# Bayesian Detection and Estimation of Cisoids in Colored Noise

Chao-Ming Cho, *Member, IEEE,* and Petar M. Djurić, *Member, IEEE*

***Abstract*—In this paper, the problem of estimating the number of cisoids in colored noise is addressed. It is assumed that the noise can be modeled by an autoregression whose order has also to be estimated. A new criterion is proposed for estimating the number of cisoids and the autoregressive model order, as well as a new algorithm for estimating the cisoidal frequencies. In the derivation, a Bayesian methodology and subspace decomposition are employed. The proposed criterion significantly outperforms the popular MDL and AIC as applied in a paper by nagesha and Kay. In addition, an algorithm that reduces the computational complexity of the solution is developed. Computer simulations that demonstrate the performance of the criterion are included.**

## I. Introduction

THE estimation of the number and parameters of close cisoids embedded in Gaussian noise is a crucial problem that arises in many fields ranging from radar, sonar, and radio communications to seismology. Recently, many so called *high-resolution* schemes have been proposed to solve this problem [11]. However, most of these methods are based on a white noise process assumption, which is frequently inadequate and leads to poor estimation performance.

When the noise process has an unknown power spectral density, it is usually assumed that the process can be approximated by a *rational transfer function* model such as autoregressive (AR), moving average (MA), or autoregressive moving average (ARMA) models [5]–[15]. In our paper, as in most of the literature, we assume that the noise process can be modeled by an AR model. The use of the AR model is motivated by its mathematical tractability and good performance for a broad spectrum of scenarios.

Under this assumption and known number of cisoids and AR model order, the maximum likelihood (ML) frequency estimator has been derived in [5], [12], [15], and its improved performance has been reported in [15], [21]. In many practical situations, however, the number of cisoids and the order of the AR process are unknown, and they have to be estimated simultaneously. We will refer to this problem as *model selection,* instead of "detection." One plausible approach for selecting the model is to apply the information criteria such as the AIC [2] and MDL [18], [20]. Unfortunately, they tend to overestimate the number of signals [15].

Manuscript received August 20, 1993; revised June 25, 1995. This work was supported by the National Science Foundation under Award MIP-9110628. The associate editor coordinating review of this paper and approving it for publication was Prof. Fu Li.

C.-M. Cho is with the Microelectronic Technology Inc. (MTI), Taiwan.

P. M. Djurić is with the Electrical Engineering Department, State University of New York at Stony Brook, New York 11794 USA.

IEEE Log Number 9415283.

In this paper, a new Bayesian solution for estimating the number of superimposed cisoids corrupted by additive AR random noise process is proposed. The solution also provides MAP estimates of the cisoidal frequencies which are identical to the ML estimates obtained in [5] and [12], [15]. We derive the marginal *a posteriori* distribution of the signal parameters of interest as well as the Bayesian predictive densities of the data conditioned on the model hypotheses. The estimated number of signals and their parameters are then obtained by minimizing a criterion function derived from these densities. The criterion significantly outperforms the AIC and MDL as used in [15]. Moreover, we propose a recursive algorithm (RA) for solving the nonlinear, multi-dimensional maximization problem in the parameter estimation. With this algorithm which resembles a dynamic programming procedure, we are able to transform the multi-variable maximization to a sequence of single-variable maximizations. This considerably reduces the computational complexity associated with the frequency estimation and model selection.

The paper is organized as follows. The signal model and the problem of interest are defined in Section II. In Section III we derive the Bayesian frequency estimator, and in Section IV we develop a model selection criterion that treats the estimates of the frequencies as true frequencies. The recursive algorithm is described in Section V. The performance of the proposed criterion is demonstrated by computer simulations, and the results are presented in Section VI. Finally, the conclusion is given in Section VII.

## II. Problem Formulation

Assume that we observe $N$ data samples of a complex time series that represent the superimposed cisoids embedded in an AR noise process. That is, the observed data $y[m]$ obey the model (also known as CARD model [12])

$$y[m] = \sum_{i=1}^{q_s} s_i e^{jm\theta_i} + v[m], \qquad m = 1, 2, \cdots, N \quad (1)$$

where $q_s$ is the number of cisoids, $s_i$ is the complex amplitude of the $i$th cisoid, and $\theta_i$ is the frequency of the $i$th cisoid. The amplitudes $s_i$ and the parameters $\theta_i$ are unknown deterministic constants. The sequence $v[m]$ is a $q_a$th order AR noise process expressed by

$$v[m] = \sum_{l=1}^{q_a} a_l \cdot v[m-l] + w[m] \quad (2)$$

where $w[m]$ is a complex zero mean white Gaussian process with the real and imaginary components identically distributed

and with an unknown variance $\sigma^2$. Furthermore, we assume that the AR noise process is covariance stationary, i.e., it does not have poles on or outside the unit circle.

For convenience, we may transform the signal model (1) to a *generalized autoregressive* (GAR) model [5], [21]. The GAR model of (1) is given by

$$y[m] = \sum_{i=1}^{q_s} s_i' e^{jm\theta_i} + \sum_{l=1}^{q_a} a_l \cdot y[m-l] + w[m] \tag{3}$$

where $m = 1, 2, \cdots, N$, and

$$s_i' = s_i \left( 1 - \sum_{l=1}^{q_a} a_l e^{-jl\theta_i} \right). \tag{4}$$

When the above assumptions hold, the problem can be stated as follows. Given a finite number of observed data samples, where $N > q_s + 2q_a + 1$, estimate the parameters of the $q_s$ cisoids, $\theta_1, \cdots, \theta_{q_s}$. When $q_s$ and $q_a$ are unknown, estimate $q_s$ and $q_a$ as well as $\theta_1, \cdots, \theta_{q_s}$.

## III. Bayesian Parameter Estimation

Let $\mathcal{H}_{(k,p)}$ denote the hypothesis that the number of cisoids is $k$ and the order of the AR processes is $p$. Clearly, to use the data model (1) (or (3)), we need the samples $y[0], y[-1], \cdots, y[-p+1]$. Since they are unknown, we express the hypothesis $\mathcal{H}_{(k,p)}$ by

$$\begin{aligned} \boldsymbol{y}_{(p+1,N)} &= \boldsymbol{D}(\boldsymbol{\theta}_{(k)})\boldsymbol{s} + \boldsymbol{v}_{(p+1,N)} \\ &= \boldsymbol{D}(\boldsymbol{\theta}_{(k)})\boldsymbol{s}' + \boldsymbol{Y}\boldsymbol{a} + \boldsymbol{w}_{(p+1,N)} \\ &= \tilde{\boldsymbol{D}}(\boldsymbol{\theta}_{(k)})\boldsymbol{s}_a + \boldsymbol{w}_{(p+1,N)} \end{aligned} \tag{5}$$

where $\boldsymbol{\theta}_{(k)} = [\theta_1 \theta_2 \cdots \theta_k]$, and

$$\boldsymbol{y}_{(p+1,N)} = [y[p+1]y[p+2]\cdots y[N]]^T \tag{6a}$$

$$\boldsymbol{D}(\boldsymbol{\theta}_{(k)}) = \begin{bmatrix} e^{j(p+1)\theta_1} & e^{j(p+1)\theta_2} & \cdots & e^{j(p+1)\theta_k} \\ e^{j(p+2)\theta_1} & e^{j(p+2)\theta_2} & \cdots & e^{j(p+2)\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ e^{jN\theta_1} & e^{jN\theta_2} & \cdots & e^{jN\theta_k} \end{bmatrix} \tag{6b}$$

$$\boldsymbol{s} = [s_1 s_2 \cdots s_k]^T \tag{6c}$$

$$\boldsymbol{s}' = [s_1' s_2' \cdots s_k']^T \tag{6d}$$

$$\boldsymbol{a} = [a_1 a_2 \cdots a_p]^T \tag{6e}$$

$$\boldsymbol{v}_{(p+1,N)} = [v[p+1]v[p+2]\cdots v[N]]^T \tag{6f}$$

$$\boldsymbol{w}_{(p+1,N)} = [w[p+1]w[p+2]\cdots w[N]]^T \tag{6g}$$

and

$$\boldsymbol{Y} = \begin{bmatrix} y[p] & y[p-1] & \cdots & y[1] \\ y[p+1] & y[p] & \cdots & y[2] \\ \vdots & \vdots & \ddots & \vdots \\ y[N-1] & y[N-2] & \cdots & y[N-p] \end{bmatrix}. \tag{6h}$$

The vector $\boldsymbol{s}_a^T = [\boldsymbol{s}'^T \boldsymbol{a}^T]$ denotes a $(k+p) \times 1$ nuisance parameter vector which is formed by the $k$ signal amplitudes $s_i'$ and $p$ AR coefficients $a_l$. The $(N-p) \times (k+p)$ data model matrix $\tilde{\boldsymbol{D}}(\boldsymbol{\theta}_{(k)})$ is defined by

$$\tilde{\boldsymbol{D}}(\boldsymbol{\theta}_{(k)}) = \boldsymbol{D}(\boldsymbol{\theta}_{(k)}) \oplus \boldsymbol{Y} = [\boldsymbol{D}(\boldsymbol{\theta}_{(k)}) \vdots \boldsymbol{Y}]. \tag{7}$$

To simplify the notation, we use $\boldsymbol{\theta}$ instead of $\boldsymbol{\theta}_{(k)}$ whenever it does not cause ambiguity.

