ESPRIT—Estimation of Signal Parameters Via Rotational Invariance Techniques

RICHARD ROY AND THOMAS KAILATH, FELLOW, IEEE

Abstract—**High-resolution signal parameter estimation is a problem of significance in many signal processing applications. Such applications include *direction-of-arrival* (DOA) estimation, system identification, and time series analysis. A novel approach to the general problem of signal parameter estimation is described. Although discussed in the context of direction-of-arrival estimation, ESPRIT can be applied to a wide variety of problems including accurate detection and estimation of sinusoids in noise. It exploits an underlying *rotational invariance* among signal subspaces induced by an array of sensors with a *translational invariance* structure. The technique, when applicable, manifests significant performance and computational advantages over previous algorithms such as MEM, Capon's MLM, and MUSIC.**

I. Introduction

In many practical signal processing problems, the objective is to estimate from measurements a set of *constant* parameters upon which the received signals depend. For example, high-resolution direction-of-arrival (DOA) estimation is important in many sensor systems such as radar, sonar, electronic surveillance, and seismic exploration. High-resolution frequency estimation is important in numerous applications, recent examples of which include the design and control of robots and large flexible space structures. In such problems, the functional form of the underlying signals can often be assumed to be known (e.g., narrow-band plane waves, cisoids). The quantities to be estimated are parameters (e.g., frequencies and DOA's of plane waves, cisoid frequencies) upon which the sensor outputs depend, and these parameters are assumed to be constant.[1]

There have been several approaches to such problems including the so-called maximum likelihood (ML) method of Capon (1969) and Burg's (1967) maximum entropy (ME) method. Although often successful and widely used, these methods have certain fundamental limitations (especially bias and sensitivity in parameter estimates), largely because they use an incorrect model (e.g., AR rather than special ARMA) of the measurements. Pisarenko (1973) was one of the first to exploit the structure of the data model, doing so in the context of estimation of parameters of cisoids in additive noise using a covariance approach. Schmidt (1977) and independently Bienvenu (1979) were the first to correctly exploit the measurement model in the case of sensor arrays of arbitrary form. Schmidt, in particular, accomplished this by first deriving a complete geometric solution in the absence of noise, then cleverly extending the geometric concepts to obtain a *reasonable* approximate solution in the presence of noise. The resulting algorithm was called MUSIC (MUltiple SIgnal Classification) and has been widely studied. In a detailed evaluation based on thousands of simulations, M.I.T.'s Lincoln Laboratory concluded that, among currently accepted high-resolution algorithms, MUSIC was the most promising and a leading candidate for further study and actual hardware implementation. However, although the performance advantages of MUSIC are substantial, they are achieved at a considerable cost in computation (searching over parameter space) and storage (of array calibration data).

In this paper, a new algorithm (ESPRIT) that dramatically reduces these computation and storage costs is presented. In the context of DOA estimation, the reductions are achieved by requiring that the sensor array possess a *displacement invariance*, i.e., sensors occur in matched pairs with identical displacement vectors. Fortunately, there are many practical problems in which these conditions are or can be satisfied. In addition to obtaining signal parameter estimates efficiently, *optimal* signal copy vectors for reconstructing the signals are elements of the ESPRIT solution as well. ESPRIT is also manifestly more robust (i.e., less sensitive) with respect to array imperfections than previous techniques including MUSIC [1].

To make the presentation as clear as possible, an attempt is made to adhere to a somewhat standard notational convention. Lowercase boldface italic characters will generally refer to vectors. Uppercase boldface italic characters will generally refer to matrices. For either real- or complex-valued matrices, $(\cdot)^*$ will be used to denote the Hermitian conjugate (or complex-conjugate transpose) operation. Eigenvalues of square Hermitian matrices are assumed to be ordered in decreasing magnitude, as are the

Manuscript received January 12, 1988; revised October 5, 1988. This work was supported in part by the Joint Services Program at Stanford University (U.S. Army, U.S. Navy, U.S. Air Force) under Contract DAAG29-85-K-0048, and the SDI/IST Program managed by the Office of Naval Research, under Contract N00014-85-K-0550.

The authors are with the Information Systems Laboratory, Stanford University, Stanford, CA 94305.

IEEE Log Number 8928125.

[1]Extensions to situations in which the parameters may be time varying can be made, however, they rely on an inherent *time-scale* or *eigenvalue separation* between the parameter dynamics and the dynamics of the signal process. Fundamentally, the assumption is made that over time intervals long enough to collect sufficient information from which to obtain accurate parameter estimates, the parameters have not changed significantly.

singular values of nonsquare matrices. Knowledge of the fundamental theorems of matrix algebra dealing with eigendecompositions and singular value decompositions (SVD) is assumed (cf. [2]).

II. The Data Model

Although ESPRIT is generally applicable to a wide variety of problems, for illustrative purposes the discussions herein focus on DOA estimation. In many practical signal processing applications, data from an array of sensors are collected, and the objective is to locate point sources assumed to be radiating energy that is detectable by the sensors (cf. Fig. 1). Mathematically, such problems are quite simply, although abstractly, modeled using *Green's functions* for the particular differential operator that describes the physics of radiation propagation from the sources to the sensors. For the intended applications, however, a few *reasonable* assumptions can be invoked to make the problem analytically tractable.

The transmission medium is assumed to be isotropic and nondispersive so that the radiation propagates in *straight lines*, and the sources are assumed to be in the *far-field* of the array. Consequently, the radiation impinging on the array is in the form of a sum of *plane waves*. For simplicity, it will initially be assumed that the problem is planar, thus reducing the location parameter space to a single-dimensional subset of $\mathfrak{R}$, i.e., $\theta_i \in [-\pi, \pi]$, where θ_i is the direction-of-arrival (DOA) of the ith source.

The signals are assumed to be *narrow-band* processes, and can be considered to be sample functions of a stationary stochastic process or deterministic functions of time. This can represent a significant difference when estimating the signals themselves, leading to entirely different estimation algorithms. However, as far as estimation of signal parameters such as DOA is concerned, both assumptions lead to the same (suboptimal) algorithm under certain conditions (e.g., *persistent excitation*).

Since the narrow-band signals are assumed to have the same *known* center frequency[2] (ω_0), the ith signal can be written as $\tilde{s}_i(t) = u_i(t) \cos(\omega_0 t + v_i(t))$, where $u_i(t)$ and $v_i(t)$ are slowly varying[3] functions of time that define the amplitude and phase of $\tilde{s}_i(t)$, respectively. For such signals, it is often more convenient to use the *complex envelope* representation [3] in which $\tilde{s}_i(t) = \text{Re}\{s(t)\}$, where $s(t) = u(t)\, e^{j(\omega_0 t + v(t))}\}$. Noting that the narrow-band assumption implies $u(t) \approx u(t-\tau)$ and $v(t) \approx v(t-\tau)$ for all possible propagation delays τ, the effect of a time delay on the received waveforms is simply a phase shift, i.e., $s(t-\tau) \approx s(t)\, e^{-j\omega_0 \tau}$. The result is that $x_k(t)$, the complex signal output of the kth sensor at time t, can be written as

$$x_k(t) = \sum_{i=1}^{d} a_k(\theta_i)\, s_i(t - \tau_k(\theta_i)) = \sum_{i=1}^{d} a_k(\theta_i)\, s_i(t)\, e^{-j\omega_0 \tau_k(\theta_i)}, \quad (1)$$

where $\tau_k(\theta_i)$ is the propagation delay between a reference point and the kth sensor for the ith wavefront impinging on the array from direction θ_i, $a_k(\theta_i)$ is the corresponding sensor element complex response (gain and phase) at frequency ω_0, and there are assumed to be d point sources present.

Employing vector notation for the outputs of the m sensors, the *data model* becomes

$$\boldsymbol{x}(t) = \sum_{i=1}^{d} \boldsymbol{a}(\theta_i)\, s_i(t), \quad (2)$$

where

$$\boldsymbol{a}(\theta_i) = [a_1(\theta_i)\, e^{-j\omega_0 \tau_1(\theta_i)}, \cdots, a_m(\theta_i)\, e^{-j\omega_0 \tau_m(\theta_i)}]^T, \quad (3)$$

often termed the *array response* or *array steering vector* for direction θ_i. Setting $\boldsymbol{A}(\boldsymbol{\theta}) = [\boldsymbol{a}(\theta_1), \cdots, \boldsymbol{a}(\theta_d)]$, $s(t) = [s_1(t), \cdots, s_d(t)]^T$, and adding measurement noise $\boldsymbol{n}(t)$, the *measurement model* for the passive sensor array narrow-band signal processing problem is

$$\boldsymbol{x}(t) = \boldsymbol{A}(\boldsymbol{\theta})\, s(t) + \boldsymbol{n}(t). \quad (4)$$

Note that $\boldsymbol{x}(t), \boldsymbol{n}(t) \in \mathbb{C}^m$, $s(t) \in \mathbb{C}^d$, and $\boldsymbol{A}(\boldsymbol{\theta}) \in \mathbb{C}^{m \times d}$, and it will be assumed that $m > d$.

