DETERMINING THE NUMBER OF SIGNALS BY INFORMATION THEORETIC CRITERIA

BY

INFORMATIO N 1

STANFORD UNIV CA DEPT OF ELECTRICAL ENGINEERING M WAX ET AL 11 NOV 83

UNCLASSIFIED AFOSR-TR-84-0515 DAAG29-R1-K-0057 F/G 9/4

NL
The determination of the number of signals in a wide class of problems, including array processing, harmonic retrieval and pole retrieval, is addressed. A new approach, based on the application of the information theoretic criteria for model identification introduced by Akaike, Schwartz and Rissanen, is presented. It is shown that the criterion introduced by Schwartz and Rissanen yields a consistent estimate of the number of signals, while the criterion introduced by Akaike yields an inconsistent estimate that tends, in the large sample limit, to overestimate the number of signals.
DETERMINING THE NUMBER OF SIGNALS BY INFORMATION THEORETIC CRITERIA

Mati Wax and Thomas Kailath
Information Systems Laboratory
Stanford University
Stanford, CA 94305

ABSTRACT

The determination of the number of signals in a wide class of problems, including array processing, harmonic retrieval, and pole retrieval, is addressed. A new approach, based on the application of the information theoretic criteria for model identification introduced by Akaike, Schwartz and Rissanen, is presented. It is shown that the criterion introduced by Schwartz and Rissanen yields a consistent estimate of the number of signals, while the criterion introduced by Akaike yields an inconsistent estimate that tends, in the large-sample limit, to overestimate the number of signals.

I. Introduction

In many problems in signal processing, the vector of observations can be modeled as a linear combination of a finite number of signals that are contaminated by additive noise. This is the case, for example, in the harmonic retrieval problem [13], in the array processing problem [12], and in the problem of retrieving the poles of a system from the natural response [15]. A key issue in these problems is that of determining the number of signals.

A promising approach to the problem is based on the observation that the number of signals can be determined from the multiplicity of the smallest eigenvalue of the covariance matrix of the observation vector. The conventional method used for determining this multiplicity is based on a procedure known as the Bartlett-Lawley test. This procedure takes the form of a sequence of hypothesis tests for the multiplicity of the smallest eigenvalue. For each hypothesis, the likelihood ratio statistic is compared to a threshold. The hypothesis accepted is the first one for which the threshold is crossed. The problem with this method is the subjective judgment required for deciding on the threshold level.

In this paper we present a new method for determining the number of signals. It is based on the application of the information theoretic criteria for model identification introduced by Akaike (AIC), and by Schwartz and Rissanen (MDL). The advantage of this method is that no subjective judgment is required in the decision process; the number of signals is determined as the value for which the chosen criterion is minimized.

The problem is formulated in section II. The information theoretic criteria for model identification are introduced in section III. The application of these criteria to the problem of determining the number of signals from the multiplicity of the smallest eigenvalue of the covariance matrix is presented in section III. The consistency of these criteria is discussed in section IV. It is shown that the criterion introduced by Schwartz and Rissanen yields a consistent estimate of the number of signals, while the criterion introduced by Akaike yields an inconsistent estimate that tends, in the large-sample limit, to overestimate the number of signals. Simulation results that illustrate the performance of the new method are described in section VI.

II. Formulation of the Problem

Certain important problems in signal processing such as harmonic retrieval, array processing, and pole retrieval from the natural response, have identical structure; their observation vector can be expressed as

\[ r(t) = \sum_{i=A}^{T} A(\theta_i) s_i(t) + n(t) \]  

(1)

where \( s_i(t) \) - the i-th signal - is a zero-mean complex random process, \( A(\theta_i) \) (\( i = 1, \ldots, q \)) is a \( px1 \) complex vector, determined by the \( dx1 \) parameter vector \( \theta_i \) associated with the i-th signal, and \( n(t) \) is a \( px1 \) complex vector of additive white noise. We use complex (analytic) signal representation since this is the natural representation for the problems of interest.

A key issue in the generic model described in (1) is that of determining the number of signals \( q \) from the observed data \( r(f_k) \) (\( k = 1, \ldots, N \)).

A promising approach to this problem is based on the analysis of the eigenstructure of the covariance matrix of the observation vector \( r(t) \). To introduce this approach, let us, first, rewrite (1) as

\[ r(t) = A s(t) + n(t) \]

(2a)

where \( A \) is the \( pxq \) matrix

\[ A = [A(\theta_1), \ldots, A(\theta_q)] \]

(2b)

and \( s(t) \) is the \( qx1 \) vector

\[ s(t) = [s_1(t), \ldots, s_q(t)] \]

(2c)

The covariance matrix of \( r(t) \), assuming that the signals and noises are uncorrelated, is given by

\[ R = \Phi + \sigma^2 I \]

(3a)

where

\[ \Phi = ASA' \]

(3b)

Approved for public release; distribution unlimited.