Given the data model (5), we want to find the maximum *a posteriori* (MAP) estimates of $\boldsymbol{\theta}$. Since $\boldsymbol{\theta}$ are the only parameters of interest, we adopt the marginal Bayesian inference strategy, that is, we integrate out the nuisance parameters $\boldsymbol{s}', \boldsymbol{a}$, and $\sigma$. The marginal MAP estimate of $\boldsymbol{\theta}$ is given by

$$\begin{aligned} \hat{\boldsymbol{\theta}} &\triangleq \arg \min_{\theta \in \Theta} \{ -\ln f(\boldsymbol{\theta} | \boldsymbol{y}_{(1,N)}, \mathcal{H}_{(k,p)}) \} \\ &= \arg \min_{\theta \in \Theta} \{ -\ln f(\boldsymbol{y}_{(p+1,N)} | \boldsymbol{\theta}, \boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)}) \\ &\quad \cdot f(\boldsymbol{\theta} | \boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)}) \} \end{aligned} \tag{8}$$

where $\hat{\boldsymbol{\theta}}$ denotes the marginal MAP estimate of $\boldsymbol{\theta}$ and $\Theta$ denotes the domain of $\boldsymbol{\theta}$. The conditional likelihood function (CLF) $f(\boldsymbol{y}_{(p+1,N)} | \boldsymbol{\theta}, \boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)})$ is derived by using the marginalization,

$$\begin{aligned} &f(\boldsymbol{y}_{(p+1,N)} | \boldsymbol{\theta}, \boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)}) \\ &\quad = \int_{\boldsymbol{\phi}} f(\boldsymbol{y}_{(p+1,N)} | \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)}) \\ &\qquad \cdot f(\boldsymbol{\phi} | \boldsymbol{\theta}, \boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)}) \, d\boldsymbol{\phi} \end{aligned} \tag{9}$$

where $\boldsymbol{\phi} = \{\boldsymbol{s}_a, \sigma\}$ are the nuisance parameters.

To deal with the marginalizations of the nuisance parameters in (9), we decompose the observed data space $\mathcal{S}_y$ into two complementary subspaces. The subspace spanned by the columns of $\tilde{\boldsymbol{D}}(\boldsymbol{\theta})$ is referred to as the *composite signal subspace*, $\mathcal{S}_{s_a}$. The orthogonal subspace to $\mathcal{S}_{s_a}$ is referred to as the *noise subspace*, $\mathcal{S}_n$. According to this decomposition, the observed data vector $\boldsymbol{y}_{(p+1,N)}$ is then split into two subspace vectors, i.e.,

$$\boldsymbol{y}_{(p+1,N)} = \boldsymbol{G}(\boldsymbol{\theta}) \begin{bmatrix} \boldsymbol{x}_s \\ \boldsymbol{x}_n \end{bmatrix} \tag{10}$$

where $\boldsymbol{x}_s$ and $\boldsymbol{x}_n$ denote the $(k+p) \times 1$ composite signal subspace vector and the $(N-k-2p) \times 1$ noise subspace vector, respectively. $\boldsymbol{G}(\boldsymbol{\theta})$ is an $(N-p) \times (N-p)$ unitary coordinate transformation matrix, and it is given by

$$\boldsymbol{G}(\boldsymbol{\theta}) = [\boldsymbol{U}_s(\boldsymbol{\theta}) \vdots \boldsymbol{U}_n(\boldsymbol{\theta})]. \tag{11}$$

It satisfies the relations,

$$\begin{aligned} \boldsymbol{P}_{\tilde{\boldsymbol{D}}(\boldsymbol{\theta})} &= \tilde{\boldsymbol{D}}(\boldsymbol{\theta})(\tilde{\boldsymbol{D}}^H(\boldsymbol{\theta})\tilde{\boldsymbol{D}}(\boldsymbol{\theta}))^{-1}\tilde{\boldsymbol{D}}^H(\boldsymbol{\theta}) \\ &= \boldsymbol{U}_s(\boldsymbol{\theta})\boldsymbol{U}_s^H(\boldsymbol{\theta}) \end{aligned} \tag{12}$$

and

$$\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{D}}(\boldsymbol{\theta})} = \boldsymbol{I}_{(N-p)} - \boldsymbol{P}_{\tilde{\boldsymbol{D}}(\boldsymbol{\theta})} = \boldsymbol{U}_n(\boldsymbol{\theta})\boldsymbol{U}_n^H(\boldsymbol{\theta}) \tag{13}$$

where $\boldsymbol{I}_{(m)}$ denotes the $m \times m$ identity matrix. $\boldsymbol{U}_s(\boldsymbol{\theta})$ is an $(N-p) \times (k+p)$ matrix whose orthonormal column vectors span the composite signal subspace, and $\boldsymbol{U}_n(\boldsymbol{\theta})$ is an $(N-p) \times (N-k-2p)$ matrix formed by a set of orthonormal vectors as basis of the noise subspace. The matrices $\boldsymbol{P}_{\tilde{\boldsymbol{D}}(\boldsymbol{\theta})}$ and $\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{D}}(\boldsymbol{\theta})}$ are two complementary projection matrices which project onto $\mathcal{S}_{s_a}$ and $\mathcal{S}_n$, respectively.

The property of unitary decomposition allows us to write the marginal CLF of $\boldsymbol{\theta}$ in (9) as [6]

$$\begin{aligned} f(\boldsymbol{y}_{(p+1,N)}|\boldsymbol{\theta},\boldsymbol{y}_{(1,p)},\mathcal{H}_{(k,p)}) &= f(\boldsymbol{x}_s,\boldsymbol{x}_n|\boldsymbol{\theta},\boldsymbol{y}_{(1,p)},\mathcal{H}_{(k,p)}) \\ &= \int_{\sigma}\int_{\tilde{\boldsymbol{s}}_a} f(\boldsymbol{x}_s,\boldsymbol{x}_n|\boldsymbol{\theta},\tilde{\boldsymbol{s}}_a,\sigma,\boldsymbol{y}_{(1,p)},\mathcal{H}_{(k,p)}) \\ &\quad \cdot f(\tilde{\boldsymbol{s}}_a,\sigma|\boldsymbol{\theta},\boldsymbol{y}_{(1,p)},\mathcal{H}_{(k,p)})\, d\tilde{\boldsymbol{s}}_a\, d\sigma \end{aligned} \tag{14}$$

where $\tilde{\boldsymbol{s}}_a = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{s}_a$ denotes a transformed nuisance parameter vector in the composite signal subspace and $\boldsymbol{R}(\boldsymbol{\theta})$ is obtained from the *QR factorization* of the matrix $\tilde{\boldsymbol{D}}(\boldsymbol{\theta})$ ($\tilde{\boldsymbol{D}}(\boldsymbol{\theta}) = \boldsymbol{U}_s(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{\theta})$) [19].

