Detection and Estimation of DOA's of Signals via Bayesian Predictive Densities

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

***Abstract*—A new criterion based on Bayesian predictive densities and subspace decomposition is proposed for simultaneous detection of signals impinging on a sensor array and estimation of their direction-of-arrivals (DOA's). The solution is applicable for both coherent and noncoherent signals and an arbitrary array geometry. The proposed detection criterion is strongly consistent and outperforms the MDL and AIC criteria, particularly for a small number of sensors and/or snapshots, and/or low SNR, without increased computational complexity. When the prior of the direction-of-arrival is a uniform distribution, the Bayesian estimator for the directional parameters coincides with the unconditional maximum likelihood estimator. Simulation results that demonstrate the performance of the proposed solution are included.**

I. Introduction

In the area of array processing, the most popular approaches for the detection of the number of signals are based on the Akaike's information criterion (AIC) [1] and the minimum description length (MDL) principle [2], [3]. For noncoherent signals, the number of signals is determined from the "multiplicity" of the smallest eigenvalue of the sample covariance matrix without estimating the directional parameters [4]–[6]. When the signals are coherent, this approach is not applicable since the rank of the signal covariance matrix is reduced. Preprocessing techniques such as "spatial smoothing" [7] and "frequency smoothing" [8] provide a partial solution to this problem, however, their applicability is limited to uniform linear arrays and wide-band signals, respectively.

Recently, Wax and Ziskind [9], [10] have proposed a subspace decomposition approach for detection and estimation of coherent signals based on the AIC and MDL criteria and the maximum likelihood method. The solution is applicable for arbitrary array geometry. However, the AIC criterion suffers two drawbacks. It tends to asymptotically overestimate the number of signals, and its probability of error cannot reach zero even at a high signal-to-noise ratio (SNR). On the other hand, the MDL criterion is consistent, but it performs poorly at low SNR and/or a small number of snapshots. Unfortunately, most frequently it might be that the energy of the signals impinging on an array is low and the number of snapshots is limited. Under those circumstances, the MDL criterion usually underestimates the number of signals.

In our paper, a new criterion is proposed for simultaneous detection of coherent or noncoherent signals impinging on a sensor array, with arbitrary geometry, and estimation of their directional parameters. The approach is based on Bayesian predictive densities (BPD) [11] and a subspace unitary decomposition [9]–[11]. Within the framework of Bayesian inference, the predictive distribution of the observed data and the *a posteriori* distribution of the signal parameters of interest are found, and by maximizing these distributions, the number and the directional parameters of the signals are estimated. The proposed DOA estimator coincides with the unconditional maximum likelihood (ML) estimator when the priors of the *nuisance parameters*[1] are chosen by using Jeffreys' invariance theory [13]. The detector is strongly consistent in detecting the number of signals and is less sensitive to variations in the prior of the parameters. Furthermore, the proposed detection criterion outperforms the MDL and AIC criteria, especially in cases where the number of snapshots is small and/or the SNR is low. This improvement is achieved without increased computational complexity. The key to the improved performance is that the penalty term for the nuisance parameters is derived without using asymptotical assumptions.

The presentation is organized as follows. The problem is formulated in Section II. The Bayesian predictive density criterion for detection and a marginal maximum *a posteriori* (MAP) estimator are derived in Section III. In Section IV simulation results are included, and comparisons with the MDL and AIC criteria are made. Finally, the conclusion is given in Section V.

II. Problem Formulation

Consider that the far-field sources emit narrow-band wavefronts (signals) centered at a known frequency, ω_0, which impinge on the sensors in a planar manner. The number of superimposed signals is q, the number of sensors is N, where $q < N$. The locations and the directional characteristics of the sensors are allowed to be arbitrary. For simplicity, the sources and the sensors are assumed to be located in the same plane. Therefore, the only parameter that characterizes the location of the source is its direction-of-arrival.

Manuscript received November 9, 1992; revised February 1, 1994. This work was supported by the National Science Foundation under Award No. MIP-9110628. The associate editor coordinating the review of this paper and approving it for publication was Prof. Daniel Fuhrmann.

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

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

IEEE Log Number 9404768.

[1] In this paper, the nuisance parameters include the covariance matrix of the signal vectors and the variance of the noise only, but not the DOA parameters.

The observed data at the pth sensor and the ith snapshot are expressed by the complex envelope representation as

$$y_{i,p} = \sum_{l=1}^{q} s_{i,l} e^{j\omega_0 \tau_p(\theta_l)} + n_{i,p} \tag{1}$$

where $i = 1, 2, \ldots, M$, and $p = 1, 2, \ldots, N; \theta_l$ is the directional parameter (DOA) of the lth signal, assumed distinct from the other signals; $\tau_p(\theta_l)$ denotes the propagation delay between the reference point and the pth sensor to a wavefront impinging from direction θ_l. $s_{i,l}$ is the complex amplitude of the lth signal as received at the reference point, and M denotes the number of snapshots. The $n_{i,p}$ is the additive complex noise at the pth sensor.

The model for the ith snapshot can be compactly described by the following vector notations:

$$\mathbf{y}_i = \mathbf{D}(\boldsymbol{\theta}_{(q)})\mathbf{s}_i + \mathbf{n}_i, \quad i = 1, 2, \ldots, M \tag{2}$$

with

$$\mathbf{y}_i = [y_{i,1} y_{i,2} \cdots y_{i,N}]^T \tag{3.a}$$
$$\mathbf{D}(\boldsymbol{\theta}_{(q)}) = [\mathbf{d}(\theta_1)\mathbf{d}(\theta_2)\cdots\mathbf{d}(\theta_q)] \tag{3.b}$$
$$\mathbf{d}(\theta_l) = [e^{j\omega_0\tau_1(\theta_l)} e^{j\omega_0\tau_2(\theta_l)} \cdots e^{j\omega_0\tau_N(\theta_l)}]^T \tag{3.c}$$
$$\mathbf{s}_i = [s_{i,1} s_{i,2} \cdots s_{i,q}]^T \tag{3.d}$$
$$\mathbf{n}_i = [n_{i,1} n_{i,2} \cdots n_{i,N}]^T \tag{3.e}$$

where $\mathbf{D}(\boldsymbol{\theta}_{(q)})$ is an $N \times q$ matrix consisting of q steering vectors $\mathbf{d}(\theta_l), l = 1, \ldots, q$, and $\boldsymbol{\theta}_{(q)} = \{\theta_1, \theta_2, \ldots, \theta_q\}$. Any q distinct steering vectors from the array manifold are linearly independent. $\mathbf{y}_i$ is an $N \times 1$ observed data vector, $\mathbf{s}_i$ is a $q \times 1$ signal vector, and $\mathbf{n}_i$ is an $N \times 1$ noise vector. T denotes transpose operation.

We assume that the signal sample vectors $\mathbf{s}_i, i = 1, 2, \ldots, M$, are statistically independent zero mean complex Gaussian random vectors. The noise samples $n_{i,p}$ are zero mean complex Gaussian random variables uncorrelated across both i and p, with uncorrelated real and imaginary components, each with variance $\sigma_n^2/2$. Furthermore, they are assumed to be uncorrelated with the impinging signals. The covariance matrix of the observed data is then given by

$$\Sigma_{yy} = \mathbf{D}(\boldsymbol{\theta}_{(q)})\mathbf{R}_{ss}\mathbf{D}^H(\boldsymbol{\theta}_{(q)}) + \sigma_n^2\mathbf{I} \tag{4}$$

where $\mathbf{R}_{ss}$ is a $q \times q$ unknown signal covariance matrix. The signals may be uncorrelated (noncoherent), partially correlated, or fully correlated (coherent). When the signals are coherent, one signal might be a scaled and delayed version of the other, e.g., in multipath propagation. H denotes conjugation and transposition.

When the above assumptions hold, the problem can be stated as follows. Given the observed data samples, it is desired to simultaneously detect the number of the coherent and noncoherent signals and estimate their directional parameters (DOA's).

