PROPER PRIOR MARGINALIZATION OF THE CONDITIONAL ML MODEL FOR COMBINED MODEL SELECTION/SOURCE LOCALIZATION

Bill M. Radich and Kevin M. Buckley

Department of Electrical Engineering
University of Minnesota
Minneapolis, MN 55455

ABSTRACT

We present a Bayesian evidence technique for the parameter estimation/model selection problem within the conditional maximum likelihood (CML) framework. The CML is chosen because of its flexibility: it allows for a wide range of source amplitude models (e.g., no unreasonable or restrictive assumptions, such as Gaussian signals are necessary). In contrast to other CML studies, we eliminate the large number of unknown amplitude parameters by marginalization with a proper (normalizable), yet very broad prior. The resulting marginal is used to derive a new model selection/parameter estimation procedure, based on the Bayesian evidence of each considered model, given the observed data. Monte Carlo simulations for a scenario consisting of two narrowband, far-field sources demonstrate the effectiveness of the proposed method in low SNR, small temporal/spatial sample situations.

1. INTRODUCTION

Detecting the number of sources impinging on an arbitrary array of sensors is widely accepted as a critical and difficult problem in signal processing. This problem is confounded when the number of temporal/spatial samples is relatively small, and few simplifying assumptions (e.g., Gaussian normal models) can be made on the signal amplitude parameters.

In view of the fact that an assumed number of sources determines the dimensionality of a parametric probabilistic model, the detection problem can be treated as a special case of model selection. Until recently, many of the most popular model selection schemes for source detection have been based on asymptotic information theoretic criteria [1],[2],[3],[4]. Within the vast literature on model selection, however, Bayesian techniques have proven to be valuable in small sample situations, and have provided objective measures for evaluating competing models given a data observation [5],[6],[7].

In this paper, a marginalization approach is used to derive a new combined model selection/source parameter estimation technique, where amplitude parameters are integrated out of the full likelihood by assuming a diffuse, normalizable multivariate prior. The primary focus of this paper is the detection performance of the proposed Bayesian method; however, the corresponding Bayesian source parameter cost function is also investigated via simulations.

This work was supported in part by *ONR* under contract number N00014-90-J-1049 and NSF under contract number MIP-9057071.

2. BACKGROUND

2.1. Conditional Maximum Likelihood Model and Notation

Assuming the standard narrowband observation model for D far-field sources (represented by the D dimensional vector θ) impinging on an array of M sensors, the array output $x(t)$ is

$$x(t) = A(\theta)s(t) + n(t),\ t = 1, ..., N, \quad (1)$$

where $s(t)$ denotes the $D \times 1$ vector of signal amplitudes, A is the $M \times D$ matrix with (assumed linearly independent) array response vectors $a(\theta_i)$, for $i = 1, 2, ..., D$. The $M \times 1$ noise vector, $n(t)$ is modeled as a zero-mean, temporally and spatially white Gaussian process of variance σ^2.

Under the conditional maximum likelihood (CML) model [8],[9], the complex amplitudes are treated as a fixed set of ND deterministic parameters. If source localization is of prime interest (estimation of θ), this set represents a large number of nuisance parameters.

The full likelihood function is readily expressed as $f(X \mid \theta, s(1), ..., s(N), \sigma^2) =$

$$\pi^{-MN}\sigma^{-2MN} \exp \{\frac{1}{\sigma^2}\sum_t^N \mid x(t) - A(\theta)s(t) \mid^2\}. \quad (2)$$

Replacing $s(1), ..., s(N)$, and σ^2 with their maximum likelihood estimates and taking the negative logarithm results in the familiar *condensed* cost function (within a constant)

$$v_{ML}(\theta) = \log(tr\{P^{\perp}_{A(\theta)}XX^H\}), \quad (3)$$

where tr { . } represents the matrix trace operator, $P^{\perp}_{A(\theta)} = I_M - A(\theta)A^{\dagger}(\theta)$ is the projection matrix that projects onto the subspace orthogonal to the column space of $A(\theta)$, and $\dagger$ denotes the pseudoinverse.

3. BAYESIAN MODEL SELECTION

Let I_d represent a model hypothesis indexed by the number, d, of assumed narrowband sources. The posterior density function for model I_d is given by

$$f(I_d \mid X) \propto f(X \mid I_d) f(I_d), \qquad (4)$$

where $f(X \mid I_d)$ is the Bayesian "evidence" for I_d [5],[6]. Assuming no prior preference of one proposed model over another, the "optimal" model is chosen as the one that maximizes

$$f(X \mid I_d) = \int_\theta f(X \mid \theta, I_d) f(\theta \mid I_d)\, d\theta. \qquad (5)$$

Evaluating the integral in (5) requires the marginal density function $f(X \mid \theta, I_d)$, and this means that the amplitude and noise variance parameters must first be integrated out of $f(X \mid \theta, s(1), \ldots, s(N), \sigma, I_d)$.