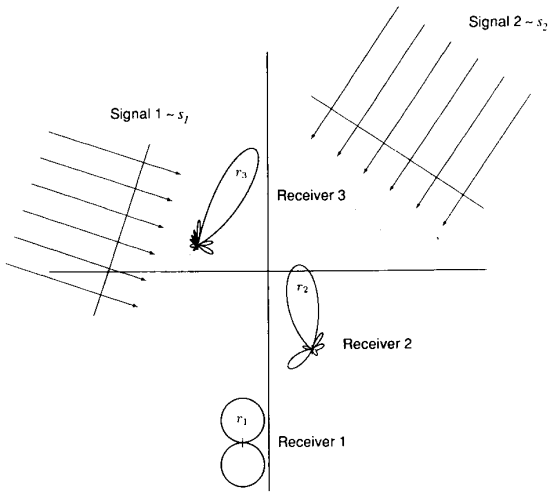


Fig. 1. Passive sensor array geometry.

III. The Geometric Approach

In 1977, Schmidt [4] developed the MUSIC (MUltiple SIgnal Classification) algorithm by taking a geometric view of the signal parameter estimation problem. One of the major breakthroughs afforded by the MUSIC algo-

[2]This condition is often termed *co-channel interference* in communication applications. If the center frequencies are not the same, the problem can be greatly simplified by first separating the signals in the frequency domain.

[3]The definition of slowly varying is taken to mean that the approximation $\tilde{s}_i(t - \tau_k(\theta_i)) \approx u_i(t) \cos(\omega_0(t - \tau_k(\theta_i)) + v_i(t))$, is valid, i.e., the amplitude and phase variations as functions of spatial position for fixed t are *negligible* over the extent of the array.

rithm was the ability to handle arbitrary arrays of sensors. Until the mid-1970's, direction finding techniques required knowledge of the array directional sensitivity pattern in analytical form, and the task of the antenna designer was to build an array of antennas with a prespecified sensitivity pattern. The work of Schmidt essentially relieved the designer from such constraints by exploiting the reduction in analytical complexity that could be achieved by *calibrating* the array. Thus, the highly nonlinear problem of calculating the array response to a signal from a given direction was reduced to that of measuring and storing the response. Although MUSIC did not mitigate the computational complexity of solution to the DOA estimation problem, it did extend the applicability of high-resolution DOA estimation to *arbitrary* arrays of sensors.

A. Array Manifolds and Signal Subspaces

To introduce the concepts of the *array manifold* and the *signal subspace*, recall the noise-free data model $\boldsymbol{x}(t) = \boldsymbol{A}(\boldsymbol{\theta})\,\boldsymbol{s}(t)$. The vectors $\boldsymbol{a}(\theta_i) \in \mathbb{C}^m$, the columns of $\boldsymbol{A}(\boldsymbol{\theta})$, are elements of a set (not a subspace), termed the *array manifold*[4] ($\mathfrak{A}$), composed of all array response (*steering*) vectors obtained as θ ranges over the entire parameter space. $\mathfrak{A}$ is completely determined by the sensor directivity patterns and the array geometry, and can sometimes be computed analytically. However, for complex arrays that defy analytical description, $\mathfrak{A}$ can be obtained by calibration (i.e., physical measurements).

For azimuth-only DOA estimation, the array manifold is a one-parameter manifold that can be viewed as a *rope* weaving through $\mathbb{C}^m$. For azimuth and elevation DOA estimation, the manifold is a *sheet* in $\mathbb{C}^m$. To avoid ambiguities, it is necessary to assume that the map from $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_d\}$ to $\mathcal{R}\{\boldsymbol{A}(\boldsymbol{\theta})\}$, the subspace spanned by the columns of $\boldsymbol{A}(\boldsymbol{\theta})$, is one-to-one. This property can be ensured by proper array design.

The key observation is that if $\boldsymbol{x}(t) = \boldsymbol{a}(\theta)\,s_\theta(t)$ is an appropriate data model (in the absence of noise) for a single signal, the data are confined to a *one-dimensional subspace* of $\mathbb{C}^m$ characterized by the vector $\boldsymbol{a}(\theta)$. For d signals, the observed data vectors $\boldsymbol{x}(t) = \boldsymbol{A}(\boldsymbol{\theta})\,\boldsymbol{s}(t)$ are constrained to the d-dimensional subspace of $\mathbb{C}^m$, termed the *signal subspace* ($\mathcal{S}_X$), that is spanned by the d vectors $\boldsymbol{a}(\theta_i)$, the columns of $\boldsymbol{A}(\boldsymbol{\theta})$.[5]

B. Intersections as Solutions

The concepts of an observed signal subspace and a calibrated array manifold permit an immediate visualization of the solution. In the absence of noise, the outputs of the sensor array lie in a d-dimensional subspace of $\mathbb{C}^m$, the *signal subspace* ($\mathcal{S}_X$) spanned by the columns of $\boldsymbol{A}(\boldsymbol{\theta})$. Once d independent vectors have been observed,[6] $\mathcal{S}_X$ is known, and intersections between the observed subspace and the array manifold yield the set of vectors from the array manifold that span the observed signal subspace. A three-sensor two-source example is graphically depicted in Fig. 2. Assuming that the sensor array has been designed such that the map from parameters to array manifold vectors is unique, the parameters are immediately determined.

Problems arise when only noisy measurements $\boldsymbol{x}(t) = \boldsymbol{A}(\boldsymbol{\theta})\,\boldsymbol{s}(t) + \boldsymbol{n}(t)$ of the array output are available, since $\mathcal{S}_X$ must be estimated. Imposing the constraint that the estimate $\hat{\mathcal{S}}_X$ be spanned by elements from $\mathfrak{A}$ and assuming unknown deterministic signals and Gaussian noise, a maximum-likelihood (ML) estimator can be formulated as described in Appendix A. However, the ML solution (of obtaining a set of vectors from the array manifold that *best fits* all the measurements) is computationally prohibitive in most practical applications.

Schmidt's idea was to employ a suboptimal *two-step* procedure instead. First, an *unconstrained* set of d vectors that best fits all the measurements is found. Then points of closest approach of the space spanned by those vectors to the array manifold are sought. This procedure, although clearly suboptimal, retains some of the key properties of the ML solution, including the fact that the *exact* answer is obtained asymptotically as the number of measurements goes to infinity.

C. Estimating the Signal Subspace

To obtain an unconstrained estimate of the signal subspace, the *least-squares* (LS) criterion is most often employed. The idea is to find a set of d vectors that span a subspace of $\mathbb{C}^m$ that *best fits*, in an LS sense, the observed data. Assuming the signals and noise are zero-mean, a method can be derived by first examining the covariance matrix of the measurements. If the signals are modeled as stationary (zero-mean) stochastic processes, they are assumed to be uncorrelated with the noise and possess a positive definite covariance matrix $\boldsymbol{R}_{SS} > 0$. If, on the other hand, a deterministic (zero-mean) signal model is chosen, a persistent excitation condition is imposed, i.e.,

$$\boldsymbol{R}_{SS} \overset{\text{def}}{=} \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{s}(t)\,\boldsymbol{s}^*(t)$$

is assumed to exist and be positive definite.

Under the conditions given above, the covariance matrix of the measurements is given by

$$\boldsymbol{R}_{XX} \overset{\text{def}}{=} E\{\boldsymbol{x}\boldsymbol{x}^*\} = \boldsymbol{A}\boldsymbol{R}_{SS}\boldsymbol{A}^* + \sigma^2\boldsymbol{\Sigma}_n. \tag{5}$$

The objective is to find a set of d linearly independent vectors that is contained in $\mathcal{S}_X = \mathcal{R}\{\boldsymbol{A}\}$, the subspace

[4]Technically, a *k-dimensional manifold* in $\mathbb{C}^m$ is a subset of points in $\mathbb{C}^m$ satisfying certain local continuity and differentiability conditions. The physics of sensor arrays guarantee the continuity and differentiability properties will be satisfied. Associated with each point on the manifold is a vector to that point from the origin in $\mathbb{C}^m$ using the standard basis.

[5]It is also convenient to define the *noise subspace* ($\mathcal{S}_X^{\perp}$) as the orthogonal complement of the signal subspace in $\mathbb{C}^m$.

[6]The problem of degenerate signal spaces, i.e., fully correlated signals, is discussed in [1]. For the purposes of this discussion, it is assumed that the signals are not fully correlated.

