

ADA 081 736

14
RADC-TR-79-63

In-House Report
October 1979

6
P.S.
REL



**PROCEEDINGS OF THE RADC
SPECTRUM ESTIMATION WORKSHOP (2nd)**

held 3, 4, & 5 October 1979

Griffiss AFB, NY.

11 Oct 79

12 300

**DTIC
ELECTE**
MAR 1 2 1980

A

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

DDC FILE COPY

**ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, New York 13441**

80 3 12 001
309 050

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-79-63 has been reviewed and is approved for publication.

APPROVED: 
PAUL VAN ETTEN
Project Engineer

APPROVED: 
FRANK J. REHM
Technical Director
Surveillance Division

FOR THE COMMANDER: 
JOHN P. HUSS
Acting Chief, Plans Office

Do not return this copy. Retain or destroy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-79-63	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PROCEEDINGS OF THE RADC SPECTRUM ESTIMATION WORKSHOP		5. TYPE OF REPORT & PERIOD COVERED In-House Report
		6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) Multiple		8. CONTRACT OR GRANT NUMBER(s) N/A
9. PERFORMING ORGANIZATION NAME AND ADDRESS Rome Air Development Center (OCTS) Griffiss AFB NY 13441		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E. 62702F
11. CONTROLLING OFFICE NAME AND ADDRESS Same		12. REPORT DATE October 1979
		13. NUMBER OF PAGES 301
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Same		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) Same		
18. SUPPLEMENTARY NOTES RADC Project Engineer: Paul Van Etten (OCTS)		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Signal Processing Spectrum Estimation		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This is the second Spectrum Estimation Workshop sponsored by RADC to provide a means for key researchers in the field to describe their work and also provide a means for comparing the work of various researchers using a common data base for representative problems of importance to the Air Force. This report is a collection of papers that were submitted for presentation at RADC's Spectrum Estimation Workshop held 3, 4, and 5 October 1979 at Griffiss Air Force Base, NY 13441. The papers were published as received by RADC and have not been edited. Further, publication of these papers does not represent approval or		

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

DTIC AL

~~DDC~~. #A054650.

7

they fall

etermined by the

Accession For	
NTIS (GAAI	☒
DND TAB	☐
Unannounced	☐
J. Location	
P.	
J. Location/	
J. Codes	
list	Arrival and/or special
A	

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

RADC's SPECTRUM ESTIMATION WORKSHOP

AGENDA

3, 4, and 5 October 1979

Page

Wednesday, 3 October

0830-0930	Registration	
0935	Welcome	
	Administrative Announcements Clarence Silfer (Co-chairman)	
0940	An Introduction to the Second RADC Spectral Estimation Workshop, Lester A. Gerhardt (Co-chairman), Rensselaer Polytechnic Institute	1
Session I	Dr. Henry Radoski, AFOSR	
1000	Minimum Cross-Entropy Spectral Analysis--Introduction and Examples, John E. Shore and Rodney W. Johnson, Naval Research Laboratory	7
1020	Coffee Break	
1100	Optimal Estimation for Bandlimited, Time-Concentrated Signals, D. P. Kolba and T. W. Parks, Rice University	23
1120	The Use of Linear Prediction for the Interpolation and Extrapolation of Missing Data and Data Gaps Prior to Spectral Analysis, Stephen B. Bowling and Shu Lai, Massachusetts Institute of Technology Lincoln Laboratory	39
1140	A New Autoregressive Spectrum Analysis Algorithm, Larry Marple, Advent Systems, Inc.	51
1200	Adjourn for lunch	
Session II	Dr. Sherman Karp, DARPA	
1330	ARMA Spectral Estimation: An Iterative Procedure, James A. Cadzow, Virginia Polytechnic Institute and State University	67

	ARMA Spectral Estimation: An Efficient Closed-Form Procedure, James A. Cadzow, Virginia Polytechnic Institute and State University	81
1350	Extrapolating Bandlimited Signals with Noise and Quantization, Kenneth Abeud and Judith R. Platt, RCA Government Systems Division	99
1410	Accuracy of Spectral Estimates of Band-Limited Signals, William B. Gordon, Naval Research Laboratory	117
1430	Coffee Break	
1500	Compensation of Autoregressive Spectral Estimates for the Presence of White Observation Noise, Steven Kay, Raytheon Company	127
1530	Order Determination for Autoregressive Spectral Estimation, M. Kaveh and S. P. Bruzzone, University of Minnesota	139
1600	Adjourn for the day	
Thursday, 4 October		
Session III	Dr. Donald Burlage, U.S.A. R&D Missile Command	
0900	Difficulties Present in Algorithms for Determining the Rank and Proper Poles with Prony's Method, Michael L. VanBlaricum, ETI, Incorporated	147
0920	A Unifying Model for Spectral Estimation, Charles Byrne, The Catholic University of America and Raymond Fitzgerald, Naval Research Laboratory	157
0940	A Comparison of the Burg and the Known-Auto-correlation Autoregressive Spectral Analysis of Complex Sinusoidal Signals In Additive White Noise, Robert W. Herring, Communications Research Centre	163
1000	Coffee Break	
1040	A Two-Dimensional Maximum Entropy Spectral Estimator, Salim Roucos and D. G. Childers, College of Engineering, University of Florida	179
1100	Spectral Estimation and Signal Extrapolation in One and Two Dimensions, Anil K. Jain, University of California	195

1120	Antenna Spatial Pattern Viewpoint of MEM, MLM, and Adaptive Array Resolution, William F. Gabriel, Naval Research Laboratory	215
1200	Lunch	
Session IV	Dr. David Kerr, NRL	
1330	Aperture Sampling Processing for Ground Reflection Elevation Multipath Characterization, James E. Evans and David F. Sun, MIT Lincoln Laboratory	229
1350	Multiple Emitter Location and Signal Parameter Estimation, Ralph O. Schmidt, ESL, Incorporated	243
1410	The Maximum Entropy Spectral Estimator Used as a Radar Doppler Processor, Simon Haykin and Hing C. Chan, McMaster University, Ontario, Canada	259
1430	Coffee Break	
1500	Applications for MESA and the Prediction Error Filter, William R. King, King Research	273
1520	The Maximum Entropy Method Applied to Radar Adaptive Doppler Filtering, J. H. Sawyers, Hughes Aircraft Company	289
1540	Complex Maximum Power Spectral Analysis of AFGL Magnetometer Data, Fougere, AFGL. (Paper not published)	

Friday, 5 October

0900	A Comparison of Solutions to the Workshop Problems, Dr. Lester A. Gerhardt (Co-chairman), Rensselaer Polytechnic Institute
1030	Coffee Break
1100	Workshop Panel Activity
1230	Adjourn the Workshop

PREFACE

This is the second Spectrum Estimation Workshop sponsored by RADC to provide a means for key researchers in the field to describe their work and also provide a means for comparing the work of various researchers using a common data base for representative problems of importance to the Air Force. This report is a collection of papers that were submitted for presentation at RADC's Spectrum Estimation Workshop held 3, 4, and 5 October 1979 at Griffiss Air Force Base, NY 13441. The papers were published as received by RADC and have not been edited. Further, publication of these papers does not represent approval or endorsement by the Rome Air Development Center or the U. S. Air Force.

Proceedings of the first workshop are available from DDC, #A054650.

Participants were also presented with a set of sample problems called the Spectral Estimation Experiment. The object of this experiment was to establish a basis for comparison of the wide variety of techniques available as a function of selected applications on both real and artificial data sets representing specialized problem classes which are of interest to the government. The common data base offers several additional advantages.

Four problems have been formulated by the workshop committee. They fall generally into the areas of radar, pattern recognition and system identification.

The detailed description of the problem and the solutions as determined by the many different algorithms employed will be published separately.

SPECTRUM ESTIMATION WORKSHOP COMMITTEE

1. Russel Brown (RADC/OCTS)
2. Edward Christopher (RADC/OCTS)
3. Lester Gerhardt (RPI/Co-chairman)
4. Clarence Silfer (RADC/OCTS/Co-chairman)
5. Paul Van Etten (RADC/OCTS)
6. Haywood Webb (RADC/ISCP)

AN INTRODUCTION TO THE SECOND RADC
SPECTRAL ESTIMATION WORKSHOP

LESTER A. GERHARDT

Professor and Chairman
Electrical and Systems Engineering Department
Rensselaer Polytechnic Institute
Troy, NY 12181

Introduction

As Co-Chairman of this Second RADC Spectral Estimation Workshop, it is a honor and pleasure for me to welcome all of you to this gathering. Being Co-Chairman of the First Workshop held last year, I have been privileged to watch this field of Spectral Estimation as it has emerged from its newly founded embryonic stage in the first workshop to one of increased maturity this year. Many previously diverse approaches and fields have coalesced and common mathematical tools as well as problems identified. (This has been further aided by other workshops such as the one scheduled for January 1980 at Arden House, Harriman, NY sponsored jointly by the IEEE and Geophysical Society.) Still some major problems remain, such as the determination of model order, and no single technique has been clearly identified as being superior. However, new techniques and algorithms are still sought, and the current Workshop offers you some of them. It also presents new and a broader range of application papers with concrete results. Finally, another comparison of methods will be made using a set of representative problems utilizing a common data base.

It should be noted that the First Workshop was for many a jumping off point to the field and served as a means of focusing attention to this class of spectral estimation problems. Since that time, many of the authors have been engaged in actively sponsored government, industrial and academic research, hardware systems using MEM are being reduced to hardware/firmware, and published papers have substantially increased relating to this subject area. Moreover, there have been other publications such as the IEEE Press publication of Modern Spectrum analysis methods, which have helped in identifying the field as significant. The Workshop Committee would like to think that our first get together aided in the increased interest shown in the field over the last year and take this opportunity to thank you for your involvement and contribution towards that end.

Last year, my paper included a brief mathematical development of each of the major techniques. This year this seems unnecessary and it should be sufficient to summarize the papers grouping some common aspects and identifying emerging trends, leaving the detailed accomplishments to the authors themselves.

Summary of Technical Papers

Twenty-one papers form the basis of the technical sessions that follow. There are eight papers authored by personnel from university or university related organizations, seven from industry, five from government, and one jointly offered from university and government. The majority of government papers originate from the Naval Research Laboratories (NRL). Compared to last year, this represents an increase in the percentage of papers originating from industry, perhaps indicative of a trend of the spectral estimation techniques towards practical implementation.

As one reviews the papers in detail, some major common threads or themes appear, and some general observations emerge.

The technical content of the papers falls broadly into the two categories of theory and application, with many papers incorporating both. At the theoretical end, the first paper by Shore and Johnson is an excellent treatment of Cross Entropy Spectral Analysis. This approach is useful to estimate power spectra given a priori estimates of the spectra and new information in the form of autocorrelation function samples, and reduces to maximum entropy spectral analysis in special cases. It should help expose unfamiliar users to the effectiveness and applicability of this approach. The next two papers are concerned with treating a limited number of discrete time domain samples available. The paper by Kolba and Park develops an implementation of a recursive estimation procedure by minimizing the maximum error given a limited number of time samples but with a priori knowledge of the bandwidth and time duration of the signal (time concentrated signals); whereas the paper by Bowling and Lai describes a linear prediction method to interpolate and extrapolate missing data using a spectrally consistent estimate - this done prior to spectral analysis. The latter paper considers applications to radar for both real and simulated data. The next paper by Marple, also is concerned with radar as the area of application, but is in major part a theoretical treatment of a new autoregressive algorithm for spectral estimation using a least squares approach which all but eliminates line splitting, a problem cited several times at last year's Workshop. Jim Cadzow, one of several principle researchers in this field in recent years, next offers an ARMA autocorrelation estimator method (AEM) for rational spectral density estimation which permits the use of poles and zeroes yielding a more robust procedure. Following this is the paper by Abend and Platt which extends Cadzow's method (presented at the 1978 Workshop) using an iterative steepest descent method to invert the sometimes ill-conditioned matrix, and treats the problem of noise and quantization.

The paper by Gordon concentrates on the topic of accuracy of spectral estimates and particularly discusses the effects of noise. Considering a bandlimited signal in additive white noise, Gordon analyzes the mean squared error of the linear spectral estimate as a function of the time bandwidth

product and signal to noise ratio. The next paper by Kay is also directed at noise problems and takes the approach of compensation of autoregressive spectral estimates when imbedded in white noise. Kay assumes the noise variance is known and with that compares his compensation technique with the ARMA approach to handling noise.

The last two papers in the basic theory group are directed at the classical problem of order determination. Kaveh and Bruzzone use Akaike's Information Criterion (AIC) to determine the order of the autoregressive spectral estimate, while Van Blaricum offers the eigenvalue method and the HFTI method as means to determine the order (or alternatively rank and poles) associated with Prony's method. This still remains a difficult and open problem with no overall solution to the selection of optimum cutoff or to how the effects of noise may be handled.

The next two papers are almost exclusively directed at developing a unifying theoretical base or model, or to a comparison of techniques. The paper by Byrne and Fitzgerald stress comparison of techniques including the implicit models of Cadzow and Figueiredo among others and attempts to establish that these models are related and in fact covered by their unifying model, which serves as both a minimum energy extrapolation and least mean square approximation of the spectrum. It serves to establish the trend of searching for commonality in approaches. This is further enhanced by Herring's paper which compares Burg's method with known autocorrelation autoregressive spectral analysis in white noise.

The two papers that follow, the first by Childers and Roucos, and the second by Jain, both deal with 2-D spectral estimation and continue a theme begun at the first Workshop. The first paper develops a 2-D estimation algorithm whereas Jain's paper deals with one and two dimensional estimation. Jain uses a minimum norm least square (MNLS) formulation and treats the problem appropriately as iterative matrix inversion using a gradient approach.

The remaining seven papers are primarily applications oriented, although include significant theoretical developments as well. The first three of this set of application papers focus on the use of spatial information. Gabriel compares MEM and MIM spectral estimation methods to their adaptive array counterparts popular in the adaptive antenna array field, both being recognized as a matrix inversion problem. This similarity reinforces points made in my introductory talk last year, and serves to bring these fields closer together. The paper by Sun and Evans is directed at multipath, and suggests the use of aperture sampling to improve angular resolution and tracking. It also applies the high resolution MLM and MEM methods to spatial data. Finally, in this category, is the paper by Schmidt which deals with multiple emitter location and again deals with a spatially distributed application.

The last four applications are quite diversified and demonstrate the increased breadth of spectral estimation. Last year for example, the applications were almost solely radar oriented. This year aside from the spatially distributed applications above, the paper by Haykin and Chan uses maximum entropy estimation to function as a Doppler processor. They show that MEM is only slightly suboptimal to conventional processors using the DFT for additive white noise, but for additive clutter (narrow spectral bandwidth) MEM is much better for low Doppler targets. The paper by King discusses several applications for MESA and his prediction error filter including clutter reduction, signal detection, etc. Sawyers paper applies the MEM to adaptive digital filtering directly tying these two fields, and finally Fougere's paper deals with estimating the dominant frequencies and polarization patterns of magnetic pulsation events using both linear and nonlinear methods which are also claimed to eliminate line splitting and shifting.

General Comments

From the previous descriptions, it should be apparent that there continues to be new theoretical developments, as well as advances on previously developed techniques to improve accuracy, reduce the effect of noise, and yield improved computational capability. Two main problems of order determination and line splitting are addressed again this year. However, in at least the two papers by Fougere and Marple, the line splitting problem is fairly well resolved whereas the establishing of the model order needed (degrees of freedom, poles, etc.) as exemplified in the papers by Van Blaricum and Kaveh remains a substantial difficulty. A mix of linear and nonlinear techniques continue also, with no one method cited as clearly superior. Work on extensions to two dimensional processing remains active and of substantial interest.

In an effort to provide some motivation in other directions, (or stimulate controversy or opposition as the case may be), let me say that I feel there is still too much emphasis on treating additive white noise exclusively. This is exemplified in papers by Gordon, Kay, Herring among others. Noting the substantial differences in performance of the MEM obtained in the paper by Haykin and Chan for white noise vs. narrowband noise, certainly some work is needed to explore the effects of non-spectrally flat noise. There also remains a continued special concentration of mean squared error criterion such as in Gordon, Marple, and Jain. Although very effective perhaps other criteria warrant study

The reader should carefully observe that representation continues from not only government, industry, and the university, but also from electrical engineering, geophysics, etc. It is rewarding to see, nonetheless, that several papers, more than in the past year, now are strongly interrelating the fundamental problems of spectral estimation with those of other fields,

a suggestion made by myself and others last year. For example, the papers by Gabriel, Bowling and Lai, Abend and Platt, and Kolba and Park all cover to one extent or another the relation between spectral estimation and adaptive techniques, including gradient methods for iteratively inverting a matrix etc. In many more, there is a solid appreciation of the basic nature of the matrix inversion problem as it dominates the field of spectral estimation.

Insofar as applications papers are concerned, there is an integral mix of theory and applications, a broader diversification of applications than before, and more concrete results. The effectiveness of the techniques, will best be measured in the comparison of results on the common data sets provided to all participants, a discussion better left to another day.

In conclusion, let me restate my welcome to the authors including those returning for a repeat performance and those here for the first time, and to the general audience. I trust the forthcoming technical sessions and subsequent comparative problem sessions will be as fruitful and rewarding to you as it has been to the committee members and myself in helping to prepare them.

6-BLANK

MINIMUM CROSS-ENTROPY SPECTRAL ANALYSIS --- INTRODUCTION AND EXAMPLES

JOHN E. SHORE
RODNEY W. JOHNSON

Naval Research Laboratory
Washington, D.C. 20375

Abstract

The principle of minimum cross entropy (minimum directed divergence) is summarized, discussed, and applied to the classical problem of estimating power spectra given samples of the autocorrelation function. This new approach reduces to maximum-entropy spectral analysis (MESA) in certain special cases but, in contrast to MESA, permits use of a prior estimate of the power spectrum. Examples of applications are given.

1. Introduction

Work reported in [1]-[2] showed that the principle of minimum cross-entropy (minimum directed divergence) provides a correct, general method of inductive inference in terms of continuous probability densities when given a prior density and information about the "true" density in the form of expected values. Subsequent work [3] showed how cross-entropy minimization can be used to estimate power spectra when given a prior estimate of the spectrum and new information in the form of autocorrelation function samples. This new technique reduces to maximum entropy spectral analysis [4]-[5] in certain special cases. In this paper we summarize the new technique and we give examples of its application.

2. Cross-Entropy Minimization

Let $\underline{x}$ denote a single state of some system that has a set D of possible system states and a probability density $q^*(\underline{x})$ of states. Let $\mathcal{D}$ be the set of all probability densities q on D such that $q(\underline{x}) \geq 0$ for $\underline{x} \in D$ and

$$\int_D d\underline{x} \, q(\underline{x}) = 1. \quad (1)$$

We assume that the existence of $q^* \in \mathcal{D}$ is known but that q^* itself is unknown. The density q^* is sometimes known as a "true" density.

Suppose $p \in \mathcal{D}$ is a prior density that is our current estimate of q^* , and suppose we gain new information about q^* in the form of a set of

expected values

$$\int_{\mathcal{D}} dx \, q^*(x) g_r(x) = \langle g_r \rangle = \bar{g}_r, \quad (2)$$

for a known set of bounded functions $g_r(x)$ and numbers $\bar{g}_r$, $r = 1, \dots, m$. Now, because the constraints (2) do not determine q^* completely, they are satisfied not only by q^* but by some subset of densities $\mathcal{Q} \subseteq \mathcal{D}$. Which single density should we choose from this subset to be our new estimate of q^* , and how should we use the prior p and the new information (2) in making this choice?

The solution to this inference problem is obtained by minimizing a functional $H(q, p)$ called cross-entropy,

$$H(q, p) = \int_{\mathcal{D}} dx \, q(x) \log(q(x)/p(x)). \quad (3)$$

Specifically, of all the densities $q' \in \mathcal{Q}$ that satisfy the constraints (2), we choose the one with the smallest cross-entropy $H(q', p)$ with respect to the prior p . Stated differently, the posterior density q satisfies

$$H(q, p) = \min_{q' \in \mathcal{Q}} H(q', p),$$

where $\mathcal{Q} \subseteq \mathcal{D}$ comprises all of the densities that satisfy the constraints (2).

Mathematically, the solution is obtained using the method of Lagrangian multipliers and standard techniques from the calculus of variations. The minimization condition is

$$\log(q(x)/p(x)) + 1 + \lambda_0 + \sum_r \beta_r g_r(x) = 0, \quad (4)$$

where the β_r are Lagrangian multipliers corresponding to the constraints (2), and where λ_0 is a Lagrangian multiplier corresponding to the normalization constraint (1). The solution of (4) is

$$q(x) = p(x) \exp(-\lambda - \sum_r \beta_r g_r(x)), \quad (5)$$

where $\lambda = \lambda_0 + 1$. It is convenient to write (5) in the form

$$q(x) = Z^{-1} p(x) \exp(-\sum_r \beta_r g_r(x)), \quad (6)$$

where Z is the "partition function",

$$Z = \exp(\lambda) = \int_{\mathcal{D}} dx \, p(x) \exp(-\sum_r \beta_r g_r(x)). \quad (7)$$

The values of the multipliers β_r are determined by the known expectation values $\bar{g}_r$ in (2). One can express the posterior q directly in terms of the values $\bar{g}_r$ by solving the equations

$$\begin{aligned}
g_r &= -Z^{-1} \frac{\partial Z}{\partial \beta_r} \\
&= - \frac{\partial}{\partial \beta_r} \log(Z)
\end{aligned} \tag{8}$$

for the β_r , or by substituting (6) into the constraint equations (2) and solving for the β_r . Such solutions are often difficult or impossible to obtain analytically, but one can obtain them computationally in general [1, Appendix B], [6].

The principle of minimum cross-entropy was first proposed by Kullback [7], who called it a principle of minimum directed divergence or minimum discrimination information. The term cross-entropy is due to Good [8]. Cross-entropy can be characterized axiomatically [9] in terms of properties that are desirable for an information measure [9], [10], and it can be argued [11] that cross-entropy measures the amount of information necessary to change a prior p into the posterior q . The principle of cross-entropy minimization then follows intuitively. This justification is somewhat indirect --- it is based on a formal description of what is required of an information measure rather than on a formal description of what is required of a method for taking new information into account.

Recently, we obtained a stronger justification [1]-[2]. Our approach was to formalize the requirements of inductive inference directly in terms of four consistency axioms that make no reference to information measures or properties of information measures. All of the axioms are based on a single fundamental principle: If a problem can be solved in more than one way, the results should be consistent. We were then able to prove that the principle of minimum cross-entropy provides a correct, general method of inductive inference in the following sense: Given a prior density and new information in the form of constraints on expected values, there is only one posterior density satisfying these constraints that can be chosen in a manner that satisfies the axioms; this unique posterior can be obtained by minimizing cross-entropy.

The principle of minimum cross-entropy is a generalization of the principle of maximum entropy [12]-[13]. When the prior density is uniform, cross-entropy minimization reduces to entropy maximization.

3. Minimum-Cross-Entropy Probability Densities for Stochastic Signals Given Expected Spectral Powers

Consider time-domain signals of the form

$$s(t) = \sum_{k=1}^n a_k \cos(\omega_k t) + b_k \sin(\omega_k t) , \tag{9}$$

with non-zero ω_k that need not be uniformly spaced. These are discrete-spectrum, band-limited signals without DC components. (The assumption of no DC term, which is reasonable for many signal processing applications, is made for mathematical convenience.) The power at each frequency is given by the variables x_k ,

$$x_k = a_k^2 + b_k^2. \quad (10)$$

If we consider the x_k to be random variables, we may describe a stochastic signal in terms of a joint probability density $q(\underline{x})$, where we write $\underline{x}$ for $x_1, x_2, \dots, x_n$. Instead of constantly referring to $q(\underline{x})$ as the spectral power probability density of a stochastic signal, we will informally refer to $q(\underline{x})$ as a "signal."

Now consider the problem of choosing $q(\underline{x})$ when we know the expected power P_k at each frequency

$$P_k = \langle x_k \rangle = \int_{\mathcal{D}} dx \, x_k q(\underline{x}), \quad (11)$$

where $dx = dx_1 dx_2 \dots dx_n$. To apply the principle of minimum cross-entropy, we need a prior density $p(\underline{x})$ to represent our state of knowledge before we learn even (11). Since in any real situation there will be a physical limit on the magnitude of the x_k , we assume that the domain of $\underline{x}$ is bounded. We may therefore use a uniform prior density. For a more detailed analysis of this assumption, see [3].

We choose $q(\underline{x})$ by minimizing cross-entropy subject to the constraints (1) and (11). The result (see (5)) is

$$q(\underline{x}) = A \exp(-\sum_k \beta_k x_k),$$

where the β_k are the Lagrangian multipliers corresponding to (11), and where the uniform prior and the Lagrangian multiplier corresponding to (1) have been absorbed into the constant A ,

$$A^{-1} = \int_{\mathcal{D}} dx_1 dx_2 \dots dx_n \exp(-\sum_k \beta_k x_k). \quad (12)$$

Provided that the P_k are much less than the maximum values of the x_k , we may use integration limits $(0, \infty)$ in (12); this leads to $A = \beta_1 \beta_2 \dots \beta_n$. In terms of the multipliers β_k , the power constraints (11) become

$$P_k = \beta_1 \beta_2 \dots \beta_n \int_{\mathcal{D}} dx_k x_k \exp(-\beta_k x_k) \prod_{m \neq k} \int dx_m \exp(-\beta_m x_m) \\ = 1/\beta_k.$$

The posterior $q(\underline{x})$ is therefore

$$q(\underline{x}) = \prod_{k=1}^n (1/P_k) \exp(-x_k/P_k) . \quad (13)$$

Thus, $q(\underline{x})$ is a multivariate exponential --- each spectral power x_k is exponentially distributed with mean P_k .

4. Minimum-Cross-Entropy Power Spectra Given Autocorrelation Information and a Prior Estimate of the Power Spectrum

Let some unknown signal $q^\dagger(\underline{x})$ have a power spectrum $G(f)$ and autocorrelation function $R(t)$. Suppose we obtain information about G in the form of a set of samples of the autocorrelation function $R(t_r)$,

$$R_r = R(t_r) = \int_{-W}^W df G(f) \exp(2\pi i t_r f) , \quad (14)$$

$r = 1, \dots, m$. We do not assume that the t_r are equally spaced. If the frequency spectrum is discrete, as we have assumed in (9), we can express $G(f)$ as

$$G(f) = \sum_{k=-n}^n G_k \delta(f - f_k) ,$$

where $f_k = -f_{-k}$, $G_k = G_{-k} = G(f_k)$, and $G_0 = 0$. Then (14) becomes

$$R_r = \sum_{k=-n}^n G_k \exp(2\pi i t_r f_k) ,$$

which we prefer to express in the non-complex form

$$R_r = \sum_{k=1}^n G_k c_{rk} , \quad (15)$$

where

$$c_{rk} = 2 \cos(2\pi t_r f_k) . \quad (16)$$

Since the G_k satisfy

$$G_k = \langle x_k \rangle = \int_{\mathcal{D}} d\underline{x} x_k q^\dagger(\underline{x}) , \quad (17)$$

we can rewrite (15) as

$$R_r = \int_{\mathcal{D}} d\underline{x} \left(\sum_k x_k c_{rk} \right) q^\dagger(\underline{x}) . \quad (18)$$

This has the form of known expected values of the unknown density $q^\dagger(\underline{x})$, and we may therefore use the principle of minimum cross-entropy to infer an estimate of $q^\dagger$. In terms of the general form (2), the functions g_r are $g_r = \sum_k x_k c_{rk}$. This minimum cross-entropy problem differs from the one discussed in Section 3 in that the Section 3 problem assumed knowledge

of the expected spectral powers in the form (17), whereas in this problem we have only the form (18). Since typically $m \leq n$, knowledge of (18) provides less information than does (17).

Now suppose we obtain the autocorrelation information (18) when we already have an estimate P_k of the power spectrum G_k (17). We reflect this prior information as a prior density with the exponential form (13)

$$p(\underline{x}) = \prod_{k=1}^n (1/P_k) \exp(-x_k/P_k) , \quad (19)$$

which itself is the minimum cross-entropy density, with respect to a uniform prior, given knowledge of the expected spectral powers P_k .

We then solve the problem of estimating G_k , given a prior estimate P_k and new autocorrelation information (18), by assuming the prior density (19) and minimizing cross-entropy subject to the constraints (18) and (1). The result is

$$q(\underline{x}) = p(\underline{x}) \exp(-\lambda - \sum_{r=1}^m \beta_r \sum_{k=1}^n x_k c_{rk}) , \quad (20)$$

where the β_r are m Lagrangian multipliers corresponding to the autocorrelation constraints (18). For convenience, we define

$$u_k = \sum_{r=1}^m \beta_r c_{rk} , \quad (21)$$

so that (20) can be written as

$$\begin{aligned} q(\underline{x}) &= p(\underline{x}) \exp(-\lambda - \sum_k u_k x_k) \\ &= e^{-\lambda} \prod_k (1/P_k) \exp(-(u_k+1/P_k)x_k) . \end{aligned} \quad (22)$$

Since λ 's value must be such that $q(\underline{x})$ satisfies the normalization constraint (1), (22) becomes

$$q(\underline{x}) = \prod_{k=1}^n (u_k+1/P_k) \exp(-(u_k+1/P_k)x_k) . \quad (23)$$

For our posterior estimate Q_k of the power spectrum, we use the density (23) to compute $Q_k = \langle x_k \rangle = 1/(u_k+1/P_k)$, or

$$Q_k = \frac{1}{(1/P_k) + \sum_r \beta_r c_{rk}} , \quad (24)$$

where the multipliers β_r are determined by the requirement that the Q_k satisfy the autocorrelation constraints (15)

$$R_r = \sum_{k=1}^n Q_k c_{rk} . \quad (25)$$

The minimum cross-entropy result (24)-(25) can also be derived by arguments concerning the cross-entropy between the input and output of linear filters [3].

Suppose that the prior estimates P_k are uniform ($P_k = P$), and suppose that one of the autocorrelation samples, say R_1 , is for zero lag ($t_1 = 0$). Then (24) reduces to

$$Q_k = \frac{1}{\sum_r \beta_r c_{rk}}, \quad (26)$$

where the constant $1/P$ has been absorbed into the multiplier β_1 since $c_{1k} = 2$ holds for all k (see (16)). This is identical to the standard result for maximum entropy spectral analysis (MESA), except that the MESA equations are usually expressed in complex form (see [5], p. 9, for example). Therefore, (26) is also identical to the results obtained by autoregressive, linear predictive, and minimum least squares techniques [5], [14]. The reduction of (24) to (26) reflects the general equivalence of cross-entropy minimization and entropy maximization in the case of uniform priors. When the prior spectral power estimate P_k is not uniform, MESA and cross-entropy minimization (24) give different results. For a more detailed comparison, see [3].

5. Examples

In this section we present some numerical examples in which conventional maximum-entropy spectral estimates are compared with minimum-cross-entropy estimates that take into account prior information about the spectrum. In each example, autocorrelations at a small number of equally spaced lags were computed from an assumed "true" spectrum; then maximum-entropy and minimum-cross-entropy spectra were computed from the autocorrelations and plotted.

For the first example, the original spectrum is the sum of a "background" term, approximating $1/f$ noise, and a "signal" term corresponding to a sinusoidal signal at a fixed frequency. The background term is given by

$$G_{+k}^{(b)} = .01/f_k, \quad (k = 1, \dots, 50)$$

for fifty equally spaced frequencies $f_k = (.005, .015, \dots, .495)$ between 0 and 0.5 (which is the Nyquist frequency: we take the spacing between autocorrelation lags to be unity). The signal term is given by

$$G_{+k}^{(s)} = \begin{cases} 2 & (f_k = .105) \\ 0 & \text{otherwise} \end{cases}$$

The sum is shown in Fig. 1; the first few corresponding autocorrelations R_r are as follows:

$t_r =$	0	1	2	3	4	5
$R_r =$	15.7511	11.6149	7.8699	4.5411	2.0145	1.1413 .

The maximum entropy spectrum computed from these six autocorrelations is shown in Fig. 2. For the minimum-cross-entropy calculation, the background term $G^{(b)}$ has been used as the prior spectral estimate; the resulting posterior is shown in Fig. 3. As one might expect, the $1/f$ background is considerably better estimated in Fig. 3 than in Fig. 2. More important, however, there is a clearly discernible peak in Fig. 3 corresponding to the sinusoidal signal at frequency .105; no such peak is evident in Fig. 2.

For the second example, spectral powers are shown at the same frequencies as for the first, autocorrelations are computed for the same lags, and the original spectrum is again the sum of a "background" term $G^{(b)}$ and a "signal" term $G^{(s)}$. In this example, the background consists of white noise plus a peak corresponding to a sinusoid at frequency .215:

$$G_{\pm k}^{(b)} = \begin{cases} 1.02 & (f_k = .215) \\ .02 & \text{otherwise} . \end{cases}$$

The signal term consists of a nearby, similar peak at frequency .165:

$$G_{\pm k}^{(s)} = \begin{cases} 1 & (f_k = .165) \\ 0 & \text{otherwise} . \end{cases}$$

The original spectrum is shown in Fig. 4, the autocorrelations are

$t_r =$	0	1	2	3	4	5
$R_r =$	6.0000	1.4544	-2.7732	-3.2248	0.2032	2.6900

and the maximum-entropy spectrum is shown in Fig. 5. For the minimum-cross-entropy calculation, the background term $G^{(b)}$ has again been taken as a prior spectral estimate. The posterior estimate is shown in Fig. 6. The information in the prior has permitted the resolution of the "expected" peak at frequency .215 from the "unexpected" peak at frequency .165. In the maximum-entropy estimate, by contrast, the two peaks are coalesced into a single peak at about the center frequency, .190.

References

1. J. E. Shore and R. W. Johnson, December, 1978, "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," NRL Memorandum Report 3898.

2. J. E. Shore and R. W. Johnson, "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," IEEE Trans. Inform. Theory, to be published.
3. J. E. Shore, January, 1979, "Minimum Cross-Entropy Spectral Analysis," NRL Memorandum Report 3921.
4. J. P. Burg, 1967, "Maximum Entropy Spectral Analysis," presented at the 37th Annual Meeting Soc. of Exploration Geophysicists, Oklahoma City, Okla.
5. J. Burg, 1975, "Maximum Entropy Spectral Analysis," Ph.D. Dissertation, Stanford University (University Microfilms No. 75-25,499).
6. R. W. Johnson, 1979, "Determining Probability Distributions by Maximum Entropy and Minimum Cross-Entropy," APL Quote Quad 9, No. 4 (APL 79 conf. proceedings), pp. 24-29.
7. S. Kullback, 1959, Information Theory and Statistics, Wiley, New York.
8. I. J. Good, 1963, "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," Annals Math. Stat. 34, pp. 911-934.
9. R. Johnson, "Axiomatic Characterization of the Directed Divergences and Their Linear Combinations," IEEE Trans. Inform. Theory, to be published.
10. A. Hobson and B. Cheng, 1973, "A Comparison of the Shannon and Kullback Information Measures," J. Stat. Phys. 7, No. 4, pp. 301-310.
11. A. Hobson, 1969, "A New Theorem of Information Theory," J. Stat. Phys. 1, No. 3, pp. 383-391.
12. E. T. Jaynes, 1957, "Information Theory and Statistical Mechanics I," Phys. Rev. 106, pp. 620-630.
13. E. T. Jaynes, "Prior Probabilities," IEEE Trans. Systems Science and Cybernetics SSC-4, 1968, pp. 227-241.
14. A. van den Bos, "Alternative Interpretation of Maximum Entropy Spectral Analysis," IEEE Trans. Inform. Theory IT-17, 1971, pp. 493-4.