3.1. Marginal for Parameter Estimation

For notational convenience, let S denote the collection of signal amplitudes, $s(1), \ldots, s(N)$. The removal of nuisance parameters from the likelihood function by integration yields the following marginal density $f(X \mid \theta, I_d) =$

$$\int_{S,\sigma} f(X \mid \theta, S, \sigma, I_d) f(S, \sigma \mid I_d) dS\, d\sigma, \qquad (6)$$

where it is assumed that $f(S, \sigma \mid I_d) = f(S \mid I_d) f(\sigma)$.

The standard noninformative prior for amplitude parameters in the context of Bayesian estimation is simply a constant [7],[10],[11],[12]. However, as noted in [6],[7], using an improper prior for $f(s(1), \ldots, s(N) \mid I_d)$ (that is, a prior that doesn't integrate to a finite va lue) invariably leads to erroneous model selection within the Bayesian evidence framework. One remedy is offered by the Bayesian predictive density method analyzed in [7]. However, this technique requires the parsing of data into validation and estimation sets, and we are interested in using CML for applications where the number of available snapshots can be $N < 10$ [13]. Another approach, proposed by Wax, et. al., [2] is to project the data onto signal and noise subspaces, and determine the minimum description length (MDL) for each projection. This method is also inappropriate for a small number of snapshots because the MDL criterion is based on asymptotic arguments (it is worth noting that MDL can be derived as a large sample approximation of Rissanen's *stochastic complexity* [14], which is essentially the negative logarithm of the Bayesian evidence).

The approach outlined briefly here results from a Gaussian prior for the amplitude parameters, and assumes that the prior variance associated with each amplitude (that is, level of uncertainty) is large with respect to the noise. In this way, the resulting marginal does not bias the likelihood. For each model, the amplitude prior uncertainties are reduced to a single Bayesian hyper-parameter, γ, which can be integrated out, along with σ, by using the Jeffrey's prior for scale parameters [12]: $f(\gamma) \propto \frac{1}{\gamma}$. Although the Jeffrey's prior for a scale parameter does not integrate to unity, the final result is not affected because each model has the same number of hyper-parameters [6]. This allows us to avoid integration of multivariate improper priors.

After performing the integration indicated in (6), and eliminating irrelevant constants (but keeping those that depend on d), we derive:

$$\begin{aligned} f(X \mid \theta, I_d) &\propto \Gamma(Nd)\, \Gamma(N(M-d)) \\ &\times \mid A^H A \mid^{-N} (tr\{P_A^\perp X X^H\})^{-N(M-d)} \\ &\times (tr\{(A^\dagger)^H A^\dagger X X^H\})^{-Nd}, \end{aligned} \qquad (7)$$

which results in the following Bayesian cost function for estimating θ :

$$\begin{aligned} v_B(\theta) &= \log\ \mid A^H A \mid \\ &+ (M-d) \log(tr\{P_A^\perp X X^H\} \\ &+ d \log(tr\{(A^\dagger)^H A^\dagger X X^H\}). \end{aligned} \qquad (8)$$

Equations (7), (8) can be derived from the analogous real-val ued formulation given in [13],[15].

The Bayesian cost function given in (8) exhibits desirable characteristics in "stressfull" circumstances, and is worthy of further study in and of itself, apart from the model selection/detection point of view stressed in this paper. In particular, as SNR or M decreases, this estimator effectively keeps overall MSE lower than ML and related methods by progressively allowing an increasing amount of asymptotic bias for the sake of substantially lower variance.

3.2. Marginal for Model Selection

In principle, the evidence for each model, I_d, can be evaluated by substituting the marginal $f(X \mid \theta, I_d)$ given by (7) into the integral in equation (5). However, this integral cannot be solved for our problem because the likelihood is in general a nonlinear function of θ. Assuming a diffuse multivariate Gaussian normal prior for θ, and using the second order Taylor series approximation for $-\log f(X \mid \theta, I_d)$, the following approximation can be derived [13]:

$$\begin{aligned} f(X \mid I_d) &\propto f(X \mid \hat{\theta}_B, I_d) \mid \hat{\Sigma}_B \mid^{-\frac{1}{2}} \\ &\times \Gamma(\tfrac{d}{2})(\tfrac{1}{2} \mid \hat{\theta}_B \mid^2)^{-\frac{d}{2}}, \end{aligned} \qquad (9)$$

where $\mid \hat{\Sigma}_B \mid$ denotes the determinant of the Hessian matrix of $-\log f(X \mid \theta, I_d)$, evaluated at $\hat{\theta}_B$.