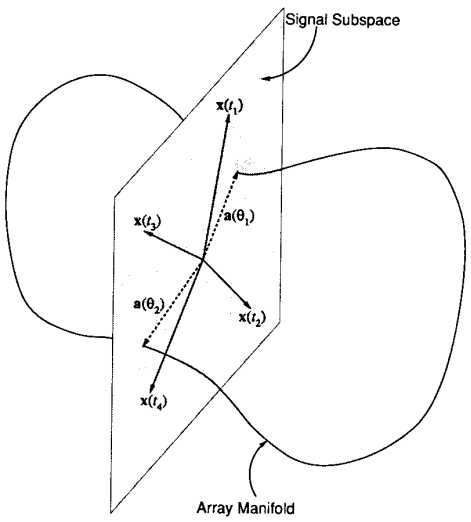


Fig. 2. The geometry of MUSIC for a three-sensor two-source example (no noise).

spanned by the columns of $\boldsymbol{A}$. One such set of vectors is given by $\boldsymbol{E}_S = \boldsymbol{\Sigma}_n[\boldsymbol{e}_1 | \cdots | \boldsymbol{e}_d]$ where $\{\boldsymbol{e}_i, i = 1, \cdots, d\}$ is the set of generalized eigenvectors (GEV's) of $\boldsymbol{R}_{XX}$ corresponding to the d largest generalized eigenvalues (GE's). This fact follows directly from the defining property of GEV's, i.e., $\boldsymbol{R}_{XX}\overline{\boldsymbol{E}} = \boldsymbol{\Sigma}_n\overline{\boldsymbol{E}}\lambda_i$, and their $\boldsymbol{\Sigma}_n$-orthogonality ($\overline{\boldsymbol{E}}^*\boldsymbol{\Sigma}_n\overline{\boldsymbol{E}} = \boldsymbol{I}$) as follows:

$$\begin{aligned}
&\boldsymbol{R}_{XX}\overline{\boldsymbol{E}} = \boldsymbol{\Sigma}_n\overline{\boldsymbol{E}}\boldsymbol{\Lambda},\\
\Rightarrow\ &\boldsymbol{A}\boldsymbol{R}_{SS}\boldsymbol{A}^*\overline{\boldsymbol{E}} + \sigma^2\boldsymbol{\Sigma}_n\overline{\boldsymbol{E}} = \boldsymbol{\Sigma}_n\overline{\boldsymbol{E}}\boldsymbol{\Lambda},\\
\Rightarrow\ &\overline{\boldsymbol{E}}^*\boldsymbol{A}\boldsymbol{R}_{SS}\boldsymbol{A}^*\overline{\boldsymbol{E}} = \overline{\boldsymbol{E}}^*\boldsymbol{\Sigma}_n\overline{\boldsymbol{E}}\boldsymbol{\Lambda} - \sigma^2\overline{\boldsymbol{E}}^*\boldsymbol{\Sigma}_n\overline{\boldsymbol{E}},\\
\Rightarrow\ &\overline{\boldsymbol{E}}^*\boldsymbol{A}\boldsymbol{R}_{SS}\boldsymbol{A}^*\overline{\boldsymbol{E}} = \boldsymbol{\Lambda} - \sigma^2\boldsymbol{I},\\
\Rightarrow\ &\boldsymbol{A}\boldsymbol{R}_{SS}\boldsymbol{A}^* = \overline{\boldsymbol{E}}^{-*}[\boldsymbol{\Lambda} - \sigma^2\boldsymbol{I}]\,\overline{\boldsymbol{E}}^{-1},\\
\Rightarrow\ &\boldsymbol{A}\boldsymbol{R}_{SS}\boldsymbol{A}^* = \boldsymbol{\Sigma}_n\overline{\boldsymbol{E}}[\boldsymbol{\Lambda} - \sigma^2\boldsymbol{I}]\,\overline{\boldsymbol{E}}^*\boldsymbol{\Sigma}_n. \qquad (6)
\end{aligned}$$

Since $\boldsymbol{A}\boldsymbol{R}_{SS}\boldsymbol{A}^*$ is rank d and positive semidefinite by construction, the d largest GE's of $(\boldsymbol{R}_{XX}, \boldsymbol{\Sigma}_n)$ are simply the d nonzero GE's of $(\boldsymbol{A}\boldsymbol{R}_{SS}\boldsymbol{A}^*, \boldsymbol{\Sigma}_n)$ augmented by σ^2, and $\mathcal{R}\{\boldsymbol{E}_S\} = \mathcal{R}\{\boldsymbol{A}\} \overset{\text{def}}{=} \mathcal{S}_X$. Similarly defining $\boldsymbol{E}_N = \boldsymbol{\Sigma}_n[\boldsymbol{e}_{d+1} | \cdots | \boldsymbol{e}_m]$, $\mathcal{R}\{\boldsymbol{E}_N\} = \mathcal{S}_X^{\perp}$. Note also that the $m - d$ smallest GE's of $(\boldsymbol{R}_{XX}, \boldsymbol{\Sigma}_n)$ are all equal to σ^2, a fact that can be used to find d and σ^2.

In most situations, the covariance matrices required above are not known and must be estimated from measurements. Practically, the noise covariance can be estimated from measurements made when signals of interest are not present. Measurements made when signals of interest are present are used to estimate $\boldsymbol{R}_{XX}$. The standard estimate[7] $\hat{\boldsymbol{R}}_{XX} = 1/N\,\Sigma_{t=1}^{N}\boldsymbol{x}(t)\boldsymbol{x}^*(t) = \hat{\boldsymbol{R}}_{XX} = 1/N\,\boldsymbol{X}\boldsymbol{X}^*$, where $\boldsymbol{X}$ is the $m \times N$ data matrix, is most often employed. Note that this estimate can be scaled by $N/(N - m)$ to obtain an unbiased estimate, but the scaling only affects the magnitudes of the eigenvalues, not the eigenvectors, so the subspace estimates remain unaffected.

When the covariances are estimated from finite data matrices, the $m - d$ smallest GE's are only clustered around, and not all equal to, σ^2. In this case, special statistical techniques based on likelihood ratio (LR) tests [1], [5] can be used to obtain an estimate $\hat{d}$ of the signal subspace dimension. In any case, in the presence of a finite amount of noisy measurements, $\mathcal{R}\{\boldsymbol{E}_S\} = \hat{\mathcal{S}}_X$ and $\mathcal{R}\{\boldsymbol{E}_N\} = \hat{\mathcal{S}}_X^{\perp}$, the *estimated signal and noise subspaces*.

D. Estimating the Signal Parameters

In the absence of noise, parameter estimates can be obtained by finding *intersections* of $\mathcal{A}$ with $\mathcal{S}_X$ or, equivalently, finding elements of $\mathcal{A}$ that are orthogonal to $\mathcal{S}_X^{\perp}$. At this point in the suboptimal signal subspace algorithms, the real computational effort[8] begins. Even with perfect knowledge of the signal subspace, searching the array manifold for d intersections with $\mathcal{S}_X$ can be quite costly, especially for multidimensional parameter spaces (e.g., azimuth, elevation, and range). The problem is further complicated in the presence of noise since, with probability one, $\hat{\mathcal{S}}_X \cap \mathcal{A} = \varnothing$; there are no intersections. Consequently, there are no elements of $\mathcal{A}$ that are orthogonal to $\hat{\mathcal{S}}_X^{\perp}$. Referring to Fig. 2, it would seem that intersections could almost always be found. In this respect, the figure is somewhat misleading. For three sensors and two sources, the signal subspace is a two-dimensional *complex* subspace of three-dimensional *complex* space. In the real field, the estimated signal subspace is actually a four-dimensional subspace of six-dimensional space, and it need not intersect the one-dimensional manifold at all. Obviously, elements of $\mathcal{A}$ that are *closest* to $\hat{\mathcal{S}}_X$ should be considered as potential solutions, but the issue of an appropriate measure of *closeness* remains.

Schmidt [4] proposed the following function as one possible measure[9] of the *closeness* of an element of $\mathcal{A}$ to $\hat{\mathcal{S}}_X$:

$$P_M(\theta) = \frac{\boldsymbol{a}^*(\theta)\,\boldsymbol{a}(\theta)}{\boldsymbol{a}^*(\theta)\,\boldsymbol{E}_N\boldsymbol{E}_N^*\boldsymbol{a}(\theta)}. \qquad (7)$$

In the absence of noise, this measure, termed the MUSIC *spectrum*, is infinite for elements of $\mathcal{A}$ belonging to $\mathcal{S}_X$. In the presence of noise, this measure is clearly peaked near points of closest approach of $\mathcal{A}$ to $\hat{\mathcal{S}}_X$, points where $\mathcal{A}$ is nearly orthogonal to $\hat{\mathcal{S}}_X^{\perp}$. This property is used to obtain parameter estimates as those parameters, or parameter vectors in the case of multidimensional array manifolds, that yield the d largest *peaks in the spectrum*. Modifications to this measure to include *eigenvalue weighting*

[7]The sample covariance is known to be a *sufficient statistic* for the estimation problem of finding the best fit of a rank d subspace to the observed data in a least-squares sense.

[8]Only for extremely large arrays such as phased array radars with hundreds of elements is the eigendecomposition of the measurement covariance matrix computationally significant in comparison to the search over $\mathcal{A}$.

[9]In [4], the numerator was not explicitly included in the measure since it was assumed that the array manifold vectors were all *normalized in some suitable fashion*.

of the eigenvectors have been proposed by several investigators, e.g., Johnson [6].

Although conceptually simple, the *one-dimensional*[10] MUSIC measure has several drawbacks. Primarily, problems in the finite measurement case arise from the fact that since d signals are known to be present, d parameter estimates, $\{\theta_1, \cdots, \theta_d\}$, should be sought *simultaneously* by maximizing an appropriate functional rather than obtaining estimates one at a time as is done in the search over $P_M(\theta)$. However, multidimensional searches are exponentially more expensive than one-dimensional searches. The price paid for the computational reduction achieved by employing a *one-dimensional search* for d parameters is that the method is finite-sample-biased in the multiple source environment (cf. [1]). Furthermore, in low SNR scenarios and in situations where even small sensor array errors are present, the ability of the conventional MUSIC *spectrum* to *resolve* closely spaced sources (i.e., observe multiple peaks in the measure) is severely degraded (cf. [1]). Nevertheless, it should be emphasized that in spite of these drawbacks, MUSIC has been shown to outperform previous techniques (cf. [7]).