Figures

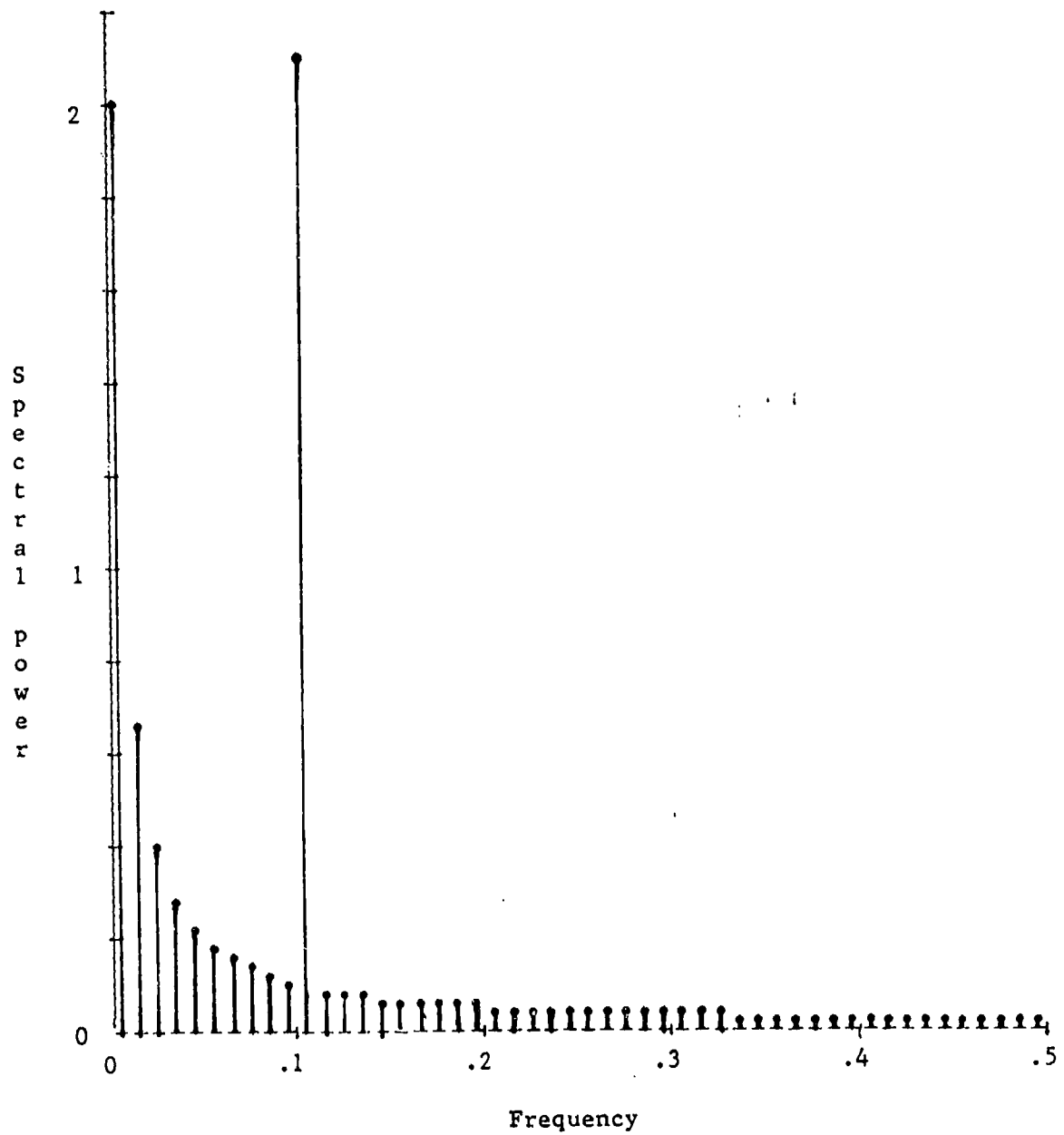


FIGURE 1. Original Spectrum for First Example

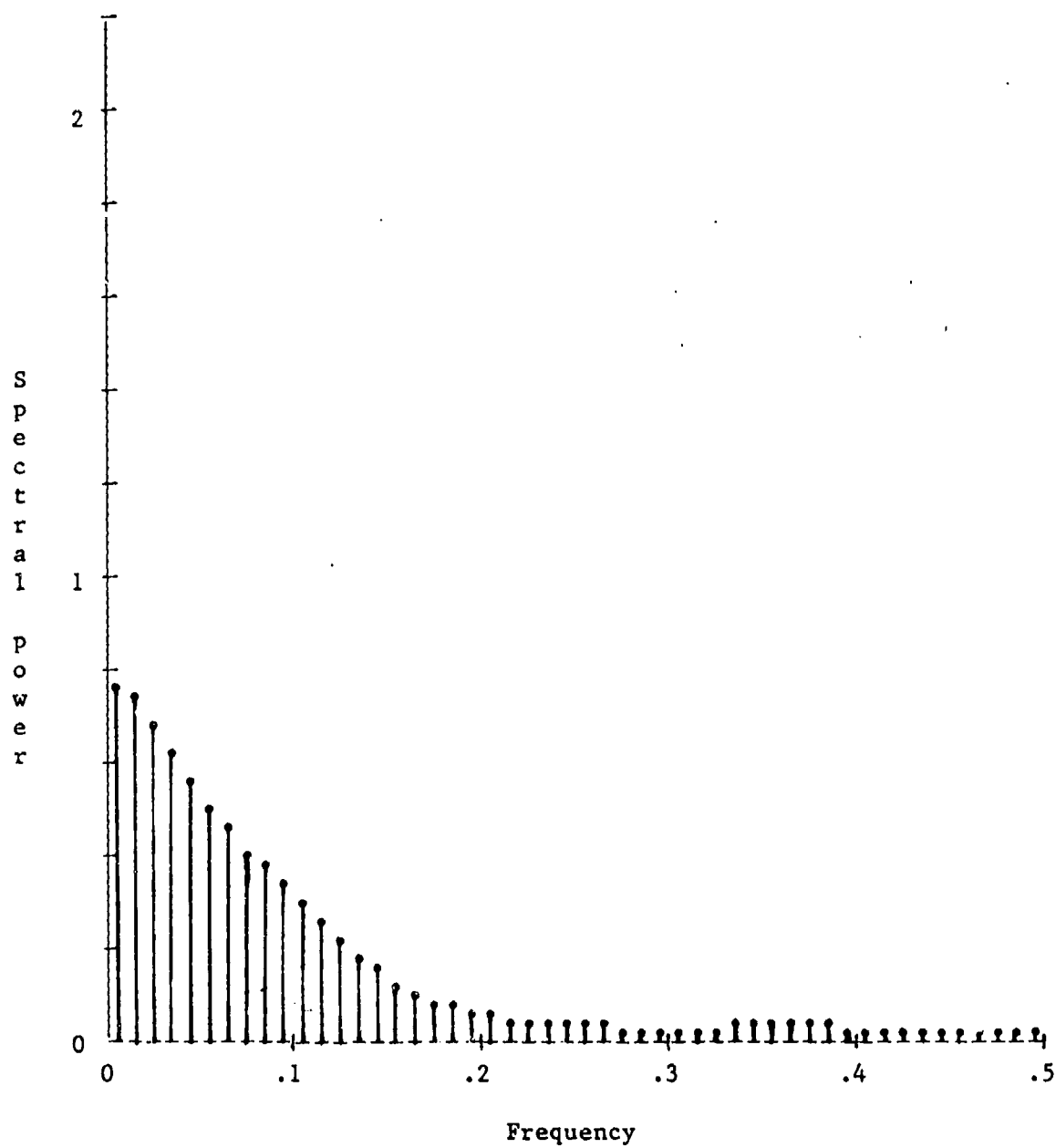


FIGURE 2. Maximum-Entropy Spectrum for First Example

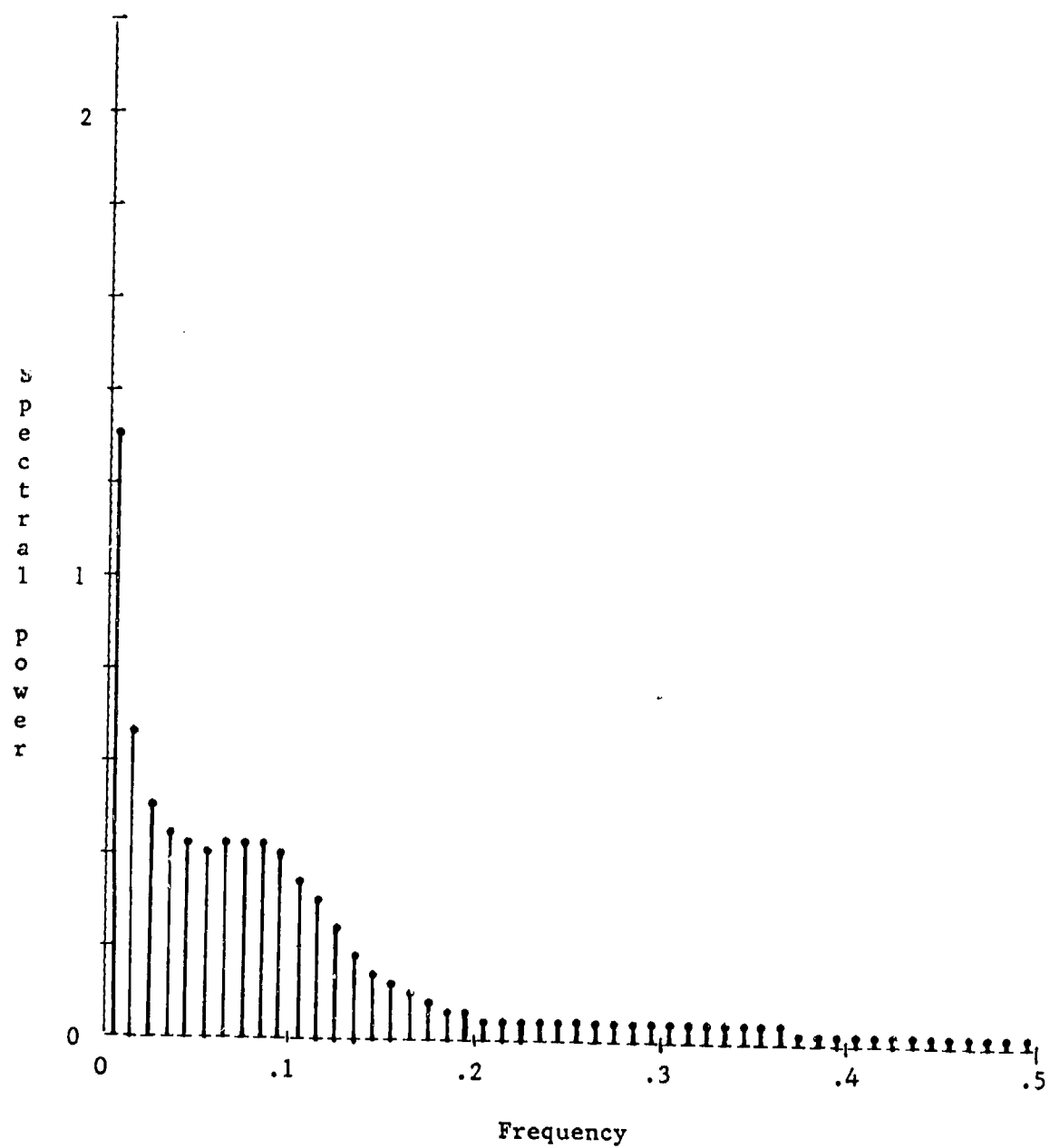


FIGURE 3. Minimum-Cross-Entropy Spectrum for First Example

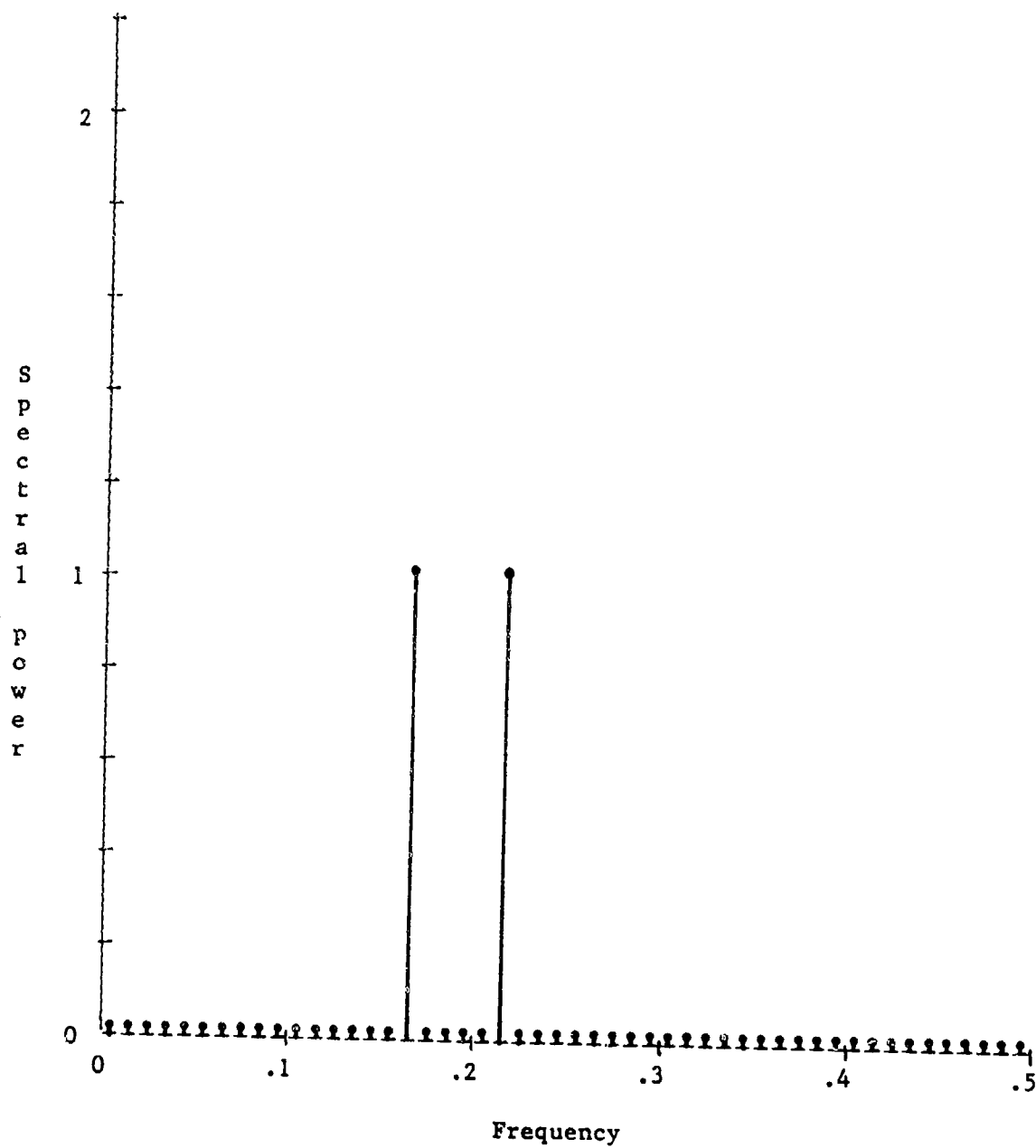


FIGURE 4. Original Spectrum for Second Example

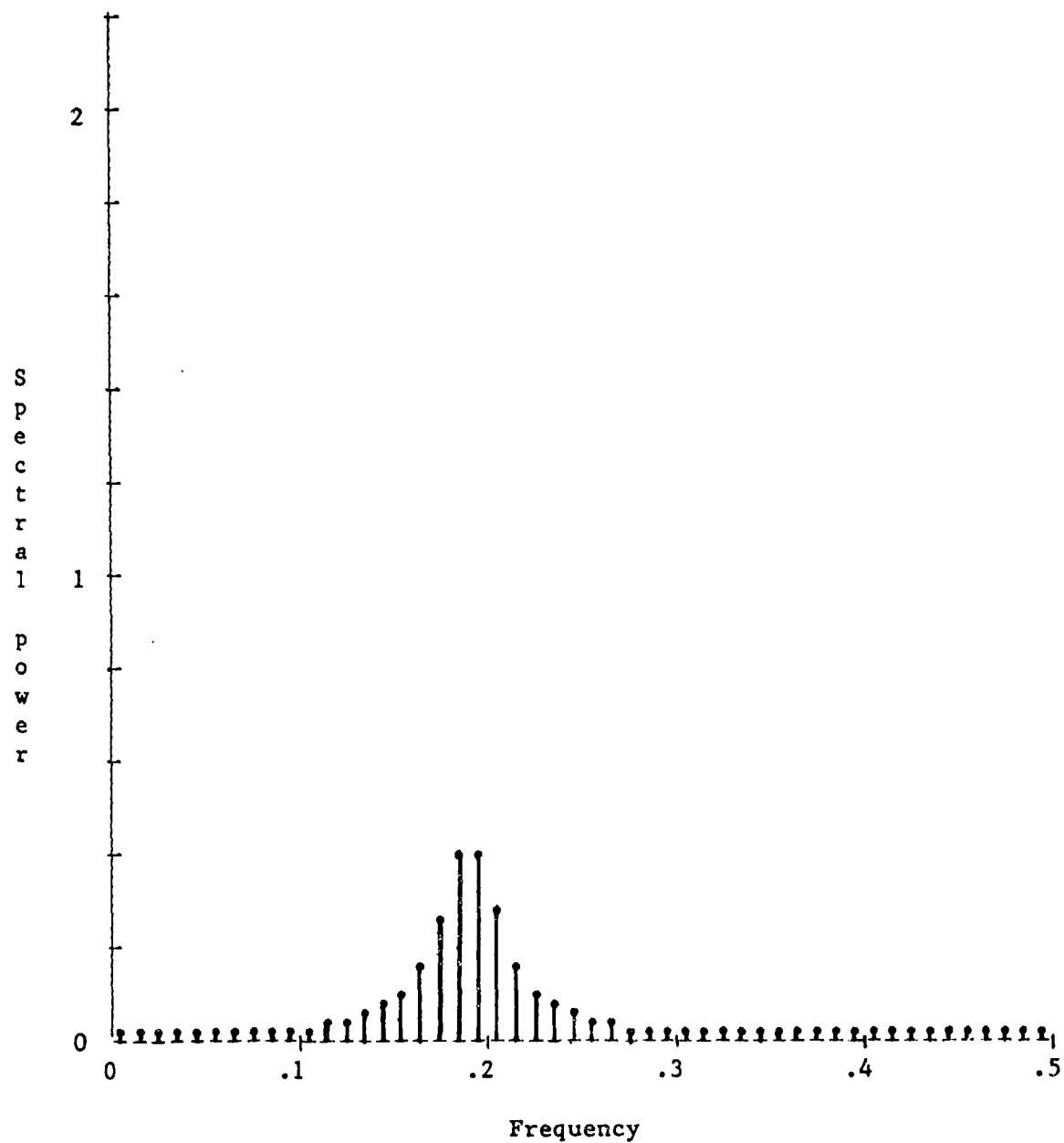


FIGURE 5. Maximum-Entropy Spectrum for Second Example

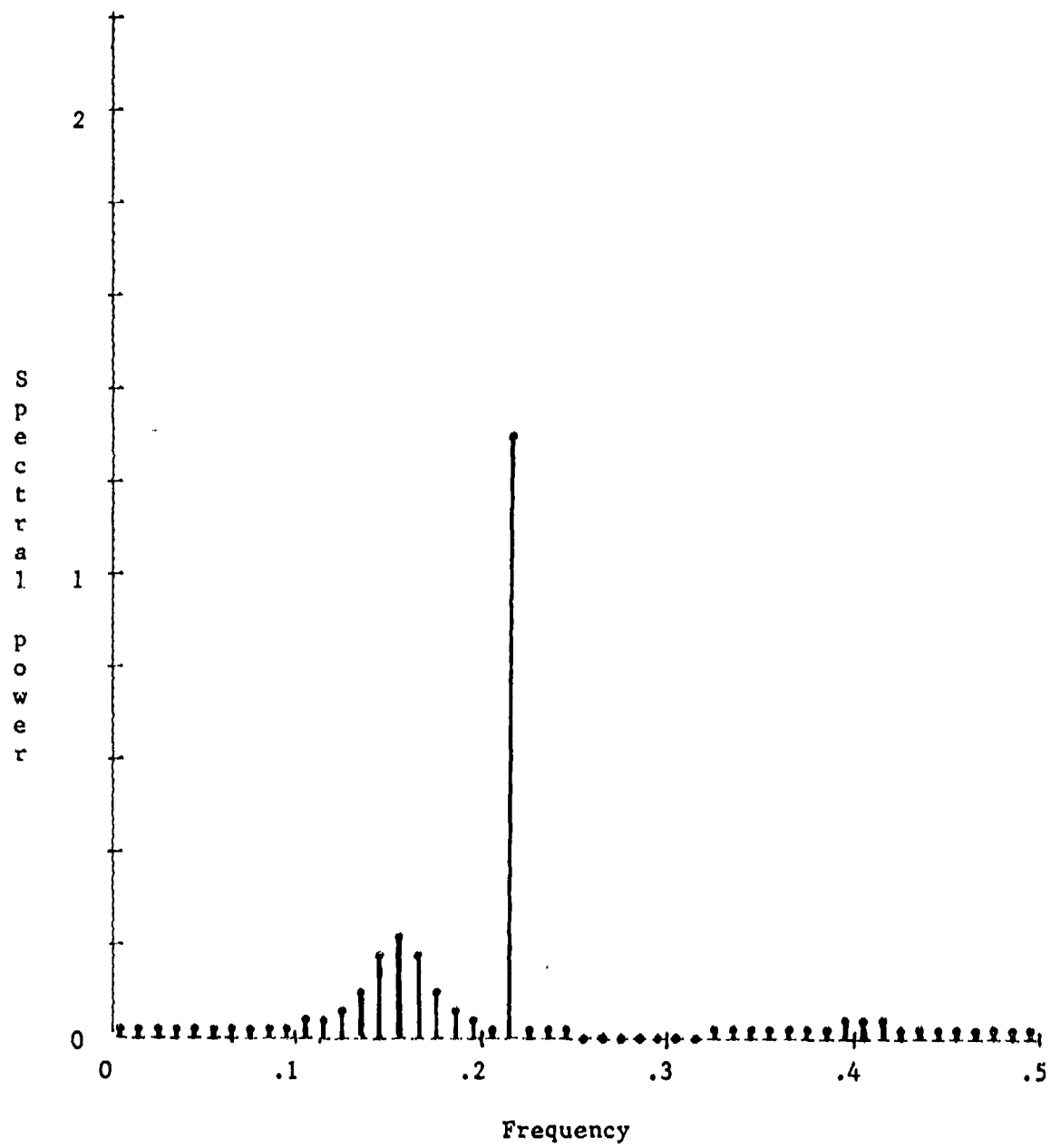


FIGURE 6. Minimum-Cross-Entropy Spectrum for Second Example

22. Blank

OPTIMAL ESTIMATION FOR BANDLIMITED, TIME-CONCENTRATED SIGNALS

D.P. KOLBA AND T.W. PARKS

Department of Electrical Engineering
Rice University
Houston, Texas

Abstract

A new estimation procedure for bandlimited, time-concentrated signals is described. The method optimally estimates desired measurements of an unknown signal by minimizing a maximum error. The information available for the estimation consists of a limited number of time samples and knowledge about the bandwidth and approximate time duration of the signal. An example will compare the new method with existing bandlimited estimation methods. The requirements on sampling rate and the effects of errors (noise) in the data are discussed. A recursive implementation of the estimation is presented.

Introduction

In the calculation of the spectrum of a discrete time signal,

$$X(f) = \sum_{n=-\infty}^{\infty} x(n) e^{-j2\pi fn}, \quad (1)$$

problems arise when only a finite portion of the signal is available. If no additional information is available, the spectral estimate is computed from a windowed version of the data [1]. If additional information about $x(n)$ is known, such as knowing $x(n)$ is bandlimited or bandlimited and time-concentrated, then this additional knowledge should be used to give a better spectral estimate. The additional information may be incorporated into a new direct estimate for the spectrum. Alternatively, the additional information can be used to extrapolate the given segment of data. The transform of this extrapolated signal is then the desired estimate of the spectrum.

Extrapolation and spectral estimation for bandlimited signals have been studied recently in [2-5]. A finite data segment, $x(-M), x(-M+1), \dots, x(M)$,

This work was supported by NSF Grant ENG78-09033.

of a bandlimited signal is known. From this given information, the extrapolation and spectral estimation are derived. In [2], the extrapolation is shown to be the minimum energy extension of the finite data segment to a bandlimited signal. This method will have the extrapolation as concentrated on the measurements as possible. Therefore, this method will be called concentrated on measurements (COM) estimation. The methods of [4] and [5] are equivalent to this COM estimation. The iterative method of [3] will converge to the COM estimate.

The COM method concentrates the estimation on the measurement interval regardless of where the true signal is concentrated. If the signal is known in advance to be concentrated on an interval larger than that over which the measurements are taken, then this additional information should be incorporated into the estimation method. In this paper, a new estimation method is described which incorporates this additional information about the signal into the solution [10]. This new method will be called concentrated on the signal (COS) estimation.

The new COS estimation method will be derived using deterministic estimation theory. A short presentation of this theory will first be made. Then, the COS estimation method will be derived. Next, the new COS method will be compared to the COM method on a practical example. A new sampling rate criterion for COS estimation will be discussed. The sensitivity of the method to errors in the data will be determined. Finally, a recursive formulation for the new estimation procedure will be presented.

Deterministic Estimation

In a deterministic estimation problem [6,7], some measurement of the deterministic signal x is desired. This desired measurement could be a frequency sample, a time sample, a derivative sample, or any other measurement which is a continuous linear functional of the signal x . The information known about x is incomplete; therefore, the desired measurement must be estimated from the limited knowledge about x . The estimate is made optimal by minimizing the maximum magnitude of the error.

The limited information about x consists of knowing x is an element in a signal space, knowing certain measurements on x , and having a bound on the size of x . In more detail, the signal x is known to be a member of a Hilbert space H_0 . The inner product in this space is denoted $(\cdot, \cdot)_0$. The known measurements on x are the N continuous linear functionals

$$(x, u_k)_0 = a_k, \quad k = 1, 2, \dots, N. \quad (2)$$

The bound on the size of x is given by

$$(x, x)_Q \leq \epsilon^2 \quad (3)$$

This given information defines the set R of possible signals,

$$R = \{x \in H_Q : (x, u_k)_Q = a_k, k = 1, 2, \dots, N \text{ and } (x, x)_Q \leq \epsilon^2\} \quad (4)$$

The desired measurement on x is also a continuous linear functional on x ;

$$(x, u_0)_Q = a_0 \quad (5)$$

A value for a_0 must be estimated from the knowledge that $x \in R$. The estimate, $\bar{a}_0$, is selected to minimize $\max_{x \in R} |a_0 - \bar{a}_0|$. The error in this estimation is bounded by

$$E(R) = \min_{\bar{a}_0} \max_{x \in R} |a_0 - \bar{a}_0|. \quad (6)$$

In [6,7], the solution to this estimation problem is derived using the geometry of H_Q . Since the bound on the size of x , (3), is a quadratic form, the optimal estimate is a linear combination of the data, (2). In fact, the optimal reconstruction of the entire signal, $\bar{x}$, is a linear combination of the data. From $\bar{x}$, any desired measurement on x is estimated by

$$\bar{a}_0 = (\bar{x}, u_0)_Q. \quad (7)$$

The optimal reconstruction of x is the unique signal which is in R and also in the subspace spanned by the linearly independent measurement signals, u_k , $k = 1, 2, \dots, N$ [6,7]. Thus, $\bar{x}$ is a linear combination of the u_k 's:

$$\bar{x} = \sum_{k=1}^N b_k u_k. \quad (8)$$

Since $\bar{x} \in R$, $\bar{x}$ must satisfy (2):

$$(\bar{x}, u_\ell)_Q = a_\ell, \ell = 1, 2, \dots, N. \quad (9)$$

Using (8) in (9) leads to

$$\sum_{k=1}^N b_k (u_k, u_l)_Q = a_l, \quad l = 1, 2, \dots, N. \quad (10)$$

The coefficients for the expansion in (8) can now be solved for from (10). Using matrix notation, $\bar{x}$ is

$$\bar{x} = \underline{a_l} \left[(u_k, u_l)_Q \right]^{-1} u_k, \quad k, l = 1, 2, \dots, N \quad (11)$$

where u_k is a column of the measurement signals. Now, replacing $\bar{x}$ in (7) by (11) gives

$$\bar{a}_0 = \underline{a_l} \left[(u_k, u_l)_Q \right]^{-1} (u_k, u_0)_Q, \quad k, l = 1, 2, \dots, N. \quad (12)$$

The optimal estimate may be calculated directly from the data, (12), or may be calculated by taking the desired measurements on the reconstructed signal using (7) and (11).

The error bound of (6) is evaluated in [7] as

$$E(R) = \left\{ (u_0, u_0)_Q - \frac{(u_l, u_0)_Q}{\left[(u_k, u_l)_Q \right]^{-1} (u_k, u_0)_Q^*} \right\}^{1/2} \quad (13)$$

$$\left\{ \epsilon^2 - \underline{a_l} \left[(u_k, u_l)_Q \right]^{-1} a_k^* \right\}^{1/2}, \quad k, l = 1, 2, \dots, N.$$

In (13), the first term specifies the error in approximating u_0 by a linear combination of the u_k 's. The second term in (13) measures how large the set R of possible signals is and depends on the data.

A property of the optimal reconstruction shown in [7] is

$$(\bar{x}, \bar{x})_Q = \min_{x \in R} (x, x)_Q. \quad (14)$$

This property will be used to formulate a new estimation procedure in which the estimated signal is concentrated on a known signal interval.

Estimation Concentrated On The Signal

Consider the problem of reconstructing a bandlimited signal from a finite data segment. In [2], the best reconstruction of x with this given information has minimum energy. Since the energy of the signal on the samples, $n=-M, -M+1, \dots, M$ is fixed, this reconstruction also has minimum energy tails outside the measurement interval. The reconstruction procedure (COM) is concentrated on the measurement interval. If the signal is known to be of longer duration than from $-M$ to M , this information should be used to extend the concentration beyond the measurement interval to the larger signal interval. The new estimation method developed in this paper (COS) is concentrated on the signal interval.

Let B denote the bandlimiting operator so that for finite energy $x(n)$, $y=Bx$ implies

$$Y(f) = \begin{cases} X(f) & \beta_1 \leq |f| \leq \beta_2 \\ 0 & \text{otherwise} \end{cases}$$

for $0 \leq \beta_1 < \beta_2 < .5$. Here, $Y(f)$ and $X(f)$ denote the transforms of $y(n)$ and $x(n)$ as given by (1). Next, let D denote the timelimiting operator so that $y=Dx$ implies

$$y(n) = \begin{cases} x(n) & |n| \leq L \\ 0 & \text{otherwise.} \end{cases}$$

The operator $(I - D)$ will retain the tails of the signal it operates on:

$$(I-D)x = \begin{cases} 0 & |n| \leq L \\ x(n) & |n| > L \end{cases}.$$

For the COS estimation, the property of concentrating the reconstruction of x on the signal interval can also be viewed as minimizing the tails outside of the signal interval. Recalling (14), leads to the selection of an inner product which deals with the tails outside of the signal interval, $n = -L, -L+1, \dots, L$. This inner product will be denoted $(\cdot, \cdot)_{I-D}$ and is defined by

$$(x, y)_{I-D} = \sum_{|n| > L} x(n) y^*(n) \quad (15)$$

Using this inner product in the deterministic estimation will result in selecting as the reconstruction that signal which fits the data and has minimum energy tails outside of the signal interval. If $L=M$, COM and COS are identical. When $L > M$ the methods are different.

The space of bandlimited signals concentrated on $[-L, -L+1, \dots, L]$ as discussed in [2] is given by

$$H = \left\{ x : x = BDy, y \text{ has finite energy} \right\}. \quad (16)$$

A deterministic estimation problem can now be formulated which will give the desired COS estimation method. The Hilbert space H from (16) with inner product $(\cdot, \cdot)_{I-D}$ given by (15) will be the known signal space. This implies that the bandwidth and time duration are assumed known.

Signals to be estimated will have bounded energy in the tails : $(x, x)_{I-D} \leq \epsilon^2$. The known measurements on x will be the N time samples $a_k = x(n_k)$ $k = 1, 2, \dots, N$. This information specifies the unknown signal as an element of the set

$R = \{x \in H : (x, x)_{I-D} \leq \epsilon^2 \text{ and } x(n_k) = a_k, k = 1, 2, \dots, N\}$. The estimation problem for any desired measurement can now be solved if x of (11) can be found. To this end, the measurement signals of (2) with respect to $(\cdot, \cdot)_{I-D}$ must be found.

From [2], a complete, orthonormal basis for H is the set of signals

$$v_i(n) = \lambda_i v_i(n+L) \quad i=0,1,\dots,2L, \quad (17)$$

where λ_i and $v_i(n+L)$ are the eigenvalues and discrete prolate spheroidal sequences of [2] extended to the bandpass case. For any $x \in H$,

$$x = \sum_{i=0}^{2L} c_i \gamma_i, \quad (18)$$

where

$$c_i = \sum_{n=-\infty}^{\infty} x(n) \gamma_i(n). \quad (19)$$

These basis signals are also orthogonal on the signal interval:

$$\sum_{n=-L}^L \gamma_i(n) \gamma_j(n) = \lambda_i \delta_{ij}, \quad i, j = 0, 1, \dots, 2L. \quad (20)$$

Using the orthonormality of the γ_i 's and (20) leads to

$$(\gamma_i, \gamma_j)_{I-D} = (1 - \lambda_i) \delta_{ij}, \quad i, j = 0, 1, \dots, 2L. \quad (21)$$

Now, define

$$K(n, m) = \sum_{j=0}^{2L} \frac{1}{1-\lambda_j} \gamma_j(n) \gamma_j(m) \quad (22)$$

For any $x \in H$, (22) gives

$$(x(m), K(n, m))_{1-D} = \sum_{j=0}^{2L} \frac{1}{1-\lambda_j} \gamma_j(n) (x(m), \gamma_j(m))_{1-D} \quad (23)$$

Substituting (18) in (23) yields

$$(x, K(n, \cdot))_{1-D} = \sum_{j=0}^{2L} \frac{1}{1-\lambda_j} \gamma_j(n) \sum_{i=0}^{2L} c_i (\gamma_i, \gamma_j)_{1-D} \quad (24)$$

Applying (21) to (24) and simplifying gives

$$(x, K(n, \cdot))_{1-D} = \sum_{j=0}^{2L} c_j \gamma_j(n) = x(n) \quad (25)$$

From (25), the measurement functions sought are seen to be

$$u_k = K(n_k, \cdot) = \sum_{i=0}^{2L} \frac{1}{1-\lambda_i} \gamma_i(n_k) \gamma_i, \quad (26)$$

since the data can then be written as

$$x(n_k) = (x, u_k)_{1-D} = a_k.$$

Now that the measurement signals have been found, the optimal reconstruction of x can be calculated using (11). The matrix to be inverted has elements $(u_k, u_l)_{1-D}$. Since $u_k \in H$,

$$(u_k, u_l)_{1-D} = u_k(n_l) = \sum_{i=0}^{2L} \frac{1}{1-\lambda_i} \gamma_i(n_k) \gamma_i(n_l). \quad (27)$$

Using (27) and (26) in (11) gives $\bar{x}$.

Now that the optimal reconstruction of the unknown signal has been found, any desired measurement can be estimated using (7). Two common

estimation problems are extrapolation and spectral estimation. For extrapolation, $x(n_0)$ is estimated as $\hat{x}(n_0)$. For spectral estimation, $X(f_0)$ is estimated as $\hat{X}(f_0)$. If error bounds from (13) are desired for these estimates, then the signal u_0 of (5) needs to be found. For extrapolation,

$$u_0(n) = K(n_0, n) . \quad (28)$$

The evaluation of (13) requires

$$(u_0, u_0)_{1-D} = K(n_0, n_0)$$

and

$$(u_k, u_0)_{1-D} = K(n_k, n_0) .$$

For spectral estimation,

$$u_0(n) = \sum_{i=0}^{2L} \frac{1}{1-\lambda_i} \psi_i^*(f_0) \gamma_i(n) , \quad (29)$$

where

$$\psi_i(f_0) = \sum_{n=-\infty}^{\infty} \gamma_i(n) e^{-j2\pi f_0 n} .$$

The evaluation of (13) now requires

$$(u_0, u_0)_{1-D} = \sum_{i=0}^{2L} \frac{1}{1-\lambda_i} |\psi_i(f_0)|^2$$

and

$$(u_k, u_0)_{1-D} = \sum_{i=0}^{2L} \frac{1}{1-\lambda_i} \psi_i(f_0) \gamma_i(n_k) .$$

With this COS solution to the estimation problem, the new method can be compared to the COM method on a practical example.

Comparison Of Methods

An application in which knowledge of signal duration is valuable consists of a problem in which signals of known duration have been over-

lapped. Consider the signal

$$x_1(t) = \begin{cases} 2000 t e^{-400t} \sin 2\pi 400t & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (30)$$

This signal has a time duration of about 12.5 msec. and is approximately bandlimited with a bandwidth of 500 Hz. Now, several of these signals are overlapped and added:

$$x(t) = 1/2 x_1(t) + x_1(t-.009) + 2x_1(t-.0177) \quad (31)$$

The components of this signal are shown in Fig. 1. From this composite signal, information about the middle component is desired. Since the duration of the first component is 12.5 msec. and the third component starts at $t=17.7$ msec., good measurements for the second component are available in the interval $12.5 \leq t \leq 17.7$. With a sample spacing of 0.5 msec., 11 measurements are made on the second component.

In the discrete time formulation, the second component is considered as a discrete time signal centered about $n=0$ and concentrated on $n=-12, -11, \dots, 12$ (corresponding to 12.5 msec. sampled every 0.5 msec.). The 11 measurements are therefore the samples at $n=-5, -4, \dots, 5$. The assumed bandwidth for the signal is $\beta_1 = .05$ Hz and $\beta_2 = .30$ Hz using normalized frequency.

Extrapolation and spectral estimation for the COS and COM methods can now be compared. The extrapolations for both methods are shown in Fig. 2. The actual desired signal is the dotted curve. Fig. 3 shows the spectral estimates obtained using the two methods. As can be seen, the COS method gives a better reconstruction of the signal and a correspondingly more accurate spectrum than the COM method.

Sampling Rate For COS Estimation

The COS estimation method deals with bandlimited, time-concentrated signals. The class of signals with bandwidth β and duration T has approximately $2 \beta T$ dimensions [2]. Therefore, approximately $2 \beta T$ measurements should give a good estimate of a signal in this class. This is verified by the error bound of (13). The first term in (13) was calculated for the COS spectral estimation problem of Fig. 3. This term of the error bound was evaluated at 9 frequencies between $f_0 = .05$ and $f_0 = .3$. The resulting RMS

value is plotted versus the number of samples used for the estimation in Fig. 4. The error bound has reached a small value near $2 \beta T$ measurements.

The need for about $2 \beta T$ measurements determines the sampling rate at which the measurements are taken. If T_M is the duration of the measurement interval and T_L is the duration of the signal, then the $2 \beta T_L$ samples must be taken in the measurement interval. This requires a sampling period of

$$T_s = \frac{T_M}{T_L} \frac{1}{2\beta} \text{ sec.} \quad (32)$$

From (32), if the measurement interval decreases with respect to the signal interval, the sampling rate must increase to retain $2 \beta T_L$ samples. This increased sampling rate will cause an increase in the sensitivity of the estimates to errors in the data measurements.

Sensitivity

The increase in sensitivity as a result of increased sampling rate affects both estimation methods. These methods are linear; and so, the effects of errors on the data can be treated separately. Let e be an error signal which is added to the true signal x . Let A be the linear estimation operator which takes the measurements on the signal and generates the desired estimate. Then,

$$A(x+e) = Ax + Ae \quad (33)$$

The term Ae is the additional error in the estimation due to the measurement errors. This can be bounded by

$$\|Ae\| \leq \|A\| \cdot \|e\| \quad (34)$$

The bound on the sensitivity of the estimation to errors is directly related to $\|A\|$. An approximation to $\|A\|$ for both COM and COS is given by

$$\|A\| \approx 0.6 [5(1-2\alpha)]^N \text{ for } .1 \leq \alpha \leq .35, \quad (35)$$

where the parameter α is the normalized bandwidth of the discrete time signal and N is the number of measurements used. From (32), the value of α can be calculated for the lowpass case as $\alpha = \frac{T_M}{2T_L}$. If N is selected as $2 \beta T_L$ and T_L is held fixed, then the sensitivity increases as T_M decreases.

Currently, estimation with smoothing (see[8]) is being investigated as a means of reducing sensitivity to noise.

Recursive Computation

The COS estimation procedure described can be calculated without the matrix inversion of (11) or (12). A recursive solution is possible using Gram-Schmidt orthonormalization (with respect to $(\cdot, \cdot)_Q$) on the measurement signals [8]. This new set of basis signals for the optimal reconstruction subspace will be denoted w_ℓ , $\ell=1, 2, \dots, N$. These signals are derived from the u_k , $k=1, 2, \dots, N$ as follows [9]:

$$\begin{aligned} v_1 &= u_1 & w_1 &= \frac{v_1}{\|v_1\|_Q} \\ v_\ell &= u_\ell - \sum_{j=1}^{\ell-1} (u_\ell, w_j)_Q w_j & \text{and } w_\ell &= \frac{v_\ell}{\|v_\ell\|_Q} \quad \ell = 2, 3, \dots, N \end{aligned} \quad (36)$$

The optimal reconstruction is a linear combination of these new basis signals:

$$\bar{x} = \sum_{\ell=1}^N c_\ell w_\ell \quad (37)$$

where

$$c_\ell = (x, w_\ell)_Q. \quad (38)$$

Since the w_ℓ are linear combinations of the u_k as determined in (36), they can be written as

$$w_\ell = \sum_{k=1}^{\ell} d_{\ell k} u_k. \quad (39)$$

Using (39) in (38) allows the calculation of the c_ℓ 's by

$$c_\ell = \sum_{k=1}^{\ell} d_{\ell k}^* (x, u_k)_Q \quad \ell = 1, 2, \dots, N. \quad (40)$$

Applying (2) to (40) finally gives

$$c_l = \sum_{k=1}^l d_{lk}^* a_k \quad (41)$$

Thus, the c_l 's required in (37) are calculated from the data using the coefficients given in (39) from (36).

For any desired measurement, the estimate is found by substituting (37) in (7) :

$$\bar{a}_0 = \sum_{l=1}^N c_l (w_l, u_0)_Q \quad (42)$$

This estimate can be calculated recursively using

$$\bar{a}_0 = \sum_{l=1}^{N-1} c_l (w_l, u_0)_Q + c_N (w_N, u_0)_Q \quad (43)$$

The first term in (43) is the estimate given the first $N-1$ measurements. The second term updates the estimate when the N th data measurement is included.

Conclusions

The application of deterministic estimation theory to signal processing problems can lead to interesting new estimation methods. The new COS estimation method described in this paper has been seen to give improved estimates compared to the COM method when the signal concentration interval is larger than the measurement interval. The COS method concentrates the estimation on the signal interval as opposed to the COM method which concentrates the estimation on the measurement interval. Consideration of the error bounds for the new method leads to a new criterion for the sampling rate to be used. The sensitivity of the estimation to errors in the data has also been presented. Finally, a recursive implementation of the new method has been described.

References

1. R.B. Blackman and J.W. Tukey, "The Measurement of Power Spectra," New York: Dover, 1958.

2. D. Slepian, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty - V: The Discrete Case," BSTJ Vol. 57, No. 5, May - June 1978, PP 1371 - 1430.
3. A. Papoulis, "A New Algorithm in Spectral Analysis and Bandlimited Extrapolation," IEEE Trans. on Circuits and Systems, Vol. CAS-22, No. 9, pp 735-742, Sept. 1975.
4. J.A. Cadzow, "An Extrapolation Procedure for Band-Limited Signals," IEEE Trans. on ASSP, Vol. ASSP-27, No. 1, pp. 4-12, Feb. 1979.
5. D.P. Kolba and T.W. Parks, "Extrapolation and Spectral Estimation for Bandlimited Signals," 1978 ICASSP Record, pp. 372-374.
6. M. Golomb, Lectures on the Theory of Approximation, Argonne National Laboratory, 1962.
7. M. Golomb and H. Weinberger, "Optimal Approximation and Error Bounds," On Numerical Approximation, ed. R. Langer, Univ. of Wisconsin Press, Madison, pp. 117-190, 1959.
8. H.L. Weinert, "A Reproducing Kernel Hilbert Space Approach to Spline Problems with Applications in Estimation and Control," Information Systems Lab., Stanford Univ., Technical Report No. 7001-6, May 1972.
9. L.E. Franks, Signal Theory, Prentice Hall Inc., Englewood Cliffs, NJ 1969.
10. D.P. Kolba and T.W. Parks, "Extrapolation and Spectral Estimation for Bandlimited, Time-Concentrated Signals," 1979 ICASSP Record, pp. 190-193.

Figures

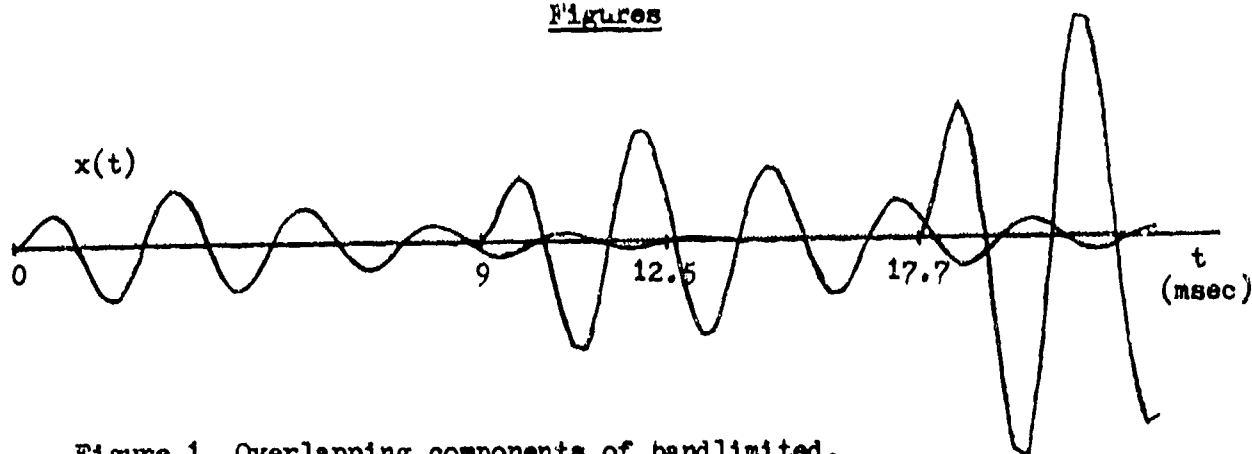


Figure 1. Overlapping components of bandlimited, time-concentrated signals.