4. EXAMPLES

To illustrate the proposed detection/estimation method, a two source case, $\theta = [5^\circ\ 8^\circ]$, was simulated for a uniform linear array of half-wavelength spacing, and $M = 6$. Minimization of the CML and Bayesian cost functions was achieved using Newton methods, and the $\hat{\Sigma}_B$ term in (9) was evaluated analytically. All results presented here are based on 200 Monte Carlo simulation trials for each data point.

Figure 1 shows the RMS error (in degrees) of the first source estimate for the Bayesian and CML techniques as a function of N. The SNR was held constant at 20db, and the two sources were equi-powered and uncorrelated. Figure 2 displays the probability of correct detection as a function of N for the Bayesian evidence, AIC, and MDL. Note that both AIC and MDL were calculated directly from the conditional maximum likelihood framework, with the parameter

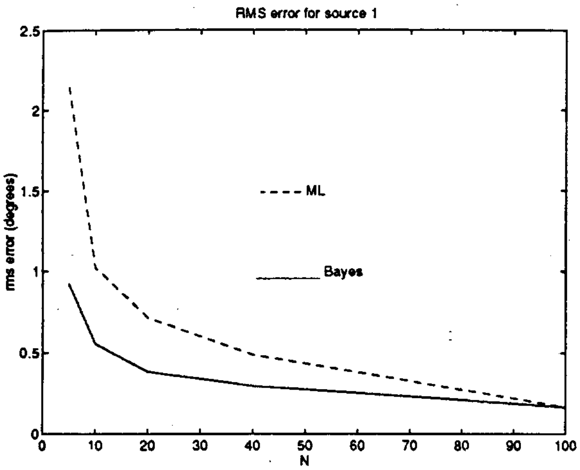

Figure 1: RMS error for estimation of θ_1 for two uncorrelated sources.

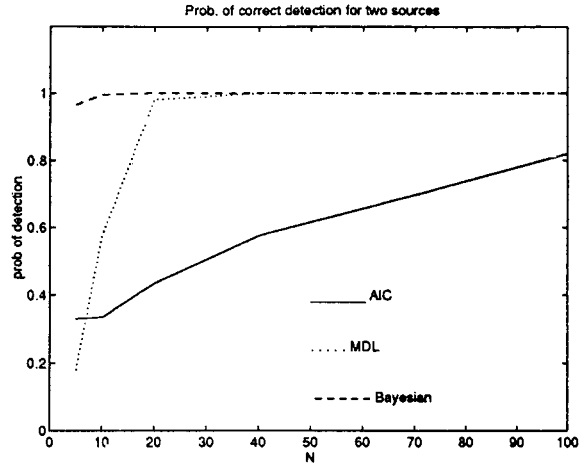

Figure 2: Detection performance vs N for case of two uncorrelated sources.

vector given by $\Theta^d = [\theta_1, ..., \theta_d, s(1), ..., s(N), \sigma]$ (in contrast to the formulation of [1], where parameterization is in terms of the eigenvalues/vectors of the observation covariance matrix). Therefore, the following forms were used:

$$\begin{aligned} \text{AIC}(d) &= 2MN\, v_{ML}(\hat{\theta}_{ML}) + 2\Psi(d,N) \qquad (10) \\ \text{MDL}(d) &= MN\, v_{ML}(\hat{\theta}_{ML}) + \frac{\Psi(d,N)}{2} \log N, \end{aligned}$$

where $\Psi(d,N) = d(2N+1)+1$, and $\hat{\theta}_{ML}$ represents the $d-$dimensional CML estimate of θ.

Finally, figures 3 and 4 show the empirical probability of detection curves versus noise variance for the case of nonzero correlation between the two sources. In figure 3 the correlation between sources was $E\{s_1(t)s_2^*(t)\} = 0.5$, and in figure 4, $E\{s_1(t)s_2^*(t)\} = 0.9$. Figures 2, 3, and 4 clearly illustrate the superior performance of the Bayesian evidence procedure, most notably for small N, and large noise variance.

