model order selection methods have been developed: *information-theoretic* criteria such as AIC [1], MDL [8], and EDC [12]; *Bayesian approaches* such as BIC [10], BPD [3], and MAP [5]; and *eigenspace-based* techniques such as Fuchs's approach [6]. These algorithms all make use of the parsimony principle; that is, they penalize higher order models in some way.

Most algorithms search for the best model order over a set of candidate model orders. The highest order is often chosen by the user in an ad-hoc fashion. In general, the detection performance of most order selection procedures depends on the number of candidate orders; more candidate model orders imply a lower probability of detection for most schemes. We propose two strongly consistent data dependent upper bounds on the model order. Further, the proposed upper bounds exhibit a high detection rate when used as estimators of model order.

Monte Carlo experiments with undamped exponential model suggest an increased rate of convergence to the asymptotic behavior (consistency for MDL and MAP) and an improved performance for finite data lengths.

In Section 2, we briefly outline the geometry of the summation models. Section 3 reports a simulation study for the undamped exponential model. The simulations are used to support the claims made in Section 4. The proposed algorithms are summarized in Section 5. In Section 6, we compare the detection performance of the proposed method with three order selection methods, MDL, AIC and MAP. Finally, we discuss the proposed methods in Section 7.

2. Topology of summation models

The models have the following structure.

$$y(t) = \sum_{i=1}^{q} s(t, \theta_{qi}) + \epsilon_t \tag{1}$$

where $t = 1, \ldots, N$. The noise sequence ϵ_j is i.i.d. $N(0,\sigma^2)$. For each i, θ_{qi} is a k dimensional parameter vector defining the signal component $s(t, \theta_{qi})$. Thus, the total number of parameters to be estimated is kq. The composite parameter θ_q belongs to the set Θ_q which is an open subset of $\mathbb{R}^{kq}$, defined as

$$\Theta_q = \{\theta_q \in \mathbb{R}^{kq} : I(\theta_q) \text{ is full rank } \} \tag{2}$$

where I is the Fisher information matrix defined as

$$I_{ij}(\theta) = -\mathcal{E}\left(\left[\frac{\partial \log p_\theta(y)}{\partial \theta_i}\right]\left[\frac{\partial \log p_\theta(y)}{\partial \theta_j}\right]\right) \quad (3)$$

Also, the function s is such that the ***identifiability criterion*** is satisfied, *i.e.*, for a given model order q, no two distinct (to permutations) θ_q and θ'_q yield the same noiseless data vector.

Note that Θ_q is an open subset and $bd(\Theta_q) = \bar{\Theta}_q \backslash \Theta_q$ contains models with order less than q. This implies that for a given model order $p > q$ and $\theta_q \in \Theta_q$, there exists a sequence of parameters in Θ_p which converges to θ_q. This property allows us to define the "consistency" of ML estimates in the case of overmodeling.

For the rest of the paper, we denote the true quantities by the superscript $*$. The true model is denoted by q^*.

Further, from the definition of Fisher information, it follows that $I(\theta_q)$ is singular on the set $bd(\Theta_q)$. Thus, if ML estimates, $\hat{\theta}_q$ for $q > q^*$ converge to a member of $bd(\Theta_q)$, then the condition number of $I(\hat{\theta}_q)$ diverges to ∞. In section 4, we prove that ML estimates converge to an element of $bd(\Theta_q)$ (with some added assumptions) as SNR $\rightarrow \infty$. Thus for model orders $q \geq q^*$, the residual error converges to 0. This leads to a "knee" in the log-likelihood curve as a function of model order.

The summation model class is often considered a nested model class, but we note that there is no unique way of embedding a lower dimensional model into a higher dimensional model. An example of nested models with unique embeddings is AR(q).