84 06 28 085
with ℍ denoting the conjugate transpose, and S denoting the covariance matrix of the sources, i.e.,

\[ S = E[s(t)s'(t)] \]

(3.6)

We assume that the following conditions hold:

(i) S is nonsingular, i.e., the signals are noncoherent.

(ii) The matrix A is of full column rank, i.e., the vectors A(θk) (k = 1,...,q) are linearly independent.

Under the above assumptions, it follows that the rank of Ψ is q, that is, its p - q smallest eigenvalues are equal to zero. Denoting the eigenvalues of R by λ1 ≥ λ2 ≥... ≥ λp it follows, therefore, that the small p - q eigenvalues of R are equal to σ2, i.e.,

\[ λ_{p+1} = λ_{p+2} = ... = λ_p = σ^2 \]

(4)

The rank q can hence be determined from the multiplicity of the smallest eigenvalue of R.

The conventional approach to the problem, proposed by Bartlett (1954) and Lawley (1956), is based on a sequence of hypothesis tests for the multiplicity of the smallest eigenvalue. The problem associated with this approach is the subjective judgement required in choosing the threshold levels.

In this paper we take a different approach. Given the observations r(t1),...,r(N), we try to determine which of the family covariance matrices

\[ R_k = Ψ_k + σ^2I \]

(5)

where Ψk is a semi-positive matrix of rank k and σ is an unknown scalar, best fits the data.

Posed in this way, it is clear that the problem is that of model identification, and therefore the information theoretic criteria for model identification can be applied.

III. Information Theoretic Criteria

The first objective procedure for model identification was proposed by Akaike (1973) (1974). When there are several competing models, this procedure selects the model which gives the minimum AIC, defined by

\[ \text{AIC} = -2\log \text{likelihood} + 2\left( \frac{\text{number of free parameters}}{\text{within the model}} \right) \]

(6)

The first term is the well-known log-likelihood of the maximum likelihood estimates of the parameters of the model. The second term is a bias correction term, inserted so as to make the AIC an estimate of the mean Kuiback-Leibler distance between the true distribution and the estimated distribution, determined by the maximum-likelihood method.

Inspired by Akaike's pioneering work, Schwartz (1978) and Rissanel (1978) approached the problem from quite different points of view. Schwartz's approach is based on Bayesian arguments. He assumed that each competing model can be assigned a prior probability, and proposed to select the model that yields the maximum posterior probability. Rissanen's approach is based on information theoretic arguments. Since each model can be used to encode the observed sequence, Rissanen proposed to select the model that yields the minimum code length of the observed data. It turns out that in the large-sample limit, both Schwartz's and Rissanen's approaches yield the same criterion, given by

\[ \text{MDL} = -\log \text{likelihood} + \frac{1}{2} \left( \frac{\text{number of free parameters}}{\text{within the model}} \right) \log N \]

(7)

where N denotes the number of observations. We call the criterion MDL (for Minimum Description Length) as done in Rissanen (1983), since Rissanen's derivation is more general: Schwartz's derivation is restricted to the case that the observations are independent and come from an exponential distribution.

IV. Determining the Number of Signals

To apply the information theoretic criteria to determine the number of signals, or equivalently, to determine the rank of the matrix Ψ, we must first parameterize the model. Using the well-known spectral representation theorem, we can express Rk as

\[ R_k = \sum_{i=1}^{k} (λ_i - σ^2) V_i V_i' + σ^2I \]

(8)

where λ1, λ2,..., λk and V1, V2,..., Vk are the eigenvalues and eigenvectors, respectively, of Rk. Denoting by Ψk the vector of the parameters of the model, it follows that

\[ Ψ_k = \left( λ_1, ..., λ_k, σ^2, V_1', ..., V_k' \right) \]

(9)

With this parameterization, we can now derive the maximum likelihood estimates of the parameters. Assuming that the observations r1,..., rN are independently and identically distributed as \( N_p(0,R) \), it follows that, up to a constant, the log-likelihood is given by

\[ L(Ψ_k) = -N \log \det R_k - tr R_k^{-1} R \]

(10.a)

where \( R_k \) is the sample-covariance matrix defined by

\[ R_k = \frac{1}{N} \sum_{i=1}^{N} r(t_i) r(t_i)' \]

(10.b)

The maximum-likelihood (ML) estimate of \( Ψ_k \) is the value of \( Ψ_k \) that maximizes \( L(Ψ_k) \). Following Anderson (1983), the ML estimates are given by

\[ Ũ_k = \hat{Λ}_k \]

(11.a)

\[ σ^2 = \frac{1}{p-k} \hat{\Lambda}_k \]

(11.b)

\[ \hat{V}_k = \hat{C}_k \]

(11.c)

where \( \hat{Λ}_k \) is a diagonal matrix of the eigenvalues of the estimated sample covariance matrix \( \hat{R} \).