III. Proposed Solution

In the sequel, a method based on Bayesian theory is used to solve the above problem. Let $\mathcal{H}_k$ denote the hypothesis that the number of signals is k, and $k \in \{0, \ldots, N-1\}$. Under the hypothesis $\mathcal{H}_k$, the data are modeled by

$$\mathbf{y}_i = \mathbf{D}(\boldsymbol{\theta}_{(k)})\mathbf{s}_i + \mathbf{n}_i, \quad i = 1, 2, \ldots, M. \tag{5}$$

We adopt the MAP criterion [3], [14] to find the estimates of q and $\boldsymbol{\theta}$, denoted by $\hat{q}$ and $\hat{\boldsymbol{\theta}}$. They are determined from[2]

$$\hat{q} = \arg \min_{k \in \{0,\ldots,N-1\}} \{-\log f(\mathcal{H}_k \mid \mathbf{y})\} \tag{6}$$

and

$$\hat{\boldsymbol{\theta}}_{(k)} = \arg \min_{\boldsymbol{\theta} \in \Theta} \{f(\boldsymbol{\theta} \mid y, \mathcal{H}_k)\} \tag{7}$$

where $f(\mathcal{H}_k \mid \mathbf{y})$ is the posterior distribution of the hypothesis $\mathcal{H}_k, \Theta$ is the "field of view" (i.e., $-\pi < \theta_l \leq \pi, l = 1, 2, \ldots, q$), and $f(\boldsymbol{\theta} \mid \mathbf{y}, \mathcal{H}_k)$ is the posterior distribution of $\boldsymbol{\theta}$ under $\mathcal{H}_k$. From Bayes' rule, the posterior distribution of $\boldsymbol{\theta}$ is found according to

$$\begin{aligned} f(\boldsymbol{\theta} \mid \mathbf{y}, \mathcal{H}_k) &= \frac{f(\mathbf{y} \mid \boldsymbol{\theta}, \mathcal{H}_k)}{f(\mathbf{y} \mid \mathcal{H}_k)} f(\boldsymbol{\theta} \mid \mathcal{H}_k) \\ &= \int_{\phi} \frac{f(\mathbf{y} \mid \boldsymbol{\phi}, \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\phi} \mid \boldsymbol{\theta}, \mathcal{H}_k)}{f(\mathbf{y} \mid \mathcal{H}_k)} d\boldsymbol{\phi} \\ &\quad \times f(\boldsymbol{\theta} \mid \mathcal{H}_k) \end{aligned} \tag{8}$$

where $f(\mathbf{y} \mid \boldsymbol{\theta}, \mathcal{H}_k)$ is the marginal likelihood function (MLF) of $\boldsymbol{\theta}$ which is obtained from the marginalization of the likelihood function (LF), $f(\mathbf{y} \mid \boldsymbol{\phi}, \boldsymbol{\theta}, \mathcal{H}_k)$. The nuisance parameters $\boldsymbol{\phi}$ include the unknown covariance matrices of the signal vectors and the noise variance, $\mathbf{R}_{ss}$ and σ_n^2, respectively. $f(\boldsymbol{\theta} \mid \mathcal{H}_k)$ is the prior distribution of $\boldsymbol{\theta}$. $f(\boldsymbol{\phi} \mid \boldsymbol{\theta}, \mathcal{H}_k)$ is the conditional prior of $\boldsymbol{\phi}$ given $\boldsymbol{\theta}$. Commonly, it is assumed that the knowledge of $\boldsymbol{\theta}$ is not related to the knowledge of $\boldsymbol{\phi}$. This implies $f(\boldsymbol{\phi} \mid \boldsymbol{\theta}, \mathcal{H}_k) = f(\boldsymbol{\phi} \mid \mathcal{H}_k)$. $f(\mathbf{y} \mid \mathcal{H}_k)$ is the marginal distribution of $\mathbf{y}$ under the hypothesis $\mathcal{H}_k$. For given $\mathbf{y}, f(\mathbf{y} \mid \mathcal{H}_k)$ is a constant, and if the prior $f(\boldsymbol{\theta} \mid \mathcal{H}_k)$ is locally uniform, the criterion (7) amounts to maximization of the MLF $f(\mathbf{y} \mid \boldsymbol{\theta}, \mathcal{H}_k)$.

Similarly, the posterior distribution of $\mathcal{H}_k$ in (6) can be obtained from

$$\begin{aligned} f(\mathcal{H}_k \mid \mathbf{y}) &= \frac{f(\mathbf{y} \mid \mathcal{H}_k)}{f(\mathbf{y})} f(\mathcal{H}_k) \\ &= \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\phi}} \frac{f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{H}_k)}{f(\mathbf{y})} f(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathcal{H}_k) d\boldsymbol{\theta}\, d\boldsymbol{\phi} \\ &\quad \times f(\mathcal{H}_k) \end{aligned} \tag{9}$$

where $f(\mathbf{y} \mid \mathcal{H}_k)$ is the MLF of $\mathcal{H}_k$, $f(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathcal{H}_k)$ is the prior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ under the hypothesis $\mathcal{H}_k$, and $f(\mathcal{H}_k)$ is the prior of the hypothesis $\mathcal{H}_k$. When the probabilities of all hypotheses $\mathcal{H}_k$ are equal, the criterion (6) amounts to maximization of the MLF $f(\mathbf{y} \mid \mathcal{H}_k)$.

[2] It should be noted that, from a Bayesian point of view, this is not strictly simultaneous detection and estimation. However, we shall see that $\hat{q}$ and $\hat{\boldsymbol{\theta}}$ may be obtained from one equation only.

The prior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ have to be chosen carefully. Unless the priors are supported by satisfactory physical or logical arguments, we prefer to use noninformative priors because they reflect our ignorance about the parameters. However, such priors are usually *improper*[3] and therefore proportional to arbitrary constants [20], [21]. Although they are convenient for representation of vague prior knowledge in the estimation problem, because of the arbitrary constants they are inappropriate for the use in the detection problem [17], [19].

We circumvent this deficiency by using a Bayesian predictive density (BPD) criterion [11]. The BPD criterion for estimating q is given by

$$\hat{q} = \arg \min_{k \in \{0,1,\ldots,N-1\}} \{-\log f(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_1, \mathcal{H}_k)\} \tag{10}$$

where $\boldsymbol{\xi}_1 = \{\mathbf{y}_1, \ldots, \mathbf{y}_L\}$, $\boldsymbol{\xi}_2 = \{\mathbf{y}_{L+1}, \ldots, \mathbf{y}_M\}$, and $1 < L < M$. The selection of L will be discussed later. The function $f(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_1, \mathcal{H}_k)$ is called a Bayesian predictive density of $\boldsymbol{\xi}_2$ according to $\boldsymbol{\xi}_1$ and the hypothesis $\mathcal{H}_k$.

Note that the original criterion (6) under the current assumptions can be rewritten as

$$\hat{q} = \arg \min_{k \in \{0,1,\ldots,N-1\}} \{-\log f(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_1, \mathcal{H}_k) - \log f(\boldsymbol{\xi}_1 \mid \mathcal{H}_k)\}. \tag{11}$$

The difference between (10) and (11) is the absence of $-\log f(\boldsymbol{\xi}_1 \mid \mathcal{H}_k)$ in (10). The reason for the absence is that the data $\boldsymbol{\xi}_1$ are used for finding proper density functions of the model parameters (recall that the model selection can be carried out only if proper density functions for the model parameters are used). Thus, by using (10) for model comparison, we lose the information that might have been gained from $\boldsymbol{\xi}_1$. Instead, we exploit $\boldsymbol{\xi}_1$ to obtain information about the model parameters and, thus, make the Bayesian procedure insensitive to the uncertainties in the parameter priors.