Since we assume that we know nothing about $\tilde{\boldsymbol{s}}_a$ and $\sigma$, we choose noninformative priors by applying the Jeffreys' invariance principle [3]. According to the principle, a noninformative prior is derived by *requiring* invariance of inference under parameter transformation, which entails that the noninformative prior for a set of parameters is proportional to the square root of the determinant of the Fisher's information matrix. Here we introduce another approximation. Namely, the prior in (14), $f(\tilde{\boldsymbol{s}}_a,\sigma|\boldsymbol{\theta},\boldsymbol{y}_{(1,p)},\mathcal{H}_{(k,p)})$, depends also on the first $p$ samples, $y[1], y[2], \cdots, y[p]$. We will ignore this dependence (a) for tractability reasons and (b) because the number of data samples is smaller than the number of unknown parameters, which precludes acquisition of significant information about $\tilde{\boldsymbol{s}}_a$, and $\sigma$ from the first $p$ samples. This assumption combined with Jeffreys' principle implies

$$f(\tilde{\boldsymbol{s}}_a,\sigma|\boldsymbol{\theta},\boldsymbol{y}_{(1,p)}\mathcal{H}_{(k,p)}) \propto \frac{1}{\sigma}. \tag{15}$$

After substituting (15) into (14), we get

$$\begin{aligned} &f(\boldsymbol{x}_s,\boldsymbol{x}_n|\boldsymbol{\theta},\boldsymbol{y}_{(1,p)},\mathcal{H}_{(k,p)}) \\ &\quad\propto \int_{\sigma}\int_{\tilde{\boldsymbol{s}}_a}\left(\frac{1}{\pi\sigma^2}\right)^{k+p} \\ &\qquad\cdot \exp\left\{-\frac{1}{\sigma^2}(\boldsymbol{x}_s-\tilde{\boldsymbol{s}}_a)^H(\boldsymbol{x}_s-\tilde{\boldsymbol{s}}_a)\right\} d\tilde{\boldsymbol{s}}_a \\ &\qquad\cdot\left(\frac{1}{\pi\sigma^2}\right)^{N-k-2p} \exp\left\{-\frac{1}{\sigma^2}\boldsymbol{x}_n^H\boldsymbol{x}_n\right\}\frac{1}{\sigma}\, d\sigma \\ &\quad= \left(\frac{1}{\pi}\right)^{N-k-2p}\frac{1}{2}\Gamma[N-k-2p]\cdot(\boldsymbol{x}_n^H\boldsymbol{x}_n)^{-(N-k-2p)}. \end{aligned} \tag{16}$$

By using (10), (11), and (13), the inner product of the noise subspace vectors in (16) can be expressed in terms of the observed data as

$$\begin{aligned} \boldsymbol{x}_n^H\boldsymbol{x}_n &= \boldsymbol{y}_{(p+1,N)}^H\boldsymbol{U}_n(\boldsymbol{\theta})\boldsymbol{U}_n^H(\boldsymbol{\theta})\boldsymbol{y}_{(p+1,N)} \\ &= \boldsymbol{y}_{(p+1,N)}^H\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{D}}(\boldsymbol{\theta})}\boldsymbol{y}_{(p+1,N)} \\ &= C(\boldsymbol{\theta}). \end{aligned} \tag{17}$$

When we assume that $f(\boldsymbol{\theta}|\mathcal{H}_{(k,p)}) \propto$ const., the marginal subspace MAP estimator of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta\in\Theta}\{C(\boldsymbol{\theta})\} \tag{18}$$

that is, the estimate of $\boldsymbol{\theta}$ is obtained by minimizing $C(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

The function $C(\boldsymbol{\theta})$ can be further simplified by using a projection matrix update formula [9]. According to this formula, the projection matrix onto the column space of the matrix $[\boldsymbol{A}:\boldsymbol{B}]$ can be decomposed into two projection operators, i.e.,

$$\begin{aligned} \boldsymbol{P}_{\boldsymbol{A}\oplus\boldsymbol{B}} &= \boldsymbol{P}_{\boldsymbol{A}} + \boldsymbol{P}_{\boldsymbol{CB}} \\ &= \boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H + \boldsymbol{CB}(\boldsymbol{B}^H\boldsymbol{CB})^{-1}\boldsymbol{B}^H\boldsymbol{C}^H \end{aligned} \tag{19}$$

where

$$\boldsymbol{C} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{P}^{\perp}_{\boldsymbol{A}}. \tag{20}$$

The matrices $\boldsymbol{P}_{\boldsymbol{A}}$ and $\boldsymbol{P}_{\boldsymbol{CB}}$ are the projection matrices onto the column spaces of the matrices $\boldsymbol{A}$ and $\boldsymbol{P}^{\perp}_{\boldsymbol{A}}\boldsymbol{B}$, respectively. With this update formula, we get

$$C(\boldsymbol{\theta}) = \boldsymbol{y}_{(p+1,N)}^H\boldsymbol{P}^{\perp}_{\boldsymbol{D}(\boldsymbol{\theta})\oplus\boldsymbol{Y}}\boldsymbol{y}_{(p+1,N)} \tag{21}$$

where

$$\begin{aligned} \boldsymbol{P}^{\perp}_{\boldsymbol{D}(\boldsymbol{\theta})\oplus\boldsymbol{Y}} &= \boldsymbol{P}^{\perp}_{\boldsymbol{D}(\boldsymbol{\theta})} - \boldsymbol{P}^{\perp}_{\boldsymbol{D}(\boldsymbol{\theta})}\boldsymbol{Y}(\boldsymbol{Y}^H\boldsymbol{P}^{\perp}_{\boldsymbol{D}(\boldsymbol{\theta})}\boldsymbol{Y})^{-1}\boldsymbol{Y}^H\boldsymbol{P}^{\perp}_{\boldsymbol{D}(\boldsymbol{\theta})} \\ &= \boldsymbol{P}^{\perp}_{\boldsymbol{Y}} - \boldsymbol{P}^{\perp}_{\boldsymbol{Y}}\boldsymbol{D}(\boldsymbol{\theta})(\boldsymbol{D}^H(\boldsymbol{\theta})\boldsymbol{P}^{\perp}_{\boldsymbol{Y}}\boldsymbol{D}(\boldsymbol{\theta}))^{-1}\boldsymbol{D}^H(\boldsymbol{\theta})\boldsymbol{P}^{\perp}_{\boldsymbol{Y}} \end{aligned}$$

and

$$\boldsymbol{P}^{\perp}_{\boldsymbol{D}(\boldsymbol{\theta})} = \boldsymbol{I}_{(N-p)} - \boldsymbol{D}(\boldsymbol{\theta})(\boldsymbol{D}^H(\boldsymbol{\theta})\boldsymbol{D}(\boldsymbol{\theta}))^{-1}\boldsymbol{D}^H(\boldsymbol{\theta}) \tag{22}$$

$$\boldsymbol{P}^{\perp}_{\boldsymbol{Y}} = \boldsymbol{I}_{(N-p)} - \boldsymbol{Y}(\boldsymbol{Y}^H\boldsymbol{Y})^{-1}\boldsymbol{Y}^H. \tag{23}$$

Not surprisingly, since we used the noninformative priors, the resulting Bayesian estimator coincides with the ML estimator derived by Chatterjee, *et al.* [5] and Nagesha and Kay [15]. However, the marginal subspace MAP estimator is instrumental in developing the Bayesian model selection criterion whose derivation is given in the sequel.

## IV. Model Selection Via Bayesian Predictive Densities

Assume the maximum number of cisoids is $Q_s$ and the maximum model order of the AR process is $Q_a$, where $Q_s+2Q_a+1 < N$. The problem is to select the optimal number of cisoids and the optimal order of the AR process. Optimality will be defined by the approximate MAP principle employed via Bayesian predictive densities (BPD) [7], [8]. The BPD criterion is given by

$$\hat{q}_s, \hat{q}_a \triangleq \arg\min_{k\in\{0,1,\cdots,Q_k\};p\in\{0,1,\cdots,Q_a\}} Z_{kp} \tag{24}$$

where

$$Z_{k,p} = \sum_{m=2}^{N} -\ln\, f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(k,p)})$$

with $\hat{q}_s$ being the estimated number of cisoids and $\hat{q}_a$ the estimated AR model order. Note that the models with no cisoid and/or zeroth order AR process (white noise) are also considered as possible hypotheses.

First we sketch the derivation of the predictive density $f(y[m]|\boldsymbol{y}_{(1,m-1)}, \mathcal{H}_{(k,p)})$ for $k>0$ and $p>0$ using the Bayesian inference scheme and the subspace decomposition. According to the subspace decomposition approach used in the previous section, the form of $f(y[m]|\boldsymbol{y}_{(1,m-1)}, \mathcal{H}_{(k,p)})$ can be written as

$$\begin{aligned} & f(y[m]|\boldsymbol{y}_{(1,m-1)}, \mathcal{H}_{(k,p)}) \\ &= \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\phi}} f(y[m]|\boldsymbol{y}_{(1,m-1)}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{H}_{(k,p)}) \\ &\quad \cdot f(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}_{(1,m-1)}, \mathcal{H}_{(k,p)})\, d\boldsymbol{\phi}\, d\boldsymbol{\theta} \\ &= \frac{f(\boldsymbol{x}_{s(m)}, \boldsymbol{x}_{n(m)}|\boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)})}{f(\boldsymbol{x}_{s(m-1)}, \boldsymbol{x}_{n(m-1)}|\boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)})} \end{aligned} \tag{25}$$

where $\boldsymbol{\phi} = \{\boldsymbol{s}_a, \sigma\}$. The last line in (25) follows easily from the first line by applying the first equality in (14). The composite signal and noise subspace vectors $\boldsymbol{x}_{s(m)}$ and $\boldsymbol{x}_{n(m)}$ are transformed from the observed data $\boldsymbol{y}_{(p+1,m)}$ by