Substituting the maximum likelihood estimates (11) in the log-likelihood (10), we obtain, after some algebra

\[ \log L(Ψ_k) = \log \left( \frac{1}{p-k} \hat{\Lambda}_k \right) \]

(12)

The number of independently adjusted parameters in the model is obtained by counting the number of degrees of freedom of the space spanned by the parameter vector \( Ψ_k \). Recalling that the eigenvalues of a complex covariance matrix are real, but that the eigenvectors are complex, it follows that \( Ψ_k \) has \( k + 1 + 2pk \) parameters. However, not all of the parameters are independently adjusted: the eigenvectors are constrained to have unit norm and to be mutually...
orthogonal. This amounts to reduction of $2k$ degrees of freedom due to the normalization and $2\frac{1}{2}(k-1)$ degrees of freedom due to the mutual orthogonalization. Thus, we obtain
\[
\text{number of free adjusted parameters within the model} = k + 1 + 2pk - \left[\frac{1}{2}k(k-1)\right] = k(2p-k)+1
\]
(13)
The form of AIC for this problem is therefore given by
\[
AIC(k) = -2\log \left[ \frac{1}{p-k} \sum \lambda_i \right] + 2k(2p-k) \tag{14}
\]
while the MDL criterion is given by
\[
MDL(k) = -2\log \left[ \frac{1}{p-k-q} \sum \lambda_i \right] + \frac{1}{2}k(2p-k)\log N \tag{15}
\]
The implementation of these criteria is as follows. First, the chosen criterion is computed for every possible $k$, that is, $k = 0, 1, \ldots, p-1$. The rank of $\Psi$ is then determined as the value of $k$ for which the criterion is minimized.

V. Consistency of the MDL Criterion

The concept of consistency is fundamental in statistical inference. In our problem, consistency means convergence of the selection criterion to the true rank $q$ in the large-sample limit. We shall show, by generalizing a method of proof given in Rissanen (1980) and Fan and Quinn (1979), that the MDL yields a consistent estimate, and that the AIC yields an inconsistent estimate that tends, in the large-sample limit, to overestimate the true rank.

The consistency of the MDL is proved by showing that in the large-sample limit, the criterion $MDL(k)$ is minimized for the true rank $k = q$. Taking first $k < q$, it follows from (15) that
\[
\frac{1}{N}[MDL(q) - MDL(k)] =
\]
\[
= \log \left[ \frac{1}{q-k} \sum \lambda_i \right]^{q-k} - \log \left[ \frac{1}{p-k-q} \sum \lambda_i \right]^{p-k-q} + \log \left[ \frac{1}{p-k-q} \sum \lambda_i \right]^{p-k-q} \tag{16}
\]
Now, it is well-known that the eigenvectors of the sample covariance matrix $X_i$ ($i = 1, \ldots, p$) are consistent estimates of the eigenvectors of the true covariance matrix $\Lambda_i$. Thus, in the large-sample limit the eigenvectors $\lambda_i$ ($i = 1, \ldots, g$) are not all equal with probability one. Therefore, by the arithmetic-mean geometric-mean inequality, it follows that in the large-sample limit
\[
\frac{1}{q-k} \sum \lambda_i > \left[ \frac{1}{q-k} \sum \lambda_i \right]^{\frac{1}{q-k}} \tag{17}
\]
Thus, the first term in (18) is negative with probability one in the large-sample limit. Similarly, by the generalized arithmetic-mean geometric-mean inequality
\[
w_1 A_1 + w_2 A_2 \geq A_1 \wedge A_2 \quad w_1 + w_2 = 1 \tag{18}
\]
it follows that
\[
\frac{1}{p-k-q} \sum \lambda_i > \left[ \frac{1}{p-k-q} \sum \lambda_i \right]^{\frac{1}{p-k-q}} \tag{19}
\]
Thus, the second term in (18) is also negative with probability one in the large-sample limit. Now, since the last term in (19) goes to zero as the sample size increases, the difference $[MDL(q) - MDL(k)]$ is negative with probability one in the large-sample limit, for $k < q$.