Using the Bayes' rule, the BPD function can be expressed as

$$\begin{aligned} f(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_1, \mathcal{H}_k) &= \int_{\boldsymbol{\theta}} f(\boldsymbol{\xi}_2 \mid \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\theta} \mid \boldsymbol{\xi}_1, \mathcal{H}_k) d\boldsymbol{\theta} \qquad (12) \\ &= \frac{\int_{\boldsymbol{\theta}} f(\mathbf{y}_{(M)} \mid \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\theta} \mid \mathcal{H}_k) d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} f(\mathbf{y}_{(L)} \mid \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\theta} \mid \mathcal{H}_k) d\boldsymbol{\theta}} \qquad (13) \\ &= \frac{\int_{\boldsymbol{\theta}} \int_{\boldsymbol{\phi}} f(\mathbf{y}_{(M)} \mid \boldsymbol{\phi}, \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\phi} \mid \mathcal{H}_k) f(\boldsymbol{\theta} \mid \mathcal{H}_k) d\boldsymbol{\phi}\, d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} \int_{\boldsymbol{\phi}} f(\mathbf{y}_{(L)} \mid \boldsymbol{\phi}, \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\phi} \mid \mathcal{H}_k) f(\boldsymbol{\theta} \mid \mathcal{H}_k) d\boldsymbol{\phi}\, d\boldsymbol{\theta}} \qquad (14) \end{aligned}$$

where $\mathbf{y}_{(L)} = \{\boldsymbol{\xi}_1\}$ and $\mathbf{y}_{(M)} = \{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2\}$. The function $f(\boldsymbol{\theta} \mid \boldsymbol{\xi}_1, \mathcal{H}_k)$ is the prior distribution of $\boldsymbol{\theta}$ conditioned on the data $\boldsymbol{\xi}_1$ and the model $\mathcal{H}_k$. As before, $\boldsymbol{\phi} = \{\mathbf{R}_{ss}, \sigma_n\}$ is the nuisance parameter vector. From (12)–(14) we observe that the use of improper priors will not cause problems any more, because in obtaining the Bayesian predictive densities the arbitrary constants cancel out [11], [17], [18].

[3] Improper priors are *not* regular probability densities. They do not integrate to one.

In order to deal with the marginalizations of the nuisance parameters in (8) and (14), the observed data space will be split into the two complementary subspaces [10], [12], [15], [16]. The subspace spanned by the columns of the matrix $\mathbf{D}(\boldsymbol{\theta}_{(k)})$ is referred to as the *signal subspace*, and the orthogonal space to the signal subspace is referred to as the *noise subspace*. According to this decomposition, the observed data vector $\mathbf{y}$ is then split into two subspace vectors by

$$\mathbf{y} = \mathbf{G}(\boldsymbol{\theta}_{(k)}) \begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_n \end{bmatrix} \tag{15}$$

where $\mathbf{x}_s$ denotes the $k \times 1$ signal subspace vector, and $\mathbf{x}_n$ denotes the $(N-k) \times 1$ noise subspace vector. $\mathbf{G}(\boldsymbol{\theta}_{(k)})$ denotes an $N \times N$ unitary coordinate transformation matrix which is given by

$$\mathbf{G}(\boldsymbol{\theta}_{(k)}) = [\mathbf{U}_s(\boldsymbol{\theta}_{(k)}) \vdots \mathbf{U}_n(\boldsymbol{\theta}_{(k)})]. \tag{16}$$

It satisfies the following identities:

$$\begin{aligned} \mathbf{P}(\boldsymbol{\theta}_{(k)}) &= \mathbf{D}(\boldsymbol{\theta}_{(k)})(\mathbf{D}^H(\boldsymbol{\theta}_{(k)})\mathbf{D}(\boldsymbol{\theta}_{(k)}))^{-1}\mathbf{D}^H(\boldsymbol{\theta}_{(k)}) \\ &= \mathbf{U}_s(\boldsymbol{\theta}_{(k)})\mathbf{U}_s^H(\boldsymbol{\theta}_{(k)}) \end{aligned} \tag{17}$$

and

$$\mathbf{P}^{\perp}(\boldsymbol{\theta}_{(k)}) = \mathbf{I} - \mathbf{P}(\boldsymbol{\theta}_{(k)}) = \mathbf{U}_n(\boldsymbol{\theta}_{(k)})\mathbf{U}_n^H(\boldsymbol{\theta}_{(k)}). \tag{18}$$

$\mathbf{U}_s(\boldsymbol{\theta}_{(k)})$ and $\mathbf{U}_n(\boldsymbol{\theta}_{(k)})$ are $N \times k$ and $N \times (N-k)$ matrices, respectively, whose columns represent sets of orthonormal vectors that span the signal and noise subspaces, respectively. The matrices $\mathbf{P}(\boldsymbol{\theta}_{(k)})$ and $\mathbf{P}^{\perp}(\boldsymbol{\theta}_{(k)})$ are two complementary projection matrices which project onto the signal and noise subspaces, respectively. To simplify the notation, we use $\mathbf{P}_{(k)}$ and $\mathbf{P}_{(k)}^{\perp}$ instead of $\mathbf{P}(\boldsymbol{\theta}_{(k)})$ and $\mathbf{P}^{\perp}(\boldsymbol{\theta}_{(k)})$.

Since the transformation (15) is linear, the two subspaces are orthogonal, and the noise is white, the signal and noise subspace vectors are independent zero mean Gaussian. The nuisance parameter $\boldsymbol{\phi}$ are also split into two nuisance parameters sets, $\boldsymbol{\phi}_s$ and $\boldsymbol{\phi}_n$, that belong to the complementary subspaces. The parameter set $\boldsymbol{\phi}_s$ denotes the nuisance parameters of the signal subspace vector. Similarly, the set $\boldsymbol{\phi}_n$ represents the nuisance parameters of the noise subspace components. Therefore, the numerator of (14) can be modified using

$$\begin{aligned} &\int_{\boldsymbol{\phi}} f(\mathbf{y}_{(M)} \mid \boldsymbol{\phi}, \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\phi} \mid \boldsymbol{\theta}, \mathcal{H}_k) d\boldsymbol{\phi} \\ &\quad = \int_{\boldsymbol{\phi}_s} \int_{\boldsymbol{\phi}_n} \frac{1}{\mathcal{J}(\mathbf{x}_s, \mathbf{x}_n; \mathbf{y})} f(\mathbf{x}_{s,(M)} \mid \boldsymbol{\phi}_s, \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\phi}_s \mid \boldsymbol{\theta}, \mathcal{H}_k) \\ &\qquad \times f(\mathbf{x}_{n,(M)} \mid \boldsymbol{\phi}_n, \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\phi}_n \mid \boldsymbol{\theta}, \mathcal{H}_k) d\boldsymbol{\phi}_s d\boldsymbol{\phi}_n \\ &\quad = f(\mathbf{x}_{s,(M)} \mid \boldsymbol{\theta}, \mathcal{H}_k) f(\mathbf{x}_{n,(M)} \mid \boldsymbol{\theta}, \mathcal{H}_k) \end{aligned} \tag{19}$$

where $\mathcal{J}(\mathbf{x}_s, \mathbf{z}_n; \mathbf{y})$ denotes the Jacobian of the transformation (15), and it is equal to 1. The denominator of (14) is manipulated in the same way.

From the assumptions in Section II and (15), it follows that the unknown signal subspace vector $\mathbf{x}_s$ is modeled as a complex Gaussian process with zero mean and an unknown covariance matrix $\mathbf{\Sigma}_{ss}$. The distribution of $\mathbf{x}_{s,(M)}$ is given by

$$f\left(\mathbf{x}_{s,(M)} \mid \mathbf{\Sigma}_{ss}^{-1}, \boldsymbol{\theta}, \mathcal{H}_k\right) = \left(\frac{1}{\pi}\right)^{kM} \left[\det\left(\mathbf{\Sigma}_{ss}^{-1}\right)\right]^{M} \times \exp\left\{-\sum_{i=1}^{M} \mathbf{x}_{s,i}^H \mathbf{\Sigma}_{ss}^{-1} \mathbf{x}_{s,i}\right\} \tag{20}$$

where $\det(\cdot)$ denotes determinant. For the white noise model, $\mathbf{x}_n$ is an $(N-k) \times 1$ complex Gaussian random vector with zero mean and a covariance matrix $\mathbf{\Sigma}_{nn} = \sigma_n^2 \mathbf{I}$, i.e.,

$$f(\mathbf{x}_{n,(M)} \mid \sigma_n, \boldsymbol{\theta}, \mathcal{H}_k) = \frac{1}{(\pi\sigma_n^2)^{(N-k)M}} \times \exp\left\{-\frac{1}{\sigma_n^2} \sum_{i=1}^{M} \mathbf{x}_{n,i}^H \mathbf{x}_{n,i}\right\}. \tag{21}$$