$$\boldsymbol{y}_{(p+1,m)} = \boldsymbol{G}(\boldsymbol{\theta}_{(k)})_{(p,m)} \begin{bmatrix} \boldsymbol{x}_{s(m)} \\ \boldsymbol{x}_{n(m)} \end{bmatrix} \tag{26}$$

where the unitary transformation matrix $\boldsymbol{G}(\boldsymbol{\theta}_{(k)})_{(p,m)}$ is defined similarly as in (10). With the use of the same noninformative prior for $\boldsymbol{\phi}$ as in (15), we get

$$\begin{aligned} & f(\boldsymbol{x}_{s(m)}, \boldsymbol{x}_{n(m)}|\boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)}) \\ &= \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\phi}} f(\boldsymbol{x}_{s(m)}, \boldsymbol{x}_{n(m)}|\boldsymbol{y}_{(1,p)}, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{H}_{(k,p)}) \\ &\quad \cdot f(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)})\, d\boldsymbol{\phi}\, d\boldsymbol{\theta} \\ &\propto \int_{\boldsymbol{\theta}} \left(\frac{1}{\pi}\right)^{m-k-2p} \frac{1}{2}\Gamma[m-k-2p] \\ &\quad \cdot \{C(\boldsymbol{\theta}_{(k)})_{(p,m)}\}^{-(m-k-2p)} \\ &\quad \cdot f(\boldsymbol{\theta}|\boldsymbol{y}_{(1,p)}, \mathcal{H}_{(k,p)})\, d\boldsymbol{\theta} \end{aligned} \tag{27}$$

where

$$C(\boldsymbol{\theta}_{(k)})_{(p,m)} = \boldsymbol{y}_{(p+1,m)}^H \boldsymbol{P}^{\perp}_{\boldsymbol{D}(\boldsymbol{\theta}_{(k)})_{(p,m)} \oplus \boldsymbol{Y}_{(p,m)}} \boldsymbol{y}_{(p+1,m)} \tag{28}$$

$$\boldsymbol{D}(\boldsymbol{\theta}_{(k)})_{(p,m)} = \begin{bmatrix} e^{j(p+1)\theta_1} & e^{j(p+1)\theta_2} & \cdots & e^{j(p+1)\theta_k} \\ e^{j(p+2)\theta_1} & e^{j(p+2)\theta_2} & \cdots & e^{j(p+2)\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ e^{jm\theta_1} & e^{jm\theta_2} & \cdots & e^{jm\theta_k} \end{bmatrix} \tag{29}$$

and

$$\boldsymbol{Y}_{(p,m)} = \begin{bmatrix} y[p] & y[p-1] & \cdots & y[1] \\ y[p+1] & y[p] & \cdots & y[2] \\ \vdots & \vdots & \ddots & \vdots \\ y[m-1] & y[m-2] & \cdots & y[m-p] \end{bmatrix}. \tag{30}$$

Since the integration of $\boldsymbol{\theta}_{(k)}$ in (27) cannot be performed in a closed form, a maximum likelihood approximation is used [1], [3], [16]. From the theory of large sample posterior distributions it is well known that the likelihood $\mathcal{L}_{(p,m)}(\boldsymbol{\theta}_{(k)})$ is approximately normal in $\boldsymbol{\theta}_k$ [17], and that this approximation may be very good even for a small number of samples. Now, let the model hypothesis be $\mathcal{H}_{(k,p)}$ and assume that the normality assumption holds, as well as that the prior $f(\boldsymbol{\theta}|\mathcal{H}_{(k,p)})$ is locally uniform in the neighborhood of $\hat{\boldsymbol{\theta}}_{(k)}$, where $\hat{\boldsymbol{\theta}}_{(k)}$ is obtained from

$$\hat{\boldsymbol{\theta}}_{(k)} = \arg \min_{\theta \in \Theta} \{C(\boldsymbol{\theta}_{(k)})_{(p,N)}\}. \tag{31}$$

Then

$$\begin{aligned} \mathcal{L}_{(p,m)}(\boldsymbol{\theta}_{(k)}) &\doteq \mathcal{L}_{(p,m)}(\hat{\boldsymbol{\theta}}_{(k)}) \\ &\quad \cdot \exp\{-\tfrac{1}{2}(\boldsymbol{\theta}_{(k)} - \hat{\boldsymbol{\theta}}_{(k)})^T \boldsymbol{J}^{(k)}_{(p,m)} (\boldsymbol{\theta}_{(k)} - \hat{\boldsymbol{\theta}}_{(k)})\} \end{aligned} \tag{32}$$

where $\boldsymbol{J}^{(k)}_{(p,m)}$ is defined by

$$\boldsymbol{J}^{(k)}_{(p,m)} = -\nabla^2_{\theta} \log f(\boldsymbol{y}_{(p+1,m)}|\boldsymbol{y}_{(1,p)}, \boldsymbol{\theta}, \mathcal{H}_{(k,p)})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{33}$$

and $\nabla_{\theta}$ denotes the gradient operator w.r.t. $\boldsymbol{\theta}$.

With the above assumptions and after some algebra, the predictive density (25) for $k>0$ and $p>0$ becomes

$$\begin{aligned} & f(y[m]|\boldsymbol{y}_{(1,m-1)}, \mathcal{H}_{(k,p)}) \\ &= \frac{1}{\pi} \frac{\Gamma(m-k-2p)}{\Gamma(m-k-2p-1)} \frac{(C(\hat{\boldsymbol{\theta}}_{(k)})_{(p,m-1)})^{(m-k-2p-1)}}{(C(\hat{\boldsymbol{\theta}}_{(k)})_{(p,m)})^{(m-k-2p)}} \\ &\quad \cdot \left(\frac{\det(\boldsymbol{J}^{(k)}_{(p,m-1)})}{\det(\boldsymbol{J}^{(k)}_{(p,m)})}\right)^{1/2} \\ &\doteq \frac{(m-k-2p-1)}{\pi} \frac{(C(\hat{\boldsymbol{\theta}}_{(k)})_{(p,m-1)})^{(m-k-2p-1)}}{(C(\hat{\boldsymbol{\theta}}_{(k)})_{(p,m)})^{(m-k-2p)}} \\ &\quad \cdot \left(\frac{m-1}{m}\right)^{3k/2} \end{aligned} \tag{34}$$

for $m > k+2p+1$. The case of $m \leq k+2p+1$ will be discussed later. We use the approximation sign in (34) because we assume that $\det(\boldsymbol{J}^{(k)}_{(p,m-1)})/\det(\boldsymbol{J}^{(k)}_{(p,m)}) \doteq ((m-1)/m)^{3k/2}$. We also assume that the terms $C(\hat{\boldsymbol{\theta}}_{(k)})_{(p,m-1)}$ and $C(\hat{\boldsymbol{\theta}}_{(k)})_{(p,m)}$ are both evaluated at $\hat{\boldsymbol{\theta}}_{(k)}$ defined by (31). These assumptions are made to allow for simplified form of the model selection rule as well as great reduction in its computational complexity. Note that the first assumption is due to the fact that $\det(\boldsymbol{J}^{(k)}_{(p,m)}) = O(m^{3k})$. The second assumption ignores the nonzero contribution of the first term of the Taylor expansion of $\mathcal{L}_{(p,m)}(\boldsymbol{\theta}_{(k)})$. However, all of these terms that appear in the overall criterion are canceled, except the first and the last one. The last one is zero, and the first one has expected value equal to zero.

Next, we consider the hypothesis of $k$ cisoids with additive white noise ($p=0$). After applying a similar approach as above, for the predictive density of $y[m]$ under $\mathcal{H}_{(k,0)}$, we obtain

$$f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(k,0)}) = \int_{\boldsymbol{\theta}}\int_{\boldsymbol{\phi}} f(y[m]|\boldsymbol{y}_{(1,m-1)},\boldsymbol{\theta},\boldsymbol{\phi},\mathcal{H}_{(k,0)}) \cdot f(\boldsymbol{\theta},\boldsymbol{\phi}|\mathcal{H}_{(k,0)})\, d\boldsymbol{\phi}\, d\boldsymbol{\theta} = \frac{(\boldsymbol{y}^H_{(1,m-1)}\boldsymbol{P}^{\perp}_{\boldsymbol{D}(\hat{\boldsymbol{\theta}}_{(k)})_{(0,m-1)}}\boldsymbol{y}_{(1,m-1)})^{(m-k-1)}}{(\boldsymbol{y}^H_{(1,m)}\boldsymbol{P}^{\perp}_{\boldsymbol{D}(\hat{\boldsymbol{\theta}}_{(k)})_{(0,m)}}\boldsymbol{y}_{(1,m)})^{(m-k)}} \cdot \frac{(m-k-1)}{\pi}\left(\frac{m-1}{m}\right)^{3k/2} \tag{35}$$

for $m > k+1$, and $\hat{\boldsymbol{\theta}}_{(k)}$ is given by

$$\hat{\boldsymbol{\theta}}_{(k)} = \arg\min_{\theta\in\Theta}\{\boldsymbol{y}^H_{(1,N)}\boldsymbol{P}^{\perp}_{\boldsymbol{D}(\hat{\boldsymbol{\theta}}_{(k)})_{(0,N)}}\boldsymbol{y}_{(1,N)}\}. \tag{36}$$

Again, we will discuss the case of $m \le k+1$ later.