Finally, as indicated above, MUSIC asymptotically yields unbiased parameter estimates since as the amount of data becomes infinite, errors in the estimate $\hat{\mathbf{S}}_X$ vanish [8]. If the noise is spatially Gaussian and temporally independent, the distribution of the eigenvectors of the sample covariance is asymptotically Gaussian, with mean equal to the true eigenvectors (assuming distinct eigenvalues[11]) and covariances that go to zero [9], [10]. Thus, the estimated signal subspace *converges in mean-square* to the true signal subspace, and the parameter estimates converge to the true values as well.

E. Summary of the MUSIC Algorithm

The following is a summary of the MUSIC algorithm based on the covariance approach described above.

1) Collect the data and estimate $\boldsymbol{R}_{XX} = E\{\boldsymbol{x}\boldsymbol{x}^*\} = \boldsymbol{A}\boldsymbol{R}_{SS}\boldsymbol{A}^* + \sigma^2\boldsymbol{\Sigma}_n$ denoting the estimate $\hat{\boldsymbol{R}}_{XX}$.

2) Solve for the eigensystem; $\hat{\boldsymbol{R}}_{XX}\overline{\boldsymbol{E}} = \boldsymbol{\Sigma}_n\overline{\boldsymbol{E}}\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} \overset{\text{def}}{=} \operatorname{diag}\{\lambda_1, \cdots, \lambda_m\}$, $\lambda_1 \geq \cdots \geq \lambda_m$, and $\overline{\boldsymbol{E}} = [\boldsymbol{e}_1 | \cdots | \boldsymbol{e}_m]$.

3) Estimate the number of sources $\hat{d}$.

4) Evaluate

$$P_M(\theta) = \frac{\boldsymbol{a}^*(\theta)\,\boldsymbol{a}(\theta)}{\boldsymbol{a}^*(\theta)\,\boldsymbol{E}_N\boldsymbol{E}_N^*\boldsymbol{a}(\theta)},$$

where $\boldsymbol{E}_N = \boldsymbol{\Sigma}_n[\boldsymbol{e}_{d+1} | \cdots | \boldsymbol{e}_m]$.

5) Find the d (largest) peaks of $\boldsymbol{P}(\theta)$ to obtain estimates of the parameters.

An alternative to first forming the measurement covariance matrix and then performing an eigendecomposition is to operate directly on the measurements using the singular value decomposition (SVD). In addition to avoiding *squaring* the data, this approach has a nice geometric interpretation. Letting $\boldsymbol{X} \in \mathbb{C}^{m\times N}$ denote the data matrix, the objective is to obtain a set of vectors spanning the column space of the rank d matrix, $\hat{\boldsymbol{X}} \in \mathbb{C}^{m\times N}$, that best approximates $\boldsymbol{X}$ in a least-squares sense. The solution is given by the d left singular vectors of $\boldsymbol{X}$ corresponding to the d largest singular values.[12]

It is easily demonstrated that the eigendecomposition and the SVD yield the same subspace estimate. If the SVD of $\boldsymbol{X}/\sqrt{N}$ is given by $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$,

$$\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^* = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^* = \hat{\boldsymbol{R}}_{XX},$$

since $\boldsymbol{\Sigma}$ is diagonal and real, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary. Thus, the left singular vectors, $\boldsymbol{U}$, of $\boldsymbol{X}$ are the right eigenvectors of $\hat{\boldsymbol{R}}_{XX}$, the sample covariance matrix. Thus, in the absence of finite precision effects in computing the decompositions, the subspace estimates from both techniques are identical.

There are, however, significant computational differences. The computation of the full SVD of $\boldsymbol{X}$ is of order mN^2 which can be significantly larger than the $O(m^3)$ operations required for an eigendecomposition of $\hat{\boldsymbol{R}}_{XX}$. The increase in computation is due to the fact that the full SVD is also obtaining a set of d vectors, the first d columns of $\boldsymbol{V} \in \mathbb{C}^{N\times N}$, that span the d-dimensional subspace of N-dimensional space spanned by the d signal vectors, vectors in $\mathbb{C}^N$ whose components are the samples of the underlying signals.[13] If the information of the full SVD is not required, *partial* SVD algorithms that compute only the left singular vectors and singular values can be employed resulting in substantial computational savings [11].

If N is the number of measurements to be processed, and m is the number of elements in each sample vector, forming the sample covariance matrix requires on the order of Nm^2 operations. Eigendecompositions of matrices in $\mathfrak{R}^{m\times m}$ require on the order of $10m^3$ operations, whereas a standard SVD of $\boldsymbol{X} \in \mathfrak{R}^{m\times N}$ requires $\approx 2Nm^2 + 4m^3$ (cf. [2, p. 175]) if only the singular values and left singular vectors are required. The computational effort is therefore on the same order for both the SVD and eigenvector approaches to reducing the measurements to the statistic $\boldsymbol{E}_S$ used by ESPRIT and MUSIC. Again, the major advantage of the SVD over the eigendecomposition is that the measurements are processed directly without *squaring* them. Thus, numerical problems associated with ill-conditioned matrices are mitigated to some extent by using the SVD.

[10]The conventional MUSIC measure is herein referred to as a *one-dimensional* measure, although the search over $\mathfrak{A}$ is potentially a multidimensional one, the dimension being that of the parameter vector (e.g., three for a parameter vector consisting of range, azimuth, and elevation).

[11]If there are repeated eigenvalues, subspace convergence is still guaranteed as long as the subspace contains all the eigenvectors associated with the repeated eigenvalues.

[12]A discussion of this problem can be found in [2, pp. 19–20], although therein the 2-norm is employed. The extension to the Frobenius norm (cf. total least-squared error) is straightforward.

[13]In the presence of noise, these d vectors are biased estimates due to the fact that for each measurement taken, another signal vector must be estimated. Thus, there is no *averaging of the noise*.

IV. ESPRIT

Although MUSIC was the first of the high-resolution algorithms to correctly exploit the underlying data model of narrow-band signals in additive noise, the algorithm has several limitations including the fact that complete knowledge of the array manifold is required, and that the search over parameter space is computationally very expensive. In this section, an approach (ESPRIT) to the signal parameter estimation problem that exploits sensor array invariances is described.[14] ESPRIT is similar to MUSIC in that it correctly exploits the underlying data model, while manifesting significant advantages over MUSIC as described in Section I.

To simplify the description of the basic ideas behind ESPRIT, much of the ensuing discussion is couched in terms of the problem of multiple source direction-of-arrival (DOA) estimation from data collected by an array of sensors. For simplicity, discussions deal only with single-dimensional parameter spaces, e.g., azimuth-only direction finding (DF) of far-field point sources, since the basic concepts are most easily visualized in such spaces. Narrow-band signals of known center frequency will be assumed. Recall that a DOA/DF problem is classified as *narrow-band* if the signal bandwidth is small compared to the inverse of the transit time of a wavefront across the array, and the array response is not a function of frequency over the signal bandwidth. The generality of the fundamental concepts on which ESPRIT is based makes the extension to higher spatial dimensions and to signals containing multiple frequencies possible.

A. Array Geometry

ESPRIT retains most of the essential features of the *arbitrary* array of sensors, but achieves a significant reduction in computational complexity by imposing a constraint on the structure of the sensor array, a constraint most easily described by an example. Consider a planar array of arbitrary geometry composed of m sensor *doublets* as shown in Fig. 3. The elements in each doublet have identical sensitivity patterns and are translationally separated by a known constant displacement vector $\mathbf{\Delta}$. Other than the obvious requirement that each sensor have non-zero sensitivity in all directions of interest, the gain, phase, and polarization sensitivity of the elements in the doublet are arbitrary. Furthermore, there is no requirement that any of the doublets possess the same sensitivity patterns although, as discussed in [1] and [14], there are advantages to employing arrays with such characteristics.

B. The Data Model

Assume that there are $d \leq m$ narrow-band sources[15] centered at frequency ω_0, and that the sources are located sufficiently far from the array such that in homogeneous isotropic transmission media, the wavefronts impinging on the array are planar. As before, the sources may be assumed to be stationary zero-mean random processes or deterministic signals. Additive noise is present at all $2m$ sensors and is assumed to be a stationary zero-mean random process with a *spatial* covariance $\sigma^2\mathbf{\Sigma}_n$.

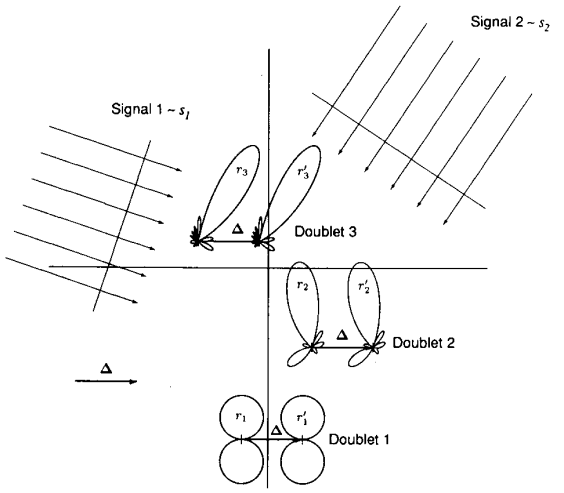


Fig. 3. Sensor array geometry for multiple source DOA estimation using ESPRIT.