Assume that the priors of $\boldsymbol{\phi}_s (= \mathbf{\Sigma}_{ss})$ and $\boldsymbol{\phi}_n (= \sigma_n)$ are independent of $\boldsymbol{\theta}$. Since information about $\mathbf{\Sigma}_{ss}$ and σ_n is not available, we will choose noninformative prior distributions for $\mathbf{\Sigma}_{ss}$ and σ_n by using Jeffreys' invariance theory [20]. According to this theory, the Jeffreys' noninformative priors are derived by *requiring* invariance of inference under parameter transformation, which entails that the noninformative prior distribution for a set of parameters is proportional to the square root of the determinant of the Fisher's information matrix. It can be shown that for $\mathbf{\Sigma}_{ss}$ and σ_n their Jeffreys' priors are given by [23]

$$f(\mathbf{\Sigma}_{ss} \mid \mathcal{H}_k) \propto \left[\det(\mathbf{\Sigma}_{ss})\right]^{-k} \tag{22}$$

$$f(\sigma_n \mid \mathcal{H}_k) \propto \frac{1}{\sigma_n}. \tag{23}$$

After substituting (20) and (22) into the signal subspace integral of (19), the integration can be carried out by changing variables of the complex Wishart distribution [24]. As shown in Appendix A, it results in

$$f(\mathbf{x}_{s,(M)} \mid \boldsymbol{\theta}, \mathcal{H}_k) = M^{-Mk} \left[\det\left(\hat{\mathbf{\Sigma}}_{ss,(M)}\right)\right]^{-M} \times \left(\frac{1}{n}\right)^{kM} \pi^{k(k-1)/2} \prod_{i=0}^{k-1} \Gamma[M-l] \tag{24}$$

where

$$\hat{\mathbf{\Sigma}}_{ss,(M)} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{x}_{s,i} \mathbf{x}_{s,i}^H \tag{25}$$

and $\Gamma[\cdot]$ is the gamma function. Similarly, the MLF of the noise subspace vectors is given by

$$f(\mathbf{x}_{n,(M)} \mid \boldsymbol{\theta}, \mathcal{H}_k) = \left(\frac{1}{\pi}\right)^{(N-k)M} \frac{1}{2}\Gamma[M(N-k)] \times M^{-M(N-k)} \cdot \left[\operatorname{tr}\left(\hat{\mathbf{\Sigma}}_{nn,(M)}\right)\right]^{-M(N-k)} \tag{26}$$

where tr denotes trace of a matrix, and

$$\hat{\mathbf{\Sigma}}_{nn,(M)} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{x}_{n,i} \mathbf{x}_{n,i}^H. \tag{27}$$

The product of (24) and (26) yields $f(\mathbf{y}_{(M)} \mid \boldsymbol{\theta}, \mathcal{H}_k)$, that is

$$f\left(\mathbf{y}_{(M)} \mid \boldsymbol{\theta}, \mathcal{H}_k\right) = \frac{1}{2}\left[C_{(M)}\left(\boldsymbol{\theta}_{(k)}\right)\right]^{-M} \left(\frac{1}{M\pi}\right)^{MN} \times (N-k)^{-M(N-k)} \pi^{k(k-1)/2} \times \Gamma[M(N-k)] \prod_{l=0}^{k-1} \Gamma[M-l] \tag{28}$$

where

$$C_{(M)}\left(\boldsymbol{\theta}_{(k)}\right) = \left(\det\left(\hat{\mathbf{\Sigma}}_{ss,(M)}\right)\right) \times \left(\frac{1}{(N-k)} \operatorname{tr}\left(\hat{\mathbf{\Sigma}}_{nn,(M)}\right)\right)^{(N-k)}. \tag{29}$$

The data term $C_{(M)}(\boldsymbol{\theta}_{(k)})$ can be directly obtained from the observed data samples. As shown in [10], we get

$$C_{(M)}\left(\boldsymbol{\theta}_{(k)}\right) = \det\left(\mathbf{P}_{(k)} \hat{\mathbf{\Sigma}}_{yy,(M)} \mathbf{P}_{(k)}\right) + \frac{1}{(N-k)} \operatorname{tr}\left(\mathbf{P}_{(k)}^{\perp} \hat{\mathbf{\Sigma}}_{yy,(M)}\right) \mathbf{P}_{(k)}^{\perp}\right) \tag{30}$$

where

$$\hat{\mathbf{\Sigma}}_{yy,(M)} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{y}_i \mathbf{y}_i^H. \tag{31}$$

Furthermore, $C_{(M)}(\boldsymbol{\theta}_{(k)})$ can also be computed in terms of the eigenvalues of the matrices involved. Using the well-known invariance properties of the unitary transformation, it becomes

$$C_{(M)}(\boldsymbol{\theta}_{(k)}) = \left(\prod_{i=1}^{k} \lambda_i^{(s)}(\boldsymbol{\theta}_{(k)})\right) \times \left(\frac{1}{(N-k)} \sum_{i=1}^{N-k} \lambda_i^{(n)}(\boldsymbol{\theta}_{(k)})\right)^{(N-k)} \tag{32}$$

where the $\lambda_i^{(s)}(\boldsymbol{\theta}_{(k)})$'s are the k nonzero eigenvalues of the rank-k matrix $\mathbf{P}_{(k)} \hat{\mathbf{\Sigma}}_{yy} \mathbf{P}_{(k)}$ and the $\lambda_i^{(n)}(\boldsymbol{\theta}_{(k)})$'s are the $(N-k)$ nonzero eigenvalues of the rank-$(N-k)$ matrix $\mathbf{P}_{(k)}^{\perp} \hat{\mathbf{\Sigma}}_{yy} \mathbf{P}_{(k)}^{\perp}$.

To find the MAP estimator of $\boldsymbol{\theta}$, we need to choose a noninformative prior for $\boldsymbol{\theta}$. We shall adopt $f(\boldsymbol{\theta}) \propto$ const. [21], [22]. This is not the Jeffreys' prior, but it is mathematically tractable and justifiable for large number of snapshots (see also Appendix B). Then the Bayesian MAP estimator of $\boldsymbol{\theta}$ is obtained by minimizing $C_{(M)}(\boldsymbol{\theta}_{(k)})$ with respect to $\boldsymbol{\theta}$ according to (8) and (28). This solution coincides with the ML estimator derived by Böhme [25], Jaffer [26], and Wax [10]. As expected, the Bayesian estimator yields the same result as the so-called unconditional ML estimator [13]. This estimator has been shown to be asymptotically unbiased (consistent) and statistically efficient, i.e., the estimation error covariance attains the Cramér-Rao bound asymptotically [27], [28].

Now we continue with the derivation of the BPD function (12) for the detection problem. To carry out the marginalization in (14), we need the form $f(\mathbf{y}_{(L)} \mid \boldsymbol{\theta}, \mathcal{H}_k)$. From (28) we deduce that

$$f(\mathbf{y}_{(L)} \mid \boldsymbol{\theta}, \mathcal{H}_k) = \frac{1}{2}\left[C_{(L)}(\boldsymbol{\theta}_{(k)})\right]^{-L}\left(\frac{1}{L\pi}\right)^{LN} \times (N-k)^{-L(N-k)}\pi^{k(k-1)/2}\Gamma \times [L(N-k)]\prod_{l=0}^{k-1}\Gamma[L-l]. \quad (33)$$

Thus, substituting (28) and (33) into (13), the distribution $f(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_1, \boldsymbol{\theta}, \mathcal{H}_k)$ is given by

$$f(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_1, \boldsymbol{\theta}, \mathcal{H}_k) = \gamma(M, L, N, k)\frac{\left[C_{(L)}(\boldsymbol{\theta}_{(k)})\right]^L}{\left[C_{(M)}(\boldsymbol{\theta}_{(k)})\right]^M} \quad (34)$$

where $\gamma(M, L, N, k)$ is a function of M, L, N, and k only. Note that for the BPD criterion we need to determine the function $f(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_1, \mathcal{H}_k)$ from (14). Since the equation (34) can not be integrated with respect to $\boldsymbol{\theta}$ analytically, a maximum likelihood approximation method will be used [17], [20]. The derivation is shown in Appendix B and the result is expressed in a logarithmic form as

$$\begin{aligned} -\log f(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_1, \mathcal{H}_k) &= (M-L)\cdot\log C_{(M)}(\hat{\boldsymbol{\theta}}_{(k)}) \\ &\quad + (M-L)(N-k)\log(N-k) + \log\frac{\Gamma[L(N-k)]}{\Gamma[M(N-k)]} \\ &\quad + \sum_{l=0}^{k-1}\log\frac{\Gamma[L-l]}{\Gamma[M-l]} + MN\log M\pi \\ &\quad - LN\log L\pi + \frac{k}{2}\log\frac{M}{L} \end{aligned} \quad (35)$$

for $1 \le k < N$. When $k = 0$

$$\begin{aligned} -\log f(\boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_1, \mathcal{H}_0) &= (M-L)\cdot\log\left(\frac{1}{N}\mathrm{tr}\hat{\boldsymbol{\Sigma}}_{yy,(M)}\right)^N \\ &\quad + (M-N+1)N\log N \\ &\quad + \log\frac{\Gamma[(N-1)N]}{\Gamma[MN]} \\ &\quad + MN\log M\pi - LN\log L\pi. \end{aligned} \quad (36)$$