3. Simulation study

In this section, we consider a widely studied example of detecting superimposed undamped exponentials in noise. Specifically, the data is assumed to come from the following model

$$y(t) = \sum_{i=1}^{q} A_i e^{j(\omega_i t + \phi_i)} + \epsilon_t \quad (4)$$

For the simulations, we consider two equal amplitude modes, half a Fourier bin apart where the true parameters are $A_1 = A_2 = 1$, $\phi_1 = \frac{\pi}{4}$, $\phi_2 = 0$, $\omega_1 = \pi$ and $\omega_2 = \frac{24\pi}{25}$. The index t ranges from 0 to 24 ($N = 25$ data points). One hundred Monte Carlo simulations were performed for SNR ranging from -10 dB to 20 dB, in steps of 1 dB. For obtaining the ML parameter estimates, a subspace-based method [7, 4] was used to initialize a gradient search method. Also, as an *ad hoc* upper bound on the model order, we consider

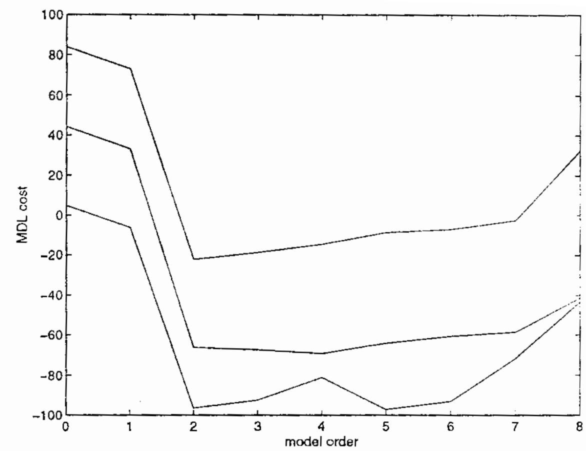

Figure 1. Three realizations of MDL cost functions at SNR = 20dB

$Q_{\max} = \lfloor \frac{N}{3} \rfloor$ so that for the highest model order, we have approximately two data points per parameter.

First, we motivate the need of a data dependent prior on model orders. In Figure 1, three typical realizations of the MDL cost function are shown (in a "waterfall" plot). The MDL cost function exhibits local minima at several model orders other than the true order, *i.e.* $q^* = 2$. The global minimum is at order 5, 4 and 2 respectively for the three cases. It is clear, if we choose a prior $[1\,1\,1\,0 \ldots 0]$ on the model order, then MDL will choose the correct model for all three cases. Since the true order is unknown, we now explore the possibility of generating a prior from data.

In Section 2, we noted that the Fisher information matrix, $I(\hat{\theta}_q)$, tends to be poorly conditioned for $q \geq q^*$. This behavior can be observed in Figure 2, where we plot inverse of condition number of the Fisher information versus model order. We could use a threshold to detect the "knee" in the curve and determine the required binary prior. But this method requires us to determine a threshold, which involves computing the statistics of inverse of condition number of Fisher information matrix. Instead, we use the properties of negative log-likelihood to propose threshold-free methods to detect the "knee".

Let $L(q)$ denote the negative log likelihood computed at ML estimates for model order q (we ignore some additive and multiplicative constants).

$$L(q) = N \log \frac{1}{N} \left\| y(t) - \sum_{i=1}^{q} s\left(t, \hat{\theta}_{qi}\right) \right\|^2 \quad (5)$$

(Note MDL minimizes $L(q) + \frac{q}{2} \log N$).

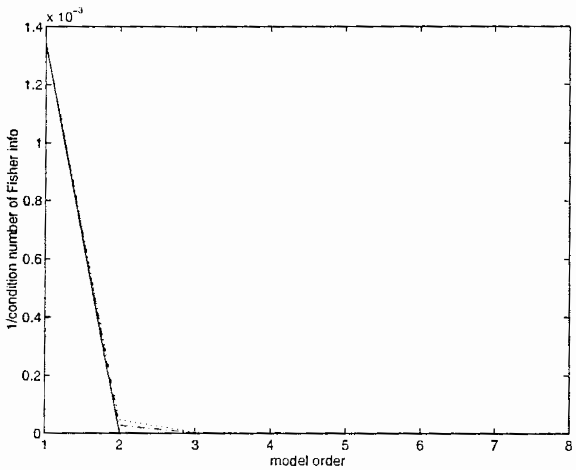

Figure 2. Inverse of condition number of Fisher information at ML estimates at SNR = 20 dB (same estimates as used in Figure 1)