To describe mathematically the effect of the *translational invariance* of the sensor array, it is convenient to describe the array as being comprised of two subarrays, Z_X and Z_Y, identical in every respect although physically displaced (not rotated) from each other by a known displacement vector $\mathbf{\Delta}$ of magnitude Δ. The signals received at the ith doublet can then be expressed as

$$x_i(t) = \sum_{k=1}^{d} s_k(t)\, a_i(\theta_k) + n_{x_i}(t),$$

$$y_i(t) = \sum_{k=1}^{d} s_k(t)\, e^{j\omega_0 \Delta \sin\theta_k / c}\, a_i(\theta_k) + n_{y_i}(t), \quad (8)$$

where θ_k is now the direction-of-arrival of the kth source relative to the direction of the translational displacement vector $\mathbf{\Delta}$.

Since the sensor gain and phase patterns are arbitrary and since ESPRIT does not require any knowledge of the sensitivities, the subarray displacement vector $\mathbf{\Delta}$ sets not only the scale for the problem, but the *reference direction* as well. The DOA estimates obtained are angles-of-arrival with respect to the direction of the vector $\mathbf{\Delta}$. A natural consequence of this fact is the necessity for a corresponding displacement vector for each *dimension* in which parameter estimates are desired.

Combining the outputs of each of the sensors in the two subarrays, the received data vectors can be written as fol-

[14] A patent has been issued on the sensor array design and concepts embodied in ESPRIT [12], [13].

[15] MUSIC imposes the requirement $d < 2m$, and can therefore handle roughly twice as many sources as ESPRIT in general. For uniform linear arrays, however, ESPRIT can handle as many sources as MUSIC [1] by employing *overlapping* subarrays.

lows:

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{n}_x(t), \tag{9}$$

$$\boldsymbol{y}(t) = \boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{s}(t) + \boldsymbol{n}_y(t), \tag{10}$$

where the vector $\boldsymbol{s}(t)$ is the $d \times 1$ vector of impinging signals (wavefronts) as observed at the reference sensor of subarray Z_X. The signals can be correlated in the sense that $E\{s_i(t)\, s_j^*(t)\} \neq 0$ for $i \neq j$, although the case of coherent (or fully correlated) sources is not considered herein (cf. [1, Section 7.11] for further discussion). The matrix $\boldsymbol{\Phi}$ is a diagonal $d \times d$ matrix of the phase delays between the doublet sensors for the d wavefronts, and is given by

$$\boldsymbol{\Phi} = \operatorname{diag}\left\{e^{j\gamma_1}, \cdots, e^{j\gamma_d}\right\}, \tag{11}$$

where $\gamma_k = \omega_0 \Delta \sin\theta_k / c$. $\boldsymbol{\Phi}$ is a unitary matrix (operator) that relates the measurements from subarray Z_X to those from subarray Z_Y. In the complex field, $\boldsymbol{\Phi}$ is a simple scaling operator. However, it is isomorphic to the real two-dimensional rotation operator and is herein referred to as a *rotation*[16] operator. The unitary nature of $\boldsymbol{\Phi}$ is a consequence of the narrow-band planewave assumption, an assumption that leads to unit-modulus cisoidal *signals* in the spatial domain. In time series analysis, the diagonal elements of $\boldsymbol{\Phi}$ are potentially arbitrary complex numbers in which case $\boldsymbol{\Phi}$ could be an *expansive* or *contractive* operator.

Defining the total array output vector as $\boldsymbol{z}(t)$, the subarray outputs can be combined to yield

$$\boldsymbol{z}(t) = \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{y}(t) \end{bmatrix} = \bar{\boldsymbol{A}}\boldsymbol{s}(t) + \boldsymbol{n}_z(t), \tag{12}$$

$$\bar{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{A}\boldsymbol{\Phi} \end{bmatrix}, \qquad \boldsymbol{n}_z(t) = \begin{bmatrix} \boldsymbol{n}_z(t) \\ \boldsymbol{n}_y(t) \end{bmatrix}. \tag{13}$$

It is the structure of $\bar{\boldsymbol{A}}$ that is exploited to obtain estimates of the diagonal elements of $\boldsymbol{\Phi}$ *without having to know* $\boldsymbol{A}$.

From (12), it is easily seen that the estimation problem posed is *scale-invariant* in the sense that absolute signal powers are not observable. For any nonsingular diagonal matrix, $\boldsymbol{D}$, the data model is invariant with respect to the transformations $\boldsymbol{s}(t) \rightarrow \boldsymbol{D}^{-1}\boldsymbol{s}(t)$ and $\bar{\boldsymbol{A}} \rightarrow \bar{\boldsymbol{A}}\boldsymbol{D}$. Thus, estimates of the signals and the associated array manifold vectors derived herein are to be interpreted modulo an arbitrary scale factor unless knowledge of the gain pattern of *one* of the sensors is available.

C. ESPRIT—The Invariance Approach

The basic idea behind ESPRIT is to exploit the *rotational* invariance of the underlying signal subspace induced by the translational invariance of the sensor array. The relevant signal subspace is the one that contains the outputs from the two subarrays described above, Z_X and Z_Y. Simultaneous sampling[17] of the output of the arrays leads to two sets of vectors, $\boldsymbol{E}_X$ and $\boldsymbol{E}_Y$, that span the same signal subspace (ideally, that spanned by the columns of $\boldsymbol{A}$).

The ESPRIT algorithm is based on the following results for the case in which the underlying $2m$-dimensional signal subspace containing the entire array output is known. In the absence of noise, the signal subspace can be obtained as before by collecting a sufficient number of measurements and finding *any set of* d *linearly independent measurement vectors*. These vectors span the d-dimensional subspace of $\mathbb{C}^m$ spanned by $\bar{\boldsymbol{A}}$. The signal subspace can also be obtained from knowledge of the covariance of the measurements $\boldsymbol{R}_{ZZ} = \bar{\boldsymbol{A}}\boldsymbol{R}_{SS}\bar{\boldsymbol{A}}^* + \sigma^2\boldsymbol{\Sigma}_n$. As discussed in Section III, if $d \leq m$ (an assumption required in later developments), the $2m - d$ smallest GE's of $(\boldsymbol{R}_{ZZ}, \boldsymbol{\Sigma}_n)$ are equal to σ^2. The d GEV's corresponding to the d largest GE's are used to obtain $\boldsymbol{E}_S = \boldsymbol{\Sigma}_n[\boldsymbol{e}_1 \mid \cdots \mid \boldsymbol{e}_d]$, where $\Re\{\boldsymbol{E}_S\} = \Re\{\bar{\boldsymbol{A}}\}$, the signal subspace.

Since $\Re\{\boldsymbol{E}_S\} = \Re\{\bar{\boldsymbol{A}}\}$, there must exist a unique (recall $d \leq m$), nonsingular $\boldsymbol{T}$ such that

$$\boldsymbol{E}_S = \bar{\boldsymbol{A}}\boldsymbol{T}. \tag{14}$$

Furthermore, the invariance structure of the array implies $\boldsymbol{E}_S$ can be decomposed into $\boldsymbol{E}_X \in \mathbb{C}^{m \times d}$ and $\boldsymbol{E}_Y \in \mathbb{C}^{m \times d}$ (cf. Z_X and Z_Y subarrays) such that

$$\boldsymbol{E}_S = \begin{bmatrix} \boldsymbol{E}_X \\ \boldsymbol{E}_Y \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}\boldsymbol{T} \\ \boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{T} \end{bmatrix} \tag{15}$$

from which it is easily seen that

$$\Re\{\boldsymbol{E}_X\} = \Re\{\boldsymbol{E}_Y\} = \Re\{\boldsymbol{A}\}.$$

Since $\boldsymbol{E}_X$ and $\boldsymbol{E}_Y$ share a common column space, the rank of

$$\boldsymbol{E}_{XY} \overset{\text{def}}{=} [\boldsymbol{E}_X \mid \boldsymbol{E}_Y] \tag{16}$$

is d, which implies there exists a unique (recall $d \leq m$) rank d matrix[18] $\boldsymbol{F} \in \mathbb{C}^{2d \times d}$ such that

$$\boldsymbol{0} = [\boldsymbol{E}_X \mid \boldsymbol{E}_Y]\,\boldsymbol{F} = \boldsymbol{E}_X\boldsymbol{F}_X + \boldsymbol{E}_Y\boldsymbol{F}_Y, \tag{17}$$

$$= \boldsymbol{A}\boldsymbol{T}\boldsymbol{F}_X + \boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{T}\boldsymbol{F}_Y. \tag{18}$$

$\boldsymbol{F}$ spans the *null-space* of $[\boldsymbol{E}_X \mid \boldsymbol{E}_Y]$. Defining

$$\boldsymbol{\Psi} \overset{\text{def}}{=} -\boldsymbol{F}_X[\boldsymbol{F}_Y]^{-1}, \tag{19}$$

[16]This is the origin of *rotational* in the acronym ESPRIT.

[17]In many practical situations, simultaneous sampling is a nontrivial hardware design issue. Although in certain of these situations it is possible to relax this condition to say one of uniform sampling in time instead, it is assumed herein that simultaneous sampling is performed. The extent to which this is not exactly achieved represents errors in the underlying model, errors to which ESPRIT is manifestly less sensitive than all other signal subspace based algorithms.

[18]This derivation, although somewhat more lengthy than at first glance seems necessary, will prove useful when noisy estimates of $\boldsymbol{E}_X$ and $\boldsymbol{E}_Y$ are available.

equation (18) can be rearranged to yield[19]

$$AT\Psi = A\Phi T \Rightarrow AT\Psi T^{-1} = A\Phi. \tag{20}$$

Assuming A to be full rank implies

$$\boxed{T\Psi T^{-1} = \Phi.} \tag{21}$$

Therefore, the eigenvalues of Ψ must be equal to the diagonal elements of Φ, and the columns of T are the eigenvectors of Ψ. This is the key relationship in the development of ESPRIT and its properties. The signal parameters are obtained as nonlinear functions of the eigenvalues of the operator Ψ that maps (rotates) one set of vectors (E_X) that span an m-dimensional signal subspace into another (E_Y).