Note that L has to be greater or equal to the dimension of the signal subspace vector to satisfy the minimum number of degrees of freedom in a complex Wishart distribution [16]. Furthermore, we know that L should be selected as small as possible to reduce the overall information loss since we only use $M - L$ snapshots of data for model comparison [11], [18]. Therefore, we choose $L = N - 1$ to allow for the maximum possible q. The final BPD criterion is thus expressed as

$$\hat{q}, \hat{\theta} = \arg\min_{k\in\{0,1,\ldots,N-1\},\theta\in\Theta}\{M\log C_{(M)}(\boldsymbol{\theta}_{(k)}) + T(k)\} \quad (37)$$

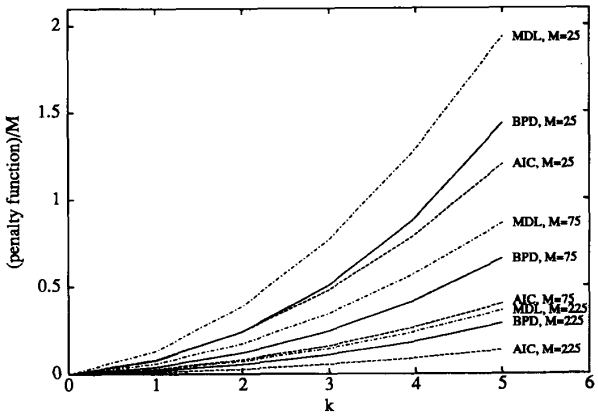


Fig. 1. Comparison of the penalty functions of the BPD, AIC, and MDL criteria for $N = 6$.

where $T(k)$ is a penalty function given by

$$\begin{aligned} T(k) &= MN\log\left(1-\frac{k}{N}\right) - Mk\log(N-k) \\ &\quad + \frac{M}{M-N+1}\left\{\log\frac{\Gamma[(N-1)(N-k)]\Gamma[MN]}{\Gamma[M(N-k)]\Gamma[N(N-1)]}\right. \\ &\quad \left. + \sum_{l=0}^{k-1}\log\frac{\Gamma[N-l-1]}{\Gamma[M-l]} + \frac{k}{2}\log\frac{M}{(N-1)}\right\}. \end{aligned} \quad (38)$$

Note that we have calibrated the penalty function so that there is no penalty when $k = 0$, i.e., $T(0) = 0$.

Under the same assumptions in Section II, the MDL and AIC criteria are [10]

$$\mathrm{MDL}(k) = \min_k\{M\log C_{(M)}(\hat{\boldsymbol{\theta}}_{(k)}) + \frac{1}{2}k(k+1)\log M\} \quad (39)$$

and

$$\mathrm{AIC}(k) = \min_k\{M\cdot\log C_{(M)}(\hat{\boldsymbol{\theta}}_{(k)}) + k(k+1)\}. \quad (40)$$

Clearly, the BPD criterion has the same data term, but a different penalty function.

Unlike the AIC criterion, the BPD criterion is strongly consistent, such that $\hat{q} \to q$ as $M \to \infty$ with probability one (see Appendix C). When compared to the rate of change of the MDL criterion's penalty function with model order, that of the BPD criterion is smaller, and is also a function of the number of sensors. We know that the MDL criterion performs very well for large M. However, it usually underestimates the number of signals in cases of low SNR and/or small M due to its overpenalization. When M is small, the relative penalty function of the BPD criterion comes close to that of the AIC. When M increases, it approaches to the MDL's penalty function. Indeed, as shown in Appendix C, the BPD and MDL criteria become equivalent when $M \to \infty$. To illustrate these relations, in Fig. 1 we compare the penalty functions $(T(k)/M)$ of the three criteria for the case when $N = 6$, and M is equal to 25, 75, and 225.

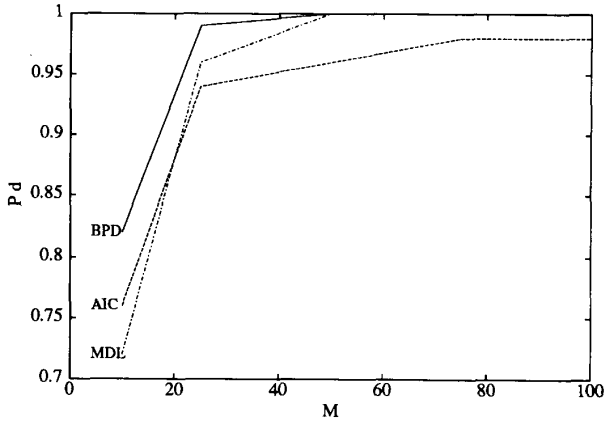


Fig. 2. Comparison of detection probabilities in terms of the number of the snapshots. Two equal power (SNR = 0 dB) coherent signals with 90° phase difference, located at 15° and 20°, impinge on a linear array with six sensors ($N = 6$).

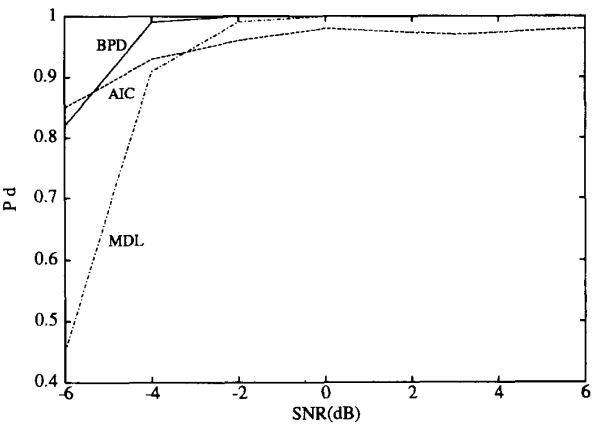


Fig. 4. Comparison of detection probabilities in terms of the SNR. Two equal power coherent signals with 90° phase difference, located at 15° and 20°, impinge on a linear array with six sensors ($N = 6$). The number of snapshots is 50.

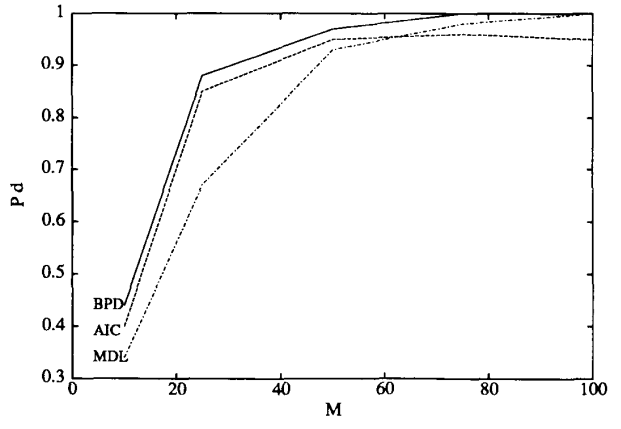


Fig. 3. Comparison of detection probabilities in terms of the number of snapshots. Two equal power (SNR = 0 dB) uncorrelated signals, located at 15° and 20°, impinge on a linear array with six sensors ($N = 6$).

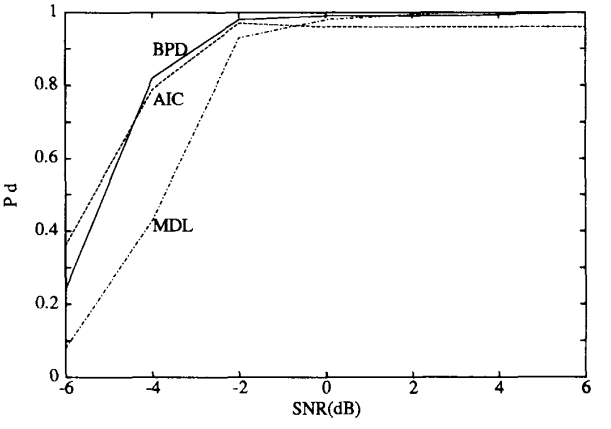


Fig. 5. Comparison of detection probabilities in terms of the SNR. Two equal power uncorrelated signals, located at 15° and 20°, impinge on a linear array with six sensors ($N = 6$). The number of snapshots is 50.