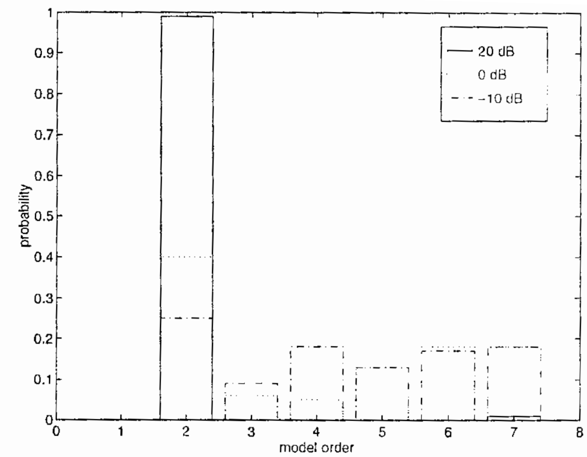

Figure 3. Histograms of the proposed estimator, $\hat{q}_1$, for SNR = 20, 0 , -5 dB

Our first proposed estimator is given by

$$\hat{q}_1 = \arg \max_{q \in [1, Q_{max}-1]} (L(q+1) + L(q-1) - 2L(q)) \quad (6)$$

We employ $\hat{q}_1$ as a data dependent upper bound on the highest model order. The proposed functional form arises by extending the definition of curvature of a function of a continuous variable to a function of a discrete variable. Since $L(q)$ exhibits a "knee" at the true model order, $\hat{q}_1$ tends to be good estimator of q^* for high SNR. For moderate SNR, the $\hat{q}_1$ tends to overestimate as seen in Figure 3. Similar behavior of $\hat{q}_2$ can be observed in Figure 4.
The second estimator

$$\hat{q}_2 = \arg \max_{q \in [1, Q_{max}]} \frac{|L(q)|}{|L(q-1)|} \quad (7)$$

uses ratios of log-likelihood instead of differences as compared to estimator (6). In both the cases, the function to be maximized, say $f(q)$, has the following asymptotic behavior.

$$f(q) \to \begin{cases} \infty & q = q^*, \\ c_q < \infty & q \neq q^*. \end{cases} \quad (8)$$

The above follows from two properties: the residual error, r_q, converges to zero as SNR increases to ∞ for $q \geq q^*$, and $\log r_q \to -\infty$ as $r_q \to 0$.

4. Motivating Results

In this section, we present the theoretical results which we use as a basis to propose the new model order selection method.

First, we consider the consistency of ML estimates in the case of overmodeling. To define the consistency for model orders $q > q^*$, we define an equivalence class using (2)as follows.

$$\mathcal{C}_{q,q^*} = \{\theta_q : \theta_q \text{ is a singular extension of } \theta_{q^*} \text{ in } \bar{\Theta}_q\} \quad (9)$$

The equivalence class belongs to the boundary of the set Θ_q, $bd(\Theta_q)$.
Assumption 1 For the true model order, ML is strongly consistent, *i.e.*, for a $\theta_{q^*} \in \Theta_{q^*}$, the ML estimates $\hat{\theta}_{q^*} \in \Theta_{q^*}$ converges *almost surely* to θ_{q^*} as the SNR $\to \infty$.

Theorem 4.1 *Let $\theta_{q^*} \in \Theta_{q^*}$ be the true parameter. If Assumption 1 holds then for any fixed $q \geq q^*$, ML estimates, $\hat{\theta}_q$ converge almost surely to a member in the equivalence class $\mathcal{C}_{q,q^*}$ as SNR $\to \infty$.*

Proof : From [11], in the case of model mismatch, the maximum likelihood estimates $\hat{\theta}_q$ converge to a θ_q such that the Kullback-Leibler (KL) distance between the true model $f_{q^*}(\theta_{q^*})$ and the misspecified model $f_q(\theta_q)$ is minimized. For $q \geq q^*$,