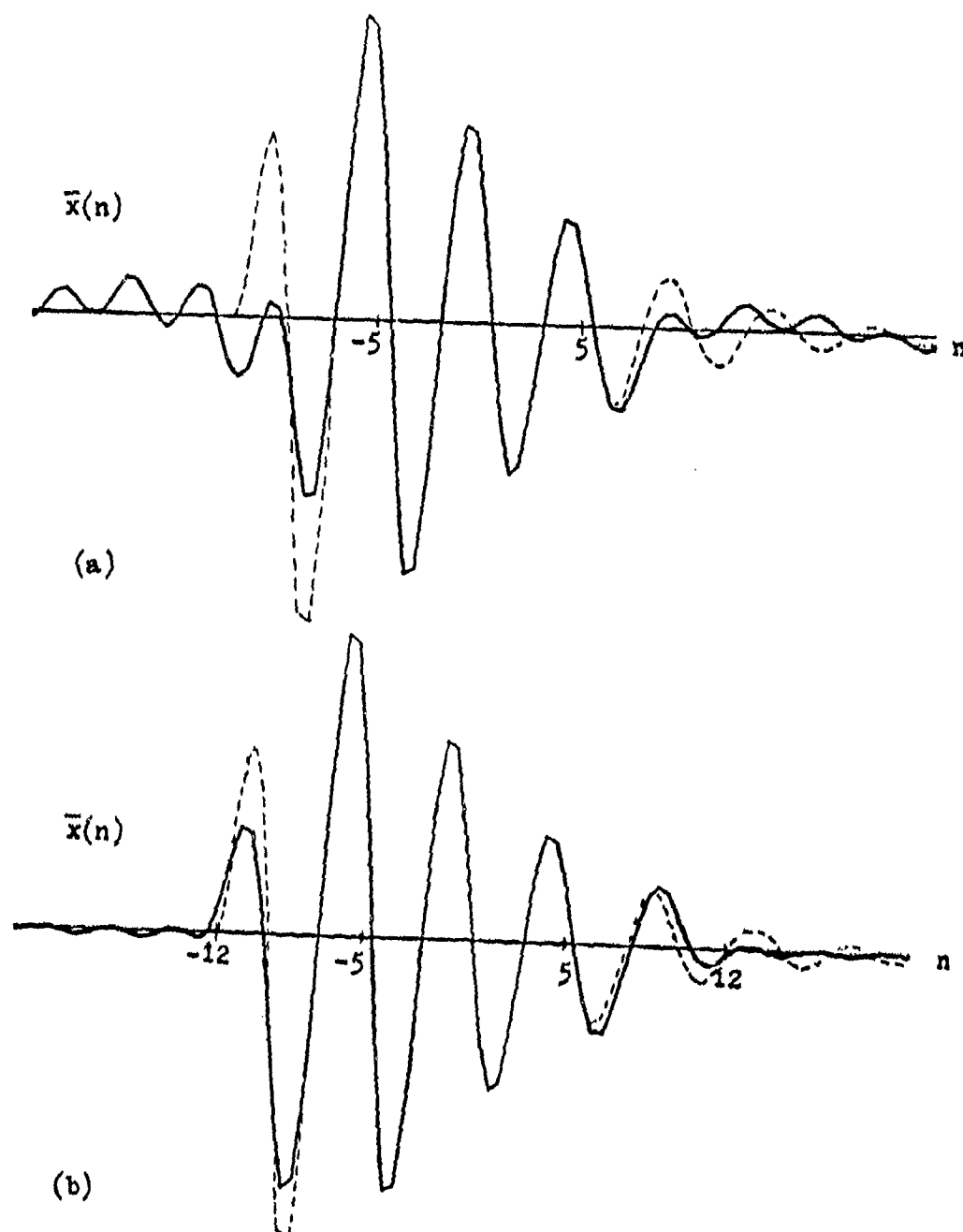


Figure 2. Comparison of extrapolation estimates. (a) COM estimate with $M=5$ (b) COS estimate with $M=5$ and $L=12$. The actual signal is the dashed curve.

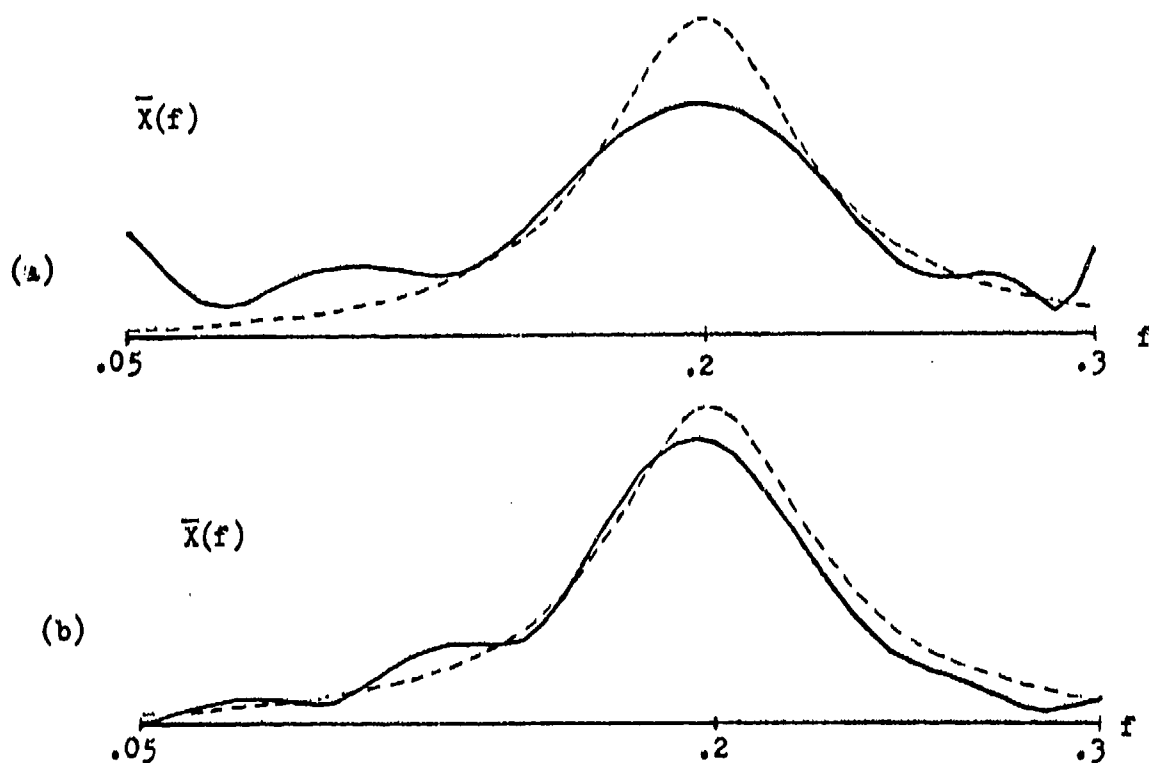


Figure 3. Comparison of spectral estimates with assumed bandwidth $\beta_1 = .05$ and $\beta_2 = .30$. (a) COM estimate (b) COS estimate
The actual spectrum is the dashed curve.

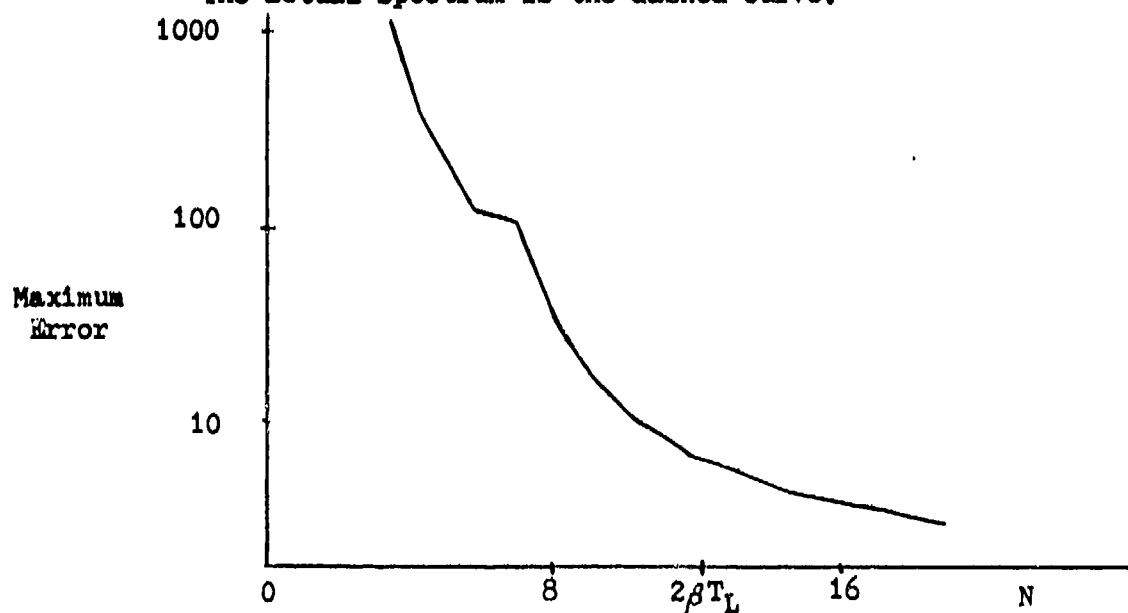


Figure 4. Maximum error bound versus number of samples for spectral estimation.

38-OSHAWB

THE USE OF LINEAR PREDICTION FOR THE INTERPOLATION
AND EXTRAPOLATION OF MISSING DATA AND DATA GAPS
PRIOR TO SPECTRAL ANALYSIS

STEPHEN B. BOWLING
SHU LAI

Massachusetts Institute of Technology
Lincoln Laboratory
Lexington, Massachusetts 02173

Abstract

The spectral analysis of a series of equally spaced samples of a coherent time-stationary process becomes difficult when samples are missing or sizable data gaps occur within the interval of interest. A linear prediction algorithm can be used to fill in the missing data with estimates that are spectrally consistent with the data that are observed. Simulated and practical radar examples demonstrate an improvement in resolution and a reduction of sidelobe interference levels.

Problem Definition

When a spectral transformation of a sampled process is performed, one must account for any samples that are missing. Assigning a value of zero to missing data prior to Fourier transformation, for example, introduces false frequencies and greatly increases sidelobe levels. Clearly, an interpolation scheme is needed that can cope with missing data and, at the same time, will not degrade the spectral information contained in the data that are observed.

Occasional missing samples, well separated from each other, can be estimated with simple interpolation procedures (polynomial or parabolic fits, spline fits, etc.). However, data may be missing in such quantity that conventional interpolation is inadequate; data gaps longer than the periods of the sinusoidal components in the data cannot be easily bridged with simple functions. A more sophisticated approach becomes necessary, and the use of a data-adaptive linear prediction filter is one feasible alternative.

In radar data processing, missing data or data gaps may occur for a variety of reasons:

- (a) hardware fails to transmit pulses or receive echoes properly;
- (b) radar transmits when it should be receiving echoes (range eclipsing);
- (c) resources are saturated by many targets that must be watched simultaneously (panic);
- (d) burst waveforms are purposely silent between bursts;
- (e) poor signal-to-noise makes detections sporadically unreliable.

In any case, the missing samples (in these examples, complex samples with amplitude and phase) must be filled in before Doppler processing can be accomplished.

Description of the Method

The use of a linear prediction filter to extend a finite complex data set before Fourier transformation was first proposed and described by Bowling (1977). Applying this original algorithm, Tomlinson and Ackerson (1978) demonstrated clutter and sidelobe reduction in the Doppler processing of a train of radar pulses.

In the application of interest here, the prediction algorithm is used to predict estimates of missing data by extrapolating from observed data. For example, suppose an observation interval contains randomly missing samples and gaps. The procedure is as follows:

- (1) Locate and designate the missing samples to be estimated.
- (2) Find the longest continuous span of data within which there are no missing samples.
- (3) Calculate an N-point linear prediction filter from the longest continuous span of data found in step (2).
- (4) Calculate an estimate of each missing sample immediately to the left and to the right of the longest continuous span of data (a total of two estimates, one on each side).
- (5) Return to step (2) until all missing data have been estimated. (Note that estimates from step (4) are to be treated as observations on an equal basis with the original data. That is, the "longest continuous span of data" is increasing in length as estimates fill in the holes, one by one, to the left and to the right.)

When the longest continuous span of data finally terminates at one of the endpoints of the observation interval, estimates continue to be made toward the other endpoint until all missing points have been filled in. The length of the prediction filter may remain a constant, or vary according to the current length of the longest span of data.

Simulated Examples

A simple example shows the improvement in the power spectrum of a data set containing missing samples and gaps.

The real and complex parts of a sampled sum of sinusoids are shown in Figs. 1(a,b). No noise has been added and no samples are missing. Figure 1(c) is the true power spectrum calculated with a standard FFT with no weighting.

Now if samples are randomly zeroed out and data gaps are introduced as shown in Figs. 2(a,b), the power spectrum in Fig. 2(c) shows increased side-lobe levels and false frequencies, both caused by processing without estimating the missing data.

Figures 3(a,b) show the data set after the linear prediction algorithm is applied, with the power spectrum shown in Fig. 3(c). Not only do Figs. 1(a,b) overlay with 3(a,b) almost exactly, but their respective power spectra are indistinguishable.

Another simple example demonstrates the performance of the linear prediction algorithm when data gaps occur periodically, such as is the case for a radar burst waveform.

Figures 4(a,b) represent the process of Figs. 1(a,b) for which three data gaps are present. Indeed, half of the data are missing from the observation interval, and the gaps are longer than any period exhibited in the data. The power spectrum of Fig. 4(c) is a very poor estimate of the true spectrum (Fig. 1(c)) because no gaps have been filled in. Transforming only one of the short spans of observed data gives a power spectrum with limited resolution, as shown in Fig. 4(d).

However, upon using the prediction algorithm on Figs. 4(a,b), we obtain Figs. 5(a,b) and the power spectrum in Fig. 5(c), which is a good estimate of the true spectrum.

In this case, the prediction algorithm has acted as a synergistic device that, by linking short pieces of data together with spectrally consistent estimates, allows a spectral transform to be performed over an effectively longer piece of data. The whole, then, has more resolving power than any of its parts.

It should be pointed out that the data gaps need not be periodic or equal in length in order for the prediction algorithm to fill them in.

Radar Example

Radar is often used to identify targets from the time history of the velocity spectrum of the target's motion about its center of mass. A series or burst of radar pulses is Fourier analyzed, and the target's velocity spectrum is observed. If not accounted for in the processing, missing pulses can introduce false velocity components and lead to an incorrect characterization or classification of the target.

For example, Fig. 6(a) shows the evolution of the velocity spectrum of a tumbling object for which missing data and data gaps exist and are set to zero in the radar pulse train. No estimation for the missing pulses has been done. It is therefore not clear if the velocities indicated are actually from the target or are an artifact of the missing data. Figure 6(b) shows the evolution of the same velocity spectrum upon using the prediction algorithm before Fourier transformation. The disappearance of some of the velocities cleans up the spectral history and indicates which velocity components actually characterize the target.

Limitations of the Method

Implicit in the use of a linear prediction filter is the assumption that the data from which the filter is derived are time-stationary. The process being sampled must be coherent during the observation interval which is being analyzed and within which the missing data and data gaps may occur.

Also, the prediction filter works best when the spectral components are approximately pure tones, confined to locally narrow bandwidths spaced within the Nyquist bounds of the spectral transform domain.

Summary

This paper proposes the use of a linear prediction algorithm to fill in missing data and data gaps that may occur within an observation interval over which a spectral transform is to be made. False frequencies and sidelobe interference, which are artifacts of the missing samples, can be suppressed or eliminated by replacing the missing samples with estimates that are spectrally consistent with neighboring observed data. Large gaps can be smoothly bridged that otherwise could not be satisfactorily interpolated by simpler schemes.

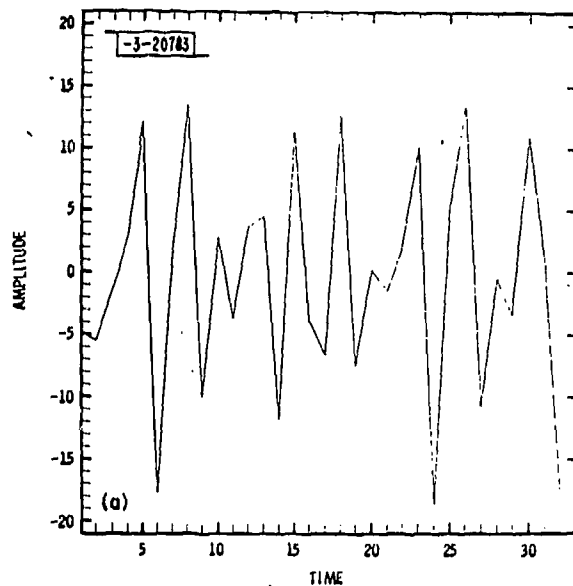
Computer programs which accomplish the interpolation and extrapolation procedures for complex data can be found in Ref.[3].

References

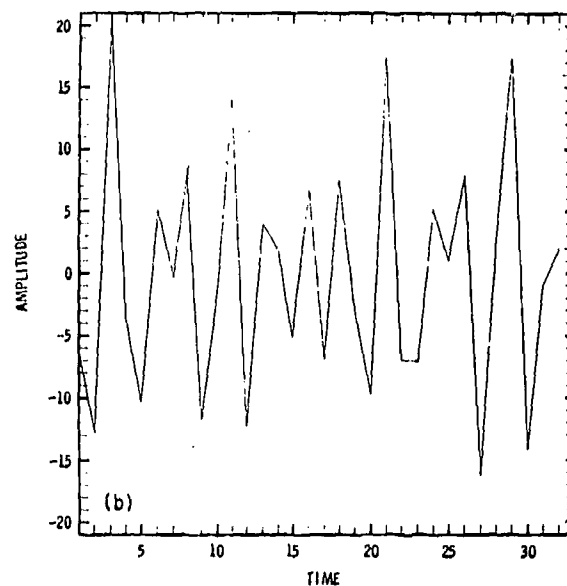
1. Bowling, S. B., "Linear Prediction and Maximum Entropy Spectral Analysis for Radar Applications," Project Report RMP-122, Lincoln Laboratory, M.I.T. (24 May 1977), DDC-AD-A042817/7.
2. Tomlinson, P. G., and G. A. Ackerson, "Air Vehicle Detection Using Advanced Spectral Techniques," Proceedings of the RADC Spectrum Estimation Workshop, Rome Air Development Center, Rome, New York (May 1978).
3. Bowling, S. B. and S. Lai, "Use of Linear Prediction for the Interpolation and Extrapolation of Missing Data and Data Gaps," Report TN-1979-46, Lincoln Laboratory, M.I.T. (to be published).

This work was sponsored by Ballistic Missile Defense Advanced Technology Center, Department of the Army.

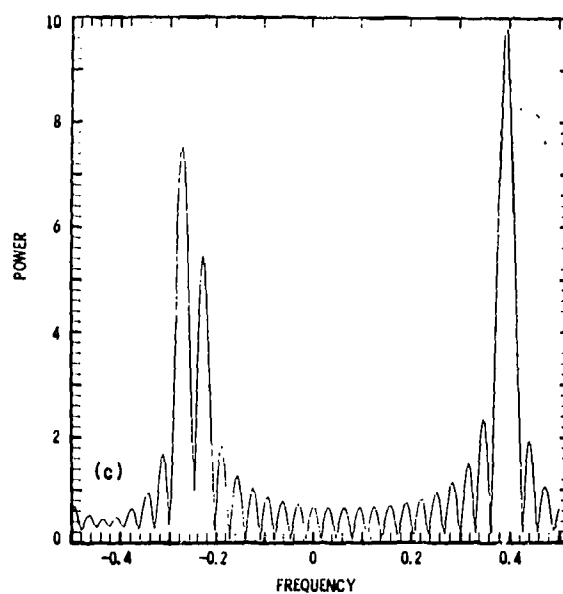
The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Government.



(A) Real part of the sum of three complex sinusoids; no samples are missing.

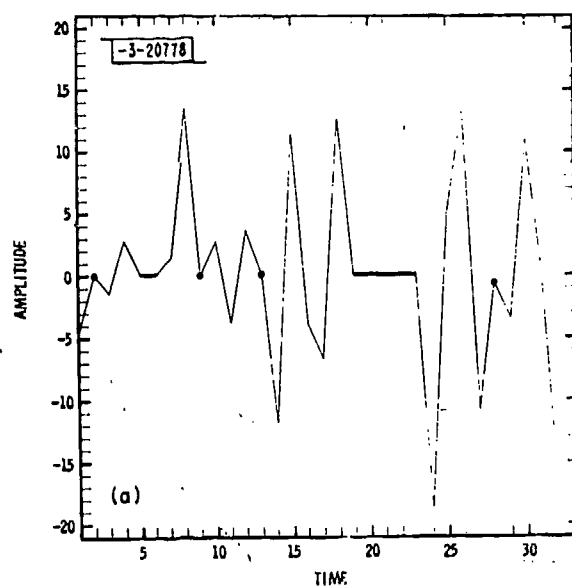


(B) Imaginary part of the sum of three complex sinusoids; no samples are missing.

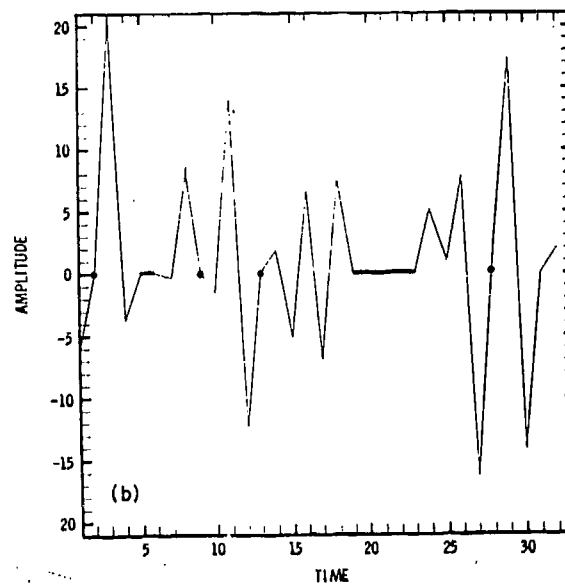


(C) Power spectrum of $l(a,b)$.

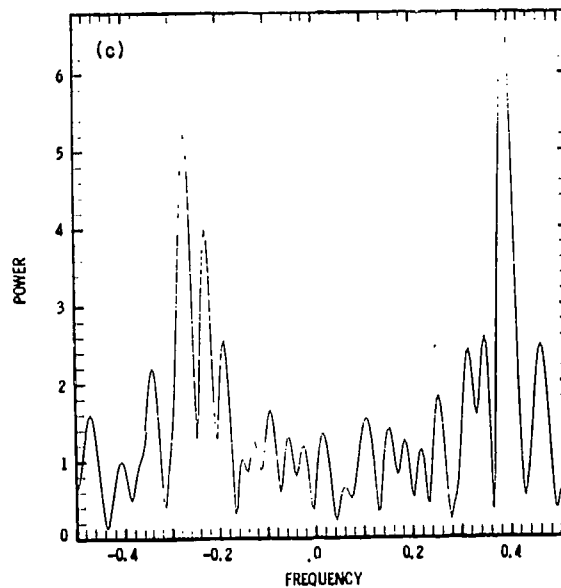
Figure 1



(A) Real part of Fig. 1 with randomly missing data and data gaps.

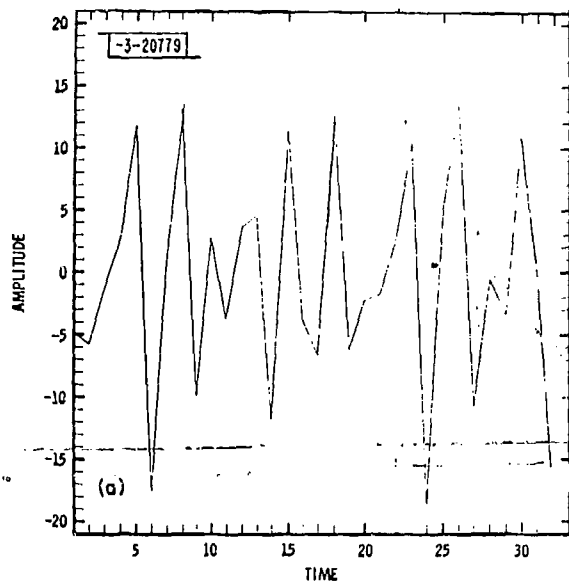


(B) Imaginary part of Fig. 1 with randomly missing data and data gaps.

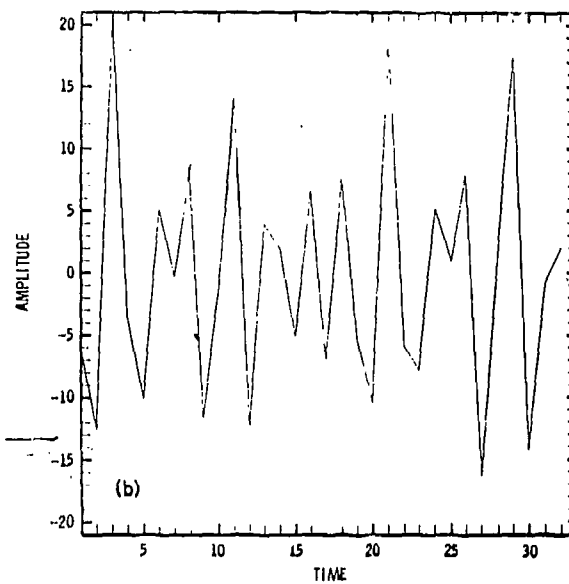


(C) Power spectrum of 2(a,b).

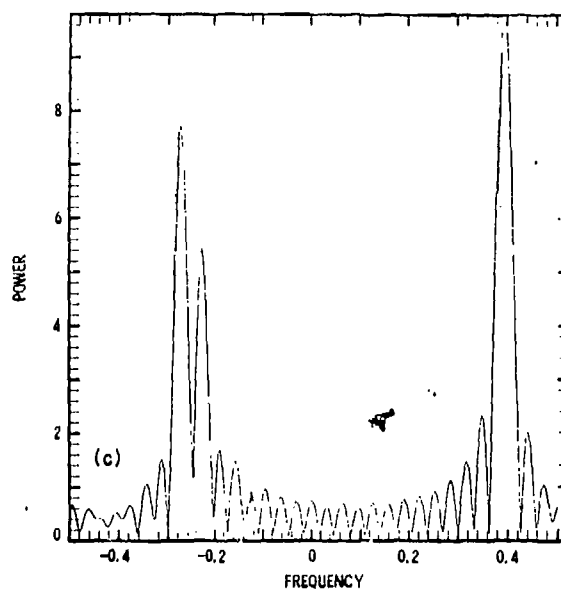
Figure 2



(A) Real part of Fig. 2 after application of the linear prediction algorithm.

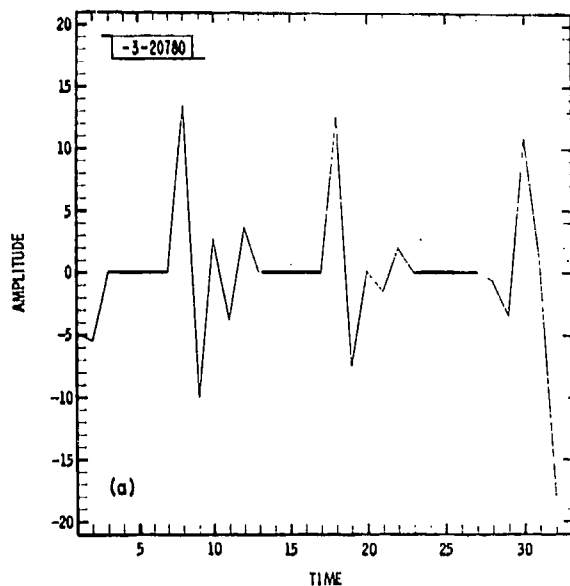


(B) Imaginary part of Fig. 2 after application of the prediction algorithm.

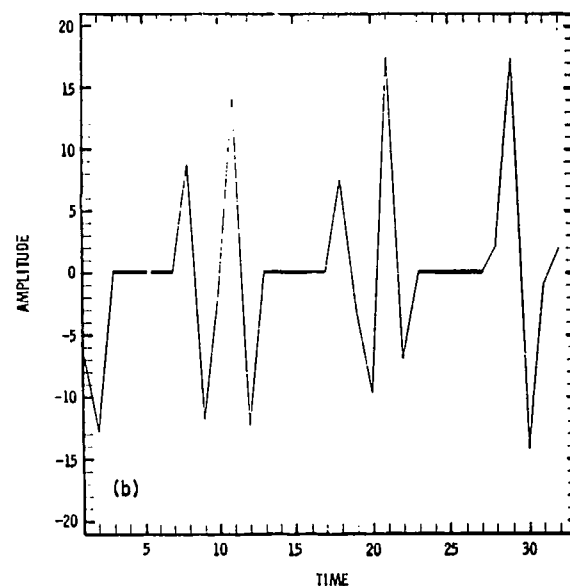


(C) Power spectrum of 3(a,b).

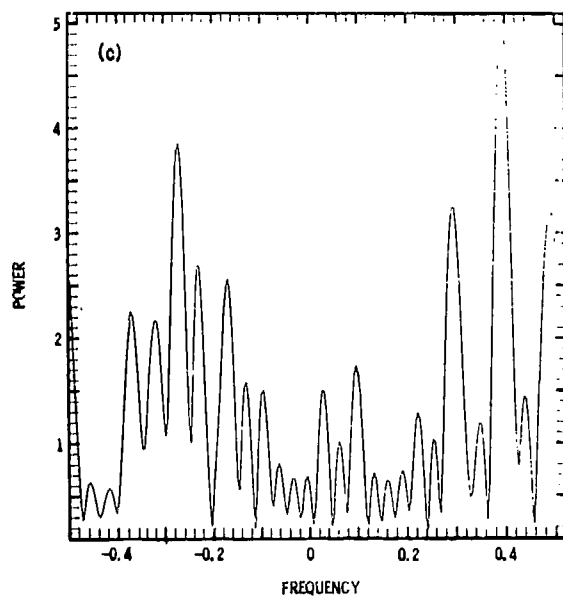
Figure 3



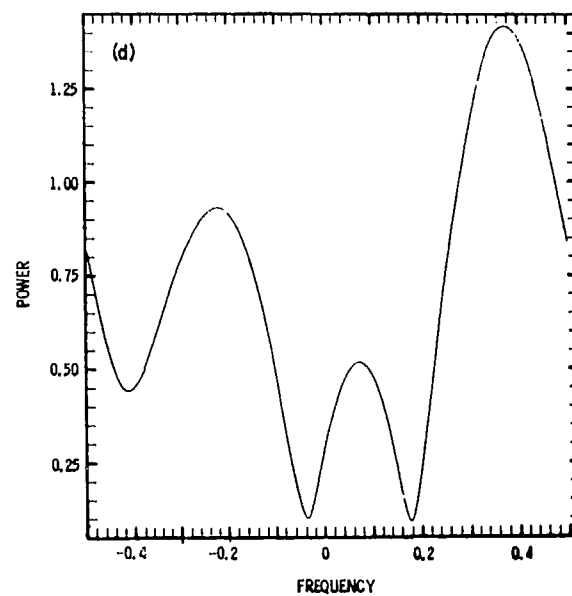
(A) Real part of Fig. 1 with data gaps.



(B) Imaginary part of Fig. 1 with data gaps.

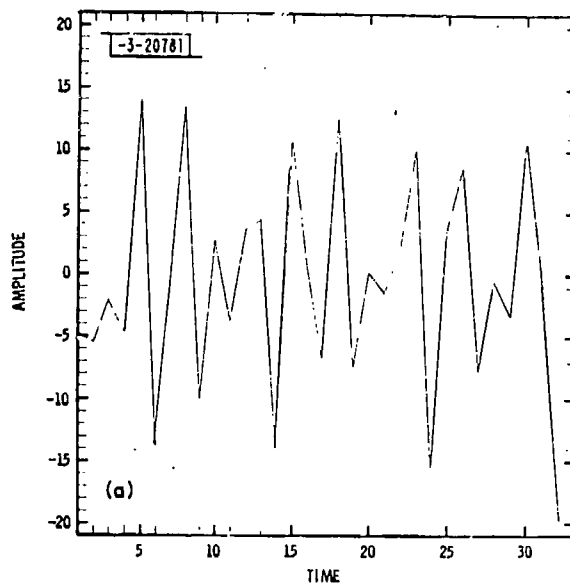


(C) Power spectrum of 4(a, b).

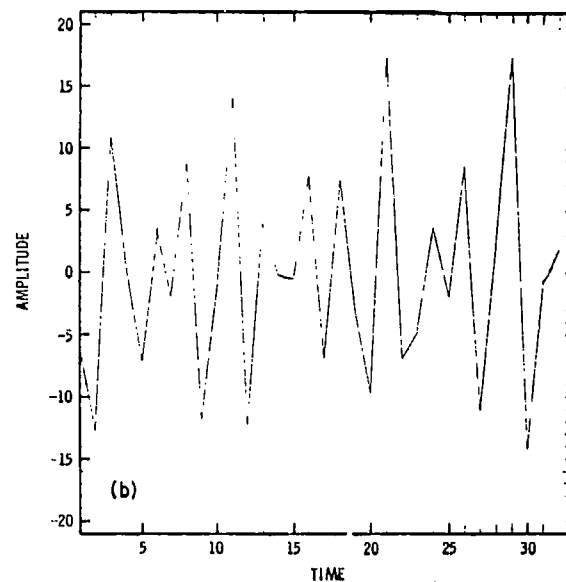


(D) Power spectrum of the center short data span in 4(a, b).

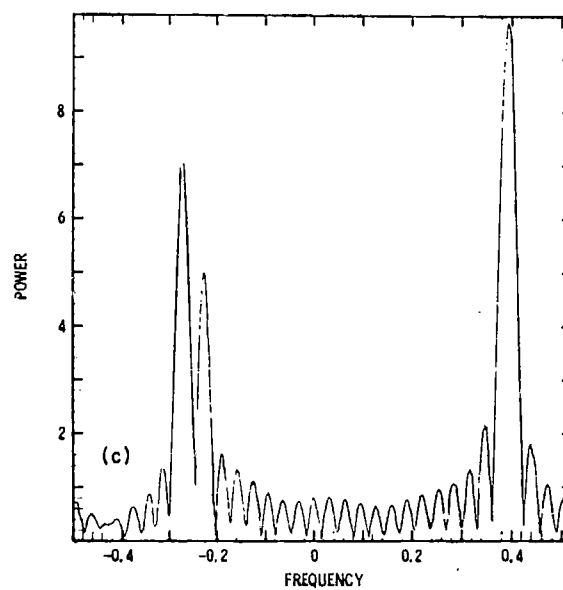
Figure 4



(A) Real part of Fig. 4 after gaps are filled in with the linear prediction algorithm.



(B) Imaginary part of Fig. 4 after gaps are filled in with the prediction algorithm.



(C) Power spectrum of 5(a,b).

Figure 5

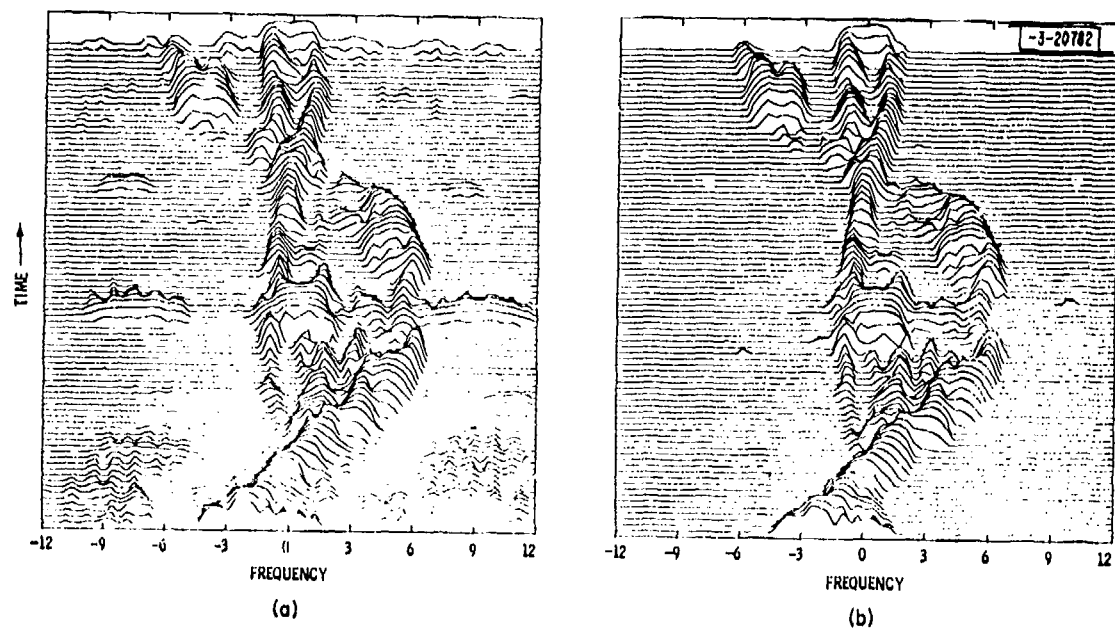


FIG. 6 (A) Doppler history of tumbling object when missing data and data gaps are not accounted for.
 (B) Doppler history after missing data are filled in with the linear prediction algorithm.

50-Blank

A NEW AUTOREGRESSIVE SPECTRUM ANALYSIS ALGORITHM

Larry Marple

Advent Systems , Inc.
355 Ravendale Drive
Mountain View, CA 94043

Abstract

A new recursive algorithm for autoregressive (AR) spectral estimation is introduced, based on the least squares solution for the AR parameters using forward and backward linear prediction. The algorithm has computational complexity proportional to the order of the process squared, comparable to that of the popular Burg algorithm. The computational efficiency is obtained by exploiting the structure of the least squares normal matrix equation, which may be decomposed into products of Toeplitz matrices. AR spectra generated by the new algorithm have improved performance over AR spectra generated by the Burg algorithm. These improvements include less bias in the frequency estimate of spectral components, reduced variance in frequency over an ensemble of spectra, and absence of observed spectral line splitting.

Introduction

Autoregressive spectrum analysis, sometimes termed maximum entropy spectrum analysis (MESA), has become a popular alternative to the periodogram as an estimate of the power spectral density (PSD) for a sampled process. For signal to noise ratios (SNRs) greater than 0 dB, the AR PSD estimate has higher frequency resolution than that of the conventional periodogram estimate [1]. AR spectral estimates also do not have the distortion produced by sidelobe leakage effects that are inherent in the periodogram approach to spectrum analysis. These are two of several attractive features of AR psectral estimation that have created interest in this technique.

The means used to estimate the autoregressive model parameters is the key to the performance of the AR technique. If $M+1$ lags of the autocorrelation function for a process are known, the M autoregressive parameters are obtained by solving the Yule-Walker normal equations using the Levinson recursion algorithm [2]. The algorithm requires a number of computational operations proportional to M^2 .

A host of techniques are available for estimating the AR parameters from data samples. The most obvious approach is to first make estimates of the autocorrelation lags with the available data, and then to apply the usual Levinson recursion with the estimated lags. This approach is rarely used primarily due to the fact that better resolution may be obtained with other estimation methods that obtain the AR parameters directly from the data. If unbiased autocorrelation estimates are used, one may also run into numerical ill-conditioning during the solution of the normal equations. Biased autocorrelation estimates reduce the risk of ill-conditioning, but at the expense

of a degradation of the AR spectral resolution and a shifting of spectral peaks from their true locations [1]. The shift effect is termed a frequency estimation bias. Another reason that has made this a seldom used technique is the problem of spectral line splitting. Spectral line splitting is the occurrence of two or more closely-spaced peaks in an AR spectral estimate when only one spectral peak should be present. The reasons for spectral line splitting in the Yule-Walker technique has been documented by Kay and Marple [3].

The most popular approach for AR parameter estimation is the Burg algorithm [4,5]. This algorithm utilizes a constrained least squares estimation procedure to obtain the M autoregressive parameter estimates from N data samples. The constraint requires the AR parameter estimates to satisfy the Levinson recursion. The Burg algorithm requires computational operations proportional to the product NM.

AR spectral estimates based on the Burg algorithm suffer from two of the same problems observed in Yule-Walker estimates of the AR spectrum. The problem of spectral line splitting in AR spectra produced by the Burg algorithm was first documented by Fougere et.al. [6]. They noted that spectral line splitting was most likely to occur when: (1) the SNR is high, (2) the initial phase of sinusoidal components is some odd multiple of 45° , (3) the time duration of the data sequence is such that sinusoidal components have an odd number of quarter cycles, and (4) the number of AR parameters estimated is a large percentage of the number of data values used for the estimation. Many spurious spectral peaks often accompany spectra that exhibit line splitting.

The connection between line splitting and the number of AR parameters estimated (model order) highlights a problem area common to all methods of AR spectrum analysis -- how to select the AR model order. Akaike [7] has suggested two popular criteria for order determination. However, this author's experience has shown that most order selection rules, including Akaike's, are not enough to be effective against the line splitting phenomenon.

A second major problem area with the Burg algorithm, as with the Yule-Walker case, is the bias in the positioning of spectral peaks with respect to the true frequency location of those peaks. If one defines the foldover frequency as $f_s = 1/2\Delta t$, where Δt is the sample rate, then it has been observed in real-valued data that spectral peaks with fractional frequencies from 0 to $.5f_s$ tend to be biased higher in frequency than their actual values. Those peaks with fractional frequencies from $.5f_s$ to $1.0f_s$ tend to be biased lower in frequency than their actual values. Swingler [8] has shown that the bias can pull the peak off frequency by as much as 16% of a resolution cell when using the Burg algorithm.

In order to alleviate the spectral line splitting problem, Fougere [9] devised a rather complicated gradient descent algorithm for AR parameter estimation. The algorithm has been shown to work for selected one and two sinusoid examples, but it is an iterative procedure that requires a much higher computational effort than the popular Burg algorithm.

This paper presents a new algorithm for AR parameter estimation that yields AR spectra with no apparent line splitting and reduced spectral peak frequency estimation biases. A set of sensitive stopping rules for order selection has been found for the algorithm. The method is based on an unconstrained least squares estimation of the AR parameters first proposed by Ulrych and Clayton [10], who termed it the least squares (LS) AR spectral estimate. In their experiments with the LS estimate, they observed, for processes with one and two sinusoids in noise, that LS generated spectra had less variation of

the spectral peaks from their actual frequencies as a function of initial phase than Burg algorithm spectra. Nuttall [11] compared the LS spectral estimate (which he termed the forward and backward prediction method) to other AR spectral estimates, including the Burg estimate, for a large ensemble of sampled AR processes. He found the LS estimate to be as good as, and often better than, the other estimators. In fact, among all AR estimation techniques examined by Nuttall, the LS method exhibited the least variation in frequency.

A straightforward matrix solution of the linear simultaneous normal equations for the LS method of AR parameter estimation has been the usual computational approach. This requires a number of computational operations proportional to NM^2 , making it computationally unattractive relative to the more efficient Burg algorithm. This paper presents an algorithm for solution of the LS equations with a computational complexity proportional to NM , making it comparable to that of the Burg algorithm.

Burg Algorithm Estimate of the AR Spectrum

The popular approach for AR parameter estimation with data samples was introduced by John Burg in 1968. The Burg algorithm may be viewed as a constrained least squares estimation procedure. Assuming an all-pole stationary stochastic process, the forward linear prediction error is given by

$$f_{M,t} = x_{t+M} + \sum_{k=1}^M a_{M,k} x_{t+M-k} = \sum_{k=0}^M a_{M,k} x_{t+M-k} \quad (1)$$

for $1 \leq t \leq N-M$ and the backward linear prediction error is given by

$$b_{M,t} = x_t + \sum_{k=1}^M a_{M,k}^* x_{t+k} = \sum_{k=0}^M a_{M,k}^* x_{t+k} \quad (2)$$

also for $1 \leq t \leq N-M$. Note that complex-valued data is assumed, a_0 is defined as unity, the $a_{M,k}$ are the AR parameters at order M , and the x_t are the data samples.