When the observed data is a $p$th order AR process only, the predictive density of $y[m]$ becomes

$$f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(0,p)}) = \frac{(\boldsymbol{y}^H_{(p+1,m-1)}\boldsymbol{P}^{\perp}_{\boldsymbol{Y}_{(p,m-1)}}\boldsymbol{y}_{(p+1,m-1)})^{(m-2p-1)}}{(\boldsymbol{y}^H_{(p+1,m)}\boldsymbol{P}^{\perp}_{\boldsymbol{Y}_{(p,m)}}\boldsymbol{y}_{(p+1,m)})^{(m-2p)}} \cdot \frac{(m-2p-1)}{\pi}. \tag{37}$$

Now, we consider the case when there are no cisoids and the additive noise is white. The result is

$$f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(0,0)}) = \frac{(\boldsymbol{y}^H_{(1,m-1)}\boldsymbol{y}_{(1,m-1)})^{(m-1)}}{(\boldsymbol{y}^H_{(1,m)}\boldsymbol{y}_{(1,m)})^{m}}\,\frac{(m-1)}{\pi} \tag{38}$$

for $m > 1$. Hence, the cost function for $\mathcal{H}_{(0,0)}$ is obtained by the summation in (24) from $m = 2$ to $N$, i.e.,

$$J_{(0,0)} = -\sum_{m=2}^{N} \ln\, f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(0,0)}). \tag{39}$$

Note that we can write the predictive density of the next sample only if there are sufficient past samples to estimate the parameters of the model [7]. Under the hypothesis $\mathcal{H}_{(0,0)}$, the only unknown is $\sigma^2$, which implies that the predictive density of $y[2]$ can be determined, and therefore the summation in (39) can start with $m = 2$.

When $k > 0$ and $p = 0$, the number of model parameters is $2k+1$. From the model, we can deduce that the least number of samples necessary to ensure a unique estimate is $k+1$. However, in order to use the same amount of data samples for comparison of different models, the procedure has to start from $m = 2$. Thus, we design an initialization policy for $m < k+1$ to allow for the derivation of the predictive densities of $y[m]$ without violating the above two requirements[1]. For instance, when $k = 1$ the cost function for $\mathcal{H}_{(1,0)}$ is given by

$$J_{(1,0)} = -\ln\, f(y[2]|y[1],\mathcal{H}_{(0,0)}) - \sum_{m=3}^{N} \ln\, f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(1,0)}). \tag{40}$$

[1]This policy, although quite logical, is not optimal in any sense.

Apparently, the evaluation of $J_{(k,0)}$ can be implemented similarly.

So is the case when $k = 0$ and $p > 0$. Then, the minimum number of samples for estimating the model parameters is $2p+1$. For example, for $p = 2$ we have

$$J_{(0,2)} = -\sum_{m=2}^{3} \ln\, f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(0,0)}) - \sum_{m=4}^{5} \ln\, f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(0,1)}) - \sum_{m=6}^{N} \ln\, f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(0,2)}). \tag{41}$$

Finally, we consider the BPD criterion for the case $k > 0$ and $p > 0$. Since the determination of the number of cisoids is usually more important than that of the AR process order, we design an initialization policy as follows: when $m < k+2p+1$, we first adopt the policy such as in (40), then continue as in (41). For instance, for $k = 2$ and $p = 1$, we have

$$J_{(2,1)} = -\ln\, f(y[2]|y[1],\mathcal{H}_{(0,0)}) - \ln\, f(y[3]|\boldsymbol{y}_{(1,2)},\mathcal{H}_{(1,0)}) - \sum_{m=4}^{5} \ln\, f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(2,0)}) - \sum_{m=6}^{N} \ln\, f(y[m]|\boldsymbol{y}_{(1,m-1)},\mathcal{H}_{(2,1)}). \tag{42}$$

In summary, we write the BPD model selection criterion as

$$\hat{q}_s, \hat{q}_a = \arg\min_{k,p}\{J_{(k,p)}\}. \tag{43}$$

Note that whenever the estimates of $\boldsymbol{\theta}$ for all the hypotheses are determined (from (31) or (36)), the computation of the cost function in the proposed BPD criterion can be found straightforwardly. The most difficult point is how to efficiently determine the estimate of $\boldsymbol{\theta}$. We discuss this in the next section.

## V. A Recursive Algorithm for Parameter Estimation

The computations of the MAP (or ML) estimator and the BPD model selection criterion proposed in the previous sections are intensive due to the multi-dimensional nonlinear parameter search. In order to reduce this intensity, we propose a recursive algorithm to estimate the nonlinear parameters. With this algorithm, the multi-dimensional optimization problem in (18) and (31) can be transformed into a multistage one-dimensional optimization problem.

First we consider the application of the recursive algorithm to the frequency estimation, and then extend it to model selection. Before we present the basic algorithm, we need to define the cost function of the MAP estimator. For a particular $p$th order AR noise process, the cost function for estimating $\boldsymbol{\theta}_{(q)}$ $(= [\theta_1 \cdots \theta_q])$ is given by

$$\hat{\boldsymbol{\theta}}_{(q)} = \arg\min_{\theta\in\Theta}\{\boldsymbol{y}^H\boldsymbol{P}^{\perp}_{\boldsymbol{D}(\boldsymbol{\theta}_{(q)})\oplus\boldsymbol{Y}}\boldsymbol{y}\} \tag{44}$$

$$= \arg\max_{\theta\in\Theta}\{\boldsymbol{y}^H\boldsymbol{P}_{\boldsymbol{D}(\boldsymbol{\theta}_{(q)})\oplus\boldsymbol{Y}}\boldsymbol{y}\} \tag{45}$$

where $\boldsymbol{y} = \boldsymbol{y}_{(p+1,N)}$, and $\boldsymbol{D}(\boldsymbol{\theta}_{(q)})$ and $\boldsymbol{Y}$ are defined in (6b) and (7), respectively.

Next, let $\boldsymbol{\theta}_{(k)} = [\theta_1 \theta_2 \cdots \theta_k]$ for $k < q$ and assume $\theta_1 < \theta_2 < \cdots < \theta_k$. According to the projection matrix update formula (19), the projection matrix onto the subspace spanned by the column vectors of $\boldsymbol{D}(\boldsymbol{\theta}_{(k+1)})$ and $\boldsymbol{Y}$ can be expressed by the sum of two projection matrices onto the orthogonal subspaces spanned by the column vectors of $\boldsymbol{D}(\boldsymbol{\theta}_{(k)}) \oplus \boldsymbol{Y}$ and the residual of $\boldsymbol{D}(\boldsymbol{\theta}_{(k)}) \oplus \boldsymbol{Y}$ projected onto $\boldsymbol{d}(\theta_{k+1})$. With this decomposition, we get

$$\boldsymbol{P}_{\boldsymbol{D}(\boldsymbol{\theta}_{(k+1)}) \oplus \boldsymbol{Y}} = \boldsymbol{P}_{\boldsymbol{D}(\boldsymbol{\theta}_{(k)}) \oplus \boldsymbol{Y}} + \boldsymbol{P}_{\boldsymbol{B}(\boldsymbol{\theta}_{(k)}) \boldsymbol{d}(\theta_{k+1})} \tag{46}$$

where

$$\boldsymbol{B}(\boldsymbol{\theta}_{(k)}) = \boldsymbol{P}^{\perp}_{\boldsymbol{D}(\boldsymbol{\theta}_{(k)}) \oplus \boldsymbol{Y}} \tag{47}$$

$$\boldsymbol{d}(\theta_{k+1}) = [e^{j(p+1)\theta_{k+1}} e^{j(p+2)\theta_{k+1}} \cdots e^{jN\theta_{k+1}}]^T. \tag{48}$$