Taking now $k > q$, it follows from (15) that
\[
2[MDL(k) - MDL(q)] =
\]
\[
= -2\log \left[ \frac{1}{p-k} \sum \lambda_i \right]^{p-k} + 2\log \left[ \frac{1}{p-k-q} \sum \lambda_i \right]^{p-k-q} + (k-q)(2p-k-q+1)\log N \tag{20}
\]
Note that the terms in the curly brackets are twice the log-likelihoods of the maximum likelihood estimator under the hypotheses that the rank of $\Psi$ is $q$ and $k$, respectively. Thus, their difference is the likelihood-ratio for deciding between these two hypotheses. From the general theory of likelihood ratios (see e.g. Cox and Hinkley (1974)) it follows that the asymptotic distribution of this statistic is $\chi^2$ with number of degrees of freedom equal to the difference of the dimensions of the parameter spaces under the two hypotheses, i.e.,
\[
[k(2p-k+1) - (q(2p-q)+1)] = (k-q)(2p-k-q+1)
\]
Thus, as the sample size increases, the probability that the term in the curly brackets in (18) exceeds the last term in (18), is given by the area in the tail from $(k-q)(2p-k-q+1)\log N$ of the mentioned $\chi^2$ distribution with $(k-q)(2p-k-q+1)$ degrees of freedom. Since the area in this tail approaches zero as the sample size increases, it follows that in the large-sample limit the difference $[MDL(k) - MDL(q)]$ is positive with probability one for $k > q$. Combining this with the previous result for $k < q$, it follows that $MDL(k)$ has a minimum at $k = q$.

Repeating the above arguments for the AIC, it follows that in the large-sample limit $[AIC(q) - AIC(k)]$ is negative with probability one for $k < q$. However, for $k > q$, $[AIC(k) - AIC(q)]$ has nonzero probability to be negative even in the large-sample limit, since the tail of the probability distribution from
VII. Simulation Results

In this section we present simulation results that illustrate the performance of the proposed method. The examples are taken from array processing, though, by the well-known duality between spatial frequency and temporal frequency, they can also be interpreted in the context of harmonic-retrieval. The examples refer to a uniform linear array of \( p \) sensors, with \( q \) incoherent sinusoidal plane-waves impinging from directions \( \phi_1, \ldots, \phi_q \). The received signal at the \( t \)th sensor (\( t = 1, \ldots, p \)) is thus given by

\[
    r_t(t) = \sum_{i=1}^{q} e^{-j[(t-1)\omega_1 + \phi_1]} + n(t) \quad (21)
\]

where

\[ p_0 = \text{random phase uniformly distributed on} \ (0, 2\pi) \]
\[ n(t) = \text{white noise with mean zero and variance} \ \sigma^2 \]

Note that this model is a special case of the generic model presented in (1).

In the first example, we considered an array with seven sensors (\( p = 7 \)) and two sources (\( q = 2 \)). The signal-to-noise ratio, defined as \( 10 \log_{10} \frac{\text{signal}}{\text{noise}} \), was \( 12\text{dB} \) and the directions-of-arrival were \( 20^\circ \) and \( 23^\circ \). Using \( N = 100 \) samples, the values of the AIC obtained by (14), are given by

<table>
<thead>
<tr>
<th>AIC(0)</th>
<th>AIC(1)</th>
<th>AIC(2)</th>
<th>AIC(3)</th>
<th>AIC(4)</th>
<th>AIC(5)</th>
<th>AIC(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1915.6</td>
<td>45.3</td>
<td>42.1</td>
<td>40.0</td>
<td>51.1</td>
<td>51.0</td>
<td>50.9</td>
</tr>
</tbody>
</table>

Clearly, the minimum value of the AIC is obtained for \( q = 2 \).

In the second example, we added another source at \( -5^\circ \) to the scenario described in the first example, and raised the signal-to-noise ratio to \( 20\text{dB} \). The results obtained in this case were

<table>
<thead>
<tr>
<th>AIC(0)</th>
<th>AIC(1)</th>
<th>AIC(2)</th>
<th>AIC(3)</th>
<th>AIC(4)</th>
<th>AIC(5)</th>
<th>AIC(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2021.8</td>
<td>102.6</td>
<td>51.0</td>
<td>48.0</td>
<td>50.0</td>
<td>51.0</td>
<td>55.0</td>
</tr>
</tbody>
</table>

Note that the minimum AIC is obtained, correctly, for \( q = 3 \). The two examples above demonstrate clearly the ability of the new procedure to detect correctly the number of sources even in relatively difficult scenario of two closely spaced sources and low signal-to-noise ratio.

VII. Concluding Remarks

The method we have described is a time-domain method since it is based on processing of the covariance matrix of the observation vector. In some cases, especially in array processing, the frequency-domain is more natural. The extension of the method to the frequency-domain, that is to the problem of determining the number of signals from the spectral-density matrix of the observation vector is presented in [14].

REFERENCES