To obtain estimates of the AR parameters, Burg minimized the sum of the forward and backward prediction error energies,

$$e_M = \sum_{t=1}^{N-M} f_{M,t} f_{M,t}^* + \sum_{t=1}^{N-M} b_{M,t} b_{M,t}^* \quad (3)$$

subject to the constraint that the AR parameters satisfy the Levinson recursion

$$a_{M,k} = a_{M-1,k} + a_{M,M}^* a_{M-1,M-k}^* \quad (4)$$

for all orders from 1 to M . This constraint was motivated by Burg's desire to have a stable AR filter (poles within the unit circle). Figure 1 is a flow-

chart of the Burg algorithm, based on a modification by Anderson [12] of the original Burg algorithm. A computational complexity analysis of the modified Burg algorithm indicates that $3NM-M^2-2N-M$ complex adds, $3NM-M^2-N+3M$ complex multiplications, and M real divisions are required. Storage of $3N+M+2$ complex words is also required.

Marple Least Squares Algorithm Estimate of the AR Spectrum

A recursive algorithm has been found by this author [13] for the exact least squares solution of the AR parameter estimates using forward and backward linear prediction. The algorithm flowchart is shown in Figure 2, although no proof is provided here.

To obtain the M normal equations for the LS algorithm, substitute (1) and (2) into (3) and determine the minimum of e_M by setting the derivatives of e_M with respect to all the AR parameters $a_{M,1}$ through $a_{M,M}$ to zero. This yields^M

$$\frac{\partial e_M}{\partial a_{M,i}} = 2 \sum_{j=0}^M a_{M,j} r_M(i,j) = 0 \quad (5)$$

for $i=1, \dots, M$, where $a_{M,0} = 1$ by definition, and

$$r_M(i,j) = \sum_{k=1}^{N-M} (x_{k+M-j} x_{k+M-i}^* + x_{k+i} x_{k+j}^*) \quad (6)$$

for $0 \leq i, j \leq M$. The minimum prediction error energy may be determined to be

$$e_M = \sum_{j=0}^M a_{M,j} r_M(0,j) \quad (7)$$

Expressions (5) and (7) can be combined into a single $(M+1)$ by $(M+1)$ matrix expression

$$R_M A_M = E_M \quad (8)$$

where

$$A_M = \begin{bmatrix} 1 \\ a_{M,1} \\ \vdots \\ a_{M,M} \end{bmatrix}, \quad E_M = \begin{bmatrix} e_M \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad R_M = \begin{bmatrix} r_M(0,0) & \dots & r_M(0,M) \\ \vdots & & \vdots \\ r_M(M,0) & \dots & r_M(M,M) \end{bmatrix} \quad (9)$$

Ulrych and Clayton [10] were the first to propose the least squares relation-

ship (5) for AR parameter estimation, in which the Levinson recursion constraint has been removed. They found the LS estimates by computing the $r_M(i,j)$ terms directly and then by solving (8) for vector A_M by matrix inversion. This requires on the order of M^3 computational operations, which places it at a computational disadvantage with respect to the Burg algorithm with M^2 operations.

Expression (9), though, has a structure that can be exploited to generate an algorithm of order M^2 operations. Although the details are not presented here, the algorithm was motivated by a similar algorithm developed by Morf et. al. [14]. Examination of R_M will show that this matrix has both hermitian symmetry [$r_M(i,j) = r_M(j,i)$] and hermitian persymmetry [$r_M(i,j) = r_M(M-i, M-j)$]. It is not Toeplitz, although it may be decomposed into a function of the Toeplitz matrix T_M ,

$$R_M = (T_M)^H T_M + (T_M^y)^H T_M^y \quad , \quad (10)$$

where

$$T_M = \begin{bmatrix} x_{M+1} & x_M & \cdot & \cdot & \cdot & x_1 \\ x_{M+2} & x_{M+1} & \cdot & \cdot & \cdot & x_2 \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ x_{2M+1} & x_{2M} & \cdot & \cdot & \cdot & x_{M+1} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ x_N & x_{N-1} & \cdot & \cdot & \cdot & x_{N-M} \end{bmatrix} \quad (11)$$

with T_M^y denoting the conjugated and reversed matrix

$$T_M^y = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_1^* & \cdot & \cdot & \cdot & \cdot & x_{M+1}^* \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ x_{N-M}^* & \cdot & \cdot & \cdot & \cdot & x_N^* \end{bmatrix} \quad (12)$$

and $*$ denoting the complex conjugate transpose operation. Thus, R_M has a structure composed of the sum of two products of Toeplitz data matrices. It is this underlying structure that allows a recursive algorithm of order M^2 operations to be generated.

The LS algorithm requires $NM+8M^2+N+7M-8$ complex additions, $NM+9M^2+N+25M-3$ complex multiplications, and $16M-4$ real divisions. The LS algorithm needs $N+4M+15$ complex-valued computer memory locations. As a typical case, consider $N=100$ samples from an $M=30$ order AR process. The total number of multiplications, adds, divisions, and storage locations for the Burg algorithm are 8181, 8059, 30, and 335 respectively. For the LS algorithm, the numbers

are 11947, 10402, 476, and 235, which is quite comparable to that required for the Burg algorithm.

Appendix A contains a FORTRAN subroutine for computation of the AR parameters via the LS algorithm. The computer version of the algorithm contains several simple checks for both numerical ill-conditioning and order selection indication. The key terms that are performance indicators of the algorithm are the error energy e_M and a divisor term called DENOM(M), which is the only divide term in the whole algorithm. Changes in these have been empirically found to be sensitive indicators of proper order selection when using the LS algorithm.

Performance of the LS Algorithm

A distinct difference in performance between the LS and Burg algorithms is illustrated by the spectra in Figure 3. A 41-point sample sequence was generated, consisting of three sinusoidal components at fractional frequencies of .3155f, .5155f, and .7655f. All sinusoids had initial phases of 45°. A gaussian white noise sequence was generated and the sinusoid amplitudes were selected to yield SNRs of 43 dB, 37 dB, and 37 dB respectively. The frequencies, initial phases, SNRs, and data segment length of 41 samples were selected based on conditions established by Fougere [6] as being the most likely to produce spectral line splitting.

Using the Final Prediction Error (FPE) criterion of Akaike [7] as the rule for order selection, the minimum FPE of the 41-point sequence with the Burg algorithm was found at order 23. The AR spectrum based on the 23 AR parameters estimated by the Burg algorithm is shown in Figure 3a. Extreme line splitting occurs at each of the three peaks of interest. In addition, many spurious low level peaks are apparent in the spectrum. This illustrates the erroneous results that may occur when an improper order for the AR estimate is selected. Using the same sample sequence, the LS algorithm selected order 7, based on the rules of order selection discussed in Appendix A. The AR spectrum of the LS-algorithm-estimated AR parameters is shown in Figure 3b. The spectrum has three sharp peaks at the correct frequencies with no spectral line splitting. For comparison, an AR spectrum using the Burg Algorithm for order 7 is shown in Figure 3c. There is no apparent line splitting, illustrating the necessity for proper order selection. Comparing the spectra of Figures 3b and 3c, it may be seen that the skirts for each spectral peak are more narrow for the LS spectrum than for the Burg spectrum. This shows that the poles have moved closer to the unit circle with the LS approach than with the Burg approach.

Ulrych and Clayton [10] have examined the sensitivity of the Burg and LS spectra to the initial phase of processes consisting of one or two sinusoids in noise. They found that the LS estimate is fairly insensitive to the initial phase and yields an accurate determination of the sinusoid frequency, whereas the Burg estimate had severe variance in the frequency location of the spectral peak as a function of initial phase.

Nuttall [11] has examined the performance of the LS approach for a non-sinusoidal process. He generated real-valued sequences from the fourth order AR process

$$x_k = \sum_{n=1}^4 a_n x_{k-n} + w_k \quad (13)$$

where $a_1=2.7607$, $a_2=-3.8106$, $a_3=2.6535$, and $a_4=-0.9238$. White gaussian noise w_k was added to the process to yield a 10 dB SNR. One hundred Burg and LS spectra were generated from independent 40-sample sequences of this AR process in steady state. Figures 4a and 4b show the overlapped spectra for the two algorithms, while Figures 4c and 4d indicate the average of all 100 spectra for each algorithm compared to the true AR spectrum. The model order was preselected at $M=4$. One observation that can be made is that the LS technique tends to have less variability in the skirts, but more spiky estimates near the peaks of the spectrum, than seen in Burg algorithm spectra. That is, the LS algorithm produces AR spectra with less frequency variability, but more power spectral density variability. The greater PSD variability can be attributed to the fact that, unlike the Burg algorithm, the LS algorithm does not restrict the poles from moving close to the unit circle. Since the area under the spectral density curve, rather than the peak height, is proportional to power, the variability in PSD amplitude is not of much concern. Rather, to obtain unbiased, accurate estimates of the spectral peak frequencies is more important for most applications.

No cases of spectral line splitting have been observed using the order selection criteria given in Appendix A. In practice, then, the LS algorithm appears to yield AR parameters that produce stable spectra, even when pole estimates fall outside the unit circle (less than 1% of the time).

Summary

A new recursive algorithm that provides AR parameters for an AR spectral estimate based on forward and backward linear prediction has been introduced. It has the same order of computational complexity as the popular Burg algorithm. Examples have been provided to illustrate the improved performance of spectra generated with the LS algorithm when compared to spectra generated with the Burg algorithm. Improvements include reduced sensitivity to initial phase, reduced bias in the frequency estimate, less frequency variability over an ensemble of spectra made from the same process, and absence of spectral line splitting. All these factors suggest that the LS algorithm is an attractive alternative to the Burg algorithm for AR spectral estimation.

References

- [1] Marple, S.L. Jr., 1978, "Frequency Resolution of High Resolution Spectrum Analysis Techniques," in *Proc. First RADC Spectrum Estimation Workshop* pp.19-35.
- [2] Ulrych, T.J. and Bishop, T.N., 1975, "Maximum Entropy Spectral Analysis and Autoregressive Decomposition," *Rev.Geophysics*, vol.13, pp.183-200.
- [3] Kay, S. and Marple, S.L. Jr., 1979, "Sources of and Remedies for Spectral Line Splitting in Autoregressive Spectrum Analysis," in *Record IEEE ICASSP*, pp. 151-154.
- [4] Anderson, N.O., 1974, "On the Calculation of Filter Coefficients for Maximum Entropy Analysis," *Geophysics*, vol. 39, pp. 69-72.

- [5] Burg, J.P., *Maximum Entropy Spectral Analysis*, 1975, PhD Thesis, Dept. of Geophysics, Stanford University.
- [6] Fougere, P.F., Zawalick, E.J., Radoski, H.R., 1976, "Spontaneous Line Splitting in Maximum Entropy Power Spectrum Analysis," *Physics Earth and Planetary Interiors*, vol. 12, pp. 201-207.
- [7] Akaike, H., 1970, "Statistical Predictor Identification," *Ann.Inst.Stat. Math.*, vol. 22, pp. 203-217.
- [8] Swingler, D.N., 1979, "A Comparison Between Burg's Maximum Entropy Method and a Nonrecursive Technique for the Spectral Analysis of Deterministic Signals," *Jour.Geo.Res.*, vol. 84, pp. 679-685.
- [9] Fougere, P.F., 1977, "A Solution to the Problem of Spontaneous Line Splitting in Maximum Entropy Power Spectrum Analysis," *Jour.Geo.Res.*, vol. 82, pp. 1051-1054.
- [10] Ulrych, T.J. and Clayton, R.W., 1976, "Time Series Modelling and Maximum Entropy," *Physics Earth and Planetary Interiors*, vol.12, pp. 188-200.
- [11] Nuttall, A.H., 26 March 1976, "Spectral Analysis of a Univariate Process with Bad Data Points, Via Maximum Entropy and Linear Predictive Techniques," *Naval Underseas Systems Command*, TR 5303.
- [12] Anderson, N.O., 1978, "Comments on the Performance of Maximum Entropy Algorithms," *Proc.IEEE*, vol. 66, pp. 1581-1582.
- [13] Marple, S.L. Jr., paper submitted for publication to *IEEE Trans.ASSP*.
- [14] Morf, M., Dickinson, B., Kailath, T., Vieira, A., 1977, "Efficient Solution of Covariance Equations for Linear Prediction," *IEEE Trans.Acoustics, Speech, & Sig.Proc.*, vol. ASSP-25, pp. 429-433.

Appendix A

Figure 5 is a FORTRAN subroutine listing for implementation of the LS algorithm with complex-valued data. The subroutine is dimensioned to accept 512 data values and is fixed to compute a maximum of 50 AR parameters. Note that arrays C and D must be dimensioned by one more than the number of maximum AR coefficients.

The following input parameters are passed to the subroutine:

X = Array of complex-valued data samples
N = Number of data samples in array X
MMAX = Maximum number of AR parameters to be estimated
TOL = Tolerance value for two of the stopping criteria. Empirically set to 10^{-3} for minicomputer implementation and to 10^{-4} for large scale computer implementation.

The following output parameters are passed from the subroutine:

M = Number of AR parameters computed when a stopping criteria was satisfied; note that $M \leq \text{MMAX}$.
A = Array of complex-valued AR parameters
P = Prediction error energy (e_M in text)
ENERGY = Twice the total signal energy in the data samples (Eq.101)
STATUS = Integer indicating stopping criteria that terminated the recursion at order M.

Five values of STATUS are possible. STATUS=1 is the normal program exit when the maximum order is reached, $M = \text{MMAX}$. STATUS=2 indicates the program terminated when $e_M/\text{ENERGY} < \text{TOL}$, that is, the residual prediction error energy is a small fraction of the total signal energy. STATUS=3 indicates the program terminated when $(e_{M-1} - e_M)/e_{M-1} < \text{TOL}$, that is, the residual prediction error energy at order M has changed by only a small fraction from the previous order M-1. This is the stopping criteria encountered most frequently. STATUS=4 occurs when $e_M \leq 0$, indicating numerical ill-conditioning or possibly a singular matrix. This was the stopping criteria encountered in Figure 4b for M=8. As a result of this condition, the order M=7 was selected for valid AR parameter estimates. STATUS=5 indicates the algorithm terminated when $\text{DENOM}(M) \leq 0$. This is also an indicator of numerical ill-conditioning within the algorithm, since DENOM must be positive-valued.

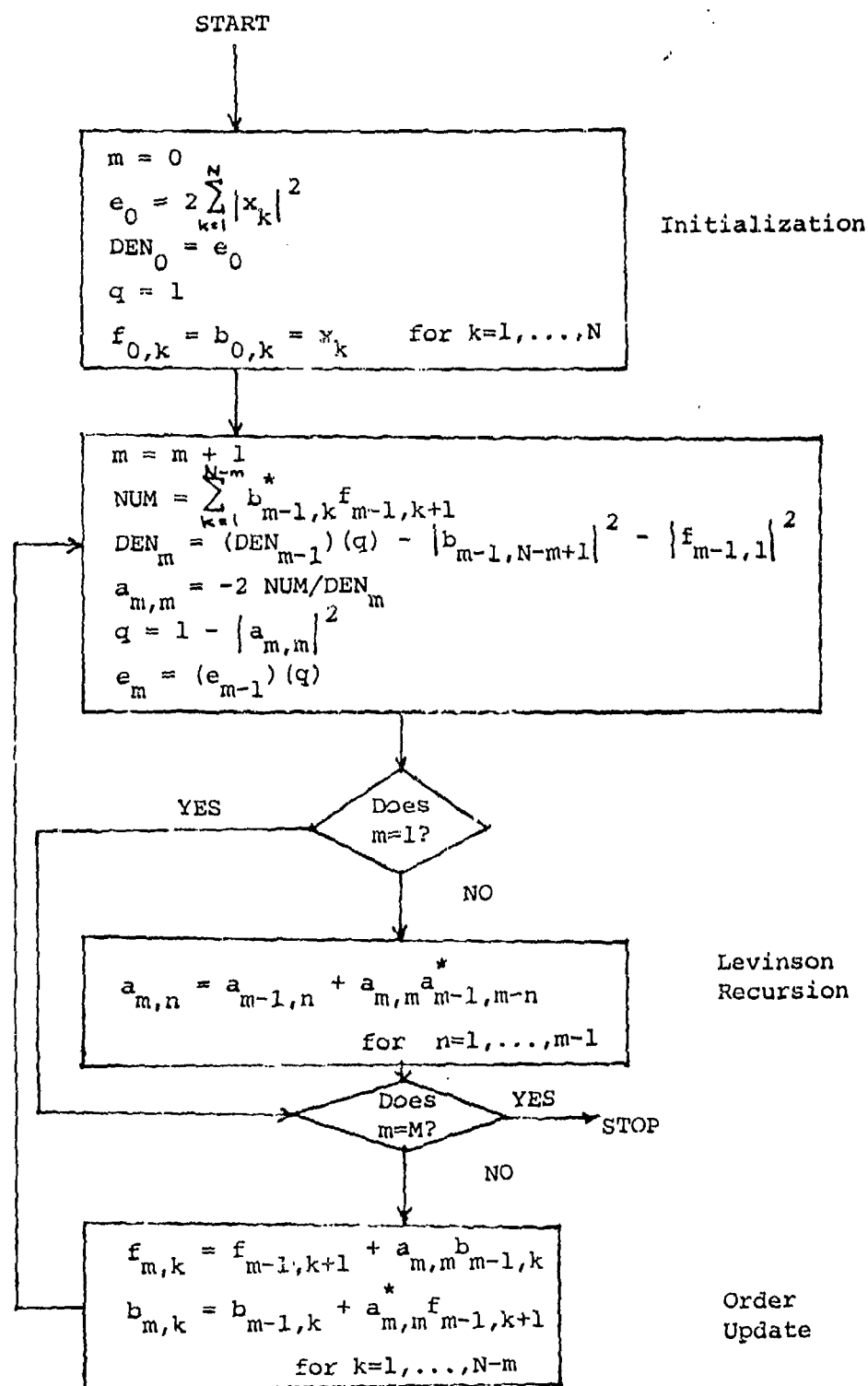
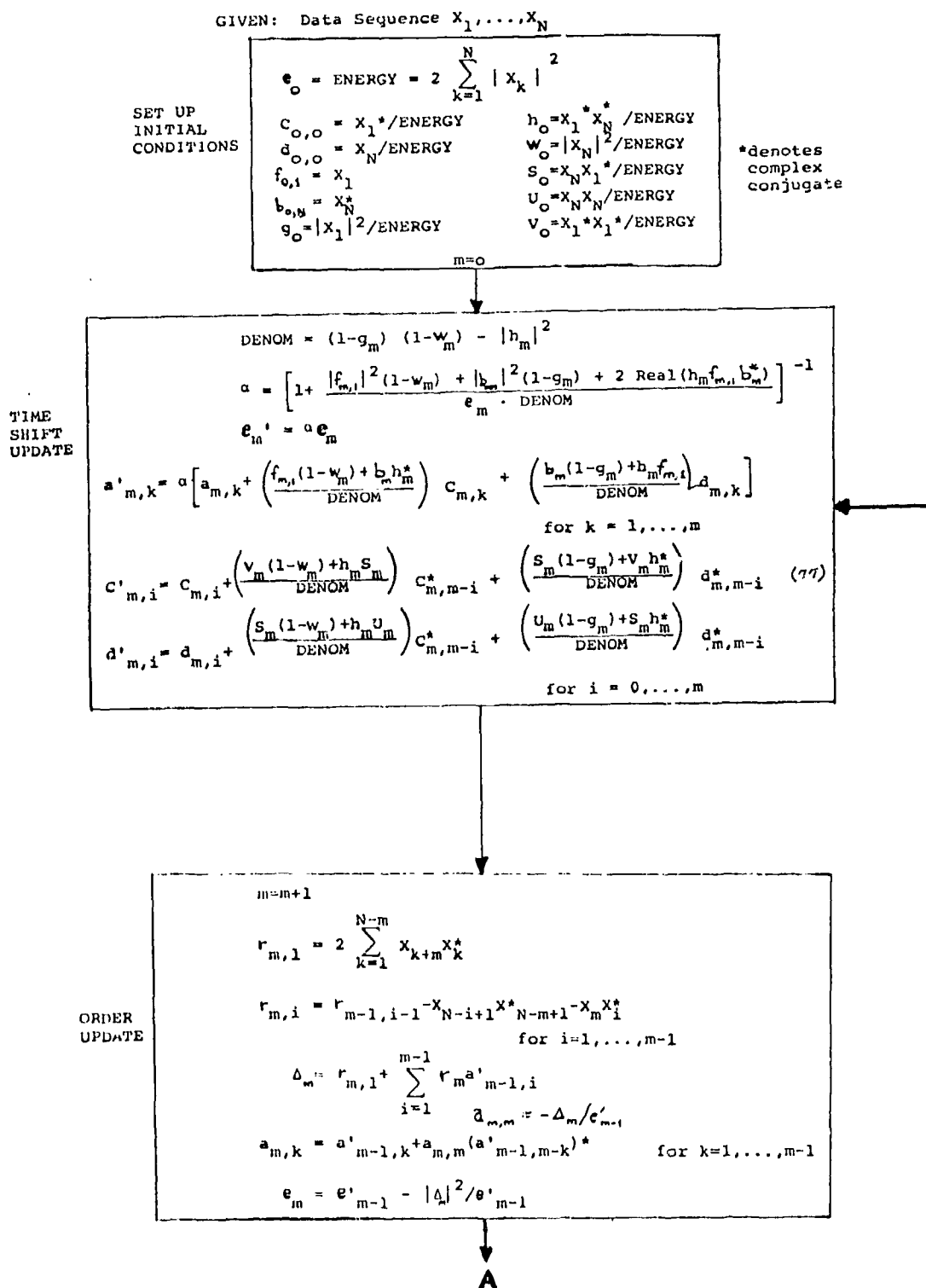
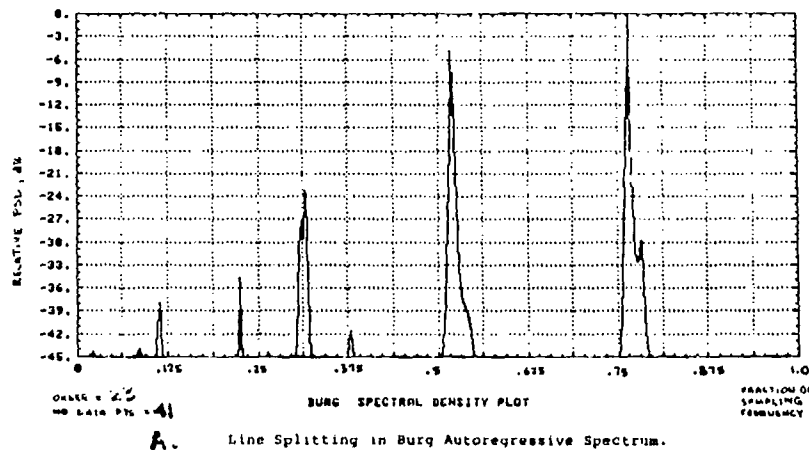


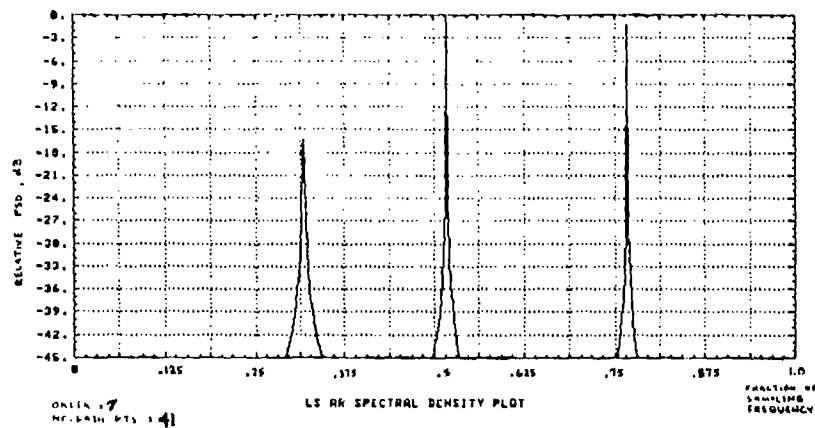
FIGURE 1. Burg Algorithm Flowchart.

FIGURE 2. Marple Least Squares Forward and Backward Linear Prediction Algorithm.

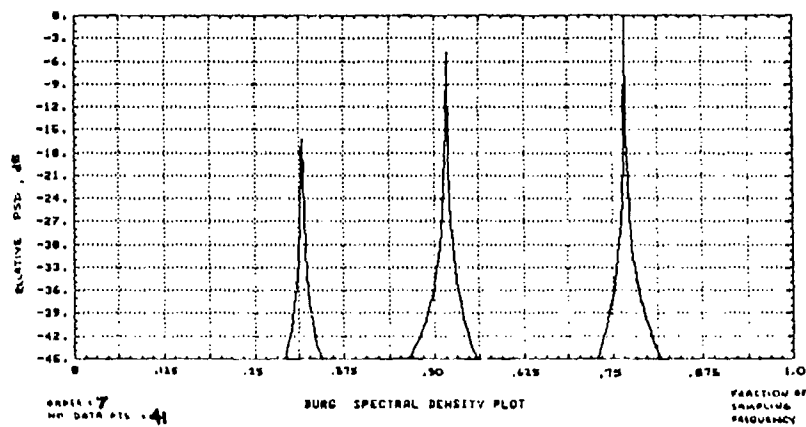




A. Line Splitting in Burg Autoregressive Spectrum.



B. Autoregressive PSD Estimate Using the Marple Algorithm.



C. Burg AR Spectrum for Three Sinusoids in White Noise.

FIGURE 3. Comparison of Burg and LS Algorithm Performances.

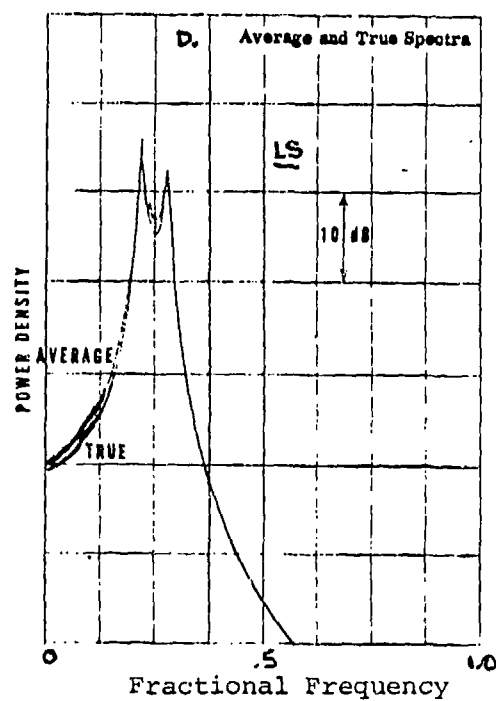
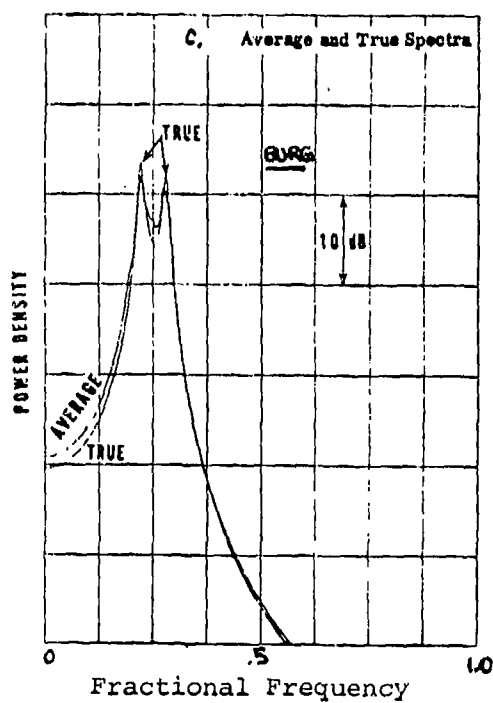
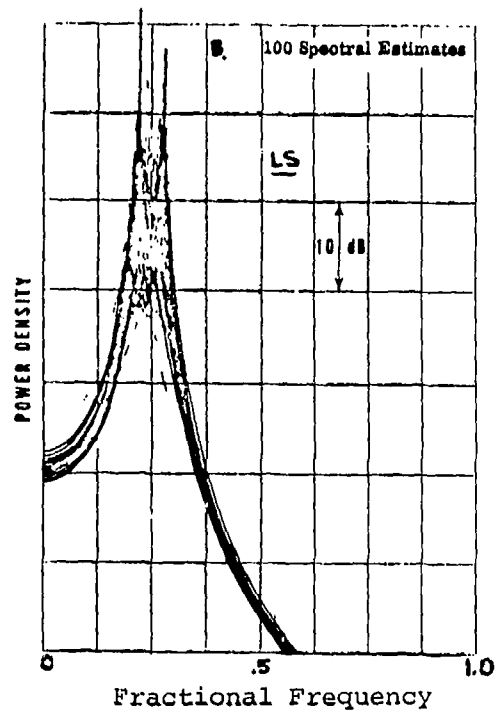
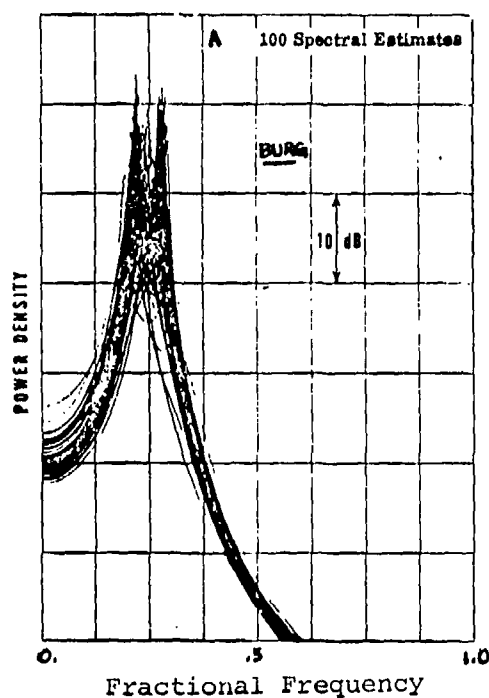


FIGURE 4. Performances of Burg and LS Algorithms to a Fourth Order AR Process (From Nuttall [11]).

FIGURE 4. Continued...

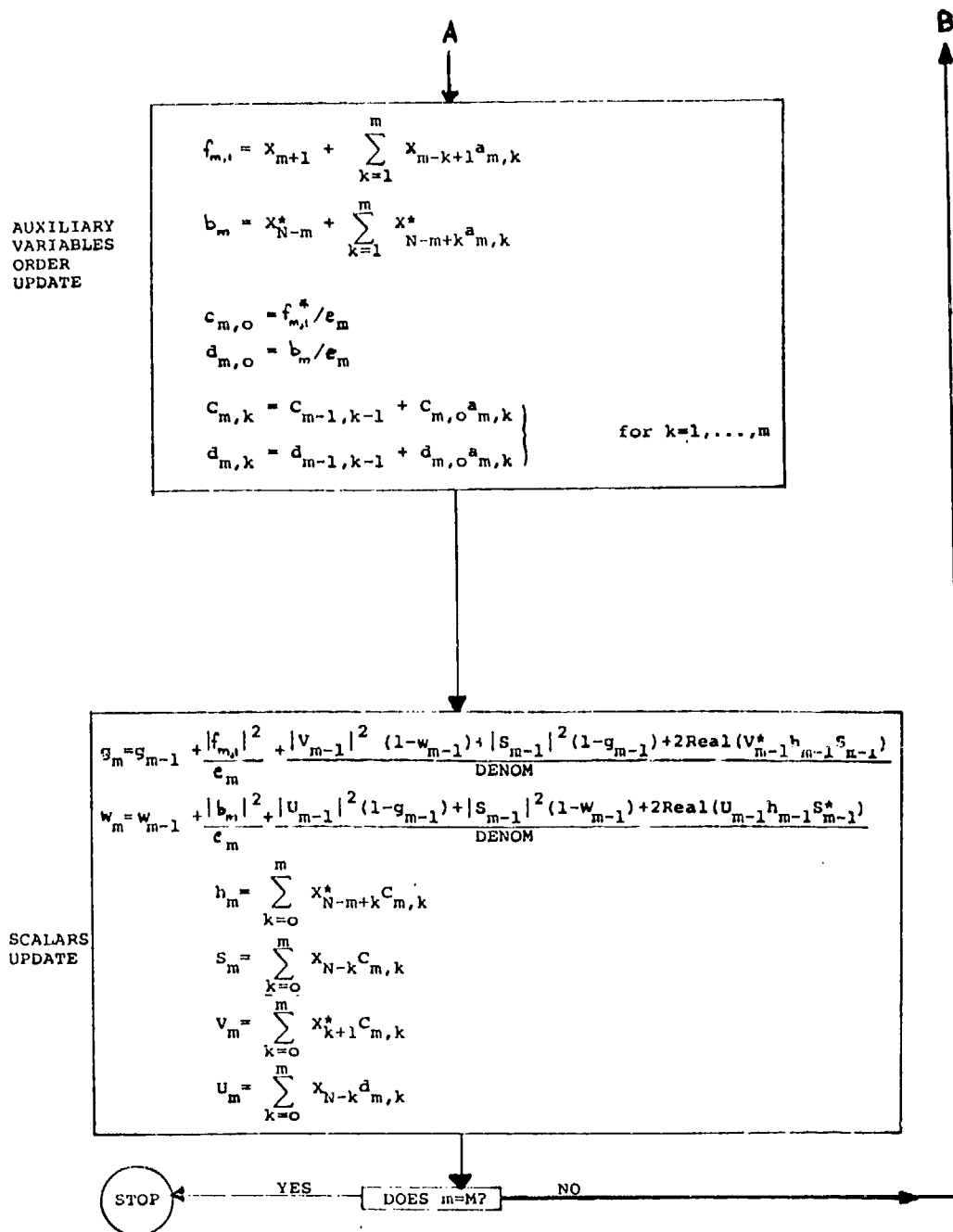


FIGURE 5. FORTRAN Program of LS Algorithm.

```

SUBROUTINE LSTSOS (N,M,MMAX,X,A,P,TOL,STATUS,ENERGY)
  COMPLEX A(50), X(512)
  COMPLEX C(51), D(51), CORR(50), E, F, H, S, U, V
  COMPLEX SAVE1, SAVE2, DELTA, C1, C2, C3, C4, C5, C6
  INTEGER STATUS

C
C  INITIALIZATION SECTION (0TH ORDER)
C
  ENERGY=0.
  DO 10 I=1,N
    ENERGY=ENERGY+CABS(X(I))**2
  10 ENERGY=2.*ENERGY
  P=ENERGY
  E=X(1)
  F=CONJG(X(N))
  H=(CONJG(X(1))*CONJG(X(N)))/ENERGY
  S=(X(N)*CONJG(X(1)))/ENERGY
  V=(CONJG(X(1))*CONJG(X(1)))/ENERGY
  U=(X(N)*X(N))/ENERGY
  G=CABS(X(1))**2/ENERGY
  R=CABS(X(N))**2/ENERGY
  C(1)=CONJG(X(1))/ENERGY
  D(1)=X(N)/ENERGY
  M=0

C
C  TIME-SHIFTED VARIABLES UPDATE
C
  1000 POLD=P
  DENOM=(1.-G)*(1.-R)-CABS(H)**2
  IF (DENOM .NE. 0) GO TO 20
  STATUS=5
  RETURN
  20 C1=H*E*CONJG(F)
  R1=2.*REAL(C1)
  R2=(1.-R)*CABS(E)**2+(1.-G)*CABS(F)**2+R1
  ALPHA=1./(1.+R2/DENOM*P)
  P=ALPHA*P
  C1=(E*(1.-R)+F*CONJG(H))/DENOM
  C2=(F*(1.-G)+H*E)/DENOM
  C3=(H*S+V*(1.-R))/DENOM
  C4=(V*CONJG(H)+S*(1.-G))/DENOM
  C5=(H*U+S*(1.-R))/DENOM
  C6=(S*CONJG(H)+U*(1.-G))/DENOM
  IF (M .EQ. 0) GO TO 40
  DO 30 I=1,M
    I1=I+1
  30 A(I1)=ALPHA*(A(I)+C1*C(I1)+C2*D(I1))
  40 M2=M/2+1
  DO 50 I=1,M2
    MI=M+2-I
    SAVE1=CONJG(C(I))
    SAVE2=CONJG(D(I))
    C(I)=C(I)+C3*CONJG(C(MI))+C4*CONJG(D(MI))
    D(I)=D(I)+C5*CONJG(C(MI))+C6*CONJG(D(MI))
    C(MI)=C(MI)+C3*SAVE1+C4*SAVE2
    D(MI)=D(MI)+C5*SAVE1+C6*SAVE2
  50 CONTINUE

C
C  ORDER UPDATE
C
  M=M+1
  DELTA=(0.,0.)
  IF (M .EQ. 1) GO TO 70
  DO 60 JJ=2,M
    J=M-JJ+1
    CORR(J+1)=CORR(J)-X(N-J+1)*CONJG(X(N-M+1))-X(M)*CONJG(X(J))
  60 DELTA=DELTA+CORR(J+1)*A(J)
  70 C1=(0.,0.)
  NM=N-M
  DO 80 K=1,NM
    C1=C1+X(K+M)*CONJG(X(K))
  80

```

THIS PAGE IS BEST QUALITY PRACTICABLE
FROM COPY FURNISHED TO DDC

```

CORR(1)=2.*C1
DELTA=DELTA+CORR(1)
C1=DELTA/P
A(M)=-C1
HOLD=P
P=HOLD-CABS(DELTA)**2/HOLD
IF (M.EQ. 1) GO TO 100
M2=M/2
DO 90 I=1,M2
  MI=M-I
  SAVE1=CONJG(A(I))
  A(I)=A(I)-C1*CCNJG(A(MI))
  IF (I.EQ. MI) GO TO 90
  A(MI)=A(MI)-C1*SAVE1
90  CONTINUE
C
C  PREDICTION COEFFICIENTS UPDATE
C
100  E=X(M+1)
      F=CONJG(X(N-M))
      DO 110 I=1,M
        MI=M+1-I
        E=E+X(MI)*A(I)
        F=F+CONJG(X(N-M+I))*A(I)
110
C
C  AUXILIARY VARIABLES ORDER UPDATE
C
      C1=CONJG(E)/P
      C2=CONJG(F)/P
      DO 120 II=1,M
        I=M-II+1
        I1=I+1
        C(I1)=C(I)+C1*A(I)
120  D(I1)=D(I)+C2*A(I)
      C(1)=C1
      D(1)=C2
C
C  SCALARS UPDATE
C
      C1=H*S*CONJG(V)
      C2=U*H*CONJG(S)
      R1=2.*REAL(C1)
      R2=2.*REAL(C2)
      R1=(1.-R)*CABS(V)**2+(1.-G)*CABS(S)**2+R1
      R2=(1.-G)*CABS(U)**2+(1.-R)*CABS(S)**2+R2
      G=G+(R1/DENOM)+(CABS(E)**2/P)
      R=R+(R2/DENOM)+(CABS(F)**2/P)
      H=(0.,0.)
      S=(0.,0.)
      V=(0.,0.)
      U=(0.,0.)
      M1=M+1
      DO 130 I=1,M1
        H=H+CONJG(X(N-M-1+I))*C(I)
        S=S+X(N+1-I)*C(I)
        V=V+CONJG(X(I))*C(I)
130  U=U+X(N+1-I)*D(I)
C
C  CHECK FOR STOPPING CRITERIA
C
      IF (P.GT. 0) GO TO 200
      STATUS=4
      RETURN
200  IF (((POLD-P)/POLD).GT. TOL) GO TO 210
      STATUS=3
      RETURN
210  IF ((P/ENERGY).GT. TOL) GO TO 220
      STATUS=2
      RETURN
220  IF (M.NE. MMAX) GO TO 1000
      STATUS=1
      RETURN
END

```

ARMA SPECTRAL ESTIMATION: AN ITERATIVE PROCEDURE

James A. Cadzow
Department of Electrical Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061
(703) 961-5694

ABSTRACT

An ARMA Autocorrelation Estimation Method (AEM) for generating the best rational spectral estimate of a stationary random discrete-time series is presented. This estimation is to be based on N contiguous observations of the infinite length time-series. As in the maximum entropy method, the AEM in effect extrapolates an autocorrelation estimate beyond the data limited range with the explicit objective of achieving improved spectral resolution. Unlike the maximum entropy method, however, the AEM spectral estimator provides for the existence of zeroes as well as poles in the resultant power spectral density and it is thereby more robust in nature. Furthermore, this method does not require an excessively large number of observations to be effective, a property not shared by most other rational ARMA spectrum estimators.

This work was supported in part by the Signal Processing Section, Surveillance Technology Branch, Rome Air Development Center through the Post Doctoral Program under Contract F30602-75-0018.

I. INTRODUCTION

In a variety of applications, it is desired to identify the spectral characteristics of a wide-sense stationary discrete-time process $\{x(n)\}$. The elements of this random process may arise naturally through a discrete-time phenomenon, or, by means of uniformly sampling a wide-sense stationary continuous-time process. Whatever the case, the processes' spectral characterization is completely determined by its corresponding autocorrelation sequence

$$r_x(n) = E\{x(k) x(k+n)\} \quad n=0, \pm 1, \pm 2, \dots \quad (1)$$

in which E denotes the expected value operator. The z -transform of this sequence is commonly referred to as the "power spectral density" associated with the process and is specified by

$$S_x(z) = \sum_{n=-\infty}^{\infty} r_x(n) z^{-n} \quad (2)$$

where z is a complex valued variable. In the spectral estimation literature, one commonly replaces the z variable by $e^{j\omega}$ to obtain the equivalent Fourier transform characterization as denoted by $S_x(e^{j\omega})$. With this Fourier representation, one is able to express the spectral characterization as a function of the real frequency variable ω .

Although relationship (2) yields an explicit procedure for determining the power spectral density, its utilization is restricted to those few situations in which one has access to the entire time history of the autocorrelation sequence. In most practical applications, one has little if any such a priori knowledge. More typically, there is available only a set of N contiguous observations

$$x(1), x(2), \dots, x(N) \quad (3)$$

of the infinite length process upon which to base a spectral estimation. This inability to observe a random process over all time reflects real world constraints which prevail in most any application. The problem of concern here is then that of using this "incomplete" set of observations to estimate the underlying power spectral density. In essence we seek to use the finite data (3) to construct a function $\hat{S}_x(z)$ which best approximates $S_x(z)$ in some fashion.