Also, a recursive form of the cost function for $\boldsymbol{\theta}_{(k+1)}$ is defined by

$$\begin{aligned} g(\boldsymbol{\theta}_{(k+1)}) &= \boldsymbol{y}^H \boldsymbol{P}_{\boldsymbol{D}(\boldsymbol{\theta}_{(k)}) \oplus \boldsymbol{Y}} \boldsymbol{y} + \boldsymbol{y}^H \boldsymbol{P}_{\boldsymbol{B}(\boldsymbol{\theta}_{(k)}) \boldsymbol{d}(\theta_{k+1})} \boldsymbol{y} \\ &= g(\boldsymbol{\theta}_{(k)}) + \boldsymbol{y}^H \boldsymbol{P}_{\boldsymbol{B}(\boldsymbol{\theta}_{(k)}) \boldsymbol{d}(\theta_{k+1})} \boldsymbol{y}. \end{aligned} \tag{49}$$

Now, we derive a recursive equation to solve the maximization problem in (45). We define a cost function, $F_{k+1}(\theta_{k+1})$, at the $(k+1)$st stage as [10]

$$F_{k+1}(\theta_{k+1}) = \max_{\theta_k} \{F_k(\theta_k) + \boldsymbol{y}^H \boldsymbol{P}_{\boldsymbol{B}(\Phi_{(k-1)}:\theta_k) \boldsymbol{d}(\theta_{k+1})} \boldsymbol{y}\} \tag{50}$$

and we define the estimate of $\theta_k$ (decision) corresponding to a particular $\theta_{k+1}$ by

$$\hat{\theta}_k(\theta_{k+1}) = \arg \max_{\theta_k} \{F_k(\theta_k) + \boldsymbol{y}^H \boldsymbol{P}_{\boldsymbol{B}(\Phi_{(k-1)}:\theta_k) \boldsymbol{d}(\theta_{k+1})} \boldsymbol{y}\}. \tag{51}$$

The vector $\Phi_{(k-1)}$ is a function of $\theta_k$ (policy function) and is given by

$$\Phi_{(k-1)}(\theta_k) = [\hat{\theta}_1(\hat{\theta}_2)\hat{\theta}_2(\hat{\theta}_3) \cdots \hat{\theta}_{k-2}(\hat{\theta}_{k-1})\hat{\theta}_{k-1}(\theta_k)]. \tag{52}$$

The residual matrix $\boldsymbol{B}(\Phi_{(k-1)}:\theta_k)$ is also a function of $\theta_k$ and is defined by

$$\boldsymbol{B}(\Phi_{(k-1)}:\theta_k) = \boldsymbol{P}^{\perp}_{\boldsymbol{D}(\Phi_{(k-1)}(\theta_k)) \oplus \boldsymbol{d}(\theta_k) \oplus \boldsymbol{Y}} \tag{53}$$

for $k = 1, 2, \cdots, q-1$. Note that $F_{k+1}(\theta_{k+1})$ consists of $F_k(\theta_k)$ resulting from the policy function $\Phi_{(k-1)}(\theta_k)$ made at all previous stages and a term dependent on the current decision $\theta_{k+1}$. In the last stage, we get the final maximum cost function,

$$\hat{\theta}_q = \arg \max_{\theta_q} \{F_k(\theta_q)\}. \tag{54}$$

Accordingly, the estimates of $\boldsymbol{\theta}_{(q)}$ are determined from $[\Phi_{(q-1)}(\hat{\theta}_q)\hat{\theta}_q]$.

Equations (50)–(54) form the basic algorithm for estimating the $\theta$s of the $q$ signals in a $p$th order AR noise process. The algorithm consists of two parts: The first part maximizes the recursive cost function based on the cost function and decision at the previous stage, and the second part makes the decision (parameter search) at the current stage. The algorithm does not guarantee the optimal solution since $\boldsymbol{B}(\boldsymbol{\theta}_{(k)})$ in (49) was substituted by $\boldsymbol{B}(\Phi_{(k-1)}:\theta_k)$ in (50). Other approximations are also possible [22], [23].

We now explain our algorithm in more detail. First, we initialize $k = 0$ with $\boldsymbol{B}(\theta_0) = \boldsymbol{P}^{\perp}_{\boldsymbol{Y}}$ and $F_0(\theta_0) = 0$. Then $F_1(\theta_1)$ is computed for $\theta_1 \in \Theta$. At stage 2, $F_2(\theta_2)$ is obtained by maximizing the sum of $F_1(\theta_1)$ and the term which is dependent on the current decision. From (51) we obtain the optimal $\theta_1$ for each state of the second stage. Next, we increase $k$ by 1 and continue along the same lines. When $k = q-1$, we compute the cost function at the last stage and stop the algorithm. The estimates of $\boldsymbol{\theta}_{(q)}$ are then sequentially determined by a backward search. For example, we substitute $\hat{\theta}_q$ into the argument of the function $\hat{\theta}_{q-1}(\theta_q)$, and substitute the resulting $\hat{\theta}_{q-1}$ into the argument of the function $\hat{\theta}_{q-2}(\theta_{q-1})$, and so on. Basically, we need a two-dimensional search at each stage to find out $F_{k+1}(\theta_{k+1})$ and $\hat{\theta}_k(\theta_{k+1})$. However, in finding these terms we can use the result of $F_k(\theta_k)$ computed from the previous stage. Furthermore, the second term of the cost function can be decomposed into a sequence of simple vector operations. This reduces the computational load considerably.

Now, we summarize this algorithm as follows:

1) Initialization:
   1.1 $k = 0$; $\boldsymbol{B}(\theta_0) = \boldsymbol{P}^{\perp}_{\boldsymbol{Y}}$; $F_0(\theta_0) = 0$;
   1.2 Compute $F_1(\theta_1)$;
2) Main Loop: For $k = 1 : (q-1)$,
   For $\theta_{k+1} \in \Theta$,
   Find $F_{k+1}(\theta_{k+1})$ from (50), and
   $\hat{\theta}_k(\theta_{k+1})$ from (51);
   Tabulate $F_{k+1}(\theta_{k+1})$ and $\hat{\theta}_k(\theta_{k+1})$;
3) Compute $\hat{\theta}_q$ from (54), and $\hat{\Phi}_{(q-1)}(\hat{\theta}_q)$
   End

For the model selection problem, the BPD model selection criterion is implemented according to

1) Initialize $p = 0$.
2) Run the basic algorithm for $Q_s$ stages.
3) Find $\hat{\theta}_1, \hat{\boldsymbol{\theta}}_{(2)}, \cdots, \hat{\boldsymbol{\theta}}_{(Q_s)}$. Substitute them into the BPD criterion.
4) Increase $p$ by 1. If $p \leq Q_a$, go back to step 2.
5) Determine the number of signals and the order of the AR process by the BPD criterion.

Finally, it should be noted that the computational load measured by the number of times we evaluate (50) and (51) is $(p+1)q_s M^2$, where $M$ is the number of grid points on the frequency axis. Clearly, this load increases linearly as we add new models to the set of competing models.

## VI. Simulation Results

We examined the performance of the subspace MAP estimator and the BPD model selection criterion by Monte Carlo simulations. Two examples with four cisoids in colored noise

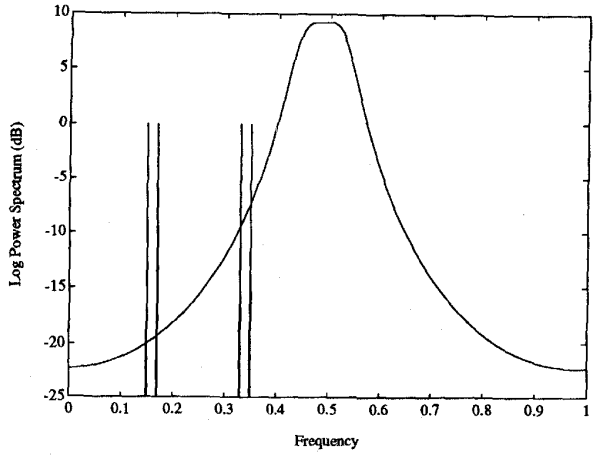



Fig. 1. True log power spectrum of the signal in example 1 (SNR = 0 dB).

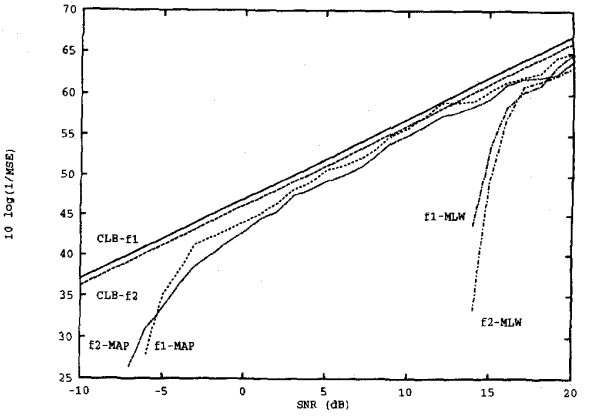



Fig. 3. Performance comparison of the MAP-C and ML-W estimators for $f_1$ and $f_2$ (example 1).