$$\min_{\theta_q} KL(f_{q^*}(\theta_{q^*}), f_q(\theta_q)) = 0 \quad (10)$$

which is achieved only by a member of $\mathcal{C}_{q,q^*}$. ■

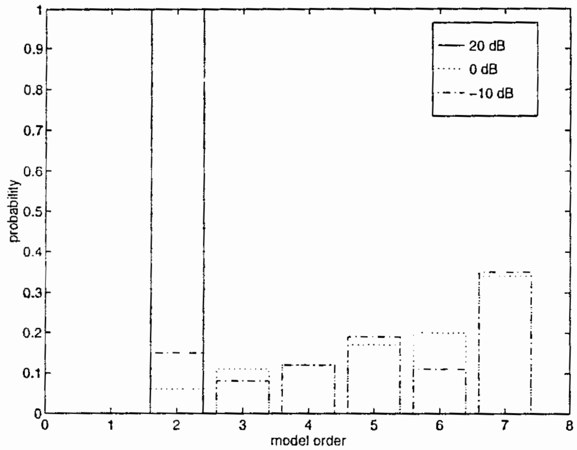

Figure 4. Histograms of the proposed estimator, $\hat{q}_2$, for SNR = 20, 0 , -5 dB

Since ML estimates are consistent for model orders $q > q^*$, the residual data fit error converges to 0. This implies that $L(q) \to -\infty$ for $q > q^*$. We state the claim about the consistency of $\hat{q}_1$ and $\hat{q}_2$ without proof.

Claim 4.2 *The estimators* $\hat{q}_1$*(and* $\hat{q}_2$*)* $\xrightarrow{a.s.} q^*$, i.e, *both the estimators are strongly consistent.*

5. Proposed Algorithms

Based on the results obtained in the previous sections, we propose following order selection procedures.

Algorithm 1 : $\hat{q} = \hat{q}_1(\text{or } \hat{q}_2)$.

The estimators $\hat{q}_1$ and $\hat{q}_2$ are given in Section 3. When $\hat{q}$ from Algorithm 1 is used as a data-dependent prior for MDL, the resulting order selection procedure can be summarized as

Algorithm 2 :

1. Compute $\hat{q}_p = \hat{q}_1(\text{or } \hat{q}_2)$.
2. Evaluate

$$\hat{q} = \arg \min_{q \in [0 \ldots \hat{q}_p]} MDL(q).$$

We label the above algorithm as *windowed MDL* (WMDL). The consistency of WMDL follows from the consistency of MDL and $\hat{q}_p$ (Claim 4.2). Note that MDL in step 2 can be replaced by any other order selection criterion, like AIC or MAP. For the methods which tend to overmodel for finite data and SNR, the above data-dependent prior tends to eliminate the overmodeling errors and in effect, increase probability of detection. It is clear that the extent of improvement is dependent on how close the upper bound is to true model order. We demonstrate the effectiveness of above procedures in the next section.

6. Simulations

For the same example as considered in Section 3, we consider the detection rates of MDL, AIC and the MAP applied to complex undamped exponentials. The model order estimates from three criteria for the undamped exponentials are given by

$$\begin{aligned}
\hat{q}_{AIC} &= \min_q L(q) + 3q \\
\hat{q}_{MDL} &= \min_q L(q) + \frac{3}{2} q \log N \\
\hat{q}_{MAP} &= \min_q L(q) + \frac{5}{2} q \log N
\end{aligned}$$

The windowed versions of MDL, AIC and MAP are obtained from Algorithm 2 using $\hat{q}_1$ given in Section 5. The improvement in detection performance for the three methods is given in Figures 5, 6 and 7. In each of the three plots, the dashed line represents the original algorithm and the solid line is the performance with the proposed data dependent upper bound. Finally, in Figure 8 we compare $\hat{q}_1$ with MAP, MDL and AIC. The performance of $\hat{q}_2$ is identical to $\hat{q}_1$, so the corresponding plots are omitted.

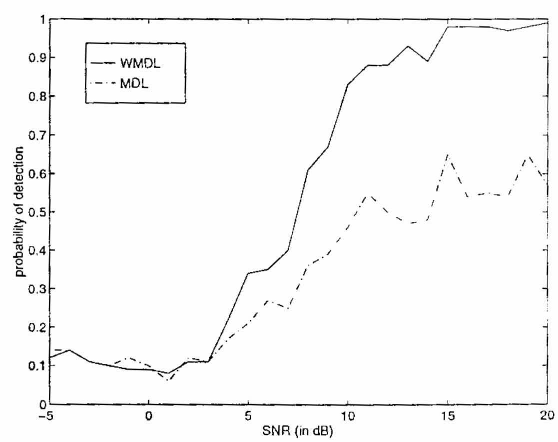

Figure 5. MDL and WMDL for the 2-mode case