The classical spectral estimation approach is to use the N process observations to compute autocorrelation estimates (i.e., $\hat{r}_x(n)$ for $n=0, \pm 1, \dots, \pm(N-1)$), and, then to take a discrete Fourier transform of a weighted version of these autocorrelation estimates (e.g., see ref. [1]). This procedure often leads to unsatisfactory results, however, since the generally erroneous assumption therein being made is that the autocorrelation

sequence is identically zero outside the data limited range $|n| \leq N-1$. This shortcoming has been recognized by investigators and a number of alternative methods that do not impose this restrictive assumption have since been developed. By in large, these methods seek to approximate the spectral density by a rational function. A rational model can be justified on the basis that any continuous power spectral density can be approximated arbitrarily closely by a rational function of sufficiently high order [2].

Undoubtably, the most widely used of these models is the "all-pole" rational spectral estimator which has given rise to the basically equivalent maximum entropy, linear predictive coding, and, autoregressive methods. In each case, one seeks to determine the coefficients of an all-pole model so as to optimize a given criteria. In the maximum entropy method, one selects the "optimum" model to be consistent with the given data observations while being simultaneously least committal about the remaining unobserved portion of the random process [3]. On the other hand, the linear predictive coding and autoregressive methods seek a data dependent whitening filter in the guise of a one-step predictor [4]. In each of these three methods, the coefficients characterizing the optimum all-pole model are obtained by solving a system of linear equations. This ease of model generation, and, the fact that all-pole models can often yield excellent spectral resolution performance for short data lengths are the primary reasons for the wide acceptance of all-pole spectral estimators. It must be mentioned, however, that these all-pole methods also have serious shortcomings. For example, if the underlying power spectral density is rational and contains zeroes as well as poles, an all-pole model can result in very poor estimates.

Conceptually, a better behaved spectral estimator would result if the rational spectrum model being used had zeroes as well as poles. In recognition of this fact, a variety of such models have been generated which typically use a whitening filter approach (e.g., see refs. [5] and [6]). These procedures have produced impressive performance when the number of data observations, N , adequately exceeds the random processes' time constant. When this is not the case, however, their spectral estimation performance falls off significantly. In order to retain the inherent advantages of using a zero-pole model while not requiring an excessively large number of data observations, a procedure which makes explicit use of the autocorrelation sequence will be now developed.

II. RATIONAL SPECTRUM MODEL

In this section, the principal implication of assuming a rational power spectral density model is investigated. The random process $\{x(n)\}$ is said to have a rational spectrum if its power spectral density can be expressed in the form

$$S_x(z) = \sigma^2 \frac{B(z) B(z^{-1})}{A(z) A(z^{-1})} \quad (4)$$

where σ^2 is a positive real scalar and the spectrum's characteristic rational function

$$\frac{B(z)}{A(z)} = \frac{1 + b_1 z^{-1} + b_2 z^{-2} + \dots + b_q z^{-q}}{1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_p z^{-p}} \quad (5)$$

is composed of polynomials $A(z)$ and $B(z)$ which have real coefficients, and, the zeroes of these polynomials all lie within the unit circle of the z -plane. This rational power spectral density is said to have order (p, q) and its zeroes and poles are seen to occur in sets of complex conjugate reciprocals.

The fact that the denominator polynomial $A(z)$ has all its zeroes located inside the unit circle enables us to provide a convenient system interpretation of this rational discrete-time process. In particular, let us consider the stable recursive linear system whose transfer function is specified by the characteristic rational function (5). This linear system is then governed by the p th order linear difference equation

$$\begin{aligned} x(n) = & \epsilon(n) + b_1 \epsilon(n-1) + b_2 \epsilon(n-2) + \dots + b_q \epsilon(n-q) \\ & - a_1 x(n-1) - a_2 x(n-2) - \dots - a_p x(n-p) \end{aligned} \quad (6)$$

in which $\epsilon(n)$ and $x(n)$ denote the excitation and response signals, respectively. It can be readily shown that if this system is excited by a stationary white noise process as statistically characterized by

$$E\{\epsilon(n)\} = 0 \quad \text{and} \quad r_\epsilon(n) = \sigma^2 \delta(n) \quad (7)$$

then the power spectral density of the response random process $\{x(n)\}$ is given precisely by expression (4). Thus, a stationary random process with a rational power spectral density can always be interpreted as being the response of a linear system to a white noise excitation. This linear system is then said to have colored the excitation process and for this reason we commonly refer to the system as a coloring filter as suggestively depicted in Figure 1.

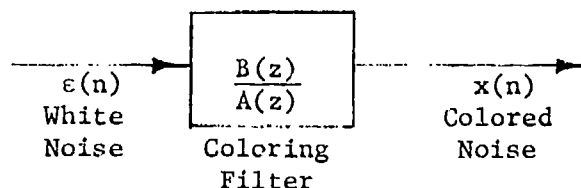


Figure 1. Generation of a Rational Spectrum

In the spectrum estimation literature, the general linear recursive system (6) is commonly referred to as an autoregressive-moving average (ARMA) model. An ARMA model is seen to give rise to a rational spectrum (4) which contains zeroes (via $B(z)$ $B(z^{-1})$) as well as poles (via $A(z)$ $A(z^{-1})$) and is often referred to as a zero-pole model. The ARMA model is the most general rational spectrum model possible. When the numerator polynomial is constrained to be one (i.e., $B(z) = 1$), the important subclass of autoregressive (AR) models is obtained. This all-pole model is the one most often used in spectral estimation primarily due to the ease with which one can determine the optimum $A(z)$ polynomial which correspond to a given finite set of observations (3). It is to be noted that an AR model arises whenever one uses the basically equivalent maximum entropy, linear predictive coding, or, autoregressive methods of spectral estimation. Another subclass of rational spectrum models is obtained by constraining the denominator polynomial to be one (i.e., $A(z) = 1$). This all-zero model is commonly referred to as the moving average (MA) model. The rational spectrum associated with each of these models is shown in Table 1.

Process	Spectrum
ARMA	$\sigma^2 B(e^{j\omega})/A(e^{j\omega}) ^2$
AR	$\sigma^2 / A(e^{j\omega}) ^2$
MA	$\sigma^2 B(e^{j\omega}) ^2$

Table 1. Rational Power Spectral Density Classes

An examination of Table 1 reveals the more robust behavior of the ARMA model in providing rational spectral estimates. This robustness was recently demonstrated empirically in reference [5] where the ARMA model was found to yield overall superior spectral estimates for a variety of problems. Unless one has a prior knowledge which would indicate otherwise, it then seems clear that the ARMA model provides the obvious choice when seeking a rational spectral estimate. Unfortunately, the practical problem of determining the optimum $A(z)$ and $B(z)$ polynomials which constitute the ARMA model is an analytically intractable one and necessitates an algorithmic solution. Moreover, unless the observed data length N is sufficiently large, the standard whitening filter approach can yield poor spectral estimates. A procedure for resolving this shortcoming will now be presented.

III. AUTOCORRELATION APPROXIMATION MODEL

In order to remove the apparent incompatibility of determining general ARMA spectral estimates from short data length observations, it will be beneficial to examine the autocorrelation sequence. In particular, our interest will be directed towards the causal segment of the autocorrelation sequence as defined by

$$r_x^+(n) = \begin{cases} r_x(n) & n \geq 0 \\ 0 & n < 0 \end{cases} \quad (8)$$

Using the fact that the autocorrelation sequence is an even function of n , the following expression is readily established

$$r_x(n) = r_x^+(n) + r_x^+(-n) - r_x(0) \delta(n) \quad (9)$$

where $\delta(n)$ denotes the Kronecker delta sequence. Upon taking the z -transform of this relationship, we obtain the associated power spectral density function

$$S_x(z) = S_x^+(z) + S_x^+(z^{-1}) - r_x(0) \quad (10)$$

in which $S_x^+(z)$ denotes the z -transform of the causal sequence $r_x^+(n)$. Clearly, there exists a one-to-one correspondence between the two z -transforms $S_x(z)$ and $S_x^+(z)$.

When the power spectral density is of the rational form as given in expression (4), a little thought will convince oneself that $S_x^+(z)$ must be of the specific form

$$S_x^+(z) = \frac{c_0 + c_1 z^{-1} + \dots + c_p z^{-p}}{1 + a_1 z^{-1} + \dots + a_p z^{-p}} \quad (11)$$

in which this representation's denominator polynomial corresponds to the $A(z)$ polynomial of expression (5). Upon multiplying both sides of this representation by $A(z)$ and then taking the inverse z -transform, the fundamental recursive relationship

$$\begin{aligned} r_x^+(n) &= c_0 \delta(n) + c_1 \delta(n-1) + \dots + c_p \delta(n-p) \\ &\quad - a_1 r_x^+(n-1) - a_2 r_x^+(n-2) - \dots - a_p r_x^+(n-p) \end{aligned} \quad (12)$$

arises in which the boundary conditions $r_x^+(n) = 0$ for $n < 0$ are imposed to reflect the causal nature of the $r_x^+(n)$ sequence. Thus, if a random process has a rational spectrum of order p , the elements of the associated autocorrelation sequence will satisfy this recursive relationship for appropriate choices of the a_k, c_k coefficients.

Upon examination of relationship (12), it is apparent that knowledge of the first $2p + 1$ values of the autocorrelation sequence will enable us to uniquely identify its characterizing a_k, c_k coefficients. This information can then, in turn, be used to construct the underlying power spectral density via expression (10). We shall now use a version of this approach to effect a data efficient means of spectral estimation. This version must take into account the fact that the autocorrelation element values are not known a priori. As such, the first step necessitates the generation of autocorrelation estimates based on the N data observations provided. A standard estimation rule for achieving this objective is given by

$$\hat{r}_x(n) = \frac{1}{N-n} \sum_{k=1}^{N-n} x(k) x(n+k) \quad \text{for } n=0, 1, \dots, N-1 \quad (13)$$

in which it is tacitly assumed that $2p + 1 < N$. One can readily verify that this autocorrelation estimate is unbiased and that its variance generally increases for increasing values of the lag index n (e.g., see ref. [7]). This statistical behavior reflects a growing lack of confidence in the autocorrelation estimate as n increases. This confidence factor will be taken into account in what is to follow.

We next seek to determine values for the a_k, c_k coefficients governing model (12) which will be most consistent with these generated autocorrelation sequence estimates. A generally accepted measure of consistency is provided by the mean squared error criterion as given by

$$I(a_k, c_k) = \sum_{n=0}^{N-1} w(n) [\hat{r}_x(n) - r_x^+(n)]^2 \quad (14)$$

where $w(n)$ is a nonnegative weighting sequence used to reflect our decreasing confidence in $\hat{r}_x(n)$ as n increases. Our objective is to then select the a_k, c_k coefficients so that the sequence $r_x^+(n)$ as generated by model (12) best approximates $\hat{r}_x(n)$ in the sense of minimizing criterion (14). This is an analytically intractable problem and its eventual solution necessitates an algorithmic approach. The ultimate success of the spectral estimation procedure here described depends critically upon the algorithm used. The linearization algorithm, as described in references [8] and [9], has proven to be a significantly more effective tool than the standard gradient method [10]. As with all algorithms, the initial coefficient selection plays an important role in regards to how quickly the linearization algorithm converges and to which relative minimum it converges.

A particularly effective initial coefficient selection method is to be found in ref. [11].

A summary of the proposed spectral estimation method is outlined in Table 2. It is important to note that the particular autocorrelation estimate, and, optimum model algorithm to be used have not been specified. Suggested procedures have been offered in this section, but the most effective selections remains a subject of future research.

Step 1:	Generate an autocorrelation estimate $\hat{r}_x(n)$ from the N data observations.
Step 2:	Determine the causal autocorrelation model (12) which is most consistent with the estimated autocorrelation sequence obtained in Step 1. One may employ an other than mean squared error criterion for measuring this consistency.
Step 3:	Construct the spectral estimate using the relationship $\hat{S}_x(e^{j\omega}) = S_x^+(e^{j\omega}) + S_x^+(e^{-j\omega}) - \hat{r}_x(0)$ $= 2 \operatorname{Re}[S_x^+(e^{j\omega})] - \hat{r}_x(0) \quad (15)$ in which the most consistent model $S_x^+(z)$ found in Step 2 is used.

Table 2. Basic Steps of the Proposed Spectral Estimation Method.

IV. NUMERICAL EXAMPLES

A spectral estimation problem which arises in a surprisingly large number of applications is that of the detection and parameter identification of sampled sinusoids from noise contaminated measurements. This particular class of problems serves as an effective means for measuring the performance of spectrum estimators relative to (i) detecting the presence of sinusoids when the additive noise is strong, and (ii) resolving two or more sinusoids whose frequencies are closely spaced. In this section, we shall apply the autocorrelation estimation method (AEM) to estimate the spectrum of the fourth order ARMA generated data as governed by

$$x(n) = A \sin[0.4\pi n] + A \sin[0.426\pi n] + w(n) \quad 0 \leq n \leq 63 \quad (16)$$

in which $\{w(n)\}$ is a zero-mean white noise Gaussian process with variance one. It is to be noted that the frequency spacing of the constituent sinusoids (i.e., 0.026π) is less than the resolution capability of the standard discrete Fourier transform (i.e., $2\pi/64 = 0.03125\pi$). This particular problem has been considered in detail in reference [12] where the performance of some of the more commonly used spectral estimators were compared. The individual sinusoidal signal-to-noise ratio for the above signal as expressed in decibels is given by $10 \log(A^2/2)$. In order to determine the effectiveness of the AEM in different noise environments we shall consider two sinusoidal amplitude parameter selections.

CASE I : $A = \sqrt{2000}$

When the sinusoid's amplitude is set at $A = \sqrt{2000}$, the prevailing SNR is 30 db. In this strong signal case, we shall be testing the spectral estimator's ability to resolve closely spaced (in frequency) sinusoids, and to accurately estimate their frequencies. Upon generation of the postulated sequence (16), the autocorrelation estimate was generated according to the unbiased rule (13). The linearization algorithm was next employed to obtain the best fourth order (i.e., $p = 4$) ARMA autocorrelation model in which the weighting sequence $w(n) = (N - n)^2$ was selected so as to reflect our decreasing confidence in the autocorrelation estimates for increasing values of n . This ARMA model was then used to generate the required spectral estimate according to

$$\hat{S}_x(e^{j\omega}) = 2\text{Re}[S_x^+(e^{j\omega})] - \hat{r}_x(0) \quad (17)$$

A plot of this AEM estimate over the normalized frequency interval is shown in Figure 2a where the frequency resolution capabilities are clearly evident. The estimated center frequencies were found to correspond almost exactly to the sinusoids used in generating the data. In this and the plots to follow, the spectral peak has been normalized to 30 db, and no special routine has been employed to determine the amplitudes of the constituent sinusoids from this spectral estimate.

For comparison purposes, a covariance AR spectral estimate of order fifteen was next generated using data (16). As demonstrated in reference [12], this particular AR spectral estimator works particularly well for this class of problems. The results of this covariance AR spectral estimate are shown in Figure 2b. As might be anticipated, the covariance AR method also yields excellent resolution capabilities in this high SNR environment. The estimated center frequencies obtained were also of good quality.

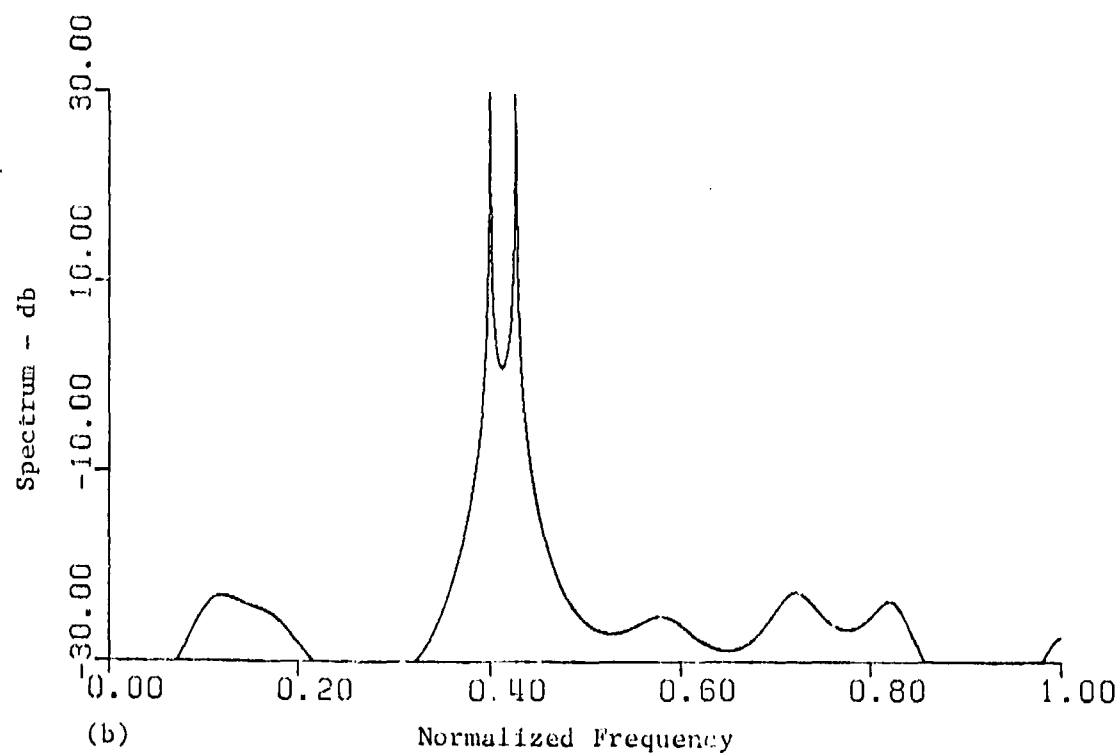
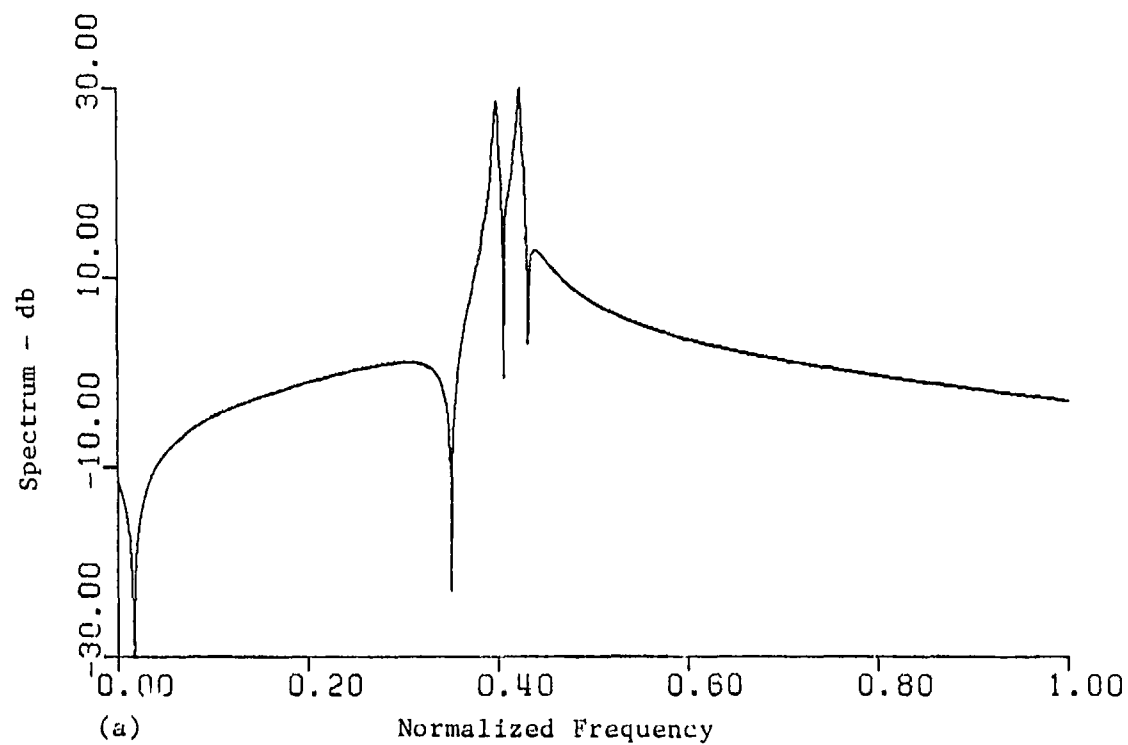


Figure 2. Plot of Spectral Estimate in Strong Signal Case SNR = 30 db.
 (a) Fourth Order AFM Spectral Estimate.
 (b) Fifteenth Order AR Spectral Estimate.

CASE II : $A = \sqrt{2}$

For this sinusoidal amplitude selection, the prevailing SNR is zero d.b. Using the same procedure as in Case I, a fourth order AEM spectral estimate for this low SNR case was found and is displayed in Figure 3a. Significantly, we are still able to detect the presence of two sinusoids and obtain reasonable estimates of the sinusoidal frequencies (i.e., 0.392 and 0.430). On the other hand, when a 15th order AR spectral estimate is generated from this data, the resultant spectrum plotted in Figure 3b demonstrates a failure to detect the two sinusoids. This inability can be attributed to the fact that the actual spectrum contains zeroes (due to the strong white noise) close to the unit circle which results in an AR spectral model mismatch. Thus, in this hostile noise environment, the AEM spectral estimator produces clearly superior results.

V. CONCLUSION

The autocorrelation estimation method (AEM) for generating an ARMA rational spectral estimate has been presented. This procedure offers the promise of achieving effective spectral estimation performance without requiring an excessively large number of data samples to do so. In order to reach its full potential, however, a number of fundamental issues have to be resolved. Perhaps the most important of these involve the specific procedure used in determining the autocorrelation estimates, and the selection of error weights used in the squared error criterion. The former is most critical since if "poor" autocorrelation estimates are used in generating the optimum ARMA model, one cannot possibly hope to achieve an accurate spectral estimate. On the other hand, even with acceptably good autocorrelation estimates, a proper weighting of model error is required in order to reflect the growing lack of confidence in the autocorrelation estimates for increasing values of n . It is felt that the weights to be used should be data dependent. Some other issues which must also be resolved are: (i) determining the order of the ARMA model, (ii) investigation of criteria other than squared error, (iii) developing fundamentally different procedures which make use of the causal autocorrelation concept herein presented for obtaining spectral estimates.

ACKNOWLEDGMENT

I would like to take this opportunity to acknowledge the contributions of Koji Ogino and Behshad Beseghi.

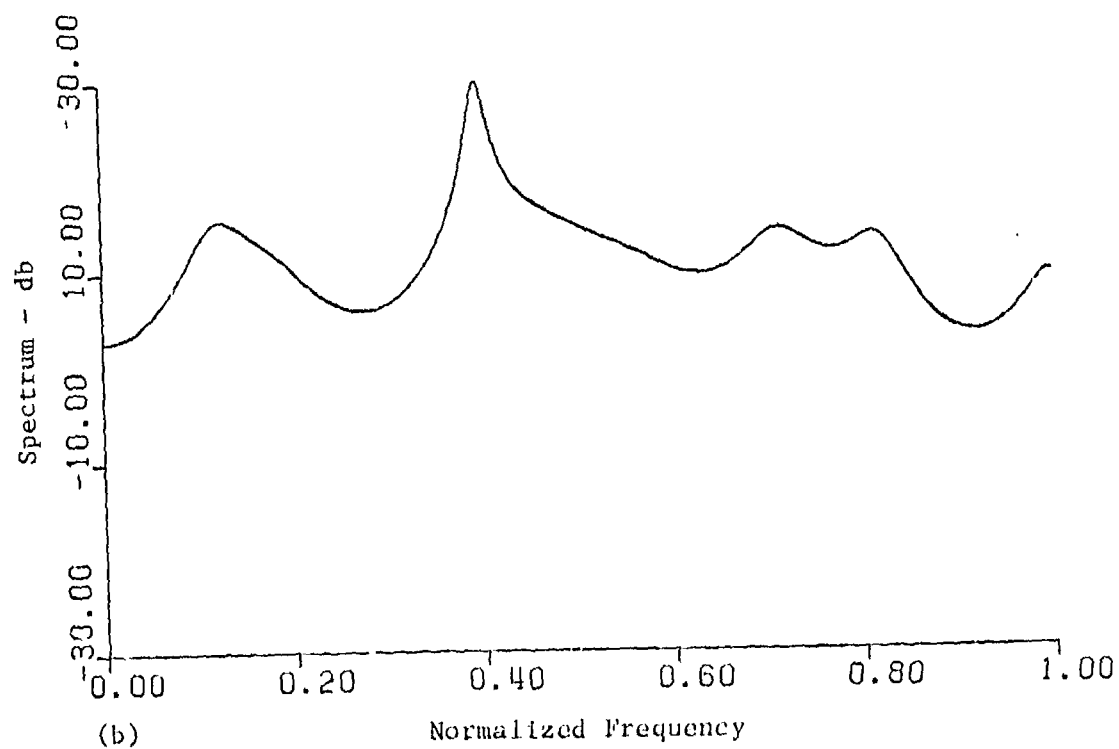
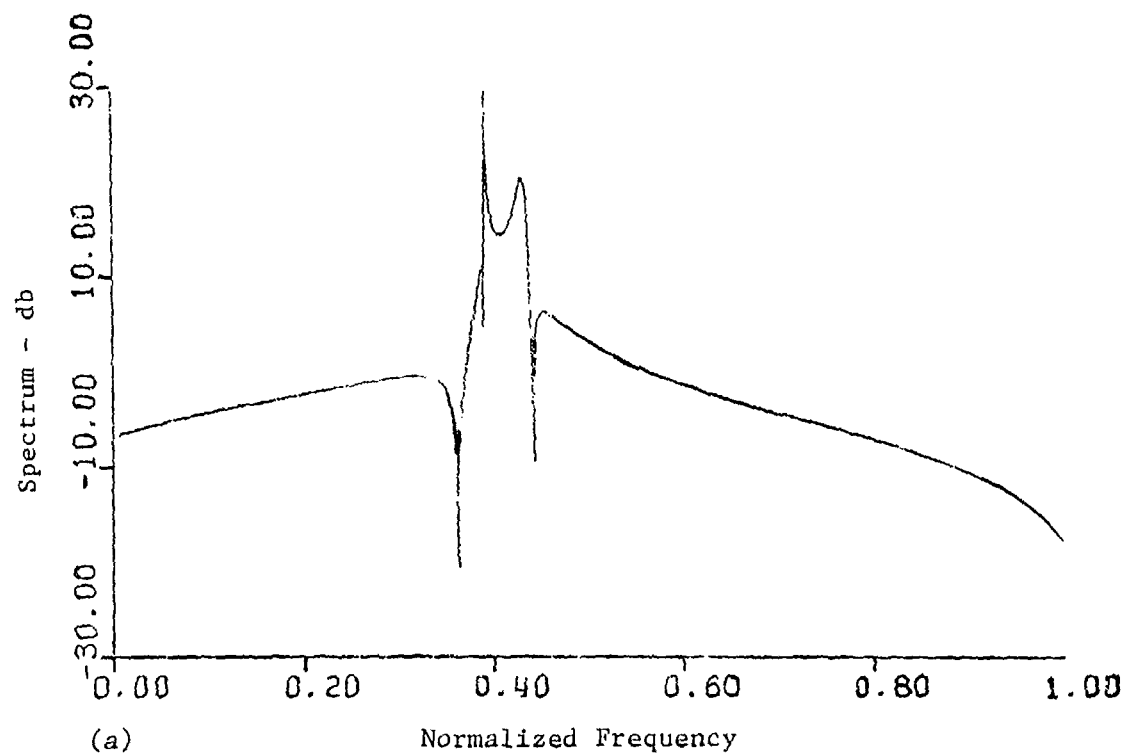


Figure 3, Plot of Spectral Estimate in Weak Signal Case SNR = 0 db.
 (a) Fourth Order AEM Spectral Estimate.
 (b) Fifteenth Order AR Spectral Estimate.

VI. REFERENCES

- [1] R. B. Blackman, and J. W. Tukey, THE MEASUREMENT OF POWER SPECTRA FROM THE POINT OF VIEW OF COMMUNICATIONS ENGINEERING, New York, Dover, 1959.
- [2] E. Cheney, INTRODUCTION TO APPROXIMATION THEORY, McGraw-Hill, New York, 1966.
- [3] J. P. Burg, "Maximum Entropy Spectral Analysis," Proceedings of the 37th meeting of the Society of Exploration Geophysicists, 1967.
- [4] J. Makhoul, "Linear Prediction, A Tutorial Review," Proc. IEEE, Vol. 63, pp. 561-580, April 1975.
- [5] P. R. Gutowski, E. A. Robinson, and S. Treitel, "Spectral Estimation: Fact or Fiction," IEEE Trans. Geoscience Electronics, Vol. GE-16, No. 2, pp. 80-84, April 1978.
- [6] S. A. Tretter and K. Steiglitz, "Power-Spectrum Identification in Terms of Rational Models," IEEE Trans. Automatic Control, Vol. AC-12, pp. 185-188, April 1967.
- [7] S. A. Tretter, DISCRETE-TIME SIGNAL PROCESSING, John Wiley and Sons, New York, 1976.
- [8] J. A. Cadzow, "Recursive Digital Filter Synthesis via Gradient Based Algorithms," IEEE Trans. Acoust., Speech, and Signal Processing, Vol. ASSP-24, No. 5, pp. 349-355, October 1976.
- [9] J. A. Cadzow, "Linear Recursive Modeling," Presented at the 1975 Pittsburgh Conf. Modeling and Simulation, University of Pittsburgh, April 1975.
- [10] G. W. Bordner, "Time Domain Design of Stable Recursive Digital Filters," Ph.D. Thesis, State University of N.Y. at Buffalo, 1974.
- [11] J. A. Cadzow, "ARMA Spectral Estimation: An Efficient Closed-Form Procedure," 1979 RADCS Spectral Estimation Workshop, October 1979.
- [12] T. M. Sullivan, O. L. Frost, J. R. Treichler, "High Resolution Signal Estimation," ARGO Systems, Inc. Tech. Rept., June 1978.

80 - Blank

ARMA SPECTRAL ESTIMATION:
AN EFFICIENT CLOSED-FORM PROCEDURE

James A. Cadzow
Department of Electrical Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061
(703) 961-5694

ABSTRACT

A closed-form procedure for generating an ARMA spectral estimate of a stationary random time series, based upon a finite set of contiguous observations, is presented. As in the maximum entropy method, this procedure in effect extrapolates an autocorrelation estimate beyond the data limited range thereby offering the possibility for improved spectral resolution in comparison to the more classical Fourier based approaches. Unlike the maximum entropy method, however, this procedure has the additional flexibility of generating a spectral model which possesses zeroes as well as poles. As such, it has a more robust behavior and therefore the capability of producing superior spectral estimation performance. This latter claim has been empirically confirmed for a number of examples in which these two methods have been applied. Significantly, the computational requirements of the two procedures are comparable. This suggests that the herein developed ARMA spectral estimator can be used as a primary tool in spectral estimation.

I. INTRODUCTION

A signal processing problem which arises in a variety of interdisciplinary applications is that of estimating the spectrum of a stationary random time series. This estimation is to be based wholly on a set of N contiguous observations of that time series as represented by

$$x(1), x(2), \dots, x(N) \quad (1)$$

The inability to monitor the entire history of the infinite length time series reflects constraints which usually prevail in virtually all real world applications. Unless some assumptions are made relative to the statistical structure of the underlying time series, the generation of a spectral estimate from

Research sponsored by the Air Force Office of Scientific Research/AFSC, United States Air Force under Contract F49620-79-C-0038. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation herein.

this finite data is a poorly posed problem. This is a direct consequence of the fact that the spectral content of a time series is completely specified by its associated autocorrelation sequence

$$r_x(n) = E\{x(k)x(k+n)\} \quad n = 0, \pm 1, \pm 2, \dots \quad (2)$$

in which E denotes the expected value operator. Clearly, there exists a basic information content incompatibility between the infinite extent autocorrelation sequence and the finite set of time series observations upon which the spectral estimate is to be based. This incompatibility is usually resolved through the process of parameterizing the underlying spectrum in some logical manner.

The power spectral density corresponding to the stationary time series $\{x(n)\}$ is defined to be the z -transform of the associated autocorrelation sequence (2), that is

$$S_x(z) = \sum_{n=-\infty}^{\infty} r_x(n) z^{-n} \quad (3)$$

where z is a complex valued variable. In the spectral estimation literature, the z variable is usually replaced by $e^{j\omega}$ thereby yielding the equivalent Fourier transform characterization as designated by $S_x(e^{j\omega})$. With this latter representation, one can interpret the spectrum as being a function of the real frequency variable ω .

In classical spectral estimation, one utilizes a standard Fourier transform based method, such as the periodogram, to effect the spectral estimate. The primary drawback in these Fourier based methods resides in the inherent assumption being there made that the time series is identically zero (or periodic) outside the observation window $1 \leq n \leq N$. Generally, this is a very unrealistic assumption to make in virtually all practical applications (e.g., radar doppler processing) and will usually result in poor spectral estimation performance. In recognition of this fact, a number of modern spectral estimation procedures have been developed over the past decade to counteract this deficiency. By in large, the typical modern spectral estimation procedure models the spectrum as a rational function. Such a model can be justified on the basis that any continuous power spectral density can be approximated arbitrarily closely by a rational function of sufficiently high order [1].

The most widely used rational spectral model is the so-called all-pole model which has given rise to the essentially equivalent autoregressive, linear predictive coding, and maximum entropy methods of spectral estimation. A set of basic papers treating these and other spectral estimation procedures is to be found in ref. [2]. All-pole spectral estimators are capable of providing increased resolution in comparison to the classical methods when only a small number of time series observations are available. It must be noted, however, that if the time series spectrum is a rational function which possesses zeroes as well as poles, then an all-pole estimator can yield poor spectral estimates. Clearly, the ability to generate a zero-pole spectrum model provides for a potentially more robust estimator in comparison with the standard all-pole model. With this in mind, a number of zero-pole spectral

estimation methods have been developed. They include estimators which utilize the so-called whitening filter concept (e.g., [3] & [4]). Unfortunately, this class of spectral estimator procedures are iterative in nature, and, typically require a relatively large number of time series observations to be effective. Another approach which makes use of the recursive nature of the time series autocorrelation sequence and does not share these liabilities was developed by Box and Jenkins [5]. A modification of this method involving a more efficient noniterative method for generating the moving average coefficients was recently proposed [6] and [7].

In this paper, a zero-pole spectral estimator is developed which also makes use of the recursive nature of the autocorrelation sequence. It distinguishes itself from the Box and Jenkins method, however, in that a least squares fit to a set of equation errors is used to generate the autoregressive coefficients, and, a noniterative procedure for generating the moving average coefficients is offered. Significantly, the proposed spectral estimator has been empirically found to produce superior estimation performance when compared with the Box and Jenkins method and its variants.

II. RATIONAL SPECTRUM MODEL

One of the most widely used models for spectral estimation is the rational model. The stochastic time series $\{x(n)\}$ is said to have a rational power spectrum if its power spectral density can be expressed in the form

$$S_x(z) = H(z) H(z^{-1}) \sigma^2 \quad (4)$$

where σ^2 is a positive constant and the characteristic rational function

$$H(z) = \frac{B(z)}{A(z)} = \frac{1 + b_1 z^{-1} + \dots + b_q z^{-q}}{1 + a_1 z^{-1} + \dots + a_p z^{-p}} \quad (5)$$

is composed of polynomials $A(z)$ and $B(z)$ which have real coefficients and have zeroes wholly contained within the unit circle. The rational power spectral density (4) is said to have order (p, q) and its zeroes and poles are seen to occur in sets of complex conjugate reciprocals. For reasons which will be shortly made clear, we shall refer to the a_k and b_k coefficients as the autoregressive and moving average coefficients, respectively.

A particularly convenient interpretation on how a stochastic time series with rational spectrum may arise follows directly from the characteristic rational function. This entails treating the characteristic rational function (5) as being the transfer function of a causal, time-invariant linear system. It then follows that this system will be characterized by the recursive equation

$$x(n) = \sum_{i=0}^q b_i \epsilon(n-i) - \sum_{i=1}^p a_i x(n-i) \quad (6)$$

where $b_0=1$ and the time series $\{\epsilon(n)\}$ and $\{x(n)\}$ are taken to be the excitation and response signals, respectively. It is well known that when this system is excited by a stationary white noise time series as statistically characterized by

$$E\{\epsilon(n)\} = 0 \quad \text{and} \quad r_\epsilon(n) = \sigma^2 \delta(n) \quad (7)$$

that the power spectral density of the response time series is given precisely by relationship (4)¹. Thus, a stationary random time series with rational power spectral density can be interpreted as being the response of a causal, time-invariant linear system to a white noise excitation. This linear system is then said to have colored the white noise excitation process (i.e., $S_\epsilon(z) = 1$) and for this reason it is commonly referred to as a coloring filter as suggestively depicted in Figure 1.

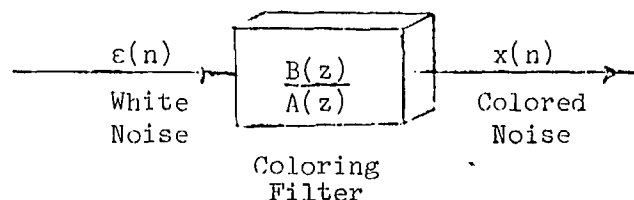


FIGURE 1: Model for a Rational Spectrum Generator

The general linear system (6) is commonly referred to as an autoregressive-moving average (ARMA) model in the spectral estimation literature. This ARMA model is said to be of order (p,q) and it gives rise to the rational spectrum (4) which possesses both zeroes (via $B(z)$) as well as poles (via $A(z)$). The ARMA model is the most general of rational spectrum models possible and its a_k and b_k coefficients uniquely characterize the spectrum.

In the spectral estimation literature, the preponderance of activity has been directed towards the special class of ARMA models known as autoregressive (AR) models. An AR model is one in which the numerator polynomial $B(z)$ is equal to the constant one (i.e., $b_k = 0$ for $k \neq 0$). As such, the AR model is also referred to as an all-pole model since its transfer function is specified by

$$H(z) = \frac{1}{A(z)} \quad (8)$$

¹ The Kronecker delta sequence is defined by

$$\delta(n) = \begin{cases} 1 & n = 0 \\ 0 & \text{otherwise} \end{cases}$$

This all-pole model is the one most often used in spectral estimation primarily due to the ease with which one can compute the a_k coefficients that correspond to a given finite set of time series observations. It should be noted that it is always possible to approximate a general ARMA model by an AR model in the following manner

$$H(z) = \frac{1}{A(z) \left[\frac{1}{B(z)} \right]} \approx \frac{1}{A(z) A_1(z)} \quad (9)$$

whereby the polynomial $A_1(z)$ is obtained by suitably truncating the power series $1/B(z)$ as generated by long division. Clearly, the effectiveness of this approach is dependent on how quickly the coefficients of the long division $1/B(z)$ converge to zero. If $B(z)$ has a zero very close to the unit circle, this convergence rate will be extremely slow thereby making impractical the approximation of an ARMA model by a reasonably low order AR model. It can be conjectured that this is one of the main factors as to why AR models fail to yield satisfactory spectral estimates of time series composed of sinusoidal samples in a strong noisy environment (an ARMA process).

Another subclass of rational spectrum models which has received attention is the so-called moving average (MA) model as characterized by $A(z) = 1$. The transfer function of a MA model is given by $B(z)$ and it is therefore also referred to as an all-zero model. With these thoughts in mind, it is apparent that a general ARMA model is composed of the cascading of an AR with an MA model. The rational spectrum associated with each of these models is displayed in Table 1.

MODEL	SPECTRUM
MA	$\sigma^2 B(e^{j\omega}) ^2$
AR	$\sigma^2 / A(e^{j\omega}) ^2$
ARMA	$\sigma^2 B(e^{j\omega}) ^2 / A(e^{j\omega}) ^2$

TABLE 1. Rational Spectrum Models

An examination of Table 1 reveals the greater flexibility which the ARMA model possesses in providing rational spectral estimates. This robustness was recently demonstrated in which the ARMA model was found to provide the overall best spectral estimates for a variety of problems [3]. Unless one has a priori knowledge which would indicate otherwise, it seems clear that the ARMA model is the one to utilize when seeking a rational spectrum model. Hereafter, we shall concern ourselves with the practical task of developing feasible procedures for determining the "optimum" coefficients of an ARMA model based on a finite set of time series measurements.

III. FUNDAMENTAL AUTOCORRELATION RECURSIVE RELATIONSHIP

For reasons alluded to in the last section, there exists a basic incompatibility in generating an ARMA spectral model, which is most consistent with a given set of time series observations, when the number of observations is small. This incompatibility can be, to a large extent, alleviated by appealing to a fundamental recursive relationship characterizing ARMA time series. This relationship is obtained by analyzing the "causal image" of the autocorrelation sequence as defined by

$$r_x^+(n) = \begin{cases} r_x(n) & n \geq 0 \\ 0 & n < 0 \end{cases} \quad (10)$$

Since the autocorrelation sequence of a real valued time series is an even function of n , it is apparent that one can reconstruct the autocorrelation sequence from its causal image according to

$$r_x(n) = r_x^+(n) + r_x^+(-n) - r_x^+(0) \delta(n) \quad (11)$$

Upon taking the z -transform of this expression, the desired power spectral density is found, that is

$$S_x(z) = S_x^+(z) + S_x^+(z^{-1}) - r_x(0) \quad (12)$$

where the function $S_x^+(z)$ denotes the z -transform of the causal image sequence (10). Thus, a power spectral density estimate may be equivalently accomplished by estimating the function $S_x^+(z)$. This will be the approach taken in this paper.