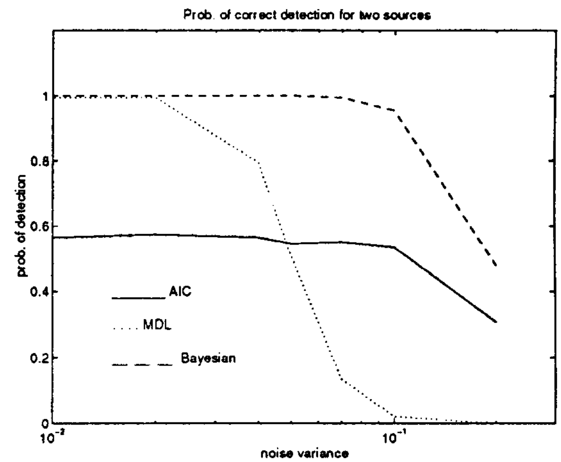

Figure 3: Detection performance vs. noise variance for case of two correlated sources. Rs(1,2) = 0.5

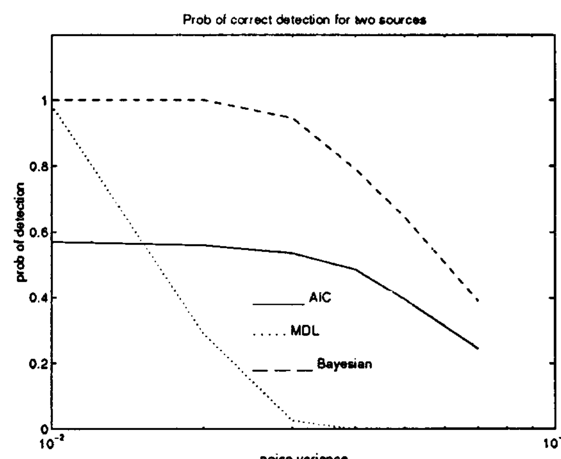

Figure 4: Detection performance vs. noise variance for case of two correlated sources. Rs(1,2) = 0.9

5. SUMMARY

We have presented a Bayesian evidence method for the combined detection/estimation of sources impinging on a general passive sensor array. Simulation results clearly demonstrate the advantages of this approach over information-theoretic techniques in small sample/low SNR situations. Although evaluation of the evidence term in equation (9) is not trivial, this technique can be used to obtain the performance improvements offered by Bayesian methods, without having to resort to Monte Carlo methods, such as importance sampling.

6. REFERENCES

[1] M. Wax and T. Kailath, "Detection of signals by information-theoretic criteria," , IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-33, pp. 387-392, 1985.

[2] M. Wax and I. Ziskind, "Detection of the number of coherent signals by the MDL principle," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 37, No. 8, pp. 1190-1196, Aug. 1989.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control,* Vol. 19, No. 6, pp. 716-722, Dec. 1974.

[4] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Annals of Statistics,* Vol. 11, No. 2, pp. 416-431, 1983.

[5] D.J.C. MacKay, "Bayesian Interpolation," in Maximum Entropy and Bayesian Methods, Seattle, pp. 39-66, Kluwer Academic Publishers, 1992.

[6] G.L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation,* Springer-Verlag, 1989.

[7] P.M. Djuric, S.M. Kay, "Model selection based on Bayesian predictive densities and multiple data records," *IEEE Transactions on Signal Processing,* Vol. 42, No. 7, pp. 1685-1699, July 1994.

[8] Y. Bresler and A. Macovski, "Exact maximum likelihood parameter estimation of superimposed exponential signals in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. ASSP-34, pp. 1081-1089, Oct. 1986.

[9] P. Stoica and A. Nehorai, "MUSIC, maximum-likelihood, and Cramer-Rao bound," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. ASSP-37, pp. 720-741, May 1989.

[10] S. Haykin, J.P. Reilly, V. Kezys, and E. Vertatschitsch, "Some aspects of array signal processing," *IEE Proceedings-F,* Vol. 139, No. 1, pp. 1-26, Feb. 1992.

[11] A.P. Quinn, "The performance of Bayesian estimators in the superresolution of signal parameters," *Proc. IEEE Int. Conf. on Acoust., Speech and Sig. Proc.* (ICASSP92), Vol. V San Francisco, 297-300.

[12] G.E.P. Box and G.C. Tiao, *Bayesian Inference in Statistical Analysis,* Addison-Wesley, 1973.

[13] B.M. Radich and K.M. Buckley, "Combined Bayesian EEG dipole localization and model selection," *in preparation, to be submitted to IEEE Trans. Biom. Eng.*

[14] J. Rissanen, *Stochastic Complexity in Statistical Inquiry,* pp. 58, World Scientific, 1989.

[15] B.M. Radich and K.M. Buckley, "A Bayesian marginalization approach for improved EEG dipole localization," *Proc. 28th Asilomar Conference on Circuits, Systems and Computers,* Asilomar, CA, 1994.