The three criteria have the same data term and their penalty functions can be precalculated. Hence, their computational complexities are the same. The computation of the DOA estimator based on (37) is complicated since a nonlinear and multimodal k-dimensional maximization has to be implemented. In order to efficiently solve this problem, reduce the computation load, and improve the convergence in the optimal search, we may use the alternating maximization technique [10], [29] or other numerical approaches [30].

IV. Simulation Results

To examine the performance of the BPD criterion, eight simulation experiments were performed, each with 100 Monte Carlo runs. The detection performance was obtained by counting the number of correctly estimated q in 100 runs. The experiments compared the detection performance of the BPD with the AIC and MDL criteria proposed by Wax in [10] ((40) and (39)) for two cases, coherent and noncoherent signals. Each case was examined in terms of M, SNR, and N. The alternating maximization technique was used for estimating the DOA parameters.

Figs. 2 and 3 show the comparisons of the detection performance in terms of the number of snapshots for two equal power coherent and noncoherent signals, respectively. We observe that the BPD criterion outperforms the AIC and MDL criteria in both experiments, especially when M is small and the signals are uncorrelated. When M is large enough, as predicted by our analysis, the performances of the BPD and MDL criteria are the same. Next, we compared the performance of the BPD, AIC, and MDL criteria for various SNR's. The results are shown in Figs. 4 and 5. We observe that the BPD outperforms the MDL criterion when the SNR is small. Although the AIC criterion yields better performance when the SNR is very small, it is inconsistent as the SNR increases. Moreover, we observe that the criterion

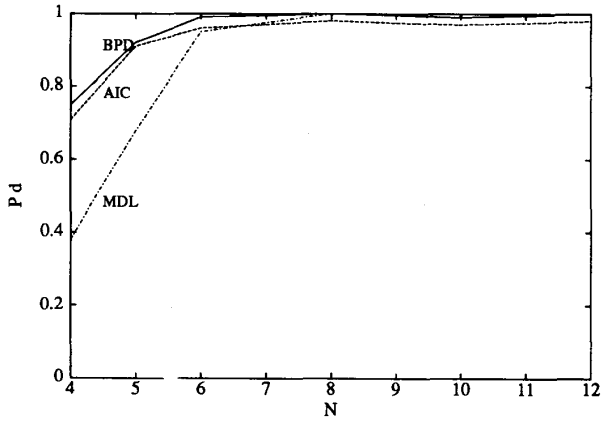


Fig. 6. Detection performance in terms of the number of sensors. Two equal power coherent signals with 90° phase difference, located at 10° and 20°,impinge on a linear array with N sensors. The SNR is -3 dB and the number of snapshots is 50.

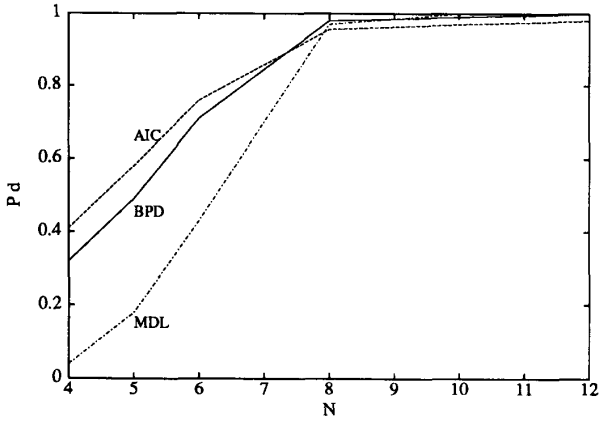


Fig. 7. Detection performance in terms of the number of sensors. Two equal power uncorrelated signals, located at 10° and 20°, impinge on a linear array with N sensors. The SNR is -3 dB and the number of snapshots is 50.

performance for low SNR is better for coherent than for noncoherent signals.

There is another issue of interest. It is the performance of the detection criteria as functions of the number of sensors. Figs. 6 and 7 show the performance in terms of N when $M = 50$ and the SNR $= -3$ dB.

Finally, we investigated the case of three signals. The numerical results that present the comparison among the criteria are given in Table I(a)-(d).

The computer simulation results show that the gain in the detection performance of the BPD approach is larger in cases of small M, SNR, and N. In addition, the gain is greater when the signals are noncoherent.

Since the Bayesian estimator (29) coincides with the unconditional ML estimator, its estimation performance is not examined here. It has been shown that this estimator outperforms the conditional ML estimator when the signals are uncorrelated or fully correlated.

TABLE I
COMPARISON RESULTS FOR THE CASES OF THREE EQUAL POWER SIGNALS LOCATED AT 15°, 20° AND 30°. PHASES OF THE SIGNALS ARE 0°, 45°, AND 90°, RESPECTIVELY. $M = 50$, AND $N = 6$. (a) COHERENT SIGNAL CASE, SNR = 3 dB; (b) NONCOHERENT SIGNAL CASE, SNR = 3 dB; (c) COHERENT SIGNAL CASE, SNR = -3 dB; (d) NONCOHERENT SIGNAL CASE, SNR = -3 dB.

k	0	1	2	3	4	5
BPD	0	0	0	100	0	0
AIC	0	0	0	96	3	1
MDL	0	0	0	100	0	0

(a)

k	0	1	2	3	4	5
BPD	0	0	1	99	0	0
AIC	0	0	1	94	4	1
MDL	0	0	1	99	0	0

(b)

k	0	1	2	3	4	5
BPD	0	0	1	96	3	0
AIC	0	0	1	91	6	2
MDL	0	3	10	87	0	0

(c)

k	0	1	2	3	4	5
BPD	0	0	3	85	11	1
AIC	0	0	1	83	14	2
MDL	0	8	49	42	1	0

(d)

V. CONCLUSION

In this paper, we have applied Bayesian inference techniques to detection and estimation of coherent and noncoherent signals. The solutions are obtained by maximization of the posterior distributions of $\mathcal{H}_k$ and $\boldsymbol{\theta}_{(k)}$. When compared to the AIC and MDL criteria derived under the same signal model, the BPD criterion has the same data term (the DOA estimator), but a different penalty term. In the derivations of the AIC and MDL criteria the asymptotical assumption (the maximum likehood approximation) is used for *all the free parameters*. Therefore, their penalty functions are less accurate when the total number of free parameters involved is large relative to the sample size. In contrast, the BPD criterion is derived using the likelihood approximation *only* for a subset of the unknown parameters, i.e., $\boldsymbol{\theta}$. The penalization for the nuisance parameters obtained from the marginalization is more accurate. This entails a remarkable property of our criterion, that is, the BPD preserves the good performance of the AIC and MDL criteria for small and large number of data snapshots, respectively. As expected, the improved detection performance is more emphasized for small M, N, and low SNR. Furthermore, we have shown that the BPD and MDL criteria are asymptotically equivalent. Unlike the AIC criterion, they are strongly consistent for estimating the number of signals.