D. Estimating the Subspace Rotation Operator

In practical situations where only a finite number of noisy measurements are available, E_S is estimated from the covariance matrices of the measurements $\hat{R}_{ZZ}$ or, equivalently, from the data matrix Z. The result is that $\mathcal{R}\{E_S\}$ is only an estimate of $\mathcal{S}_Z$, and with probability one, $\mathcal{R}\{E_S\} \neq \mathcal{R}\{\bar{A}\}$. Furthermore, $\mathcal{R}\{E_X\} \neq \mathcal{R}\{E_Y\}$. Thus, the objective of finding a Ψ such that $E_X\Psi = E_Y$ is no longer achievable.[20] A criterion for obtaining a suitable estimate must be formulated. The most commonly employed criterion for problems of this nature is the *least-squares* (LS) criterion.

The standard LS criterion applied to the model $AX = B$ to obtain an estimate of X assumes A is known and the error is to be attributed to B. Assuming the set of equations is overdetermined, the columns of A are linearly independent, and the noise in the elements of B is zero-mean and $E\{b_{ij}b_{kl}\} = \sigma^2\delta_{ik}\delta_{jl}$, the LS solution is

$$\hat{X} = [A^*A]^{-1}A^*B.$$

It is easily verified that the estimate is unbiased and minimum variance. The extension to arbitrary, but known, covariance of the rows of B is straightforward and leads to the *weighted least-squares* (WLS) solution. If both A and B are *noisy*, however, the LS solution is known to be biased.

Since it is not difficult to argue that the estimates E_X and E_Y are *equally* noisy, the LS criterion is clearly inappropriate. A criterion that takes into account noise on both A and B is the *total least-squares* (TLS) criterion. The TLS criterion can be stated [2] as finding *residual* matrices R_A and R_B of minimum Frobenius norm, and $\hat{X}$ such that

$$[A + R_A]\hat{X} = B + R_B. \tag{22}$$

This criterion is easily shown to be equivalent to replacing the zero matrix in (17) by a matrix of errors the Frobenius norm of which is to be minimized (i.e., *total* least-squared error). If the covariance of the errors, specifically the rows of $[R_A | R_B]$, is known to within a scale factor, the TLS estimate of X is *strongly consistent* [8].

Appending a nontriviality constraint $F^*F = I$ to eliminate the zero solution and applying standard Lagrange techniques leads to a solution for F given by the eigenvectors corresponding to the d smallest eigenvalues of $E_{XY}^*E_{XY}$. The eigenvalues of Ψ as defined above and calculated from the estimates F_X and F_Y are taken as estimates of the diagonal elements of Φ.

E. Summary of the TLS ESPRIT Covariance Algorithm

The TLS ESPRIT algorithm based on a covariance formulation can be summarized as follows.

1) Obtain an estimate of R_{ZZ}, denoted $\hat{R}_{ZZ}$, from the measurements Z.

2) Compute the generalized eigendecomposition of $\{\hat{R}_{ZZ}, \Sigma_n\}$

$$\hat{R}_{ZZ}\bar{E} = \Sigma_n\bar{E}\Lambda,$$

where $\Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_{2m}\}$, $\lambda_1 \geq \cdots \geq \lambda_{2m}$, and $\bar{E} = [e_1 | \cdots | e_{2m}]$.

3) Estimate the number of sources[21] $\hat{d}$.

4) Obtain the signal subspace estimate $\hat{\mathcal{S}}_Z = \mathcal{R}\{E_S\}$, and decompose it to obtain E_X and E_Y, where

$$E_S \overset{\text{def}}{=} \Sigma_n[e_1 | \cdots | e_{\hat{d}}] \Rightarrow \begin{bmatrix} E_X \\ E_Y \end{bmatrix}.$$

5) Compute the eigendecomposition ($\lambda_1 > \cdots > \lambda_{2\hat{d}}$),

$$E_{XY}^*E_{XY} \overset{\text{def}}{=} \begin{bmatrix} E_X^* \\ E_Y^* \end{bmatrix}[E_X | E_Y] = E\Lambda E^*,$$

and partition E into $\hat{d} \times \hat{d}$ submatrices,

$$E \overset{\text{def}}{=} \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}.$$

[19]The same argument used in deriving (14) can be used to derive (20) directly. However, the derivation only ensures the existence and uniqueness of such a full-rank Ψ. The advantage of the preceding derivation is that implicitly a prescription for obtaining Ψ is given. Note that the existence and uniqueness of a full-rank Ψ guarantees the invertibility of F_Y.

[20]Previous implementations of ESPRIT [15], [16] exploited the invariance by finding m-dimensional operators that *best mapped* one of the spanning sets, either E_X or E_Y, into the other using a least-squares (LS) criterion. The objective was to find a $\Psi \in \mathbb{C}^{m\times m}$ such that

$$\Psi E_X = E_Y.$$

Since the problem is underdetermined by construction (cf. typically $d < m$), there is no unique solution, although the parameter estimates, the d eigenvalues of Ψ on the unit circle that are associated with the d-dimensional subspace being *rotated*, are unique. Imposing a minimum norm constraint on Ψ leads to a unique LS solution in which $m - d$ eigenvalues are equal to zero. See [1] for further details.

[21]See [1] and [5] for details on various techniques for estimating the number of sources.

6) Calculate the eigenvalues of $\Psi = -E_{12}E_{22}^{-1}$,

$$\hat{\phi}_k = \lambda_k(-E_{12}E_{22}^{-1}), \qquad \forall k = 1, \ldots, \hat{d}.$$

7) Estimate $\hat{\theta}_k = f^{-1}(\hat{\phi}_k)$; e.g., for DOA estimation, $\hat{\theta}_k = \sin^{-1}\{c \arg(\hat{\phi}_k)/(\omega_0\Delta)\}$.

For arrays with multiple invariances, such as uniform linear arrays, the decomposition of E_S into E_X and E_Y is not unique. See [14] and [17] for more details concerning multiple invariance ESPRIT.

In many instances, it is preferable to avoid forming covariance matrices, and instead to operate directly on the *data* as discussed in Section III. This approach leads to (generalized) singular value decompositions (GSVD's) of data matrices, and a GSVD variant of ESPRIT discussed in detail in [1].

From the key relation (21), several other quite striking results can be derived. For example, not only is knowledge of the array manifold not required, but the elements thereof associated with the estimated signal parameters (DOA's) can be estimated if desired. The same is true of the source correlation matrix, knowledge of which is not needed in ESPRIT.

F. Array Calibration

Using the TLS formulation of ESPRIT, the array manifold vectors associated with each signal (parameter) can be estimated (to within an arbitrary scale factor). From (21), the right eigenvectors of Ψ are given by $E_\Psi = T^{-1}$. This result can be used to obtain estimates of the array manifold vectors as

$$E_S E_\Psi = \bar{A}TT^{-1} = \bar{A}. \tag{23}$$

No assumption concerning the source covariance is required.

Although simple to compute, this estimate will not in general conform to the invariance structure of the array in the presence of noise. In low SNR scenarios, the deviation from the assumed structure $\bar{A} = [A^T|(A\Phi)^T]^T$ may be significant. In such situations, improved estimates of the array manifold vectors can be obtained by employing the formulation discussed in [18].

G. Signal Copy

In many practical applications, not only the signal parameters, but the signals themselves, are of interest. Estimation of the signals as a function of time from an estimated DOA is termed *signal copy*. The basic objective is to obtain estimates $\hat{s}(t)$ of the signals $s(t)$ from the array output, $z(t) = \bar{A}s(t) + n(t)$. Employing a linear estimator, a squared-error cost criterion in the metric of the noise (which is ML if the noise is Gaussian), and conditioning on knowledge of $\bar{A}$, leads to the estimate $\hat{s}(t)$, the vector of coefficients resulting from the *oblique projection* of $z(t)$ onto the space spanned by the columns of $\bar{A}$ (cf. the Appendix). The resulting weight matrix W (i.e., the *linear* estimator) whose ith column is a weight vector that can be used to obtain an estimate of the signal from the ith estimated DOA and reject those from the other DOA's is given by

$$W = \Sigma_n^{-1}\bar{A}[\bar{A}^*\Sigma_n^{-1}\bar{A}]^{-1}. \tag{24}$$

In terms of quantities already available, (24) can be written as

$$W = \Sigma_n^{-1}E_S[E_S^*\Sigma_n^{-1}E_S]^{-1}E_\Psi^*, \tag{25}$$

using (23) to estimate $\bar{A}$. This equivalence is easily established since from (21) it follows that the right eigenvectors of Ψ equal T^{-1}. Combining this fact with $E_S = \bar{A}T$ and substituting in (25) yields

$$\begin{aligned} W^* &= E_\Psi^{-1}[E_S^*\Sigma_n^{-1}E_S]^{-1}E_S^*\Sigma_n^{-1} \\ &= [\bar{A}^*\Sigma_n^{-1}\bar{A}]^{-1}\bar{A}^*\Sigma_n^{-1}. \end{aligned} \tag{26}$$

Note that the *optimal* copy vector is a vector that is Σ_n^{-1} orthogonal to all but one of the vectors in the columns of $\bar{A}$ since $W^*\bar{A} = I$.