When the underlying power spectral density is of the rational form (4), a little thought should convince oneself that the function $S_x^+(z)$ must be of the specific rational form

$$S_x^+(z) = \frac{c_0 + c_1 z^{-1} + \dots + c_p z^{-p}}{1 + a_1 z^{-1} + \dots + a_p z^{-p}} \quad (13)$$

in which the denominator polynomial is identical to the $\Lambda(z)$ polynomial that in part characterizes $S_x(z)$ ¹. Upon multiplying both sides of this equation by the polynomial $\Lambda(z)$ and then taking the inverse z -transform, one readily arrives at the following fundamental recursive relationship

$$r_x^+(n) = \sum_{i=0}^p c_i \delta(n-i) + \sum_{i=1}^p a_i r_x^+(n-i) \quad (14)$$

¹It is here assumed that the ARMA model of order (p,q) is such that $p \geq q$. When this is not the case, the degree of the numerator polynomial $C(z)$ must be increased to q .

where the natural boundary conditions $r_x^+(n) = 0$ for $n < 0$ are imposed to reflect the causality of sequence $r_x^+(n)$. Thus, the causal image of an ARMA autocorrelation sequence of order (p, q) is seen to be governed by a linear difference equation of order p .

Upon examining fundamental relationship (14), it is apparent that a knowledge of the a_i , c_i coefficients will enable one to generate the entire autocorrelation sequence. If it were somehow possible to accurately estimate these coefficients from the given time series observations, a particularly effective method of spectral estimation is suggested. Namely, these coefficient estimates, when substituted into equation (13), will provide an estimate for $S_x(z)$. Using this estimate in relationship (12), the desired power spectral density estimate is then obtained

$$S_x(e^{j\omega}) = 2\text{Re} \left[\frac{\sum_{k=0}^p c_k e^{-jk\omega}}{1 + \sum_{k=1}^p a_k e^{-jk\omega}} \right] - \hat{r}_x^+(0) \quad (15)$$

where use of the fact that $S_x^+(e^{j\omega})$ and $S_x^-(e^{-j\omega})$ are complex conjugates has been made. We shall now present a procedure for estimating the a_i , c_i coefficients with the ultimate goal of using relationship (15) for the spectral estimate.

IV. ARMA MODEL COEFFICIENT SELECTION PROCEDURES

The most critical step of the proposed spectral estimation method involves estimating the a_i and c_i coefficients. In this section, the so-called direct and indirect procedures for accomplishing this task will be described. The direct approach makes explicit use of the fundamental autocorrelation relationship derived in the previous section. On the other hand, the more effective indirect approach uses an alternate approach which provides a solution procedure that is consistent with the fundamental autocorrelation relationship.

Direct Method

In the direct method, one first generates estimates of the autocorrelation sequence from the given time series observations using some convenient method.¹ These estimates, denoted as $\hat{r}_x(n)$, are then substituted into fundamental relationship (14). In recognition that the autocorrelation estimates will be generally in error, and that the ARMA model order parameter p may be incorrect, it follows that this substitution will give rise to the following "equation error" sequence

$$e(n) = \hat{r}_x(n) + \sum_{i=1}^p a_i \hat{r}_x(n-i) - \sum_{i=0}^p c_i \delta(n-i) \quad 0 \leq n \leq N-1 \quad (16)$$

in which $\hat{r}_x(n) = 0$ for $n < 0$.

¹As an example, one might use the biased estimator.

$$\hat{r}_x(n) = \frac{1}{N-n} \sum_{k=1}^{N-n} x(k)x(k+n)$$

Our objective will that of selecting the models a_i , c_i coefficients so as to minimize these equation errors in some sense. For reasons of mathematical tractability and subsequently demonstrated effectiveness, the equation error criterion to be minimized is taken to be the quadratic functional

$$f(\underline{a}, \underline{c}) = \sum_{n=0}^{N-1} w(n) e^2(n) \quad (17)$$

The nonnegative weights, $w(n)$, are usually selected to be monotonically non-increasing (i.e., $w(n) \geq w(n+1)$) so as to reflect an anticipated degradation in equation error accuracy for increasing n . This degradation behavior arises primarily from a loss in autocorrelation estimate fidelity for increasing lags (i.e., n).

In minimizing this functional with respect to the c_i coefficients, it is apparent from relationship (16) that the c_i coefficients have no effect whatsoever on the $e(n)$ for $n > p$. This being the case, it follows that the optimum c_i coefficients must be given by

$$c_n^0 = \hat{r}_x(0) + \sum_{i=1}^n a_i \hat{r}_x(n-i) \quad 0 \leq n \leq p \quad (18)$$

since such a selection will render the equation errors, $e(n)$, identically zero over $0 \leq n \leq p$ for "any" choice of the a_i autoregressive coefficients. It then follows that the optimum autoregressive coefficients must render the remaining terms (i.e., $p < n < N$) of the quadratic functional a minimum. With this in mind, let us express these specific set of equation errors in the matrix format

$$\begin{bmatrix} e(p+1) \\ e(p+2) \\ . \\ . \\ . \\ e(N-1) \end{bmatrix} = \begin{bmatrix} \hat{r}_x(p) & \hat{r}_x(p-1) & . & . & . & \hat{r}_x(1) \\ \hat{r}_x(p+1) & \hat{r}_x(p) & . & . & . & \hat{r}_x(2) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \hat{r}_x(N-2) & \hat{r}_x(N-3) & . & . & . & \hat{r}_x(N-p-1) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_p \end{bmatrix} + \begin{bmatrix} \hat{r}_x(p+1) \\ \hat{r}_x(p+2) \\ . \\ . \\ . \\ \hat{r}_x(N-1) \end{bmatrix} \quad (19a)$$

where use of relationship (16) for $n > p$ has been made. This matrix system of equations can be conveniently expressed as

$$\underline{e} = \underline{R}\underline{a} + \underline{r} \quad (19b)$$

in which $\underline{a}$ is the $p \times 1$ autoregressive coefficient vector with elements a_i , $\underline{e}$ and $\underline{r}$ are each $(N-p-1) \times 1$ vectors with elements $e(p+n)$ and $r(p+n)$, respectively, and $\underline{R}$ is a $(N-p-1) \times p$ Toeplitz matrix.

A little thought will convince oneself that for the optimum c_i coefficient selection given by relationship (18), the quadratic functional (17) may be equivalently expressed as

$$f(\underline{a}, \underline{c}^0) = [\underline{Ra} + \underline{r}]^T W [\underline{Ra} + \underline{r}] \quad (20)$$

in which W is a positive semidefinite $(N-p-1) \times (N-p-1)$ diagonal matrix whose diagonal elements are given by $w_{nn} = w(p+n)$ for $n=1, 2, \dots, N-p-1$. The minimization of this quadratic function with respect to the autoregressive coefficient vector is straightforwardly carried out and results in the following system of p linear equations for the required optimum autoregressive coefficient vector

$$[R^T W R] \underline{a}^0 = -R^T W \underline{r} \quad (21)$$

One then solves this system of linear equation to obtain the desired optimum autoregressive coefficients. Upon substitution of these autoregressive coefficients into relationship (18), the optimum c_i^0 coefficients are next determined. Finally, the desired power spectral density estimate is obtained by substituting these optimum a_i^0 , c_i^0 coefficients into relationship (15).

It is of interest to note that the system of equations for the autoregressive coefficients (21) reduces to the Box-Jenkins method for a weighting selection of $w(n+p) = 1$ for $1 \leq n \leq p$ and zero otherwise. Unfortunately, this particular weighting selection implicitly assumes that the equation errors $e(n)$ for $p+1 \leq n \leq 2p$ all have the same statistical behavior. More realistically, one would presume that the equation errors become more random as n increases. It is then conjectured that the primary reason as to why the Box-Jenkins method does not provide adequate spectral estimates for certain problems is due to this particular weighting choice and the fact that it makes no use of the fundamental autocorrelation relationship (14) for $n > 2p$ whatsoever.

Indirect Method

Although the direct method has been found to provide satisfactory spectral estimation performance, the indirect approach to be now briefly described has yielded significantly better performance. Its development is based on the coloring filter's characteristic equation (6), and the fact that the random variables $x(n)$ and $\varepsilon(m)$ are uncorrelated for $m > n$. To begin this development, we shall first replace the variable n appearing in relationship (6) by k . Next, each side of this characteristic equation is multiplied by $x(k-n)/(N-n)$ to obtain

$$x(k) \frac{x(k-n)}{N-n} = \left[\sum_{i=0}^q b_i \varepsilon(k-i) - \sum_{i=1}^p a_i x(k-i) \right] \frac{x(k-n)}{N-n}$$

If both sides of this equality are then summed over the index range $n < k \leq N$, after rearrangement one obtains

$$\bar{e}(n) = \sum_{i=1}^p \left[\frac{1}{N-n} \sum_{k=n+1}^N x(k-i)x(k-n) \right] a_i + \left[\frac{1}{N-n} \sum_{k=n+1}^N x(k)x(k-n) \right]$$

for $p < n < N$ (22)

where the pseudo equation error term is specified by

$$\bar{e}(n) = \sum_{i=0}^q b_i \left[\frac{1}{N-n} \sum_{k=n+1}^N \varepsilon(k-i)x(k-n) \right] \quad p < n < N$$

Upon examination of this expression, it is clear that the expected value of the term $\varepsilon(k-i)x(k-n)$ will be zero. This would indicate that the general pseudo equation error term $\bar{e}(n)$ will itself tend to be close to zero (this is reinforced by the division by $N-n$). With this in mind, a logical choice for the a_i coefficients used in expression (22) would be one which tended to minimize the pseudo equation error sequence.

If one compares the pseudo equation error relationship (22) with the equation error relationship (16), a similarity is in evidence. Namely, the elements within the brackets of expression (22) are recognized as unbiased autocorrelation estimates. If these estimates are substituted for the entries of matrix R and vector r in relationship (19), a new system of equations (21) for the optimum autoregressive coefficients arises. These new system of equations distinguish themselves from the former in that a genuinely different autocorrelation estimate formula is used for each equation. Once this modified system of equations have been solved for the a_i coefficients, the c_i coefficient estimates are obtained according to

$$c_n^0 = \left[\frac{1}{N-p} \sum_{k=p+1}^N x^2(k) \right] + \sum_{i=1}^n a_i \left[\frac{1}{N-p} \sum_{k=p+1}^N x(k)x(k-i) \right] \quad 0 \leq n \leq p \quad (23)$$

The required power spectral density estimate is then given by relationship (15).

V. NUMERICAL EXAMPLES

To test the effectiveness of the proposed ARMA spectral estimator method, the classical problem of detecting the presence of sinusoids in additive noise will be considered. In particular, we will investigate the specific case in which the time series observations are generated according to

$$x(n) = A_1 \sin(\pi f_1 n) + A_2 \sin(\pi f_2 n) + w(n) \quad 1 \leq n \leq N \quad (24)$$

where $w(n)$ is a zero mean Gaussian time series with variance one. This particular problem serves as an excellent vehicle for measuring a spectral estimator's performance relative to: (i) detecting the presence of sinusoids in a strong noisy background, and (ii) resolving two sinusoids whose frequencies f_1 and f_2 are nearly equal. The individual sinusoidal signal-to-noise ratios (SNR) for the above signal are given by $20\log(A_k/\sqrt{2})$ for $k=1,2$. In order to consider the effectiveness of the proposed ARMA spectral estimator in different noise environments, we shall consider two cases. These cases have been examined in reference [8] where the performance of many modern spectral estimators are empirically compared.

$$\text{CASE I: } A_1 = \sqrt{20}, f_1=0.4 \text{ and } A_2 = \sqrt{2}, f_2=0.426$$

In this example, we have two closely spaced (in frequency) sinusoids in which the stronger sinusoid has a SNR of 10 dB while the weaker sinusoid has a SNR of 0 dB. For this relatively low SNR case, the spectral estimator's ability to resolve two closely spaced sinusoids, and simultaneously identify the frequencies will be tested. Upon generating the sequence (24) for $N=64$, the indirect ARMA spectral estimator method was used for a selection of weights $w(n)=N-n$ and $p=15$. The resultant spectrum is displayed in Fig. 2a where the frequency resolving capability of this method is in evidence. The frequency identification accuracy was also excellent in that the sinusoid frequency estimates were $\hat{f}_1 = 0.398$ and $\hat{f}_2 = 0.425$.

For comparison purposes, the covariance AR spectral estimate (basically the maximum entropy method) and the revised Box-Jenkins [6] ARMA estimate of order 15 were generated using the same data. The results of these estimations are displayed in Figures 2b and 2c where an inability to resolve the two sinusoids is apparent. This gives evidence of the inherently superior performance capability of the herein described ARMA spectrum estimator over standard AR estimator procedures and other ARMA methods

$$\text{CASE II: } A_1 = \sqrt{2}, f_1 = 0.32812, A_2 = \sqrt{2}, f_2 = 0.5$$

We are now examining the ability of the ARMA spectral estimator to detect sinusoids in a low SNR environment (i.e., 0 dB). For a selection of $N=64$, $w(n) = N-n$ and $p = 5$, the resultant ARMA spectral estimation is displayed in Figure 3a. Clear, one is able to detect the presence of the two sinusoids, and the frequency estimates $\hat{f}_1 = 0.3202$ and $\hat{f}_2 = 0.5012$ are of good quality considering the prevailing SNR environment. A 15th order covariance AR spectral estimator was then found to generate the spectral estimate displayed in Figure 3b. Although the two sinusoids were properly detected, a number of false peaks are in evidence.

Digital Filter Design

It is possible to use the proposed ARMA method for synthesizing digital filters. To illustrate the approach that is taken, let us consider the specific case of designing a low-pass filter of normalized, cutoff frequency f_c . One may readily show that the impulse response of an idealized version

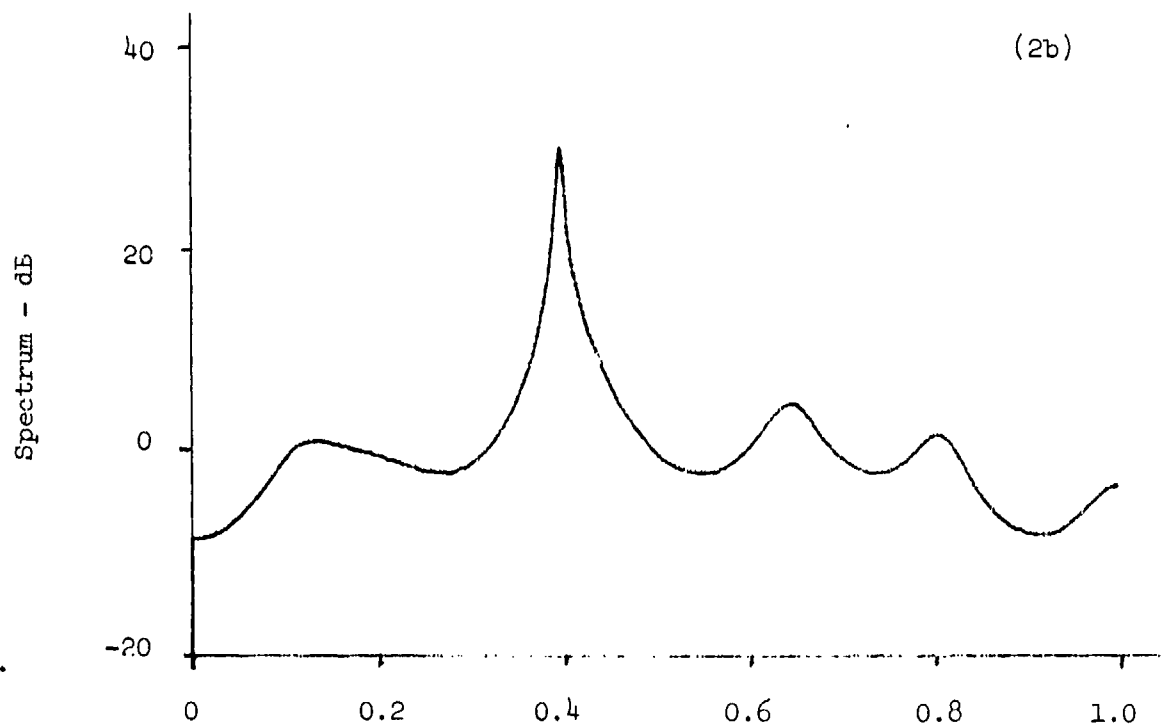
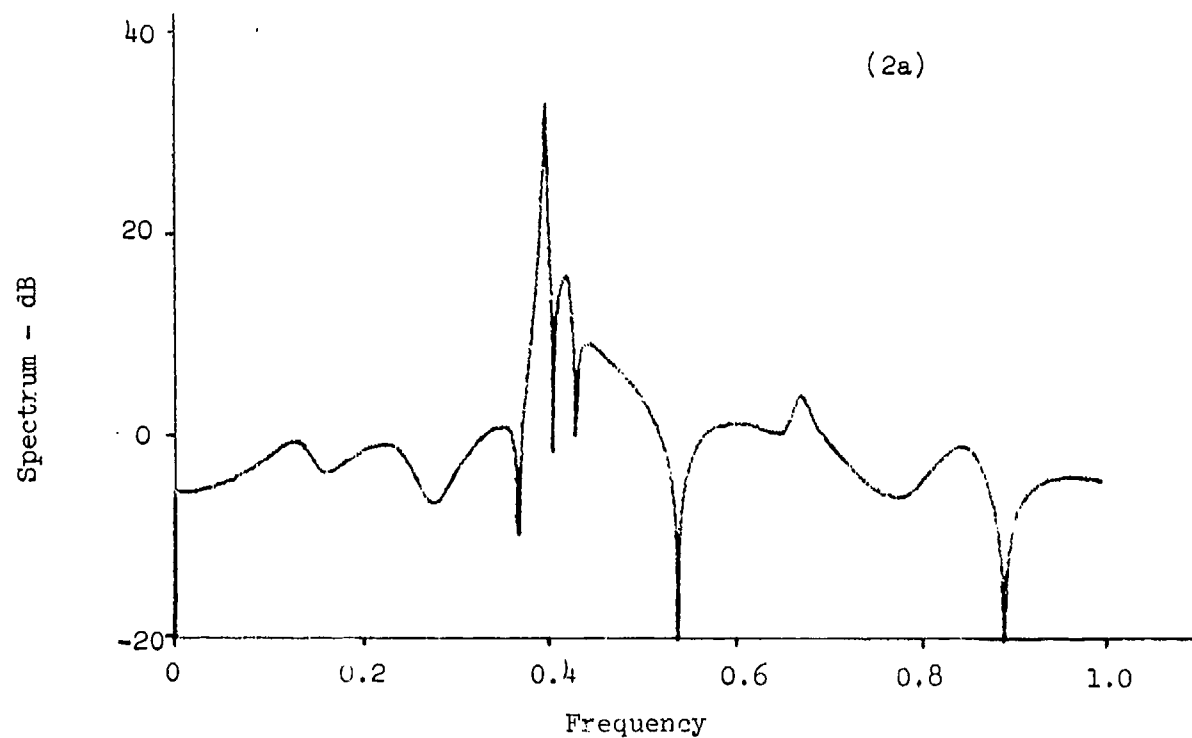
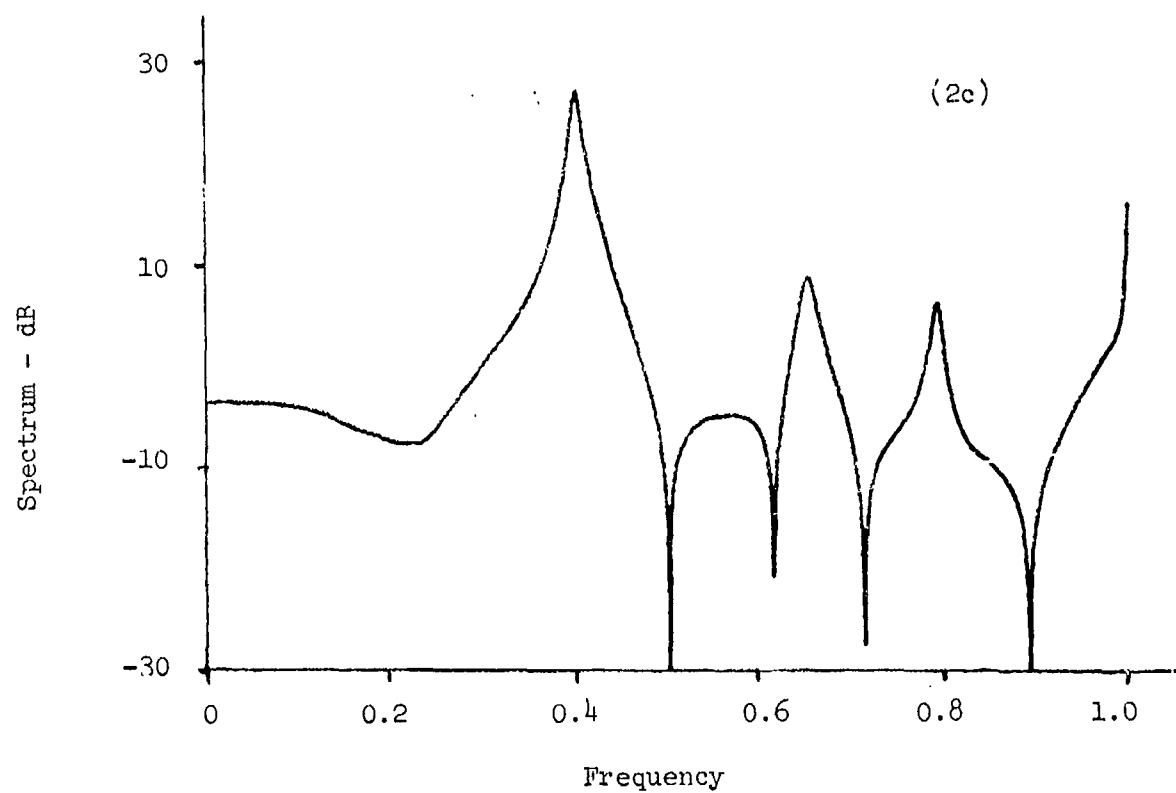


FIGURE 2: Spectral Estimation of two sinusoids with one of 10 dB SNR ($f=0.4$) and the other at 0 dB SNR ($f=0.426$) (a) 15th Order ARMA estimate, (b) 15th Order AR Estimate, (c) 15th Order ARMA Box-Jenkins Estimate



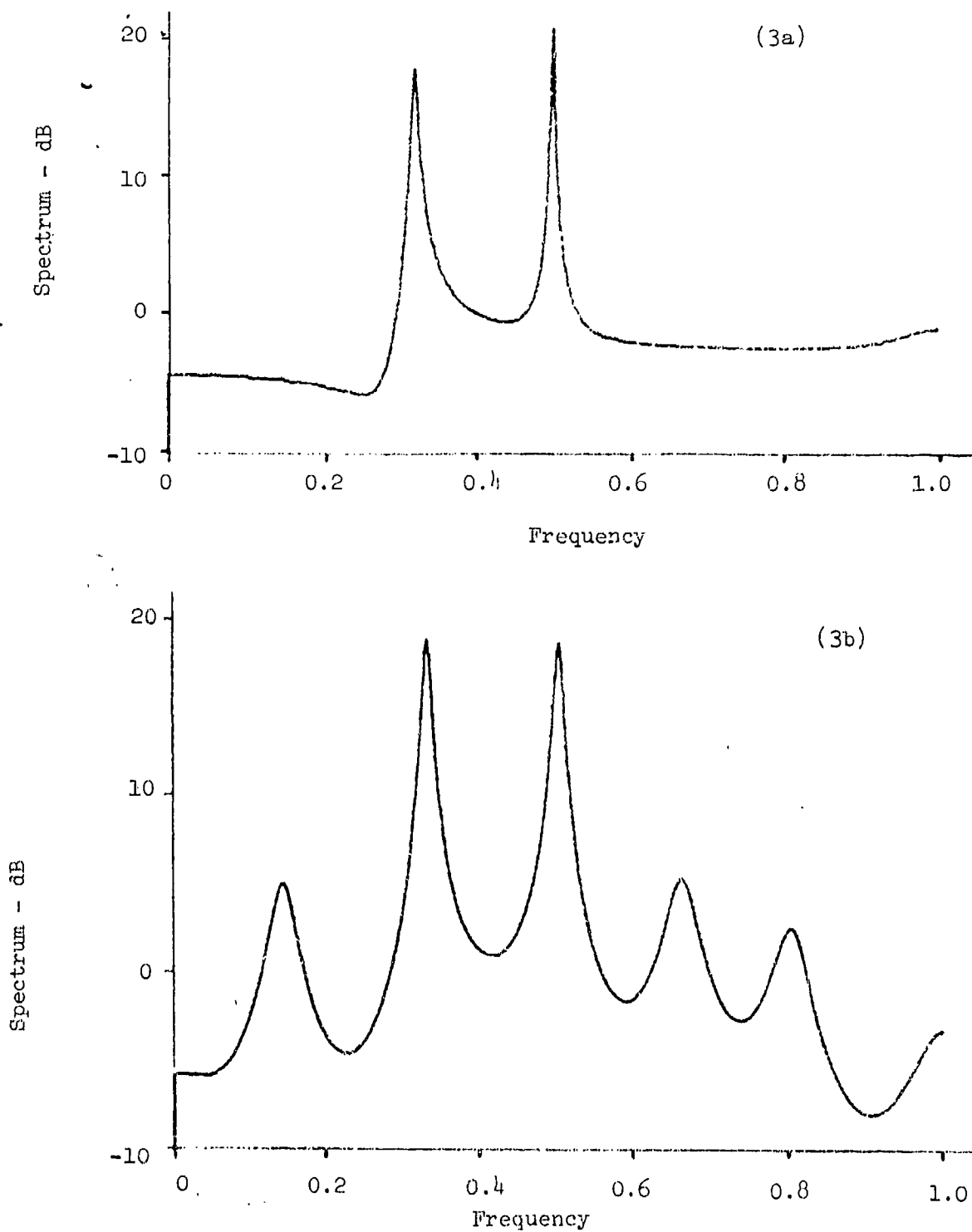


FIGURE 3: Spectral Estimation Plots for Two Sinusoids at 0 dB SNR with frequencies 0.32812 & 0.5
(a) 5th Order ARMA estimate, (b) 15th Order AR estimate

of this low pass filter is given by $\sin[\pi f_c n]/\pi n$. With this in mind, one then applies the herein developed ARMA procedure to the specific sequence

$$x(n) = \sin[\pi f_c (n-0.5N)]/\pi(n-0.5N) \quad 1 \leq n \leq N$$

The resultant ARMA model obtained in this manner will have the attenuation characteristics of the desired low-pass filter. To illustrate this, a 15th order ARMA spectral estimate of this sequence was made for $f_c = 0.2$, $N = 256$ and $w(n) = [N-n]^2$. The resultant filter's magnitude characteristics are displayed in Figure 4 where the low-pass characteristics are in evidence. In a paper now in preparation, a detailed description of this filter synthesis procedure will be made and compared to an alternate method [9].

VI. CONCLUSION

A computationally efficient closed form method for generating ARMA spectral estimates has been presented. Conceptually, the method offers the promise of producing superior spectral estimation performance in comparison to such AR spectral estimators as the autoregressive, linear predictive coding, and maximum entropy methods. Empirical results have substantiated this conjecture.

In order for this method to achieve its full potential, a number of important considerations need further investigation. They include determination of the most effective autocorrelation estimation procedure to use since an inferior procedure will generally result in poor spectral estimations. Another important consideration is the choice of error weights. This weighting selection should reflect, in some manner, our growing lack of confidence in the autocorrelation estimates for increasing lags (n). Since no statistical assumptions on the time series are being made (other than that it is an ARMA time series), it is apparent that the weighting sequence should be data dependent. One further consideration is that of determining a procedure for obtaining the best choice of the ARMA ordering parameter p .

As a final point, it should be noted that the herein presented procedure can be used to generate ARMA spectral estimates from basic AR methods. In particular, one could use any standard AR method (e.g., the maximum entropy method) to generate the autoregressive coefficient estimates. Using these coefficient estimates and suitable autocorrelation estimates (often byproducts of an AR method), one then uses relationship (18) and then finally expression (15) to generate an ARMA spectral estimate. This will result in little additional computational cost over the "pure" AR method due to the simplicity of relationships (15) and (18). The effectiveness of this hybrid approach will be subsequently reported upon.

VII. ACKNOWLEDGEMENT

This opportunity is taken to acknowledge the contributions of Koji Ogino and Behshad Baseghi in this effort.

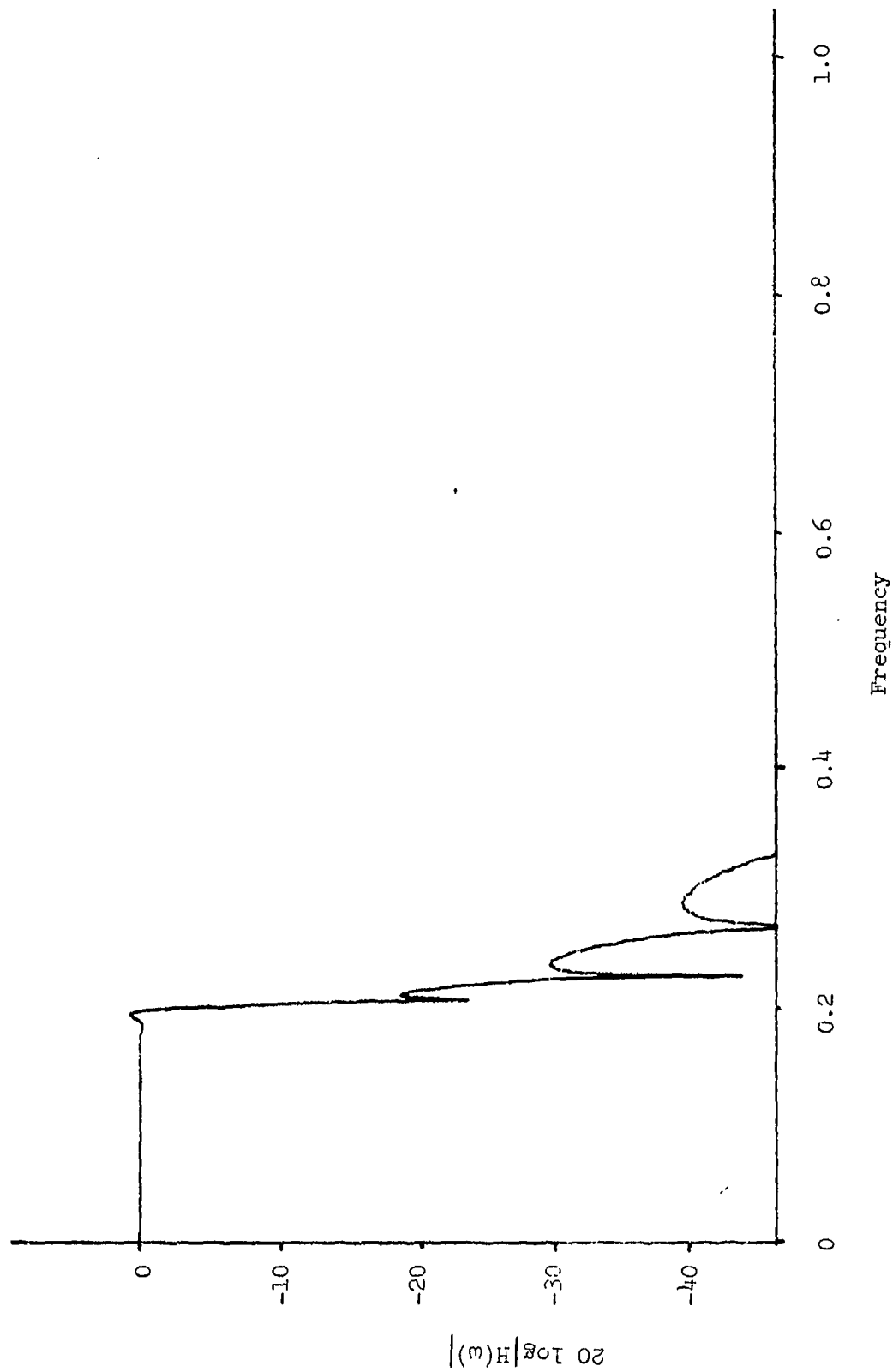


FIGURE 4. Low-pass filter with cutoff frequency 0.2 synthesized using ARMA method of order 15.

VIII. REFERENCES

- [1] L. H. Koopmans, THE SPECTRAL ANALYSIS OF TIME SERIES, Academic Press, New York, 1974.
- [2] D. G. Childers, MODERN SPECTRUM ANALYSIS, IEEE Press, 1978.
- [3] P. R. Gutowski, E. A. Robinson, and S. Treitel, "Spectral Estimation, Fact or Fiction", IEEE Trans. Geoscience Electronics, Vol. GE-16, No. 2, pp. 80-84, April 1978.
- [4] S. A. Tretter and K. Steiglitz, "Power-Spectrum Identification in Terms of Rational Models", IEEE Trans. Automatic Control, Vol. AC-12, pp. 185-188, April, 1967.
- [5] G. Box and G. Jenkins, TIME SERIES ANALYSIS; FORECASTING AND CONTROL, Revised Edition, Holden-Day, San Francisco, 1976.
- [6] M. Kaveh, "High Resolution Spectral Estimation for Noisy Signals", IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No. 3, pp. 286-287.
- [7] J. F. Kinkel, J. Perl, L. Scharf, A. Stubberud, "A Note on Covariance - Invariant Digital Filter Design and Autoregressive Moving Average Spectrum Analysis", IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No. 2, pp. 200-202, April 1979.
- [8] T. M. Sullivan, O. L. Frost, J. R. Treichler, "High Resolution Signal Estimation", ARGO Systems Inc. Tech Report, June, 1978.
- [9] L. L. Scharf and J. C. Luby, "Statistical Design of Autoregressive-Moving Average Digital Filters", IEEE Trans on Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No. 3, pp. 240-247, June, 1979.

98-Blank

EXTRAPOLATING BANDLIMITED SIGNALS WITH NOISE AND QUANTIZATION

KENNETH ABEND AND JUDITH R. PLATT

RCA Government Systems Division
Missile and Surface Radar
Moorestown, New Jersey 08057

Abstract

In many applications of spectral analysis (e.g., short radar dwells) the need arises to obtain spectral resolution from an extremely short time segment of a bandlimited signal. Instead of applying a window and assuming the function to be zero outside of the observed segment, the modern approach is to extrapolate. Knowledge of the bandwidth of the signal allows for accurate extrapolation over a limited time interval that is many times the length of the given segment, provided that the signal is sampled at many times the Nyquist rate. Of several alternate ways to find a bandlimited signal of minimum energy that fits the observed samples, Cadzow's method is the simplest because it utilizes a matrix whose size is determined by the number of samples before extrapolation. However, the ill conditioned nature of the Cadzow matrix makes it extremely difficult to extrapolate coarsely quantized, noisy signals. We solve this problem with an iterative procedure that preserves the small-matrix advantage of Cadzow's method. Examples involving non-ideal signals quantized to as few as five bits are investigated in order to determine the extent of reliable extrapolation.

Introduction and Summary

The purpose of this paper is to demonstrate that a time limited set of samples of a bandlimited signal can be extrapolated with limited computer resources and realistic signals. When Cadzow's one-step extrapolation procedure^[1-3] is applied to finely quantized samples of a bandlimited signal in bandlimited noise, the signal and noise are satisfactorily extrapolated. When the noise bandwidth is increased and/or when the noisy samples are coarsely quantized, problems arise due to limited computer accuracy because a set of linear equations is ill conditioned. We show that by solving these equations iteratively by the method of steepest descent, reliable extrapolation can be performed in the presence of noise with the input quantized to as few as five bits plus sign.

By using signals consisting of the sum of sine waves with unequal amplitudes and phases, we obtain a more realistic picture of the limitations of the algorithm. Specifically, while the input may be limited to a small or a large fraction of a Nyquist interval, the extrapolated signal is seldom reliable for much more than three Nyquist intervals.

Bandlimited Extrapolation

Let the signal $g(t)$ be bandlimited to $|f| < B$, i.e., its Fourier transform $G(f) = \int_{-\infty}^{\infty} g(t) \exp(-j2\pi ft) dt$ satisfies

$$G(f) \equiv 0 \text{ for } |f| > B. \quad (1)$$

If $g(t) = 0$ for $0 < t < T$ with $T > 0$, then $g(t) = 0$ for all t . By considering $g(t) - g_1(t)$ with $g_1(t) = g(t)$ for $0 < t < T$, we see that if $g(t)$ is given for $0 < t < T$, it is uniquely determined for all t . Cadzow's one-step extrapolation procedure is based on finding a signal $z(t)$ with the following properties.

- (a) $z(t) = 0$ for $t \leq 0$ and $t > T$
- (b) Bandlimiting $Z(t)$ to B produces a signal that agrees with $g(t)$ for $0 < t < T$ (and hence for all t), i.e.,

$$\begin{aligned} g(t) &= \int_{-B}^B \left[\int_0^T z(\tau) \exp(-j2\pi f\tau) d\tau \right] \exp(j2\pi ft) df \\ &= 2B \int_0^T z(\tau) \text{sinc}[2B(t-\tau)] d\tau \end{aligned} \quad (2)$$

where

$$\text{sinc } u \equiv \frac{\sin(\pi u)}{\pi u} \quad (3)$$

Let $g(t)$ be sampled at r times the Nyquist rate, i.e., $\Delta < 1/(2B)$, where Δ is the sampling interval and $1/2B$ is the Nyquist interval, so that:

$$r = \frac{1}{2B\Delta} > 1 \quad (4)$$

If we are given only M samples of $g(t)$: $g(m\Delta)$, $m=1,2,\dots,M$, we approximate equation (2) by

$$g(m\Delta) = \sum_{k=1}^M \frac{1}{r} \text{sinc} \left(\frac{m-k}{r} \right) z(k\Delta) \quad (5)$$

Since $z(t)$ is not bandlimited, Δ must be very small. In this sampled-data low-pass case, Cadzow's algorithm simplifies to:

1. Solve M linear equations in M unknowns for $z(k\Delta)$, $k=1,2,\dots,M$ from equation (5) with $m = 1,2,\dots,M$.
2. Using these M values of z , determine $g(m\Delta)$ for $m \leq 0$ and $m > M$ from equation (5).

More generally, we can simultaneously interpolate¹ and extrapolate by using

$$g(t) = 2B\Delta \sum_{k=1}^M \text{sinc}[2B(t-k\Delta)]z(k\Delta) \quad (6)$$

For the bandpass case ($F_1 < |f| < F_2$) Cadzow develops equations similar to (2) and (5), with the sinc function replaced by the difference of two sinc functions. However, if we use the complex envelope, $g(t)$, where

$$g_{BP}(t) = \text{Re}[g(t)\exp(j2\pi f_0 t)] \quad (7)$$

with $f_0 = (F_2 + F_1)/2$, then equation (1) through (6) remain valid with $B = (F_2 - F_1)/2$ (the bandwidth is $2B$). We can sample the complex signal² $g(t)$ at r times the Nyquist rate ($2B$) and utilize equation (5) in steps 1 and 2 above to extrapolate both the real and imaginary parts of $g(t)$. Preliminary results, obtained to date, seem to be nearly independent of whether the real or complex formulation is used. The results presented in this paper were obtained on the Hewlett-Packard System 45 desktop computer using the real formulation.

We express the sampled signal in the form

$$g_m = \sum_{k=1}^M h_{mk} z_k \quad (8)$$

-
1. In contrast to Nyquist interpolation by either $g(t) = \sum_{m=-\infty}^{\infty} \text{sinc}(t/\Delta - m)g(m\Delta)$ or $g(t) = \sum_{m=-\infty}^{\infty} \text{sinc}(2Bt - m)g(m/2B)$.
 2. This is commonly referred to as I and Q sampling.

and for $m=1,2,\dots,M$; in the vector-matrix form

$$G = H Z \quad (9)$$

Here G and Z are M -dimensional column vectors with elements $g_m=g(m\Delta)$ and $z_k=z(k\Delta)$, and H is an $M \times M$ Toeplitz matrix with elements $h_{mk}=\text{sinc}((m-k)/r)/r$ in the low-pass case. Jain and Ranganath^[4] have shown that given G , (8) and (9) produce a minimum norm least squares solution for N values of $g(m\Delta)$ with $N > M$.

The most direct solution involves inversion of the matrix H . This is especially useful when the bandwidth is known a priori and H can be inverted off line. Because of the ill conditioned nature of H , Jain and Ranganath propose iterative solutions that involve $N \times N$ matrices. By utilizing only $M \times M$ matrices, we reduce the computational complexity by an order of magnitude. We therefore consider the solution of (9) alone, and use that solution in (8).

Method of Steepest Descent^[5]

This technique which progresses geometrically to the true solution from an initial guess proved to offer valid results for all examples whether ill-conditioned or well-conditioned. The iterative procedure is initiated with an estimate of the vector values

$$Z^{(0)} = [z_1^{(0)}, z_2^{(0)}, \dots, z_m^{(0)}]^T$$

and continues with the single step iterative equation for a symmetric positive definite matrix H :

$$Z^{(k+1)} = Z^{(k)} + \alpha^{(k)} R^{(k)} \quad (10)$$

Here $R^{(k)}$ is the vector

$$R^{(k)} = [r_1^{(k)}, r_2^{(k)}, \dots, r_m^{(k)}]^T$$

with elements

$$r_i^{(k)} = g_i - \sum_{j=1}^M h_{ij} z_j^{(k)}$$

and the scalar $\alpha_0^{(k)}$ is given by

$$\alpha_0^{(k)} = \frac{\left(\sum_{j=1}^M r_j^{(k)2} \right)}{\left(\sum_{i,j=1}^M h_{ij} r_i^{(k)} r_j^{(k)} \right)}$$

The notation $R^{(k)}$ indicates that this is the k^{th} iteration of the vector R . If many iterations are expected a multistep general form of the iterative equation is

$$Z^{(k+1)} = Z^{(k)} + \sum_{\ell=0}^{p-1} \alpha_{\ell}^{(k)} H^{\ell} R^{(k)}$$

with the scalar values $\alpha_{\ell}^{(k)}$ determined from

$$\sum_{\ell=0}^{p-1} \alpha_{\ell}^{(k)} (H^i R^{(k)}, H^{\ell+1} R^{(k)}) = (H^i R^{(k)}, R^{(k)}) \quad i=0,1,\dots,p-1$$

where p is the number of iterative steps per computation, H^{ℓ} is the product of the matrix with itself ℓ times, and the notation (A,B) represents the dot product of A and B .

The iteration procedure is terminated when the differences between values of the Z vector on successive iterations change by miniscule amounts.

Results With Inaccurate Matrix Inversion

To demonstrate that results are obtainable for specific examples by merely inverting H in (9), in spite of the fact that H is ill-conditioned and the inversion was extremely inaccurate, the following data is presented. This data also illustrates conditions under which the more accurate method of steepest descent is required.