APPENDIX A

In this appendix, we derive the MLF's (24), (26). The MLF of the M signal subspace vectors in (19) is rewritten as

$$\begin{aligned} f(\mathbf{x}_{s,(M)} \mid \boldsymbol{\theta}, \mathcal{H}_k) &= \int_{\boldsymbol{\Sigma}_{ss}^{-1}} f\left(\mathbf{x}_{s,(M)} \mid \boldsymbol{\Sigma}_{ss}^{-1}, \boldsymbol{\theta}, \mathcal{H}_k\right) f\left(\boldsymbol{\Sigma}_{ss}^{-1} \mid \mathcal{H}_k\right) d\boldsymbol{\Sigma}_{ss}^{-1} \\ &= \int_{\boldsymbol{\Sigma}_{ss}^{-1}} \left(\frac{1}{\pi}\right)^{kM} \left[\det\left(\boldsymbol{\Sigma}_{ss}^{-1}\right)\right]^{M} \\ &\quad \times \exp\left\{ -\operatorname{tr}\left(M\hat{\boldsymbol{\Sigma}}_{ss,(M)}\boldsymbol{\Sigma}_{ss}^{-1}\right)\right\} \\ &\quad \times \left[\det\left(\boldsymbol{\Sigma}_{ss}^{-1}\right)\right]^{-k} d\boldsymbol{\Sigma}_{ss}^{-1}. \end{aligned} \tag{A-1}$$

If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ are M complex sample vectors from a k-variate zero mean complex Gaussian distribution, then the joint distribution of the distinct elements of the matrix $\mathbf{A}$ with

$$\mathbf{A} = \sum_{i=1}^{M} \mathbf{x}_i \mathbf{x}_i^H = M \cdot \hat{\boldsymbol{\Sigma}}_x \tag{A-2}$$

is called a complex Wishart distribution [24]. It is defined by

$$f_w(\mathbf{A}) = \{[\det(\mathbf{A})]^{M-k}/\mathbf{I}_w(\boldsymbol{\Sigma}_x)\} \exp\{-\operatorname{tr}(\boldsymbol{\Sigma}_x^{-1}\mathbf{A})\} \tag{A-3}$$

where $\boldsymbol{\Sigma}_x$ is the covariance matrix of $\mathbf{x}_i$, and

$$\mathbf{I}_w(\boldsymbol{\Sigma}_x) = \pi^{k(k-1)/2}[\det(\boldsymbol{\Sigma}_x)]^M \prod_{l=0}^{k-1} \Gamma(M-l) \cdot \tag{A-4}$$

Let $\mathbf{A} = \boldsymbol{\Sigma}_{ss}^{-1}$ and $\boldsymbol{\Sigma}_{ss}^{-1} = M \cdot \hat{\boldsymbol{\Sigma}}_{ss,(M)}$ in (A-1). Since the integration of (A-3) w.r.t. $\mathbf{A}$ equals 1, the MLF of the signal subspace vectors results in (24).

Next, the MLF of the noise subspace vectors is obtained from

$$\begin{aligned} f(\mathbf{x}_{n,(M)} \mid \boldsymbol{\theta}, \mathcal{H}_k) &= \int_{\sigma_n} f(\mathbf{x}_{n,(M)} \mid \sigma_n, \boldsymbol{\theta}, \mathcal{H}_k) f(\sigma_n \mid \boldsymbol{\theta}, \mathcal{H}_k) d\sigma_n \\ &= \int_{\sigma_n} \left(\frac{1}{\pi}\right)^{(N-k)M} \sigma_n^{-2(N-k)M} \\ &\quad \times \exp\left\{-\frac{M}{\sigma_n^2}\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{nn,(M)})\right\} \sigma_n^{-1} d\sigma_n. \end{aligned} \tag{A-5}$$

Using the identity [20]

$$\int_0^\infty x^{-(\nu+1)} e^{-ax^{-2}} dx = \frac{1}{2} a^{-\nu/2} \Gamma\left(\frac{\nu}{2}\right) \tag{A-6}$$

where $a > 0, \nu > 0$, for $a = M \cdot \operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{nn,(M)})$ and $\nu = 2(N-k)M$, we get (26).

APPENDIX B

Let the hypothesis of the number of signals be $\mathcal{H}_k$ and assume that the likelihood $\mathcal{L}_{(M)}(\theta) (= f(\mathbf{y}_{(M)} \mid \boldsymbol{\theta}))$ is normal in the parameters, i.e., that the sample size is large compared with the number of parameters, k, and that the prior $f(\boldsymbol{\theta})$ is locally uniform in the neighborhood of $\hat{\boldsymbol{\theta}}_{(ML)}$. Then [17], [20]

$$\mathcal{L}_{(M)}(\boldsymbol{\theta}) \doteq \mathcal{L}_{(M)}(\hat{\boldsymbol{\theta}}) \exp\left\{-\frac{M}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T \mathbf{J}_{(M)} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} \tag{B-1}$$

where $\mathbf{J}_{(M)}$ is defined as

$$\mathbf{J}_{(M)} = \frac{1}{M} \sum_{i=1}^{M} (-\nabla_{\boldsymbol{\theta}}^2 \log f(\mathbf{y}_i \mid \boldsymbol{\theta}, \mathcal{H}_k)) \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} . \tag{B-2}$$

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

With this approximation and substituting (28) into (13), the numerator of (13) is then rewritten as

$$\begin{aligned} \mathcal{N}_k &= \int_\theta f(\mathbf{y}_{(M)} \mid \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\theta} \mid \mathcal{H}_k) d\boldsymbol{\theta} \\ &\doteq \int_\theta \frac{1}{2}[C_{(M)}(\hat{\boldsymbol{\theta}})]^{-M} \left(\frac{1}{M\pi}\right)^{MN} (N-k)^{-M(N-k)} \\ &\quad \times \pi^{k(k-1)/2} \Gamma[M(N-k)] \prod_{l=0}^{k-1} \Gamma[M-l] \\ &\quad \times \exp\left\{-\frac{M}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T \mathbf{J}_{(M)} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} f(\boldsymbol{\theta} \mid \mathcal{H}_k) d\boldsymbol{\theta}. \end{aligned} \tag{B-3}$$

Under the condition for which the approximation (B-1) holds, a reasonable noninformative prior for $\boldsymbol{\theta}$ is taken to be proportional to an unknown constant c_k. Then,

$$\begin{aligned} \mathcal{N}_k &\propto \frac{c_k}{2}[C_{(M)}(\hat{\boldsymbol{\theta}})]^{-M} \left(\frac{1}{M\pi}\right)^{MN} (N-k)^{-M(N-k)} \\ &\quad \times \pi^{k(k-1)/2} \Gamma[M(N-k)] \prod_{l=0}^{k-1} \Gamma[M-l] \\ &\quad \times \int_\theta \exp\left\{-\frac{M}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T \mathbf{J}_{(M)} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \end{aligned} \tag{B-4}$$

$$\begin{aligned} &= \frac{c_k}{2}[C_{(M)}(\hat{\boldsymbol{\theta}})]^{-M} \left(\frac{1}{M\pi}\right)^{MN} (N-k)^{-M(N-k)} \pi^{k^2/2} \\ &\quad \times \Gamma[M(N-k)] \prod_{l=0}^{k-1} \Gamma[M-l] \cdot [\det(M \cdot \mathbf{J}_{(M)})]^{-\frac{1}{2}} \cdot \end{aligned} \tag{B-5}$$

Note that the result (B-5) is an approximation of (B-4) because we assumed that the DOA θs may extend from $-\infty$ to ∞, instead of extending between finite limits determined by the "field of view." However, the approximation is reasonable whenever the assumption (B-1) holds [21]. Similarly to (B-5), we get

$$\begin{aligned} &\int_\theta f(\mathbf{y}_{(L)} \mid \boldsymbol{\theta}, \mathcal{H}_k) f(\boldsymbol{\theta} \mid \mathcal{H}_k) d\boldsymbol{\theta} \\ &\propto \frac{c_k}{2} \cdot [C_{(L)}(\hat{\boldsymbol{\theta}})]^{-L} \left(\frac{1}{L\pi}\right)^{LN} (N-k)^{-L(N-k)} \pi^{k^2/2} \\ &\quad \times \Gamma[L(N-k)] \prod_{l=0}^{k-1} \Gamma[L-l] \cdot [\det(L \cdot \mathbf{J}_{(L)})]^{-\frac{1}{2}} \end{aligned} \tag{B-6}$$

where $\mathbf{J}_{(L)}$ is defined as (B-2) by changing M to L. After substituting (B-5) and (B-6) into (13), we get

$$f(\xi_2 \mid \xi_1, \mathcal{H}_k) \doteq \frac{[C_{(L)}(\hat{\boldsymbol{\theta}})]^L (L\pi)^{LN}}{[C_{(M)}(\hat{\boldsymbol{\theta}})]^M (M\pi)^{MN}} \times (N-k)^{-(M-L)(N-k)} \frac{\Gamma[M(N-k)] \prod_{l=0}^{k-1} \Gamma[M-l]}{\Gamma[L(N-k)] \prod_{l=0}^{k-1} \Gamma[L-l]} \times \left(\frac{L}{M}\right)^{\frac{k}{2}} \left(\frac{\det(\mathbf{J}_{(L)})}{\det(\mathbf{J}_{(M)})}\right)^{\frac{1}{2}}. \quad \text{(B-7)}$$

Assume now that the approximations in (B-5) and (B-6) are around the same $\hat{\boldsymbol{\theta}}$, and $C_{(M)}(\hat{\boldsymbol{\theta}}) \doteq C_{(L)}(\hat{\boldsymbol{\theta}})$ and $\det(\mathbf{J}_{(M)}) \doteq \det(\mathbf{J}_{(L)})$. With these assumptions, (35) follows. At a first glance, it may seem that these approximations imply relinquishment of the obtained information from the first L snapshots. This is not the case, for it turns out that the gained information is already reflected in the resulting penalty function of the criterion. In general, the vaguer this information is, the more stringent the penalty for more complex models, and vice versa. Also, these approximations will significantly reduce the computational complexity of our criterion, make our criterion independent of the particular ξ_1 set of data snapshots, and statistically improve the criterion's detection performance [16]. A general form for this likelihood approximation is derived by O'Hagan and Atkinson [18], [19].