There is, of course, a *total least-squares* alternative to conditioning on knowledge of $\bar{A}$. Since only estimates of $\bar{A}$ are available, in low SNR scenarios where accurate signal estimates are desired, the TLS approach yields improved estimates at the cost of increased computation. Although not derived herein, $\hat{s}(t)$ can be obtained by performing a (generalized) singular value decomposition of $[\bar{A}|z(t)]$. The right singular vector corresponding to the *smallest* singular value yields $\hat{s}(t)$ as the first d elements after normalizing the last element to unity.[22]

H. Source Correlation Estimation

There are several approaches that can be used to estimate the source correlations. The most straightforward is to simply note that the *optimal signal copy* matrix W obtained above removes the *spatial correlation* in the observed measurements [cf. (25)]. Thus, $W^*C_{ZZ}W = DSD^*$ where S is the source correlation (not covariance) matrix, $C_{ZZ} = R_{ZZ} - \sigma^2\Sigma_n$, and the diagonal factor D accounts for arbitrary normalization of the columns of W. Note that when R_{ZZ} must be estimated, a manifestly rank d estimate $\hat{C}_{ZZ} = E_S[\Lambda_S^{(d)} - \hat{\sigma}^2 I_d]E_S^*$ can be used [cf. (6)], where $\Lambda_S^{(d)} = \text{diag}\{\lambda_1, \cdots, \lambda_d\}$ and λ_i is a GE of (R_{ZZ}, Σ_n). Combining this with $E_S = \bar{A}T$ gives

$$DSD^* = T[\Lambda_S^{(d)} - \hat{\sigma}^2 I_d]T^*. \tag{27}$$

If a gain pattern for one of the elements is known, specifically if the gain $|a_1(\theta_k)|$ is known for all θ_k associated with sources whose power is to be estimated, then source power estimation is possible since the array manifold vectors can now be obtained with proper scaling.

V. Simulation Results

Many simulations have been conducted exploring different aspects of ESPRIT and making comparisons to

[22]This approach is clearly suboptimal if the sampled signals are temporally correlated in the sense that $E\{s_i(t)s_i(t+\tau)\} \neq 0$ for $\tau \neq 0$. If, for example, the signals are known to be sinusoidally modulated RF and uniform temporal sampling is employed, then estimating the underlying signals requires only the estimation of the modulation frequency, another problem well suited to ESPRIT. Note that in general the modulation frequency must be a small fraction of the carrier to satisfy the narrow-band assumption.

other techniques (cf. [1]). Herein, only one of the scenarios, but one that addresses several issues that arise in a practical implementation of ESPRIT, is presented. Thus, sensor gain and phase errors, as well as sensor spacing errors, are included. Furthermore, unequal source powers and a high degree of source correlation are assumed.

More specifically, the array chosen was a ten-element array with doublet spacing $\lambda/4$ and the five doublets *randomly spaced* on a line resulting in an aperture of approximately 4λ. Two sources were located at 24° and 28° (approximately 0.3 Rayleigh or 3 dB beamwidth separation), and were of unequal powers, 20 dB and 15 dB, respectively. Sensor errors were introduced by zero-mean normal random additive errors with sigmas of 0.1 dB in amplitude and 2° in phase[23] (independent of angle). Sensor location errors (along the axis of the array) with sigma 0.005 ($\lambda/2$) were included as well. The sources were 90 percent temporally correlated and 5000 trials were run. A histogram of the results is given in Fig. 4.

The number of sources was assumed to be known in the implementation of both MUSIC and ESPRIT. The indicated failure rate for MUSIC of 37 percent is the percentage of trials in which the conventional MUSIC spectrum did not exhibit two peaks in the interval [20°, 32°]. This, of course, is not an issue in ESPRIT, where two parameter estimates are obtained every time. The sample means and sigmas of the ESPRIT estimates were 23.93° ± 1.07° and 28.06° ± 1.37°, while those of the 3175 *successful* MUSIC trials were 24.35° ± 0.28° and 27.48° ± 0.38°. Note that with reference to Fig. 4, there is an overlap in the distributions of the ESPRIT estimates. This has an effect on the statistics calculated, since a simple *angle-ordering* scheme was used wherein the larger of the two angle estimates in each trial was associated with the 28° source.[24] The effect is presumed to be small in this case.

The results indicate the presence of a bias even in the *successful* conventional MUSIC estimates, the source of which is described in detail in [1]. On the other hand, the ESPRIT estimates are unbiased, although of larger variance since less information concerning the array geometry is being utilized. Note also that in comparing the estimate variances, there is no attempt to account for the 1825 trials in which MUSIC failed[25] to provide two DOA estimates! However, as the subarray separation increases, the ESPRIT parameter estimate variances approach those of MUSIC. The same experiment was run for a subarray separation of 4λ, and the resulting ESPRIT estimates were 24.003° ± 0.062° and 28.002° ± 0.089°, and the corresponding MUSIC estimates were 24.011° ± 0.056° and 27.986° ± 0.078°. Due to the increased subarray spacing (array aperture), there were no MUSIC failures. Again, ESPRIT is unbiased, but now the sample parameter estimate sigmas are nearly equal to those obtained with MUSIC. More details concerning these and many other simulations can be found in [1] (see also [18]).

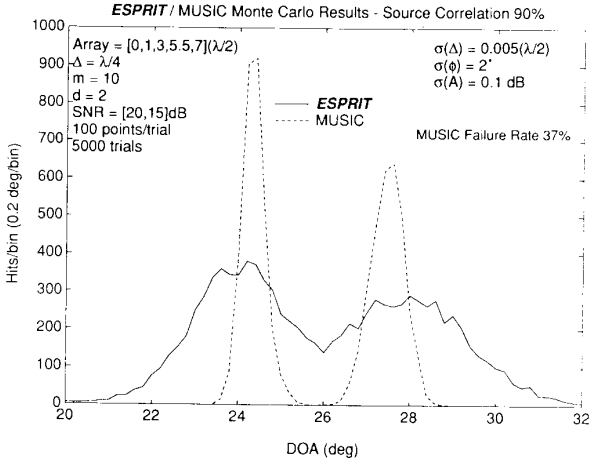


Fig. 4. Histogram of MUSIC and ESPRIT results—random 10-element linear array, source correlation 90 percent, small array aperture ($\Delta = \lambda/4$).

VI. Discussion

In this paper, a new technique (ESPRIT) for signal parameter estimation has been introduced. The algorithm differs from its predecessors in that a *total least-squares* rather than a standard least-squares (LS) criterion is employed. The earlier versions of ESPRIT described in [15] and [16] can be seen [1] to be least-squares estimators of an $m \times m$ operator whose action is restricted to a d-dimensional subspace. The fact that this LS operator is a restricted $m \times m$ operator leads to some concern over potential numerical difficulties in solving the generalized eigenproblem. Imposing the subspace restriction as a constraint prior to solving the generalized eigenproblem leads to a well-conditioned $d \times d$ generalized eigenproblem, thereby mitigating these numerical concerns; however, the least-squares property of the estimate is retained. In cases where the SNR is sufficiently large, the difference between the LS and TLS parameter estimates is small. The difference is notable at low SNR's, however; the LS estimates are biased as predicted, while the TLS estimates are relatively unbiased and have been shown [8] to be strongly consistent (converge *with probability one* to the true values).

A. ESPRIT and Null-Steering

Many of the previous high-resolution parameter estimation techniques are based on *steering beams* toward signal directions and, in some cases, simultaneously at-

[23]In DOA estimation applications, errors of these magnitudes are considered small; however, it is apparently possible to construct arrays meeting these specifications.

[24]This does not imply that the statistics are compiled by *splitting the histogram* down the middle and computing the center of mass and second moments of the truncated distributions.

[25]Unfortunately, this is often referred to as a failure of MUSIC to *resolve* the two sources. In fact, the *detection* of the number of sources was made in a prior stage of the algorithm. The failure of the conventional MUSIC spectrum (measure) to provide the appropriate number of *peaks* (estimates) indicates only the inappropriateness of the measure! Having *detected* d sources, an algorithm for finding d estimates *simultaneously* is *appropriate*.

tempting to otherwise minimize the power in some weighted combination of sensor outputs. Parameter estimates are associated with DOA's at which peaks in the power output occur. Although intuitively appealing at first glance, there is a much more powerful alternative philosophy, that of *null-steering*. It is well known that deep, sharp notches in directivity patterns and filter gain functions are much easier to achieve than sharp peaks. Interferometers exploit this fact to obtain accurate estimates of source parameters by finding the relative phase required to cancel signal components in two channels. In this context, ESPRIT can be interpreted as a *multidimensional null-steering* parameter estimation algorithm. Calculation of the eigenvalues of the (rotation) operator Ψ, which are the roots of its characteristic polynomial, can be interpreted as multidimensional null-steering. Instead of steering broad beams, ESPRIT steers sharp nulls at all sources simultaneously and does so without relying on knowledge of the array manifold!