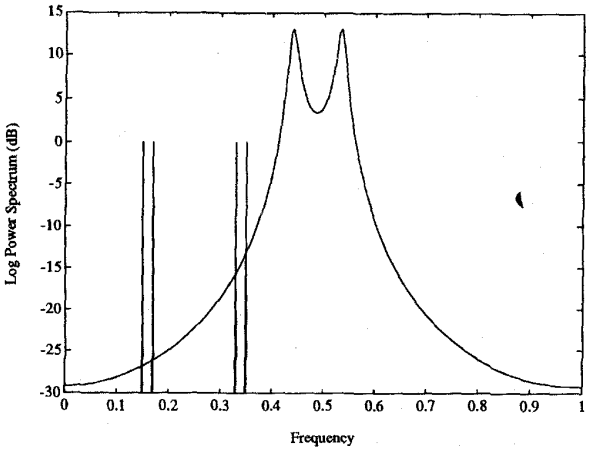



Fig. 2. True log power spectrum of the signal in example 2 (SNR = 0 dB).

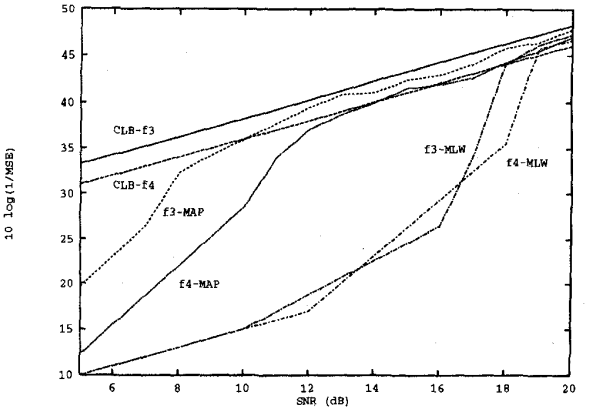



Fig. 4. Performance comparison of the MAP-C and ML-W estimators for $f_3$ and $f_4$ (example 1).

were examined. The normalized frequencies of the four equal power cisoids were $f_1 = 0.15$, $f_2 = 0.17$, $f_3 = 0.33$, and $f_4 = 0.35$, and their phases were $\phi_1 = 0$, $\phi_2 = \pi/4$, $\phi_3 = 0$, and $\phi_4 = -\pi/4$, respectively. The number of samples was 25 and the number of Monte Carlo runs was 100. In the first example, the colored noise was modeled by a second order wide-band AR process whose poles were located at $0.8\ \exp(-j1.05\pi)$ and $0.8\ \exp(-j0.9\pi)$. In the second example, the noise was a second order narrow-band AR process with poles located at $0.95\ \exp(-j1.07\pi)$ and $0.95\ \exp(-j0.88\pi)$. The true log power spectra of these examples (SNR = 0 dB) are shown in Figs. 1 and 2.

To illustrate the performance of the proposed estimator, we compare it with the ML estimator derived for the white noise case for a wide range of SNR's [11], [4][2]. For brevity, we refer to the MAP estimator for the colored noise case as MAP-C, and the ML estimator for the white noise case as ML-W. The estimates of the parameters were computed via the recursive algorithm. The search resolution used in the recursive algorithm was 1/100. The SNR is determined from

$$\text{SNR} = 10 \cdot \log_{10} \left( \frac{|s|^2 \prod_{i=1}^{q_a} (1 - |\alpha_i|^2)}{\sigma^2} \right) \tag{55}$$

where $\alpha_i$ for $i = 1, \cdots, q_a$, is the $i$th *reflection coefficient* of the $q_a$th order AR process [11].

For the first example, the performance comparison of the MAP-C (or ML-C) estimator (see (21) or (31)) and the ML-W estimator (see (36)) are shown in Figs. 3 and 4. We observe that the performance improvement of the proposed estimator is significant, especially when the cisoids are located in the regions with low noise power, i.e., $f_1$ and $f_2$ are better estimated than $f_3$ and $f_4$. The performance of the proposed estimator degrades when the cisoid is very close to the poles of the AR process. The observed nonmonotonic degradation in performance with decrease of the SNR is due to the statistical fluctuations of the Monte Carlo experiment.

Next, we investigated the performance of the proposed estimator for the narrow-band AR noise case. The simulation

[2]This comparison will show the degradation in performance of frequency estimation due to incorrect white noise assumption.

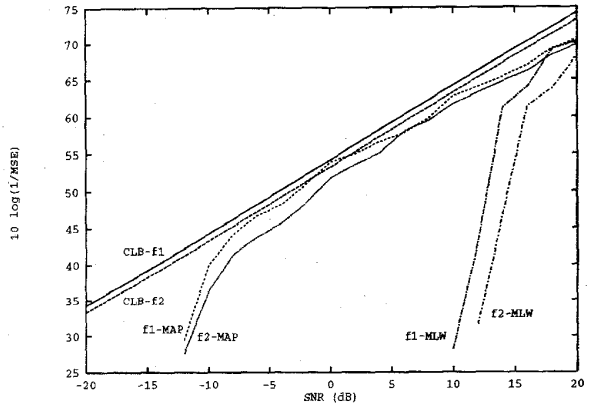



Fig. 5. Performance comparison of the MAP-C and ML-W estimators for $f_1$ and $f_2$ (example 2).

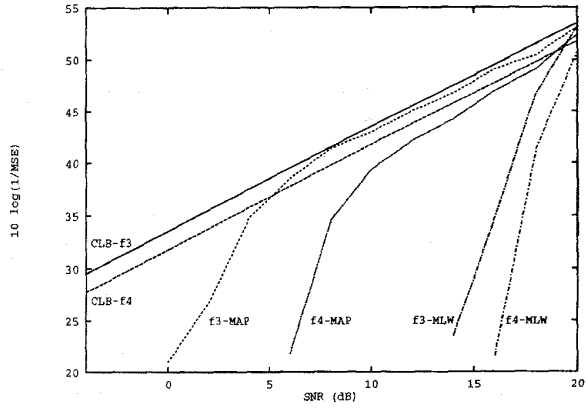



Fig. 6. Performance comparison of the MAP-C and ML-W estimators for $f_3$ and $f_4$ (example 2).

results are shown in Figs. 5 and 6. These results lead to the same conclusions as the ones for the wide-band AR noise case. The performance of the ML-W estimator is relatively poor when the noise is colored, and the MAP-C estimator significantly outperforms the ML-W estimator. The gain in the estimation performance of the MAP-C estimator is greater when the cisoids are far from the poles of the AR process.

For the model selection, we compared the performance of the BPD, MDL, and AIC criteria. According to the derivation in [15], the MDL and AIC criteria are given by

$$\mathrm{MDL}(\hat{q}_s, \hat{q}_a) = \arg \min_{k,p} \left\{ (N-p) \cdot \ln \frac{\pi C(\hat{\boldsymbol{\theta}}_{(k)})_{(p,N)}}{N-p} + \frac{3k+2p+1}{2} \ln N \right\} \tag{56}$$

and

$$\mathrm{AIC}(\hat{q}_s, \hat{q}_a) = \arg \min_{k,p} \left\{ (N-p) \cdot \ln \frac{\pi C(\hat{\boldsymbol{\theta}}_{(k)})_{(p,N)}}{N-p} + (3k+2p+1) \right\} \tag{57}$$

TABLE I
PERFORMANCE COMPARISON OF THE THREE CRITERIA FOR EXAMPLE 1, SNR = 0 DB

| | | $k=0$ | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ | $k=6$ |
|---|---|---|---|---|---|---|---|---|
| BPD | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 7 | 5 | 0 |
| | $p=2$ | 0 | 0 | 0 | 1 | 77 | 9 | 1 |
| | $p=3$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MDL | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 1 | 3 | 8 |
| | $p=2$ | 0 | 0 | 0 | 0 | 49 | 17 | 19 |
| | $p=3$ | 0 | 0 | 0 | 0 | 1 | 2 | 0 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AIC | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 2 | 1 | 11 |
| | $p=2$ | 0 | 0 | 0 | 0 | 38 | 21 | 23 |
| | $p=3$ | 0 | 0 | 0 | 0 | 0 | 1 | 3 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

TABLE II
PERFORMANCE COMPARISON OF THE THREE CRITERIA FOR EXAMPLE 1, SNR = 6 DB

| | | $k=0$ | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ | $k=6$ |
|---|---|---|---|---|---|---|---|---|
| BPD | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 8 | 3 | 0 |
| | $p=2$ | 0 | 0 | 0 | 1 | 81 | 7 | 0 |
| | $p=3$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MDL | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 1 | 3 | 11 |
| | $p=2$ | 0 | 0 | 0 | 0 | 58 | 9 | 18 |
| | $p=3$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AIC | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 2 | 5 | 11 |
| | $p=2$ | 0 | 0 | 0 | 0 | 49 | 12 | 21 |
| | $p=3$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

where $C(\hat{\boldsymbol{\theta}}_{(k)})_{(p,N)}$ is defined in (31). The performance comparisons for the wide-band AR noise process are presented in Tables I and II. We observe that the BPD criterion has the best performance for correctly choosing the signal model. The MDL and AIC criteria tend to overestimate the signal model because the penalty terms in (56) and (57) are not stringent enough. The simulation results for the narrow-band AR noise process are shown in Tables III and IV. Again, the best performance is obtained by the BPD criterion.