APPENDIX C

In order to prove the consistency of the BPD criterion, we use the lemma:

Lemma C.1: Any information criteria given by

$$IC(k) = M \log C_{(M)}(\hat{\boldsymbol{\theta}}_{(k)}) + \frac{1}{2}k(k+1)\alpha(M) \quad \text{(C-1)}$$

is strongly consistent if $\alpha(M)/\log\log M \to \infty$ and $\alpha(M)/M \to 0$ as $M \to \infty$.

Proof: See [6] and [10].

To apply this lemma, first we find the asymptotical form of $T(k)$ in (38). $T(k)$ is rewritten as

$$T''(k) = T_1 + T_2 + T_3 + T_4 + T_5 \quad \text{(C-2)}$$

where

$$T_1 = MN \log\left(1 - \frac{k}{N}\right) \quad \text{(C-3)}$$

$$T_2 = -Mk \log(N-k) \quad \text{(C-4)}$$

$$T_3 = \frac{M}{(M-N+1)} \log \frac{\Gamma[(N-1)(N-k)]\Gamma[MN]}{\Gamma[M(N-k)]\Gamma[N(N-1)]} \quad \text{(C-5)}$$

$$T_4 = \frac{M}{(M-N+1)} \sum_{l=0}^{k-1} \log \frac{\Gamma[N-l-1]}{\Gamma[M-l]} \quad \text{(C-6)}$$

$$T_5 = \frac{M}{(M-N+1)} \log \frac{M}{(N-1)}. \quad \text{(C-7)}$$

Using the asymptotical relationship

$$\log\Gamma(x) \simeq \left(x - \frac{1}{2}\right)\log x - x \quad \text{as } x \to \infty \quad \text{(C-8)}$$

when $M \to \infty$, we obtain

$$T_3 \to MN \log N - M(N-k)\log M(N-k) + Mk \log M - Mk \quad \text{(C-9)}$$

$$T_4 \to \frac{k^2}{2}\log M + Mk - Mk \log M \quad \text{(C-10)}$$

and

$$T_5 \to \frac{k}{2}\log M. \quad \text{(C-11)}$$

Summing up (C-3), (C-4), (C-9), (C-10), and (C-11), we get

$$T(k) \to \frac{1}{2}k(k+1)\log M \quad \text{as } M \to \infty. \quad \text{(C-12)}$$

Therefore,

$$\alpha(M) \to \log M \quad \text{as } M \to \infty \quad \text{(C-13)}$$

and $\alpha(M)/\log\log M \to \infty$ and $\alpha(M)/M \to \infty$ as $M \to \infty$. According to Lemma C.1, the BPD criterion is shown to be strongly consistent. Furthermore, we also showed that the BPD and MDL criteria are asymptotically equivalent. Q.E.D.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their helpful comments and suggestions.

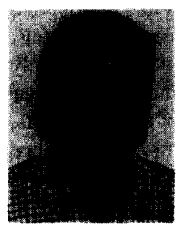
REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, no. 6, pp. 716–722, Dec. 1974.
[2] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Stat.*, vol. 11, no. 2, pp. 416–431, 1983.
[3] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.
[4] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 2, pp. 387–392, Apr. 1985.
[5] K. M. Wong *et al.*, "On information theoretic criteria for determining the number of signals in high resolution array processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 11, pp. 1959–1971, Nov. 1990.
[6] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of the number of signals in presence of white noise," *J. Multivariate Anal.*, vol. 20, no. 1, pp. 1–25, Oct. 1986.
[7] T. Shan, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation of coherent signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 4, pp. 806–811, Aug. 1985.
[8] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 4, pp. 823–831, Aug. 1985.
[9] M. Wax and I. Ziskind, "Detection of the number of coherent signals by the MDL principle," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 8, pp. 1190–1196, Aug. 1989.
[10] M. Wax, "Detection and localization of multiple sources via the stochastic signals model," *IEEE Trans. Signal Processing*, vol. 39, no. 11, pp. 2450–2456, Nov. 1991.
[11] P. M. Djurić, "Selection of Signal and System Models by Bayesian Predictive Densities," Ph.D. dissertation, Univ. of Rhode Island, 1990.
[12] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Reading, MA: Addison-Wesley, 1991.
[13] T. Leonard, "Comment on a paper by M. Lejeune and G. D. Faulkenberry," *J. Amer. Stat. Assn.*, vol. 77, no. 379, pp. 657–658, Sept. 1982.

[14] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1968, Part I.
[15] K. M. Wong *et al.*, "Estimation of the directions of arrival of signals in unknown correlated noise, part I: The MAP approach and its implementation," *IEEE Trans. Signal Processing*, vol. 40, no. 8, pp. 2007–2017, Aug. 1992.
[16] C. Cho, "Bayesian detection and estimation of superimposed signals via subspace approach," Ph.D. dissertation, SUNY at Stony Brook, NY, 1993.
[17] M. Aitkin, "Posterior Bayes factor," *J. R. Stat. Soc. B*, vol. 53, no. 1, pp. 111–142, 1991.
[18] A. O'Hagan, "Comment on a paper by M. Aitkin," *J. R. Stat. Soc. B*, vol. 53, no. 1, pp. 136, 1991.
[19] A. C. Atkinson, "Posterior probabilities for choosing a regression model," *Biometrika*, vol. 65, pp. 39–48, 1978.
[20] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley, 1973.
[21] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation* (Lecture Notes in Statistics). New York: Springer-Verlag, 1988.
[22] J. Lasenby and W. J. Fitzgerald, "A Bayesian approach to high-resolution beamforming," *Proc. Inst. Elec. Eng.-F*, vol. 138, no. 6, pp. 539–544, Dec. 1991.
[23] W. Y. Tan, "On the complex analogue of Bayesian estimation of a multivariate regression model," *Ann. Inst. Stat. Math.*, vol. 25, pp. 135–152, 1973.
[24] N. R. Goodman, "Statistical analysis based on a certain multi-variate complex Gaussian distribution (an introduction)," *Ann. Math. Stat.*, vol. 34, pp. 152–177, 1963.
[25] J. F. Böhme, "Estimation of spectral parameters of correlated signals in wavefields," *Signal Processing*, vol. 11, pp. 329–337, 1986.
[26] A. G. Jaffer, "Maximum likelihood direction finding of stochastic sources: A separable solution," in *Proc. ICASSP 88*, pp. 2893–2896, 1988.
[27] B. Ottersten, M. Viberg, and T. Kailath, "Analysis of subspace fitting and ML techniques for parameter estimation from sensor array data," *IEEE Trans. Signal Processing*, vol. 40, no. 3, pp. 590–600, Mar. 1992.
[28] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 10, pp. 1783–1795, Oct. 1990.
[29] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 10, pp. 1553–1560, Oct. 1988.
[30] D. Starer and A. Nehorai, "Newton algorithm for conditional and unconditional maximum likelihood estimation of the parameters of exponential signals in noise," *IEEE Trans. Signal Processing*, vol. 40, no. 6, pp. 1528–1534, June 1992.

Chao-Ming Cho (S'91-M'93) was born in Taichung, Taiwan, Republic of China, on Jan. 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, in 1990, all in electrical engineering.

From 1981 to 1986, he was a Research Associate at the Institute of Nuclear Sciences, Vinča, Belgrade, Yugoslavia, where his main activities were in the analysis and processing of communication signals. Currently, he is an Assistant Professor in the Department of Electrical Engineering, State University of New York at Stony Brook, NY. His main research interests are in the broad area of statistical signal processing and its applications to signal and system modeling.