B. Computational Advantages of ESPRIT

The primary computational advantage of ESPRIT is that it eliminates the search procedure inherent in all previous methods (ML, ME, MUSIC). ESPRIT produces signal parameter estimates directly in terms of (generalized) eigenvalues. As noted previously, this involves computations of the order d^3. On the other hand, MUSIC and the other high-resolution techniques require a search over $\mathfrak{A}$, and it is this search that is computationally expensive. The significant computational advantage of ESPRIT becomes even more pronounced in multidimensional parameter estimation where the computational load grows linearly with dimension in ESPRIT, while that of MUSIC grows exponentially. If r_l is the resolution (i.e., number of vectors) required in the calibration of $\mathfrak{A}$ for the lth dimension in $\boldsymbol{\Theta}$, the computation required to search over L dimensions for d parameter vectors is proportional to $\Pi_{l=1}^{L} r_l$. For $r_l = r$, the computational load is r^L.

Appendix
The Maximum Likelihood Estimator

For the class of problems considered herein, the maximum likelihood estimator is simple to derive analytically although, in most practical real-time applications, computationally prohibitive. For nonrandom signals in Gaussian noise with covariance $\boldsymbol{\Sigma}_n$, $z(t) = \boldsymbol{A}(\boldsymbol{\theta})\, s(t) + \boldsymbol{n}(t)$, the likelihood function is easily written [5], [1]:

$$\mathfrak{L}[z(t)] = -\ln\left[P\{z(t)|z(t) = \boldsymbol{A}(\boldsymbol{\theta})\, s(t) + \boldsymbol{n}\}\right] \tag{28}$$

$$\propto -[z(t) - \boldsymbol{A}(\boldsymbol{\theta})\, s(t)]^* \cdot \boldsymbol{\Sigma}_n^{-1}[z(t) - \boldsymbol{A}(\boldsymbol{\theta})\, s(t)]. \tag{29}$$

The maximization of $\mathfrak{L}$ is over $\{s(t); t \in [0, N]\} \times \{\boldsymbol{\theta} \in \Theta\}$, and is therefore a nonlinear optimization problem. It belongs, however, to the class of *separable* nonlinear optimization problems. Golub and Pereyra [19] prove that the optimization can be carried out in two steps. A solution for the optimal $s(t)$ is sought as a function of $\boldsymbol{\theta}$, then the maximization over $\boldsymbol{\theta}$ is performed. Employing this procedure gives $\hat{s}(t) = \boldsymbol{w}^*(\boldsymbol{\theta})z(t)$, where $\boldsymbol{w}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_n^{-1}\boldsymbol{A}(\boldsymbol{\theta})[\boldsymbol{A}^*(\boldsymbol{\theta})\,\boldsymbol{\Sigma}_n^{-1}\boldsymbol{A}(\boldsymbol{\theta})]^{-1}$. Substituting the expression for $s(t)$ back into (29) and using standard properties of the trace operator,

$$\mathfrak{L}(\boldsymbol{\theta}) \propto -\operatorname{Tr}\left\{P_{A^\perp(\theta)}\, \boldsymbol{R}_{ZZ}\boldsymbol{\Sigma}_n^{-1}\right\}, \tag{30}$$

where $P_{A^\perp(\theta)}$ is the *oblique* projection operator onto the complement of the space spanned by $\boldsymbol{A}(\boldsymbol{\theta})$ (in the metric $\boldsymbol{\Sigma}_n$). Maximization of this criterion is equivalent to finding

$$\max_{\boldsymbol{\theta}} \operatorname{Tr}\left\{P_{A(\theta)}\boldsymbol{R}_{ZZ}\boldsymbol{\Sigma}_n^{-1}\right\}, \tag{31}$$

as can be easily verified.[26] Although easy to describe analytically, the computational burden of actually carrying out the multidimensional projection and maximization over $\boldsymbol{\theta}$ is generally prohibitive, resulting in the need for *reasonable* approximate solutions such as MUSIC and ESPRIT.

Acknowledgment

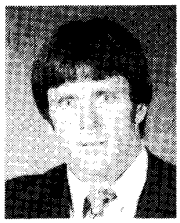
The authors would like to express their gratitude to the referees for their time and effort in carefully reviewing the preliminary manuscript. The authors also thank J. Speiser of NOSC and Prof. G. Golub of Stanford who suggested the potential application of TLS concepts to this problem. The incorporation of their suggestions and comments significantly improved the clarity of the final manuscript.

References

[1] R. H. Roy, "ESPRIT—Estimation of signal parameters via rotational invariance techniques," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1987.
[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins University Press, 1984.
[3] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1971.
[4] R. O. Schmidt, "A signal subspace approach to multiple emitter location and spectral estimation," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1981.
[5] M. Wax, "Detection and estimation of superimposed signals," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1985.
[6] D. H. Johnson and S. R. DeGraaf, "Improving the resolution of bearing in passive sonar arrays by eigenvalue analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 638–647, Aug. 1982.
[7] A. J. Barabell, J. Capon, D. F. Delong, J. R. Johnson, and K. Senne, "Performance comparison of superresolution array processing algorithms," Tech. Rep. TST-72, Lincoln Lab., M.I.T., 1984.
[8] L. J. Gleser, "Estimation in a multivariate 'errors in variables' regression model: Large sample results," *The Annals of Statistics*, vol. 9, pp. 22–44, 1981.
[9] T. W. Anderson, "Asymptotic theory for principal component analysis," *Ann. Math. Statist.*, vol. 34, pp. 122–148, 1963.
[10] ——, *An Introduction to Multivariate Statistics Analysis*, 2nd ed. New York: Wiley, 1985.
[11] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, *LINPACK Users' Guide*. Philadelphia, PA: SIAM, 1979.
[12] A. Paulraj, R. Roy, and T. Kailath, "Methods and means for signal reception and parameter estimation," Stanford Univ., Stanford, CA, 1985, Patent Application.

[26] This development assumes that the number of signals is known. When the number of signals is not known *a priori*, the maximum likelihood estimator must be redefined [5], [1].

[13] R. Roy, A. Paulraj, and T. Kailath, "Method for estimating signal source locations and signal parameters using an array of signal sensor pairs," U.S. Patent, 4 750 147, June 7, 1988.
[14] R. Roy, B. Ottersten, L. Swindlehurst, and T. Kailath, "Multiple invariance ESPRIT," in *Proc. 22nd Asilomar Conf. Signals, Syst., Comput.*, Asilomar, CA, Nov. 1988.
[15] R. Roy, A. Paulraj, and T. Kailath, "ESPRIT—A subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1340–1342, Oct. 1986.
[16] A. Paulraj, R. Roy, and T. Kailath, "A subspace rotation approach to signal parameter estimation," *Proc. IEEE*, pp. 1044–1045, July 1986.
[17] R. Roy, B. Ottersten, L. Swindlehurst, and T. Kailath, "Multiple invariance ESPRIT," *IEEE Trans. Acoust., Speech, Signal Processing*, in preparation.
[18] R. Roy and T. Kailath, "ESPRIT—Estimation of signal parameters via rotational invariance techniques," *Opt. Eng.*, 1988, in review.
[19] G. H. Golub and A. Pereyra, "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate," *SIAM J. Numer. Anal.*, vol. 10, pp. 413–432, 1973.



Richard Roy was born in Los Angeles, CA, in 1951. He received the B.S. degree in physics and the B.S.E.E. degree in 1972 from the Massachusetts Institute of Technology, and the M.S. degree in physics, M.S.E.E. degree, and Ph.D.E.E. degree in 1974, 1976, and 1987, respectively, all from Stanford University.

From 1975 to 1984 he was employed by ESL, Inc., as a senior member of the Technical Staff involved in research and development of techniques for information processing applied to intelligence data analysis. From 1983 to 1985 he was also employed as a member of the Technical Staff of a start-up company in the oil-well services business and was involved in research into and development of algorithms for real-time signal processing and information extraction from data collected in the oil-well drilling environment. As a Research Scientist for Integrated Systems, Inc. from 1985 to 1987, he was involved in the development of estimation, identification, and adaptive control algorithms for various aerospace applications, including helicopter vibration control and real-time flutter analysis among others. He is currently a Research Associate with the Information Systems Laboratory at Stanford University, and a consultant to various firms involved in statistical signal processing and systems analysis, design, and control. His current research interests are in the area of multidimensional statistical signal processing, system identification and control, and in the VLSI design and implementation of associated algorithms for real-time applications.

Thomas Kailath (S'57–M'62–F'70) was educated in Poona, India, and received the S.M. and Sc.D. degrees in 1959 and 1961, respectively, from the Massachusetts Institute of Technology.

After a year at the Jet Propulsion Laboratories, Pasadena, CA, he joined Stanford University as an Associate Professor of Electrical Engineering in 1963. He was promoted to Full Professor in 1968, served as Director of the Information Systems Laboratory from 1971 through 1980, as Associate Department Chairman from 1981 to 1987, and currently holds the Hitachi America Professorship in Engineering. He has worked in a number of areas including information and communication theory, signal processing, linear systems, linear algebra, operator theory, and control theory; his recent research interests include array processing, fast algorithms for nonstationary signal processing, and the design of special purpose computing arrays. He is the author of *Linear Systems* (Englewood Cliffs, NJ: Prentice-Hall, 1980), and *Lectures on Wiener and Kalman Filtering* (New York: Springer-Verlag, 1981).

Dr. Kailath has held Guggenheim and Churchill fellowships, among others, and received awards from the IEEE Information Theory Group, the IEEE Signal Processing Society, and the American Control Council. He is a Fellow of the Institute of Mathematical Statistics and a member of the National Academy of Engineering.