Finally, we have also compared the performance of the selection rules in a white noise case scenario. There were two sinusoids whose frequencies and phases were $f_1 = 0.45$, $f_2 = 0.5$, $\phi_1 = \pi/4$, and $\phi_2 = 0$. The SNR was varied between 0 and 10 dB in steps of 1 dB. For each SNR there were 100 Monte Carlo runs. The results displayed in Table V show the number of times the correct hypothesis $\mathcal{H}_{(2,0)}$ was selected. Again, the BPD had the best performance.

## VII. CONCLUSION

In this paper, we have presented a marginal MAP estimator for frequency estimation and a Bayesian predictive density

TABLE III
PERFORMANCE COMPARISON OF THE THREE CRITERIA FOR EXAMPLE 2, SNR = 0 DB

| | | $k=0$ | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ | $k=6$ |
|---|---|---|---|---|---|---|---|---|
| BPD | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 3 | 1 | 0 |
| | $p=2$ | 0 | 0 | 0 | 1 | 91 | 4 | 0 |
| | $p=3$ | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MDL | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 1 | 1 | 5 |
| | $p=2$ | 0 | 0 | 0 | 0 | 56 | 21 | 15 |
| | $p=3$ | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AIC | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 1 | 4 | 6 |
| | $p=2$ | 0 | 0 | 0 | 0 | 50 | 19 | 27 |
| | $p=3$ | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

TABLE IV
PERFORMANCE COMPARISON OF THE THREE CRITERIA FOR EXAMPLE 2, SNR = 6 DB

| | | $k=0$ | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ | $k=6$ |
|---|---|---|---|---|---|---|---|---|
| BPD | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| | $p=2$ | 0 | 0 | 0 | 1 | 95 | 3 | 0 |
| | $p=3$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MDL | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | $p=2$ | 0 | 0 | 0 | 0 | 76 | 9 | 12 |
| | $p=3$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AIC | $p=0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=1$ | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| | $p=2$ | 0 | 0 | 0 | 0 | 70 | 11 | 17 |
| | $p=3$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $p=4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

TABLE V
PERFORMANCE COMPARISON OF THE THREE CRITERIA FOR EXAMPLE 3

| SNR (dB) | BPD | MDL | AIC |
|---|---|---|---|
| 0 | 68 | 32 | 63 |
| 1 | 75 | 52 | 67 |
| 2 | 87 | 67 | 72 |
| 3 | 94 | 73 | 75 |
| 4 | 98 | 81 | 78 |
| 5 | 99 | 84 | 82 |
| 6 | 100 | 87 | 83 |
| 7 | 100 | 88 | 81 |
| 8 | 100 | 91 | 85 |
| 9 | 100 | 93 | 87 |
| 10 | 100 | 95 | 86 |

criterion for model selection of multiple cisoids in additive AR noise process. To choose properly the noninformative priors of the nuisance parameters and proceed with marginalization, we transformed the observed data by employing subspace decomposition. With this decomposition and applying a uniform prior for the frequencies, we obtained a MAP estimator that coincides with the ML estimator derived under the same signal and noise models. The derived BPD criterion for model selection significantly outperforms the MDL and AIC from [15]. The improvement in performance is greater when the SNR is small and/or the AR noise is narrow-band.

To reduce the computational complexity of the proposed criteria, we have developed an algorithm that reduces the multivariate maximization problem to a simple multistage single-variate maximization. From the simulation results, we observe that the proposed MAP estimator when implemented by our algorithm provides excellent estimation performance without involving tedious computations.

## REFERENCES


[1] M. Aitkin, "Posterior Bayes factor," *J. R. Stat. Soc. B,* vol. 53, no. 1, pp. 111–142, 1991.

[2] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.,* vol AC-19, pp. 716–722, Dec. 1974.

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

[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, pp. 1081–1089, Oct. 1986.

[5] C. Chatterjee, R. L. Kashyap, and G. Boray, "Estimation of close sinusoids in colored noise and model discrimination," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-35, pp. 328–337, Mar. 1987.

[6] C. M. Cho, "Bayesian detection and estimation of superimposed signals by a subspace approach," Ph.D. dissertation, State Univ. New York, Stony Brook, 1993.

[7] P. M. Djurić and S. M. Kay, "Order selection of autoregressive models," *IEEE Trans. Signal Processing,* vol. 40, pp. 2829–2833, Nov. 1992.

[8] S. Geisser and W. F. Eddy, "A predictive approach to model selection," *J. Amer. Stat. Assoc.,* vol. 74, pp. 153–160, 1979.

[9] M. L. Honig and D. G. Messerschmitt, *Adaptive Filters: Structures, Algorithms, and Applications.* Boston, MA: Kluwer, 1984.

[10] Y. D. Huang and M. Barkat, "A dynamic programming algorithm for the maximum likelihood localization of multiple sources," *IEEE Trans. Antennas Propagat.,* vol. 40, no. 9, pp. 1023–1030, Sept. 1992.

[11] S. M. Kay, *Modern Spectral Estimation: Theory and Application.* Englewood Cliffs, NJ: Prentice Hall, 1988.

[12] S. Kay and V. Nagesha, "Spectral analysis based on the canonical autoregressive decomposition," in *Proc. IEEE ICASSP-91,* 1991, pp. 3137–3140.

[13] ——, "Broadband detection based on two-dimensional mixed autoregressive models," *IEEE Trans. Signal Processing,* 1993.

[14] J. P. Le Cadre, "Parametric methods for spatial signal processing in the presence of unknown colored noise fields," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 37, pp. 965–983, 1989.

[15] V. Nagesha and S. M. Kay, "Spectral analysis based on the canonical autoregressive decomposition," submitted to *IEEE Trans. Signal Processing*

[16] A. O'Hagan, "Comment on a paper by M. Aitkin," *J. R. Stat. Soc. B,* vol. 53, no. 1, p. 136, 1991.

[17] S. J. Press, *Bayesian Statistics Principles, Models, and Applications.* New York: Wiley, 1989.

[18] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Stat.,* vol. 11, no. 2, pp. 416–431, 1983.

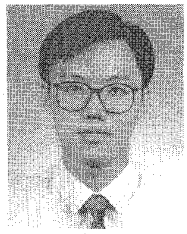
[19] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis.* Reading, MA: Addison-Wesley, 1991.

[20] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.,* vol. 6, no. 2, pp. 461–464, 1978.

[21] H. Tsuji and A. Sano, "Separable estimation of discrete and continuous spectra of signal with mixed spectrum," in *Proc. IEEE ICASSP-89,* vol. 5, 1989, pp. 2218–2221.

[22] S. F. Yau and Y. Bresler, "Maximum likelihood parameter estimation and subspace fitting of superimposed signals by dynamic programming—An approximate method," *Signal Processing,* vol. 29, pp. 283–298, 1992.

[23] S. F. Yau and Y. Bresler, "Maximum likelihood parameter estimation of superimposed signals by dynamic programming," *IEEE Trans. Signal Processing,* vol. 41, pp. 804–820, 1993.



**Chao-Ming Cho** (S'91–M'93) was born in Taichung, Taiwan, Republic of China, on January 26, 1965. He received the electrical engineering diploma from National Kaohsiung Institute of Technology, Taiwan, in 1985 and the M.S. and Ph.D. degrees in electrical engineering from the State University of New York at Stony Brook in 1989 and 1993, respectively.

He is currently a staff engineer at Microelectronic Technology Inc. (MTI), Taiwan. His research interests are in statistical signal processing, digital communications, image processing, and neural networks.

**Petar M. Djurić** (S'86–M'90) was born in Strumica, Yugoslavia, in 1957. 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, Kingston, 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, 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 of the American Statistical Association. Currently, he serves as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.