As an arbitrary example, $g(t)$ was chosen to be the sum of two cosine waves:

$$g(t) = A_1 \cos(2\pi f_1 t + \phi_1) + A_2 \cos(2\pi f_2 t + \phi_2)$$

with $A_1 = 0.5$, $A_2 = 1.0$, $f_1 = 0.4$, $f_2 = 1/6$, $\theta_1 = 0$, and $\theta_2 = \pi/3$ (Figure 1a). We assume we know only that the signal is low-pass limited to $B=0.5$. Given seven samples at eight times the Nyquist rate over a duration of $3/4$ of a Nyquist interval (Figure 1), a 4:1 extrapolation was obtained (Figure 1b for $-1 < t < 2$). With uniformly distributed independent noise samples (40 dB down) extrapolation was not achieved. With bandlimited Gaussian noise ($|f| < 0.5$), however (Figure 2a), extrapolation was no problem (Figure 2b). On the other hand, using simple matrix inversion required at least 14 bits quantization of the input samples without noise (Figure 3), and more than 32 bits with noise (Figure 4).

Though some improvements in the noisy quantized case were obtained either by using Levinson's algorithm^[6,7] or by adding a small constant to the diagonal elements of H before inverting^[4], accurate extrapolation with coarse quantization was obtained only by the method of steepest descent.

The noise samples for Figures 2a, 2b, and 4 were obtained by adding 12 uniformly distributed random variables to produce each of seven independent normal noise samples, and then multiplying that seven dimensional noise vector by the lower triangular decomposition of the desired covariance matrix, H . When this same example was run using the steepest descent algorithm 4:1 extrapolation was obtained even at 10 dB S/N and 6 bits quantization (simultaneously). Only at 4 bits quantization did the problems observed with the inaccurate algorithm reappear.

Results With Method of Steepest Descent

A variety of examples were used to test the extrapolation algorithm incorporating the method of steepest descent. To demonstrate the results using the real bandpass formulation of the algorithm, another arbitrary sum of two cosine waves was selected. Its parameters are $A_1=0.5$, $A_2=1$, $f_1=1/25$, $f_2=1/30$, $\theta_1=0$, and $\theta_2=\pi/3$; this is shown in Figure 5. Using the passband $1/36 < |f| < 1/24$ and nine samples at intervals of $\Delta=0.5$ starting at $t=1$ and terminating at $t=5$; the signal was extrapolated over the interval -50 to $+65$ yielding 251 points of which 150 appear to be valid.

The effect of quantization is illustrated in Figure 6 using the same example but quantizing the samples at various levels. Figure 6 shows the results for 9 bits plus sign, 7 bits plus sign and for 5 bits plus sign. Valid extrapolations with quantizations as low as 4 bits plus sign were obtained. Extrapolation with bandlimited noise added to signal samples is illustrated in Figure 7 for rms noise amplitudes of -25 db, -20 db, -15 db, -10 db and 0 db. Figure 8 illustrates the effect of noise and quantization; Figure 8a shows the effect of noise and quantization for a particular noise level and Figure 8b shows a particular sample of noise at various signal to noise ratios.

Each alteration of the signal samples - quantization, bandlimited noise, bandlimited noise and quantization, etc - creates a different signal. The extrapolation procedure produces a minimum norm least squares estimate^[4] of the altered signal, i.e., the bandlimited signal matching the given samples that has the minimum energy.^[8]

Valid estimation of the signal passband is crucial to the results obtained. This is demonstrated in Figure 9 where the extrapolated signal using various passbands is compared to the true signal. It is evident that the better the estimate of the passband the better the extrapolation and the longer its extent.

Another parameter of interest is the choice of sampled time interval.^[1] For the same arbitrary signal different time intervals were selected, all other parameters are unchanged, the results are shown in Figure 10.

Conclusions

Different degrees of extrapolation are obtained with different signals and different sampling rates. In the lowpass example we used seven samples at eight times the Nyquist rate, spanning $3/4$ of a Nyquist interval. The extrapolated signal is valid for three Nyquist intervals, giving a 4:1 extrapolation. In the bandpass example we used nine samples at 75 times the Nyquist rate, spanning $1/9$ of a Nyquist interval. The extrapolated signal is valid for two Nyquist intervals, giving an 18:1 extrapolation. By increasing the spacing between samples we reduce the extrapolation ratio and therefore can increase the duration of validity of the extrapolated signal only slightly. By examining many other examples we are led to the conclusion that, with a small number of samples of a non-ideal signal, reliable extrapolation is limited to a very few Nyquist intervals. Thus spectral analysis of the extrapolated signal must be performed by techniques other than the Fourier transform (e.g., maximum entropy).

Cadzow's one step extrapolation procedure is by far the most easily implemented, in that for extrapolating from M points to N points only an $M \times M$ matrix need be inverted. The problem that arises is that this matrix is ill conditioned and thus difficult to invert accurately enough to obtain reasonable results with imperfect signals (having noise and quantization). We solve this problem by solving the M equations in M unknowns iteratively by the method of steepest descent.

Jain and Ranganath^[4] have recently described two iterative procedures, also designed to overcome this same problem. The first is to use steepest descent to improve the rate of convergence of Papoulis^[9] iterative extrapolation procedure. The second, a conjugate gradient algorithm, is aimed directly at the final answer, as is our method. However, both of

their iterative procedures work with $N \times N$ matrices rather than $M \times M$ matrices, where $N > M$. As they point out [4, Sect. 7.5], Cadzow's method (involving an $M \times M$ matrix, H) produces a minimum norm least squares solution. Its shortcoming, the ill conditioned nature of H , we have overcome.

Bibliography

1. Cadzow, J.A., 1979, "An Extrapolation Procedure For Band-Limited Signals", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-27 No. 1, pp 4-12.
2. Cadzow, J.A., 1978, "Reconstruction of Signals From Their Linear Mapping Image", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 646-650.
3. Cadzow, J.A., 1978, "Improved Spectral Estimation From Incomplete Sampled Data Observations," Proceedings of the RADC Spectrum Estimation Workshop, pp. 109-123.
4. Jain, A.K. and Ranganath, S., 1978, "Extrapolation and Spectral Estimation Techniques For Discrete Time Signals," Department of Electrical Engineering, University of California, Davis, California. Final Report under Contract F30602-75-C-0122, Rome Air Development Center, Griffiss A.F.B., N.Y.
5. Berenzin, I.S. and Zhidkov, N.P., 1975, Computing Methods, Vol. II, Pergamon Press, Oxford, Great Britain, Chapter 6.
6. Zohar, Shalav, 1974, "The Solution of a Toeplitz Set of Linear Equations," Journal of the Association for Computing Machinery, Vol. 21, No. 2, pp. 272-276.
7. Levinson, N., 1947, "The Weiner RMS (root mean square) Error Criterion in Filter Design and Prediction", J. Math Phys., Vol. 26, pp. 261-278.
8. Wiley, R.G., 1979, "Concerning the Recovery of a Bandlimited Signal or Its Spectrum from a Finite Segment", IEEE Transactions on Communications, Vol. COM-27, No. 1, pp. 251-252.
9. Papoulis, A., 1975, "A New Algorithm in Spectral Analysis and Bandlimited Extrapolation", IEEE Transactions on Circuits and Systems, Vol. CAS-22, No. 9, pp. 735-742.

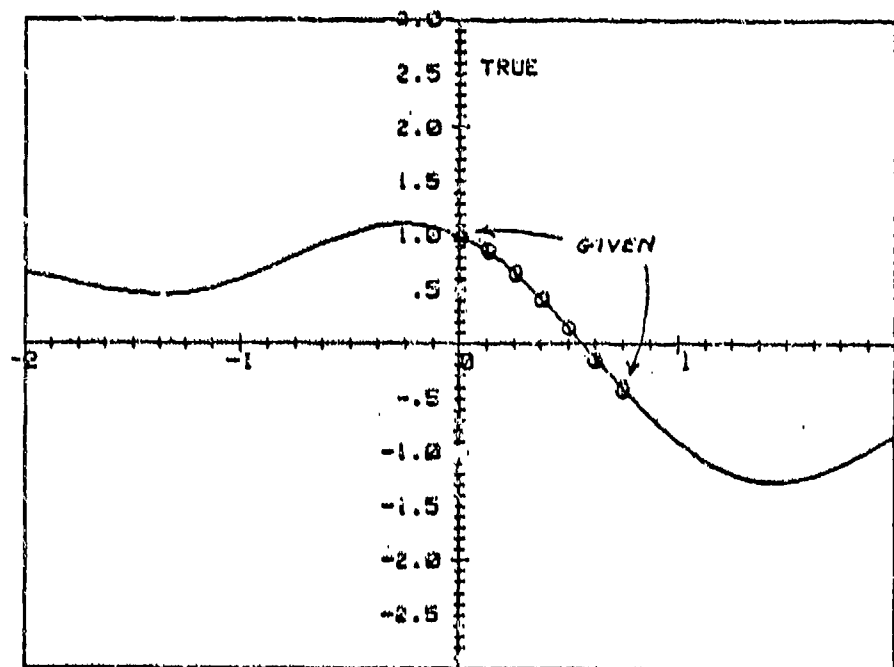


Figure 1. Sum of Cosines (a) True Signal and Given Samples

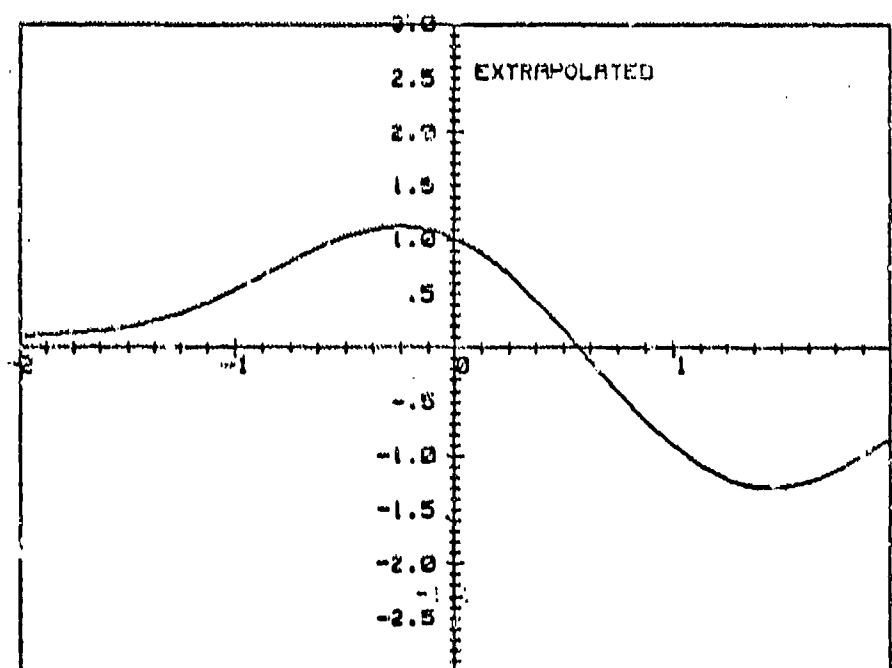


Figure 1. Sum of Cosines. (b) Extrapolated Signal.

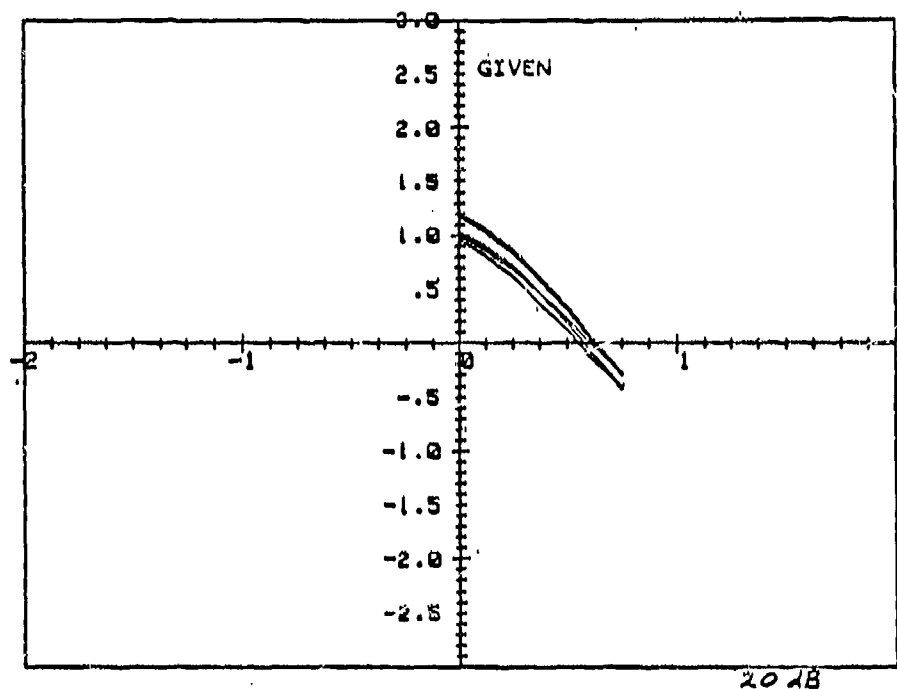


Figure 2. Sum of Cosines with Bandlimited Noise at 20 dB
Peak Signal to RMS Noise Ratio - several
independent runs. (a) Given Segment.

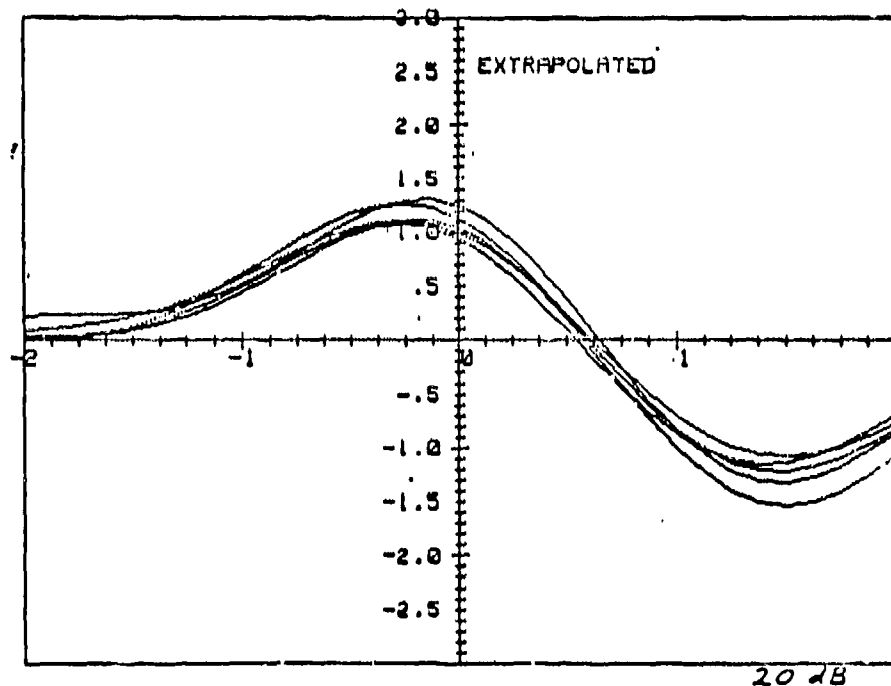


Figure 2. Sum of Cosines with Bandlimited Noise at 20dB
Peak Signal to RMS Noise Ratio - several
independent runs. (b) Extrapolated signal.

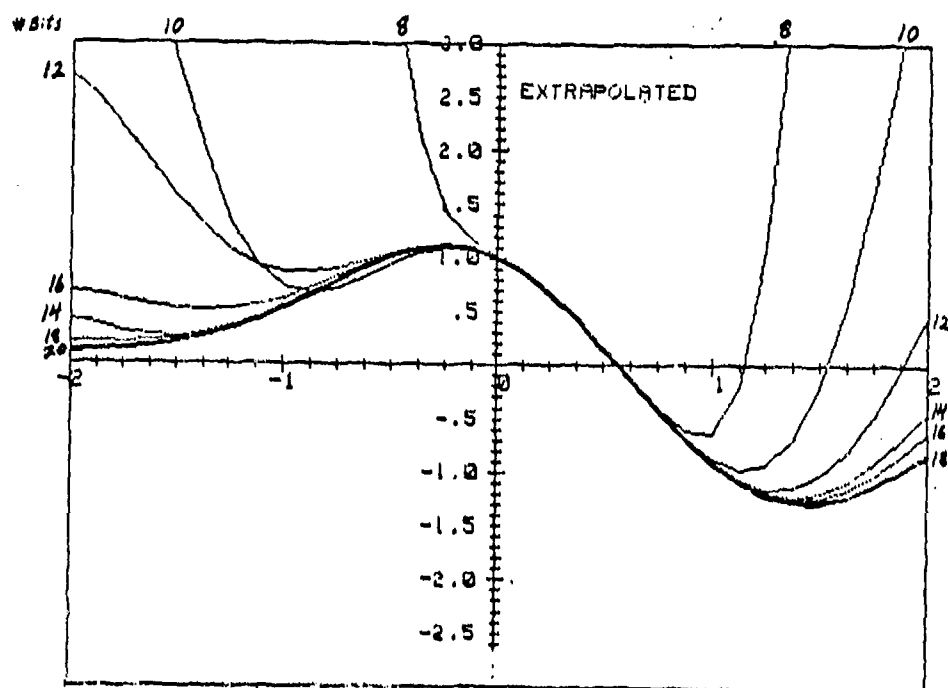


Figure 3. Effect of Quantization when Inaccurate Matrix Inversion is Used (Number of bits includes sign).

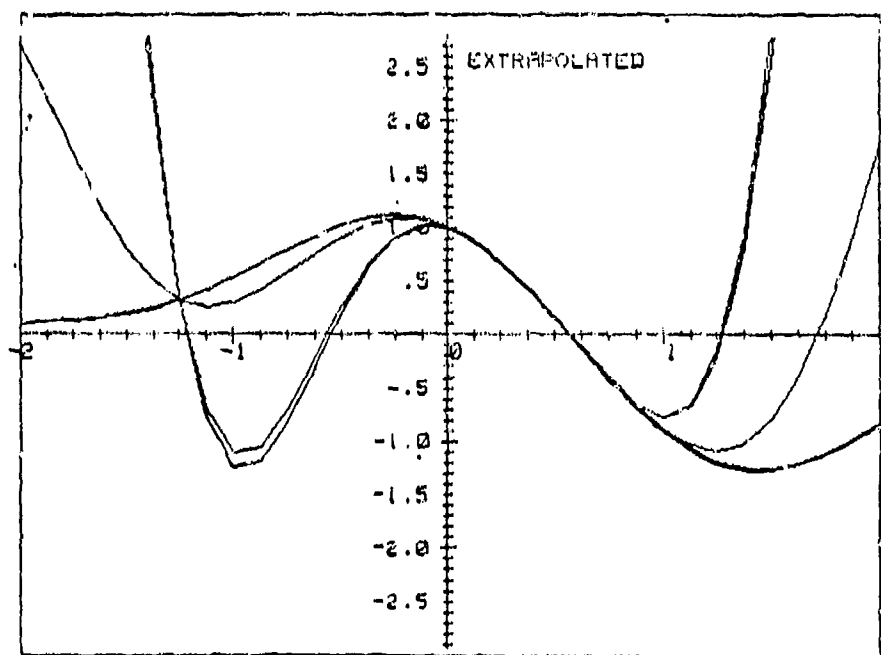


Figure 4. Effect of Bandlimited Noise at 40 dB SNR and 32 Bits Quantization (several independent runs) using inaccurate matrix inversion.

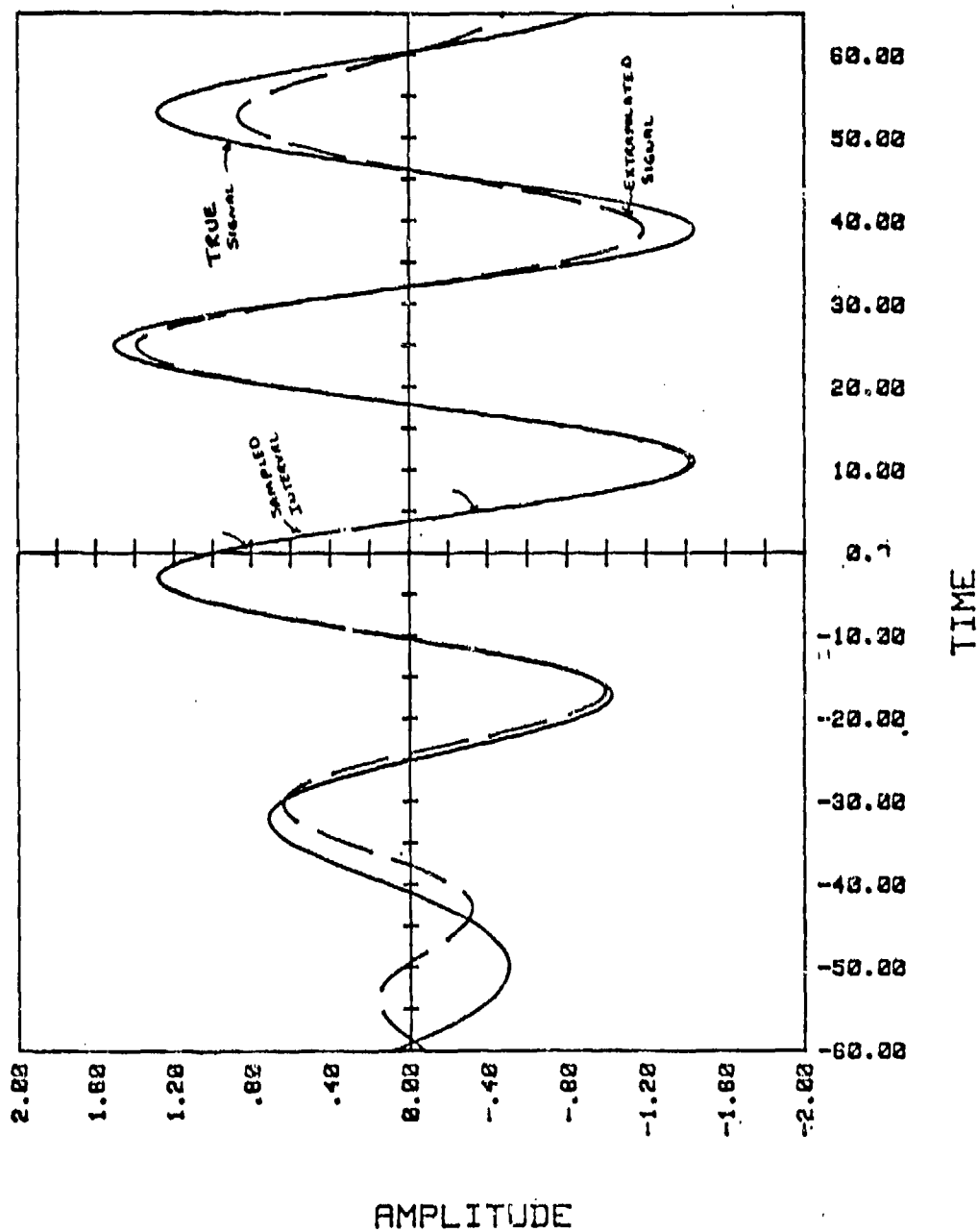


FIGURE 5. Extrapolated Sum of Cosines

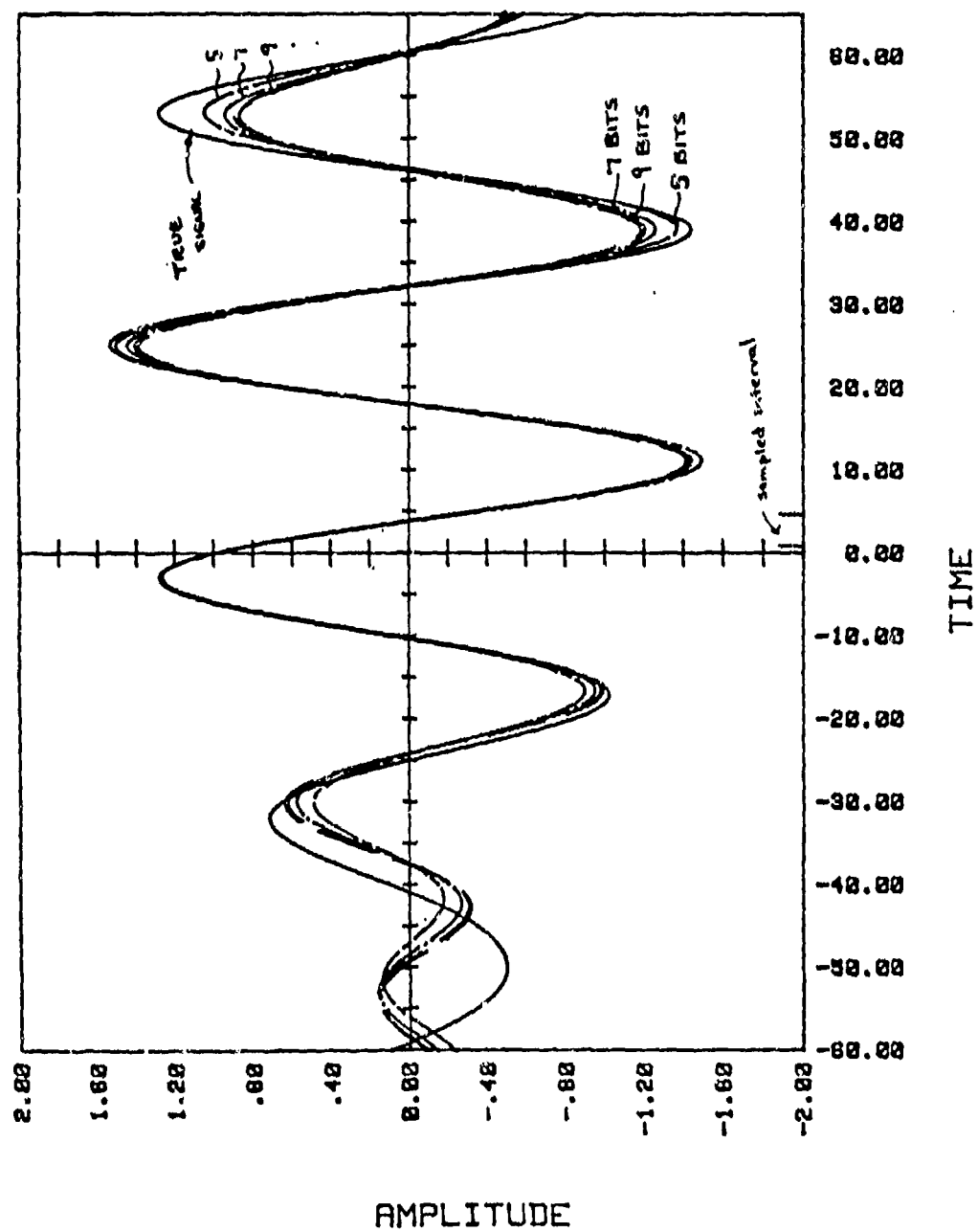


FIGURE 6. Quantized Signals
(Number of bits does not include sign)

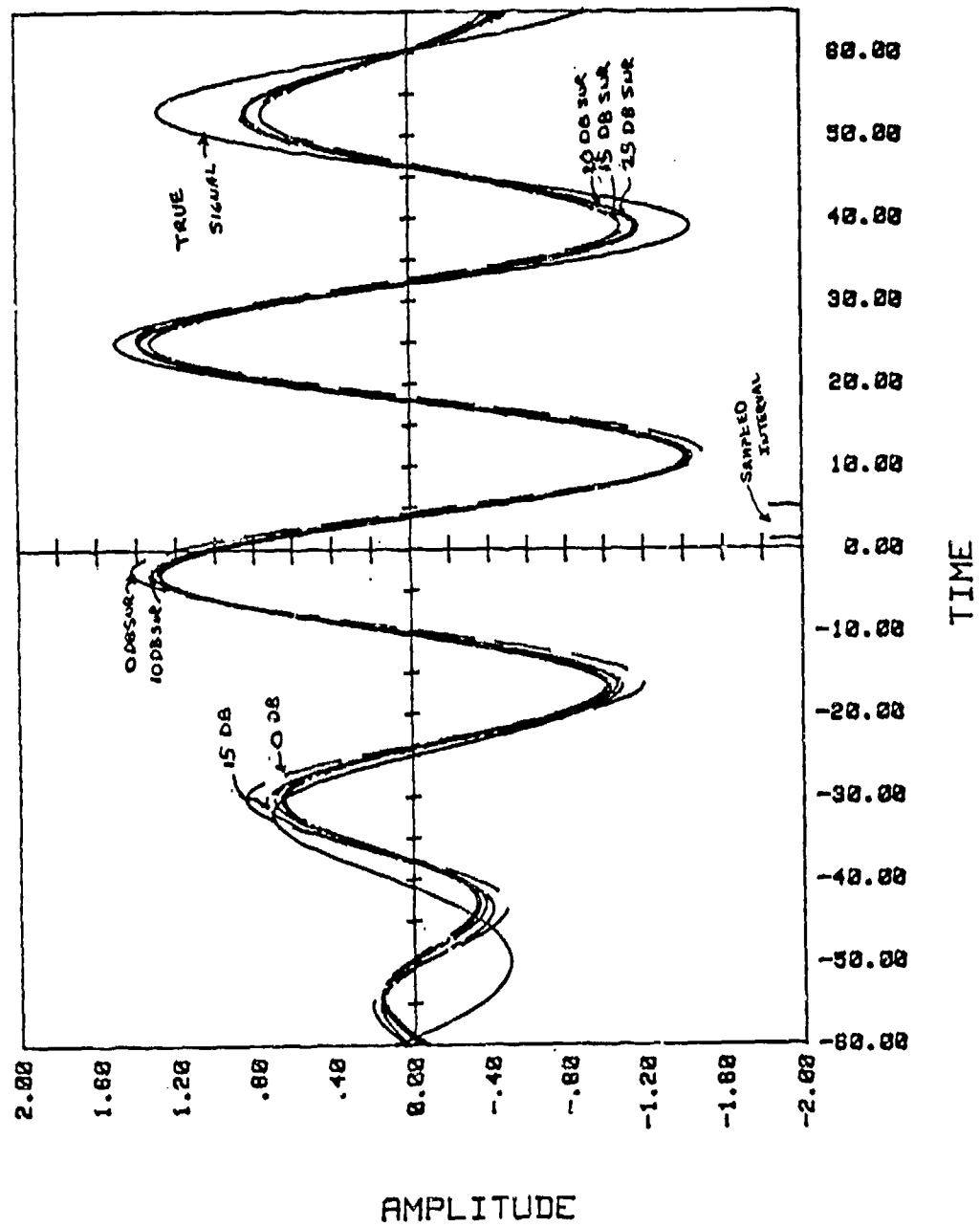


FIGURE 7. Signals In Bandlimited Noise

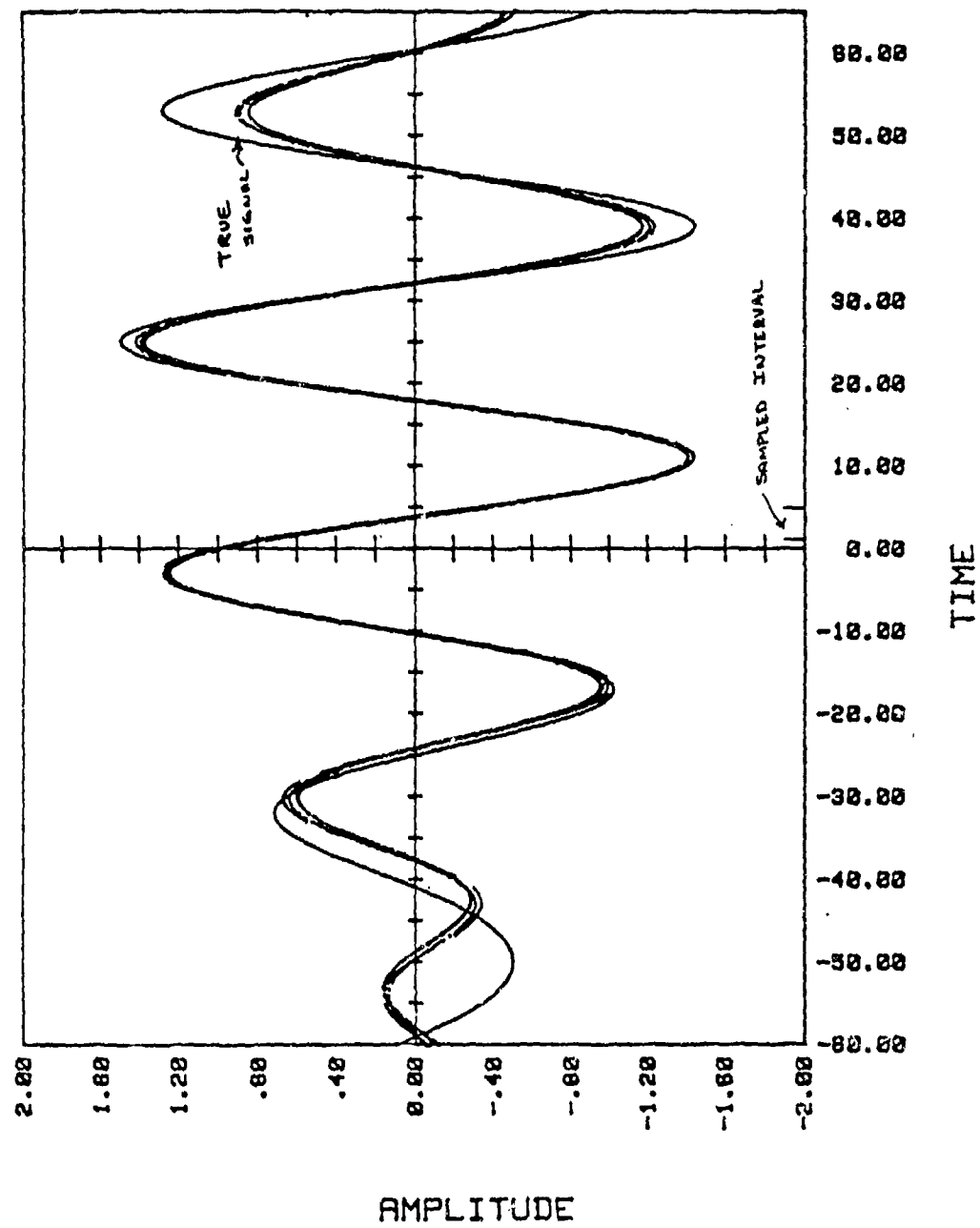


FIGURE 8A. Noise And Quantization
(7 Bits Plus Sign; 20 DBS SNR)

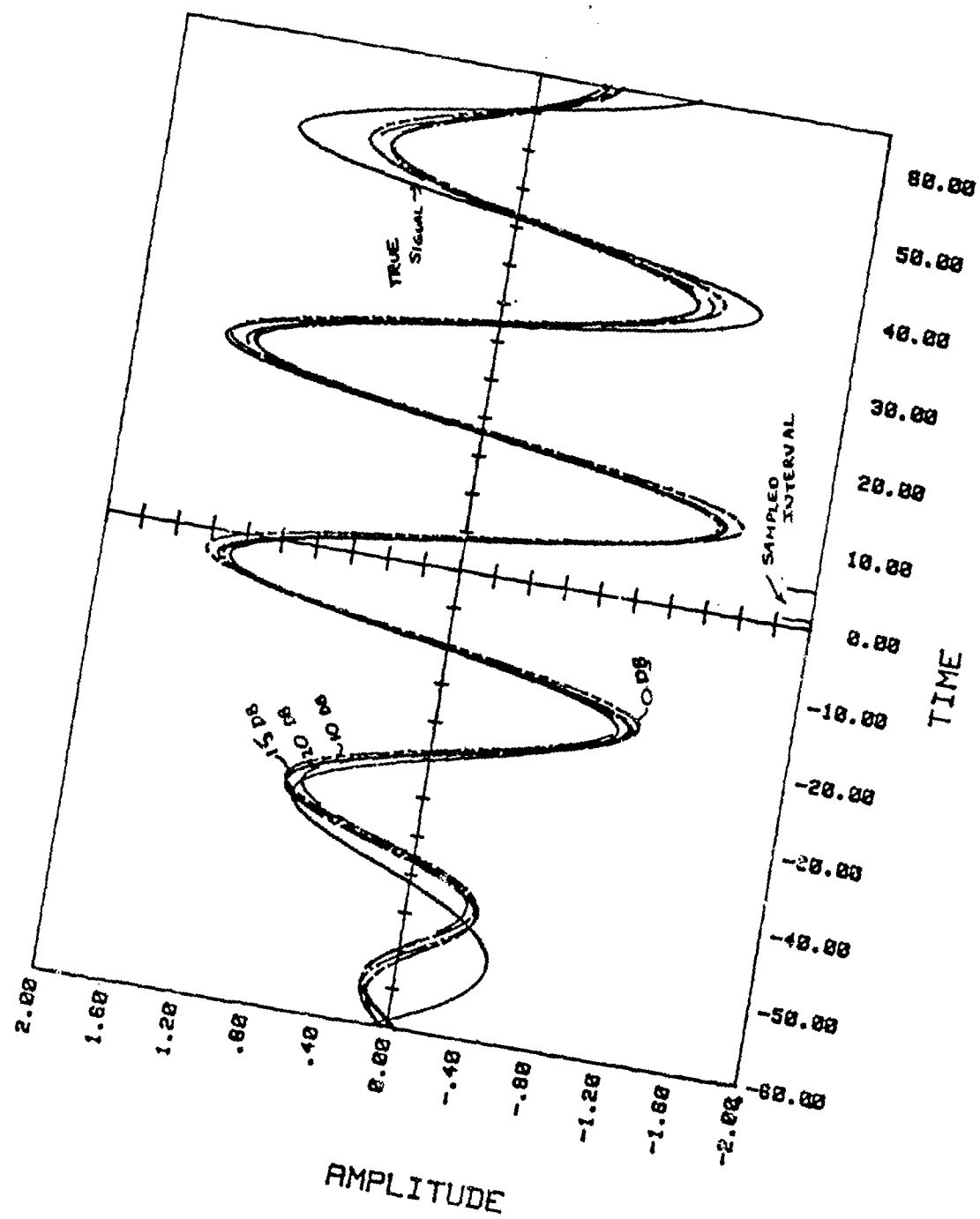


FIGURE 8B. Noise And Quantization
(7 Bits Plus Sign)

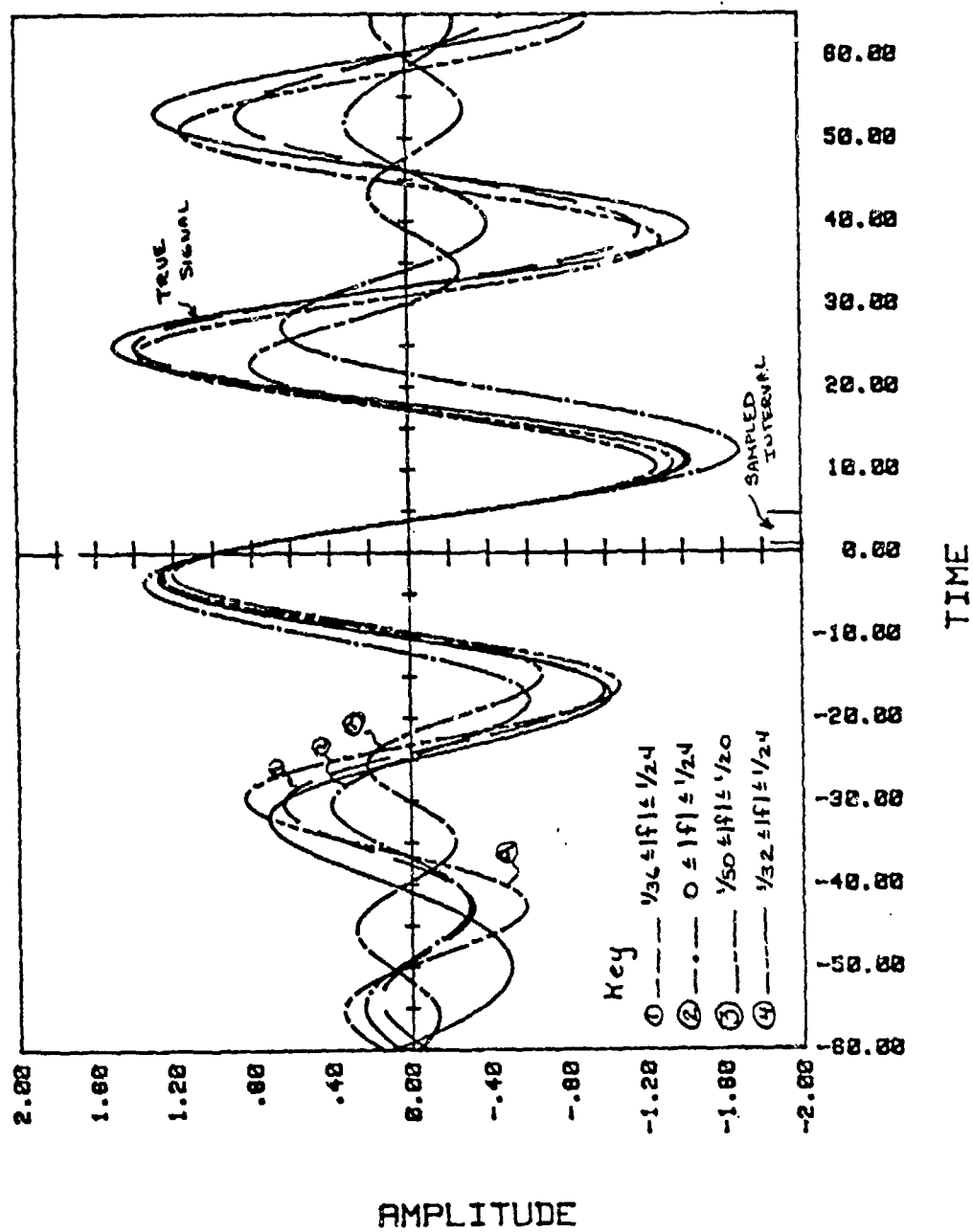


FIGURE 9. Passband Estimation

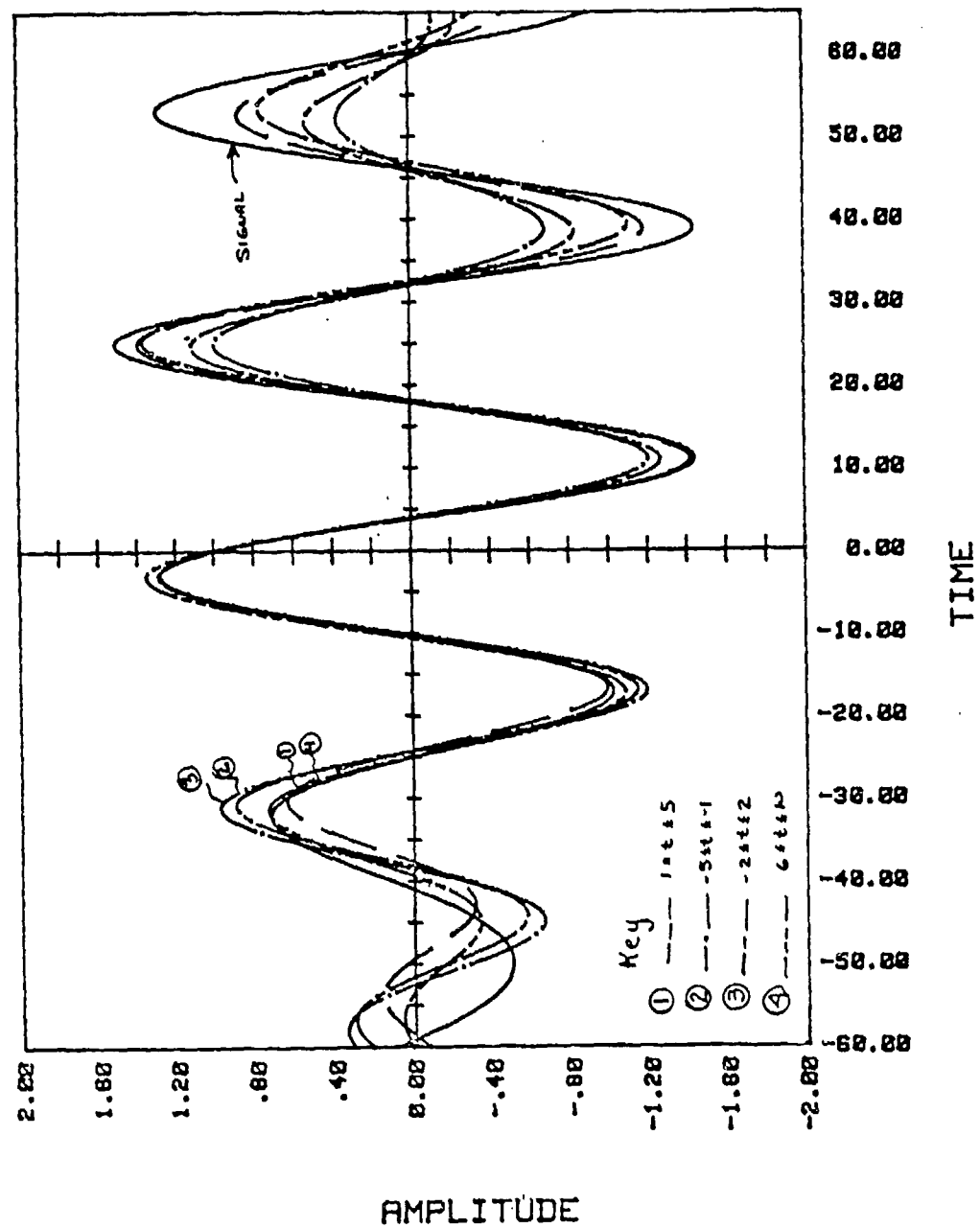


FIGURE 10. Various Time Intervals

ACCURACY OF SPECTRAL ESTIMATES
OF BAND-LIMITED SIGNALS

WILLIAM B. GORDON

Radar Division, Code 5308
Naval Research Laboratory
Washington, D.C. 20375

ABSTRACT

We consider the problem of estimating the spectrum of a band-limited signal perturbed by additive white noise. Sharp bounds on the mean square errors of linear spectral estimates are computed and expressed as functions of time-bandwidth products and signal-to-noise power ratios.

1. INTRODUCTION

A central problem in the theory of stationary time series is to estimate the spectrum of a stationary process $N(t)$ when the given data consists of samples of $s(t) = f(t) + N(t)$, where $f(t)$ is a deterministic trend of known functional form. In this paper we shall consider the dual problem: the functional form of $f(t)$ is unknown, the second order statistics of $N(t)$ are known, and the problem is to estimate the spectrum of $f(t)$. The signals $f(t)$ will be assumed to be band-limited functions having continuous Fourier transforms $\hat{f}(\nu)$ which vanish for $|\nu| \geq W$. Such signals are pulse-like, and the amount of useful information contained in an observation time window depends as much on the position of the window as it does on its length. Accurate spectral estimation requires that the observation time window capture a significant portion of the signal energy.

We shall find that when the data is sampled at the Nyquist rate $2W$ consistent spectral estimators do not exist, in the sense that infinitely accurate estimates cannot be obtained from infinitely long data records. For signals with effective time duration T_e and signal-to-noise power ratio (S/N) , most of the useful information is contained in a time window of length $N_c/(2W)$, where $N_c = (1/\pi) \sqrt{(2WT_e)^3 (S/N)}$. For any linear spectral estimator there exist signals whose corresponding spectral estimates have relative mean square errors on the order of $(1/N_c)$ and absolute mean square errors which are almost as large as the largest produced by the conventional

transform with N_c data points.

If the data is sampled at a rate higher than $2W$, longer time windows can be effectively used, and the conventional Fourier transform provides a consistent spectral estimator as the data rate increases without bound. If, however, the time window is fixed, consistent linear spectral estimators do not exist. Hence, to summarize, consistent linear spectral estimators exist only if both the data rate and the length of the time window increase without bound.

The problem of spectral estimation is essentially different from the problem of spectral peak detection and location, and hence our pessimistic results concerning the former do not preclude the possibility of high resolution (supergain) spectral peak detectors, such as have been recently proposed for band-limited signals [5,6,11,12]. However, improved resolution has as a necessary consequence a decrease in accuracy and detectability, and we hope to discuss this matter in subsequent papers.

2. THE SPACE $H(W)$

We shall consider a class $H(W)$ of complex-valued band-limited functions $f = f(t)$ whose Fourier transforms $\hat{f} = \hat{f}(v)$ are continuous and vanish for $|v| \geq W$, so that

$$f(t) = \int_{-W}^{+W} \hat{f}(v) \exp [2\pi i v t] dv. \quad (2.1)$$

$$\hat{f}(v) = \int_{-\infty}^{+\infty} f(t) \exp [-2\pi i v t] dt. \quad (2.2)$$

We shall also require the functions f in $H(W)$ to satisfy $M_2(f) < \infty$ where the moment $M_2 = M_2(f)$ is defined by

$$M_2 = 4\pi^2 \int_{-\infty}^{+\infty} t^2 |f(t)|^2 dt = \int_{-W}^{+W} |\hat{f}(v)|^2 dv. \quad (2.3)$$

We follow the standard terminology of radar and communications theory [7,14] in defining the signal energy $E = E(f)$, mean time $\tilde{t} = \tilde{t}(f)$, and effective time duration $T_G = T_G(f)$ by

$$E = \int_{-\infty}^{+\infty} |f(t)|^2 dt = \int_{-W}^{+W} |\hat{f}(v)|^2 dv.$$

$$\tilde{t} = (1/E) \int_{-\infty}^{+\infty} t |f(t)|^2 dt.$$

$$T_G^2 = (4\pi^2/E) \int_{-\infty}^{+\infty} (t - \tilde{t})^2 |f(t)|^2 dt.$$

Using Sobolev space techniques we get the fundamental inequalities

$$|f(t)|^2 \leq \frac{WM_2(f)}{2\pi^2 t^2} \left\{ 1 - \frac{\sin^2(2\pi Wt)}{(2\pi Wt)^2} \right\}, \quad (2.4)$$

$$|\hat{f}(v)|^2 \leq \frac{1 - (v/W)^2}{2} W M_2(f). \quad (2.5)$$

The time-bandwidth product $2WT_G$ will appear frequently in our subsequent discussion, one can easily establish the "uncertainty relation"

$$2W T_G \geq \pi,$$

which is the sharpest possible inequality of this type for $H(W)$.

3. THE ERROR FUNCTIONS Q^2 , R^2 , R_0^2

When the data is unperturbed by noise, $\hat{f}(v)$ will be estimated by discrete transforms of the type

$$\hat{f}_*(v) = \sum \bar{\beta}_n f(t_n),$$

and we shall derive inequalities of the type

$$|\hat{f}(v) - \hat{f}_*(v)|^2 \leq M_2(f) Q^2(v, \underline{t}, \underline{\beta}) \quad (3.1)$$

where $Q^2 = Q^2(v, \underline{t}, \underline{\beta})$ is a certain explicit function of the frequency v , the sample point set $t = \{t_1, t_2, \dots, t_N\}$, and the weights $\underline{\beta} = \{\beta_1, \beta_2, \dots, \beta_N\}$. These inequalities are sharp in the sense that for every given set of values for v , $\underline{t}$, $\underline{\beta}$ there exists a function f in $H(W)$ for which the inequality (3.1) becomes an equality. Hence the function Q^2 can be interpreted as the largest possible squared error in $\hat{f}_*(v)$ (when noise is absent) normalized by the moment M_2 .

Suppose now that the data is perturbed by additive white noise $N(t)$ with zero mean. The data consists of samples of $s(t) = f(t) + N(t)$, and the problem is to estimate $\hat{f}(v)$ by sums of the type $\sum \bar{\beta}_n s(t_n)$. This estimate has mean and bias equal to $\hat{f}_*(v)$ and $[\hat{f}(v) - \hat{f}_*(v)]$, resp., where $\hat{f}_*$ has the same meaning as before. The variance is independent of f and is given by

$$\text{Variance } \{\bar{\beta}_n s(t_n)\} = \sigma^2 \sum |\bar{\beta}_n|^2$$

where $\sigma^2 = E[|N(t)|^2]$. For any f in $H(W)$ the mean square error in the spectral estimate $\sum \bar{\beta}_n s(t_n)$ is given by $R^2(f) = (\text{bias})^2 + \text{variance}$, or

$$R^2(f) = |\hat{f}(v) - \hat{f}_*(v)|^2 + \sigma^2 \sum |\bar{\beta}_n|^2. \quad (3.2)$$

We define $R_0^2 = \sup \{R^2(f)\}$, where the supremum is taken over the set of all f in $H(W)$ having a given value of M_2 . Then from the sharpness of the inequality (3.1) we have

$$R_0^2 = M_2 Q^2(v, \underline{t}, \underline{\beta}) + \sigma^2 \sum |\bar{\beta}_n|^2, \quad (3.3)$$

and for every $\varepsilon > 0$ there exist f in $H(W)$ for which $R^2(f) \geq R_0^2 - \varepsilon$. Explicit formulas for $Q^2(v, t, \beta)$ will be given in Section 7 for arbitrary values of v, t , and β .

4. THE PARAMETER N_G

For simplicity we now assume that the sample point set $\{t_n\}$ is given by $t_n = n/(2W)$ where the integer n varies over the set $\{n: -N \leq n \leq N\}$. The number of samples will be denoted by N_g ($= 2N + 1$), and T will denote the length of the time window. Hence $N_g = 2WT$ and the time window is $[-T/2, T/2]$. When the conventional Fourier transform is used $\beta_n = (1/2W) \exp [i\pi n v/W]$ and it turns out that Q^2 is independent of v and is very closely approximated by

$$Q^2 \approx 2W/(\pi^2 N_g).$$

Hence, when the data is sampled at the Nyquist rate and the conventional Fourier transform is used we have

$$R_0^2 = \frac{2WM_2}{\pi^2 N_g} + \frac{\sigma^2 N_g}{4W^2}, \quad (4.1)$$

and R_0^2 is a convex function of N_g , whose minimum value is attained at $N_g = N_G$ where

$$N_G = \left[\frac{8W^3 M_2}{\pi^2 \sigma^2} \right]^{1/2} \quad (4.2)$$

For functions f satisfying $\bar{L}(f) = 0$ we have $M_2 = T_0^2 E$ and therefore

$$N_G = (1/\pi) \left[(2WT_0)^3 (S/N) \right]^{1/2} \quad (4.3)$$

where the signal-to-noise power ratio is defined by

$$(S/N) = \frac{\text{time-averaged signal power}}{\text{average noise power}} = \frac{E/T_0}{\sigma^2}$$

Although R_O^2 becomes an increasing function of N_g when $N_g > N_C$, we have no right to assume that the same is true for $R^2(f)$ for any particular f . However, it can be shown that for any f in $H(W)$, $R^2(f)$ eventually becomes an increasing function of N_g , and using some gross estimates this can be shown to be the case when $N_g > N_C^2$.

5. THE MAIN RESULT

We shall now present our main result (equation (5.1) below) which shows that consistent spectral estimators do not exist when the noise is white and the data is sampled at the Nyquist rate. In the last paragraph we saw that when the conventional fourier transform is used the error function R_O^2 eventually becomes an increasing function of N_g as $N_g \rightarrow \infty$. For each value of N_g we shall now choose a set of weights β which is "optimal" in the sense that it minimizes the right-hand side of (3.3) for given values of M_2 and σ^2 . These weights will be called " σ - optimal", and when they are used R_O^2 becomes a monotonically decreasing function of N_g . We define

$$R_\infty^2 = \lim_{N_g \rightarrow \infty} R_O^2$$

and from the definitions it is evident that

$$R_\infty^2 = \inf \{R_O^2\}$$

where the infimum is taken over the set of all linear spectral estimates obtained by sampling at the Nyquist rate. Hence, for any such linear spectral estimate and for any $\delta > 0$ there exist functions f in $H(W)$ satisfying

$$R^2(f) > R_\infty^2 - \delta.$$

The calculation of R_∞^2 will be described below. It turns out that

$$R_\infty^2 = WM_2 \frac{\cosh^2 [\pi N_C/2] - \cosh^2 [\pi N_C v/(2W)]}{(\pi N_C/2) \sinh \pi N_C} \quad (5.1)$$

This expression vanishes only when $v = \pm W$, which is a reflection of the fact that $\hat{f}(\pm W) = 0$ for all f in $H(W)$.

Let X_c denote the smallest possible value of R_c^2 when the conventional transform is used, i.e., the value obtained by setting $N_s = N_c$ in (4.1), and let X_0 denote the right-hand side of (2.5). Then X_0 is the mean square error of the "trivial" estimate $\hat{f}(v) \equiv 0$, and it is reasonable to compare R_∞^2 with X_0 in a neighborhood of $v = \pm W$. The ratio R_∞^2/X_0 can also be interpreted as a bound on the relative errors $R^2(f)/|\hat{f}(v)|^2$. For there exist f for which $R^2(f) > R_\infty^2$, and for any f we have $|\hat{f}(v)|^2 \leq X_0$ since X_0 is the right-hand side of (2.5). Hence, there are always f for which $R^2(f)/|\hat{f}(v)|^2$ exceeds R_∞^2/X_0 . From an examination of the ratios R_∞^2/X_c and R_∞^2/X_0 we draw the following conclusions.

Conclusion #1. When the data is sampled at the Nyquist rate every linear spectral estimator produces mean square errors having the same order of magnitude as the largest produced by a conventional Fourier transform with N_c data points, except near $v = \pm W$ where the errors have the same order of magnitude as those produced by the trivial estimate $\hat{f}(v) \equiv 0$.

Conclusion #2. When the data is sampled at the Nyquist rate every linear spectral estimator produces relative mean square errors which are on the order of $1/N_c$ near $v = 0$ and unity near $v = \pm W$.

6. OVERSAMPLING

In this section we consider the effects of oversampling. That is, we now suppose that functions of class $H(W)$ are sampled at a rate $2kW$, $k \geq 1$. The derivations of (4.1) and (5.1) require the closed-form inversion of certain matrices, which, unfortunately, we have been unable to effect for the case of oversampling. Hence, we are presently unable to give a quantitative description of how much useful information is contained in a fixed time window when the data rate is increased without bound. However, it can be shown that consistent spectral estimators exist only if both the data rate $2kW$ and the length T of the time window are allowed to increase without bound. (Cf. the discussion in Section B.1 of Blackman and Tukey [4] which suggests the existence of results of this nature.) Moreover, because of the pulse-like nature of the functions f in $H(W)$, it is also necessary to prolong

the time window in both directions, since otherwise R_0^2 will not converge to its infimum as $T \rightarrow \infty$.

These results can be proved by a reduction to the "previous case" of Nyquist sampling. For by Shannon's Sampling Theorem every finite linear combination of values $f[n/(2kW)]$ can be expressed as an infinite linear combination of values $f[n/(2W)]$.

7. DERIVATIONS

We shall now write $||f||^2$ for the norm $M_2(f)$, and we let $(\cdot, \cdot)$ denote the corresponding inner product, so that

$$(f, g) = \int_{-W}^{+W} D\hat{f}(v) \overline{D\hat{g}(v)} dv$$

where D denotes the differentiation operator. This norm is chosen because it can be expressed in terms of physically meaningful parameters, and because the maps $f \mapsto f(t)$ and $f \mapsto \hat{f}(v)$ are continuous in this norm. (This last property is not enjoyed by the usual L^2 norm.) Hence, there exist functions $K_t = K_t(s)$ and $e_v = e_v(s)$ in $H(W)$ which satisfy

$$f(t) = (f, K_t), \quad \hat{f}(v) = (f, e_v).$$

Hilbert function spaces H for which the maps $f \mapsto f(t)$ are continuous are called reproducing kernel hilbert spaces, and the function $K(s, t) = K_t(s) = (K_t, K_s)$ is called the reproducing kernel. When $H = H(W)$ we have

$$K_s(t) = \frac{1}{4\pi^3} \left\{ \frac{\sin [2\pi W(t-s)]}{st(t-s)} - \frac{\sin(2\pi Ws)\sin(2\pi Wt)}{2\pi W s^2 t^2} \right\}, \quad (s, t, s-t \neq 0).$$

$$K_t(t) = \frac{1}{4\pi^3 t^3} \left\{ 2\pi Wt - \frac{\sin^2(2\pi Wt)}{2\pi Wt} \right\}, \quad (t \neq 0).$$

$$K_0(t) = K_t(0) = \frac{1}{4\pi^3 t^3} \left\{ \sin(2\pi Wt) - (2\pi Wt) \cos(2\pi Wt) \right\}, \quad (t \neq 0).$$

$$K_0(0) = \frac{2}{3} W^3. \quad (7.1)$$

To derive these results we integrate (f, K_t) by parts and compare the results to (2.1). It is easily seen that $\hat{K}_t(v)$ is the (unique) solution to $D^2 \hat{K}_t(v) = -\exp[2\pi i v t]$ which satisfies the boundary conditions $\hat{K}_t(\pm W) = 0$. In a similar fashion, one establishes that

$$e_v(t) = \frac{1}{4\pi^2 t^2} \left\{ e^{2\pi i v t} - \cos(2\pi Wt) - (iv/W) \sin(2\pi Wt) \right\}, \quad (t \neq 0).$$

$$e_v(0) = (1/2) (W^2 - v^2) \quad (7.2)$$

$$||e_v||^2 = (W^2 - v^2)/(2W).$$

From hilbert space generalities we get

$$Q^2 = \sup_f \left\{ \frac{|\hat{f}(v) - \hat{f}_*(v)|^2}{||f||^2} \right\} = ||e_v - \sum \beta_n K_{t_n}||^2, \text{ or,} \quad (7.3)$$

$$Q^2 = \langle K \underline{\beta}, \underline{\beta} \rangle - \langle \underline{\beta}, \underline{J} \rangle - \langle \underline{J}, \underline{\beta} \rangle + ||e_v||^2$$

where the matrix $K = K(\underline{t})$, and the vector $\underline{J} = \underline{J}(v, \underline{t})$, are defined by

$$K_{nm} = (K_{t_m}, K_{t_n}) = K_{t_m}(t_n) \quad (7.4)$$

$$J_n = (e_v, K_{t_n}) = e_v(t_n) \quad (7.5)$$

The results (7.1) - (7.5) can now be substituted into (3.3), and it is easily seen that the minimization of R_0^2 with respect to $\underline{\beta}$ requires the inversion of the matrix $K + (\sigma^2/M_2)I$.

REFERENCES

1. A. Aronszajn, "Theory of Reproducing Kernels", Trans. AMS 68 (1950), p. 337-404.
2. J. Barros-Neto, "An Introduction to the Theory of Distributions", Marcel-Dekker, New York 1973.
3. L. Bers, F. John, M. Schecter, "Partial Differential Equations", Interscience, New York, 1964.
4. R. B. Blackman and J. W. Tukey, "The Measurement of Power Spectras", Dover, New York, 1959.
5. J. A. Cadzow, "Improved Spectral Estimation from Incomplete Sampled Data Observations", RADC Spectrum Analysis Workshop, Rome Air Development Center, Rome, N.Y., May 1978, p. 85-96.
6. J. A. Cadzow, "An Extrapolated Procedure for Band-Limited Signals", IEEE Trans. on Acoustics, Speech, and Signal Processing ASSP-27 (1979), p. 4-12.
7. G. W. Deley, "Waveform Design", in Radar Handbook (M. I. Skolnik, ed.), McGraw-Hill, New York, 1970.
8. Haber, "Numerical Estimation of Multiple Integrals", SIAM Review 12 (1970), p. 481-526.
9. R. S. Palais, "Foundations of Global Non-Linear Analysis", Benjamin, New York, 1968.
10. A. Papoulis, "Limits on Band-Limited Signals", Proc. IEEE 55 (1976), p. 1677-1686.
11. A. Papoulis, "A New Algorithm in Spectral Analysis and Band-Limited Extrapolation", IEEE Trans. on Circuits and Systems CAS-22 (1975), p. 735-742.
12. A. Papoulis and C. Chamzas, "Adaptive Extrapolation and Hidden Periodicities", RADC Spectrum Analysis Workshop, Rome Air Development Center, Rome, N.Y., May 1978, p. 85-96.

COMPENSATION OF AUTOREGRESSIVE SPECTRAL ESTIMATES FOR THE PRESENCE OF WHITE OBSERVATION NOISE

STEVEN KAY

Raytheon Company
Submarine Signal Division
Portsmouth, RI 02871

Abstract

The autoregressive spectral estimator possesses excellent resolution properties for time series which satisfy the "all-pole" assumption. When noise is added to the time series under analysis, the resolution of the spectral estimator degrades rapidly. The usual approach to this problem is to model the resulting time series by the more appropriate autoregressive-moving average process and to use standard time series analysis techniques to identify the autoregressive parameters. This standard technique, however, does not result in a positive-definite autocorrelation matrix. Thus, the resulting spectral estimator may exhibit a large increase in variance. An alternative approach, termed the noise compensation technique, is proposed. It attempts to correct the estimated reflection coefficients for the effect of white noise assuming the noise variance is known. Simulation results indicate that a significant decrease in the degrading effects of noise may be effected using the noise compensation technique.

I. Introduction

Autoregressive (AR) spectral estimation has received much attention lately in many diverse fields. Although based upon different theoretical foundations, Maximum Entropy Spectral Estimation, [1] used in seismic signal processing, and Linear Spectral Prediction, [2] used in speech signal processing, are in practice identical to AR spectral estimation. The principal advantage of the AR estimate over conventional Fourier-based spectral estimators is its enhanced resolution properties. [3] However, it has been shown that much of this increased resolution is lost when observation noise is added to the AR time series. [4] The reason for the degradation of the spectral estimate in the presence of noise is that the AR assumption, i.e., that the time series can be represented as the output of an all-pole filter excited by white noise is no longer valid. [5] Thus, the lower the signal-to-noise ratio (SNR), the more the "all-pole" assumption is violated, and the poorer the spectral estimate obtained.

The usual approach to this problem is to model the noise corrupted time series by the appropriate autoregressive-moving average (ARMA) process and to use standard time series analysis techniques to identify the autoregressive parameters. [6], [7]

This approach estimates the AR parameters $\{a_k, k = 1, 2, \dots, p\}$ as the solution of the equations:

$$\begin{bmatrix} \hat{R}_Y(p) & \hat{R}_Y(p-1) & \dots & \hat{R}_Y(1) \\ \hat{R}_Y(p+1) & \hat{R}_Y(p) & \dots & \hat{R}_Y(2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{R}_Y(2p-1) & \hat{R}_Y(2p-2) & \dots & \hat{R}_Y(p) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = - \begin{bmatrix} \hat{R}_Y(p+1) \\ \hat{R}_Y(p+2) \\ \vdots \\ \hat{R}_Y(2p) \end{bmatrix} \quad (1)$$

where $\hat{R}_Y(k)$ is the autocorrelation function estimate of the ARMA (p,p) process, Y_t . This standard technique, however, does not result in a positive-definite autocorrelation matrix. Thus the spectral estimator may exhibit a large increase in variance. [8] An alternative method, termed the noise compensation technique, is proposed. It attempts to correct or compensate the estimated reflection coefficients for the effects of the white observation noise. Via this method the autocorrelation matrix may be easily checked for positive-definiteness by assuring $|\hat{K}_i^c| < 1^{(1)}$, $i = 1, 2, \dots, p$, where $\hat{K}_i^c$ denotes the noise compensated reflection coefficient estimate.

II. Noise Compensation Technique

The noise compensation technique is now derived.

Assume that we are given a data record $\{Y_t, t = 1, 2, \dots, N\}$ where $Y_t = X_t + W_t$. X_t is an AR process of order p and W_t is white noise with variance σ_W^2 . The Burg estimate of K_i , the i^{th} reflection coefficient, $^{(1)}$ is

$$\hat{K}_i = \frac{-2 \sum_{t=1}^{N-i} e_{t+i}^{(i-1)} b_t^{(i-1)}}{\sum_{t=1}^{N-i} \left[e_{t+i}^{(i-1)2} + b_t^{(i-1)2} \right]} \quad (2a)$$

where

$$e_{t+i}^{(i-1)} = Y_{t+i} + \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} Y_{t+i-k} \quad (2b)$$

$$b_t^{(i-1)} = Y_t + \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} Y_{t+k}$$

Since K_i will be biased due to the noise, one is more interested in obtaining an estimate of:

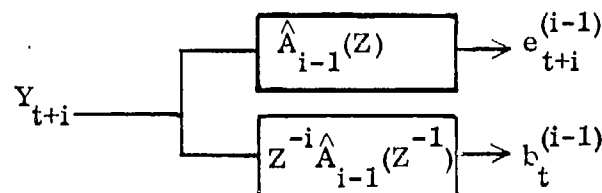
$$K_i^* = \frac{-E \left[e_{t+i}^{(i-1)} b_t^{(i-1)} \right]^*}{\frac{1}{2} \left\{ E \left[e_{t+i}^{(i-1)2} \right]^* + E \left[b_t^{(i-1)2} \right]^* \right\}}$$

where * indicates the value of the quantity for

$$Y_t = X_t, \text{ i.e., no observation noise present.}$$

If we let $\hat{A}_{i-1}(Z) = 1 + \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} Z^{-k}$,

we have



Assuming that the W_t process is uncorrelated with the X_t process and $E(W_t) = E(X_t) = 0$, it can be shown that

$$E \left[e_{t+i}^{(i-1)2} \right] = E \left[e_{t+i}^{(i-1)2} \right]^* + \sigma_W^2 \left[1 + \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)2} \right] = E \left[b_t^{(i-1)2} \right]$$

$$E \left[e_{t+i}^{(i-1)} b_t^{(i-1)} \right] = E \left[e_{t+i}^{(i-1)} b_t^{(i-1)} \right]^* + \sigma_W^2 \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} \hat{a}_{i-k}^{(i-1)}.$$

Therefore,

$$K_i^* = - \frac{E \left[e_{t+i}^{(i-1)} b_t^{(i-1)} \right] - \sigma_W^2 \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} \hat{a}_{i-k}^{(i-1)}}{\frac{1}{2} \left\{ E \left[e_{t+i}^{(i-1)2} \right] + E \left[b_t^{(i-1)2} \right] \right\} - \sigma_W^2 \left[1 + \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)2} \right]}.$$

In order to relate K_i^* to the Burg estimate, we take as estimators of the expected values the following:

$$\widehat{E \left[e_{t+i}^{(i-1)} b_t^{(i-1)} \right]} = \frac{1}{N-i-(i-1)} \sum_{t=1}^{N-i} e_{t+i}^{(i-1)} b_t^{(i-1)}$$

$$\widehat{E \left[e_{t+i}^{(i-1)2} \right]} = \frac{1}{N-i-(i-1)} \sum_{t=1}^{N-i} e_{t+i}^{(i-1)2}$$

$$\widehat{E \left[b_t^{(i-1)2} \right]} = \frac{1}{N-i-(i-1)} \sum_{t=1}^{N-i} b_t^{(i-1)2}$$

Note that the estimates have accounted for the $(i-1)$ degrees of freedom lost in estimating $\{a_k^{(i-1)}\}$, which are needed to generate $e_{t+i}^{(i-1)}, b_t^{(i-1)}$.

Then,

$$\hat{K}_i = \frac{-\frac{1}{N-i-(i-1)} \sum_{t=1}^{N-i} e_{t+i}^{(i-1)} b_t^{(i-1)}}{\frac{1}{2} \left[\frac{1}{N-i-(i-1)} \sum_{t=1}^{N-i} e_{t+i}^{(i-1)2} + \frac{1}{N-i-(i-1)} \sum_{t=1}^{N-i} b_t^{(i-1)2} \right]}$$

Letting $\hat{K}_i^c$ be our compensated estimate, we have,

$$\hat{K}_i^c = - \frac{\frac{1}{N-i-(i-1)} \sum_{t=1}^{N-i} e_{t+i}^{(i-1)} b_t^{(i-1)} - \sigma_W^2 \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} \hat{a}_{i-k}^{(i-1)}}{\frac{1}{2} \left[\frac{1}{N-i-(i-1)} \sum_{t=1}^{N-i} e_{t+i}^{(i-1)2} + \frac{1}{N-i-(i-1)} \sum_{t=1}^{N-i} b_t^{(i-1)2} \right] - \sigma_W^2 \left[1 + \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)2} \right]}$$

or finally,

$$\hat{K}_i^c = - \frac{2 \sum_{t=1}^{N-i} e_{t+i}^{(i-1)} b_t^{(i-1)} - 2(N-2i+1) \sigma_W^2 \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} \hat{a}_{i-k}^{(i-1)}}{\sum_{t=1}^{N-i} \left[e_{t+i}^{(i-1)^2} + b_t^{(i-1)^2} \right] - 2(N-2i+1) \sigma_W^2 \left[1 + \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)^2} \right]}.$$

To be more realistic, we replace σ_W^2 by $\alpha_i \sigma_W^2$ where $0 < \alpha_i < 1$. Thus,

$$\hat{K}_i^c = - \frac{2 \sum_{t=1}^{N-i} e_{t+i}^{(i-1)} b_t^{(i-1)} - 2(N-2i+1) \alpha_i \sigma_W^2 \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} \hat{a}_{i-k}^{(i-1)}}{\sum_{t=1}^{N-i} \left[e_{t+i}^{(i-1)^2} + b_t^{(i-1)^2} \right] - 2(N-2i+1) \alpha_i \sigma_W^2 \left[1 + \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)^2} \right]} \quad (3)$$

It is seen that $\hat{K}_i^c$ is not constrained to be between -1 and +1. To maintain this range, we will need to choose $\{\alpha_i\}$ carefully. In the simulation examples to follow, we choose α_i as:

$$\alpha_i = \frac{(p-i) \alpha_{MAX} + \alpha_{MIN} (i-1)}{p-1} \quad (4)$$

where $0 < \alpha_{MIN} < \alpha_{MAX} < 1$.

Thus, α_i decreases linearly with i , which reflects our confidence in the estimates of the higher order reflection coefficients.

To demonstrate the capability of the noise compensation technique several simulation examples are given. The first example utilizes data composed of an AR(4) process [9] with power spectral density given in Figure 1 and white Gaussian noise. The SNR is 15 dB. 100 realizations of the conventional ARMA approach, which uses (1) with the autocorrelation estimate

$$\hat{R}_Y(k) = \frac{1}{N} \sum_{t=1}^{N-k} Y_t Y_{t+k},$$

are shown in Figure 2. In Figure 3 the noise compensated estimates are plotted. Note that $p = 8$ was used since $p = 4$ did not result in adequate resolution of the spectral peaks. Comparing the figures, we see that not only is the variance of the noise compensated spectral estimator less than that of the ARMA approach but the spectral peaks are clearly visible whereas in the ARMA case only one peak is seen.

Finally, a simulation was conducted for two equi-amplitude sinusoids in white Gaussian noise. The peak of the spectral estimate was normalized to 0 dB and the true spectral lines are indicated by arrows. The results are shown in Figures 4 and 5. A large improvement is noted.

III. Conclusions

The noise compensation technique described in this paper offers an alternative and possibly better method than the conventional ARMA approach of reducing the effects of white observation noise on the autoregressive spectral estimator. When used properly, the large bias error introduced by the noise can be significantly reduced.

References

1. Burg, J.P., 1975, "Maximum Entropy Spectral Analysis", Ph.D. dissertation, Stanford Univ.
2. Makhoul, J., April 1975, "Linear Prediction: a Tutorial Review", Proc. IEEE, Vol. 63, pp 561-580.
3. Lacoss, R.T., August 1971, "Data Adaptive Spectral Analysis Methods", Geophysics, Vol. 36, pp 661-675.
4. Marple, S.L., Jr., 1977, "Resolution of Conventional Fourier, Autoregressive, and Special ARMA Methods of Spectral Analysis", IEEE Int. Conf. on ASSP, Hartford, CN.
5. Kay, S.M., "The Effects of Noise on the Autoregressive Spectral Estimator", to be published in the IEEE Trans. on Acoustics, Speech, and Signal Processing.
6. Gersch, W., October 1970, "Estimation of the Autoregressive Parameters of a Mixed Autoregressive Moving-Average Time Series", IEEE Trans. on Automatic Control, pp 583-588.
7. Box, G.E.P., Jenkins, G.M., 1970, Time Series Analysis—Forecasting and Control, Holden-Day, San Francisco, CA.
8. Kay, S.M., "Noise Compensation for Autoregressive Spectral Estimates", submitted to the IEEE Trans. on Acoustics, Speech, and Signal Processing.
9. Ulrych, T.J., Bishop, T.N., February 1975, "Maximum Entropy Spectral Analysis and Autoregressive Decomposition", Reviews of Geophysics and Space Physics, Vol. 13, pp 183-200.

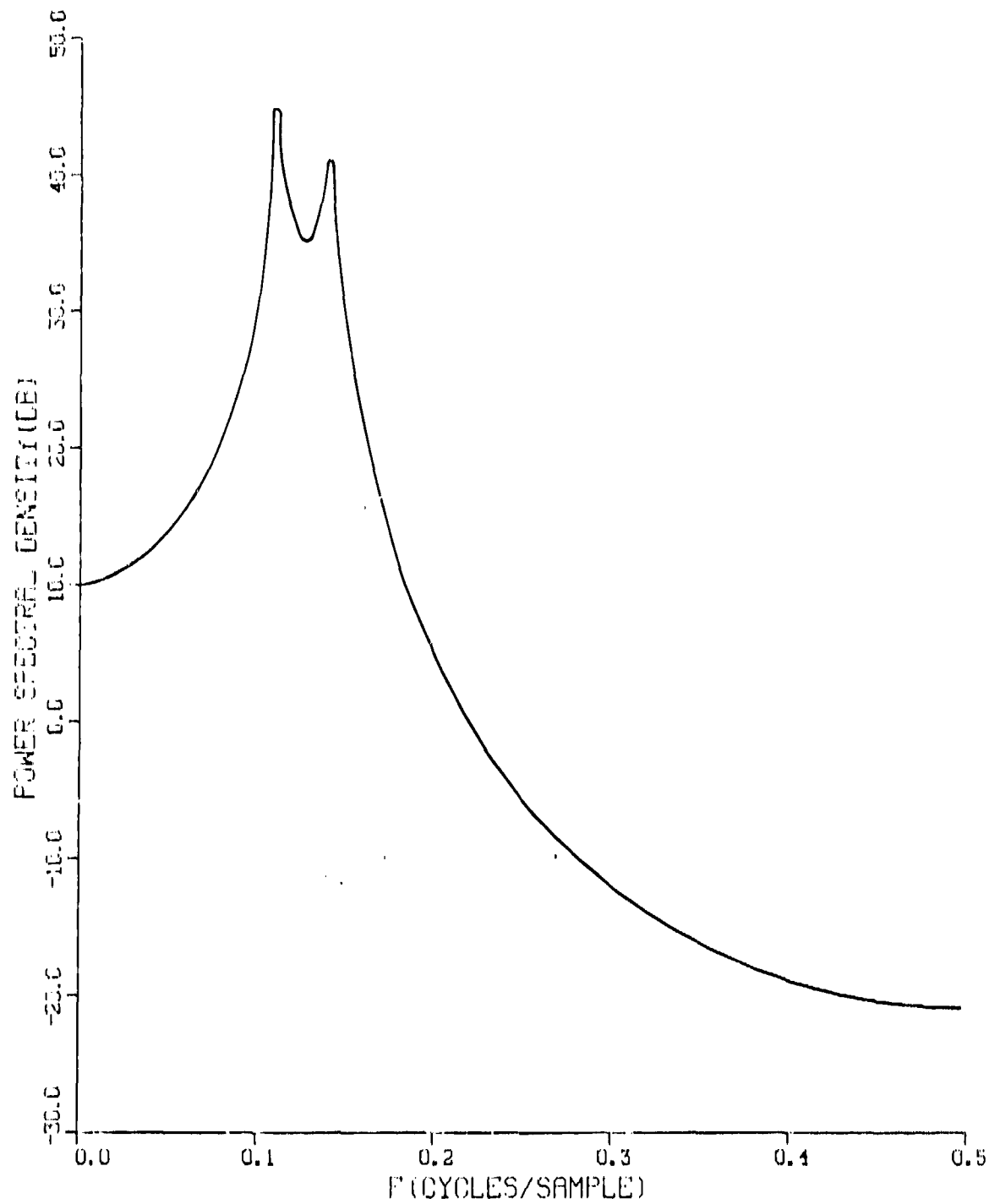


FIGURE 1. True Power Spectral Density of AR(4) Process

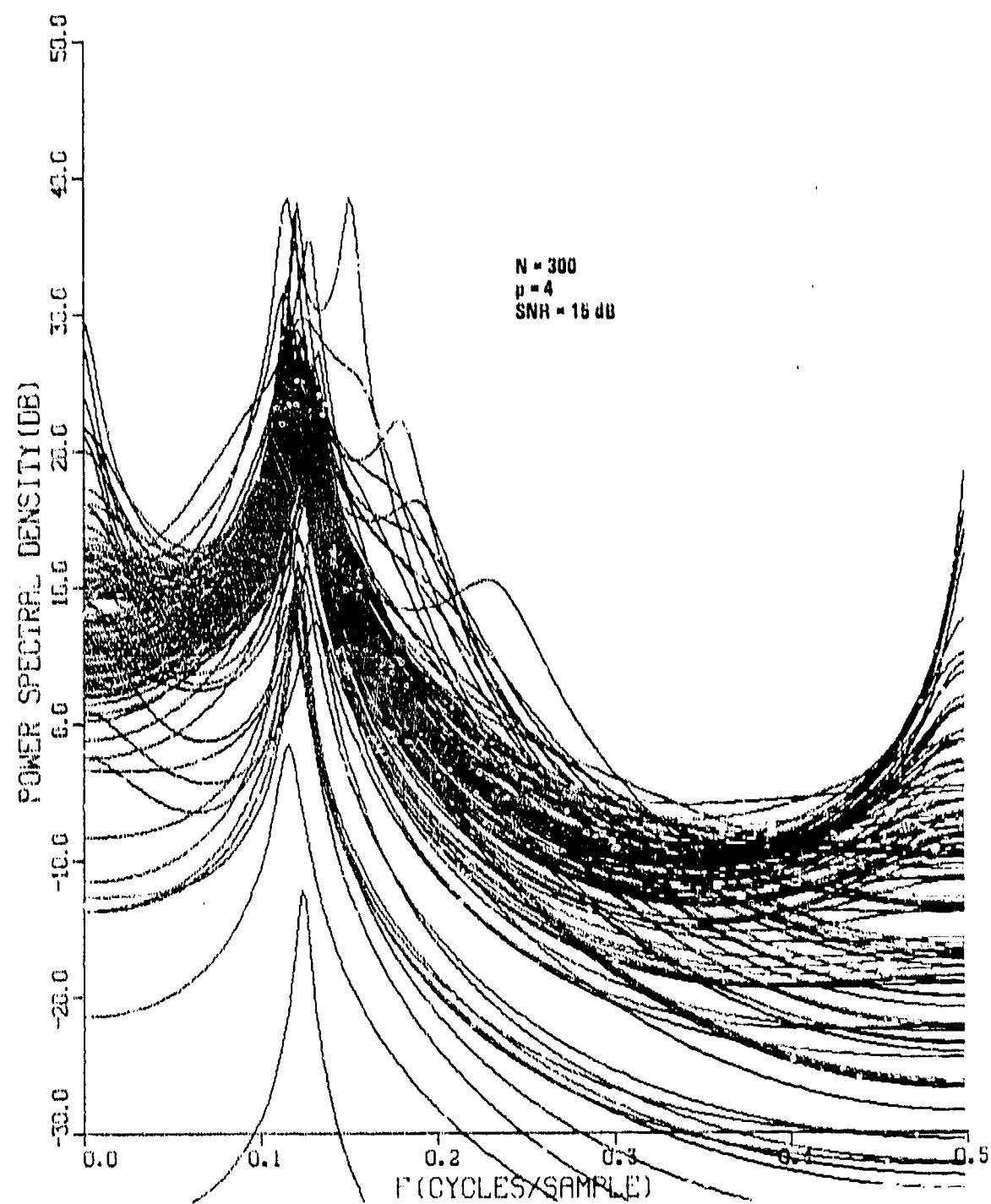


FIGURE 2. ARMA Approach Spectral Estimate of
Noise Corrupted AR(4) Process

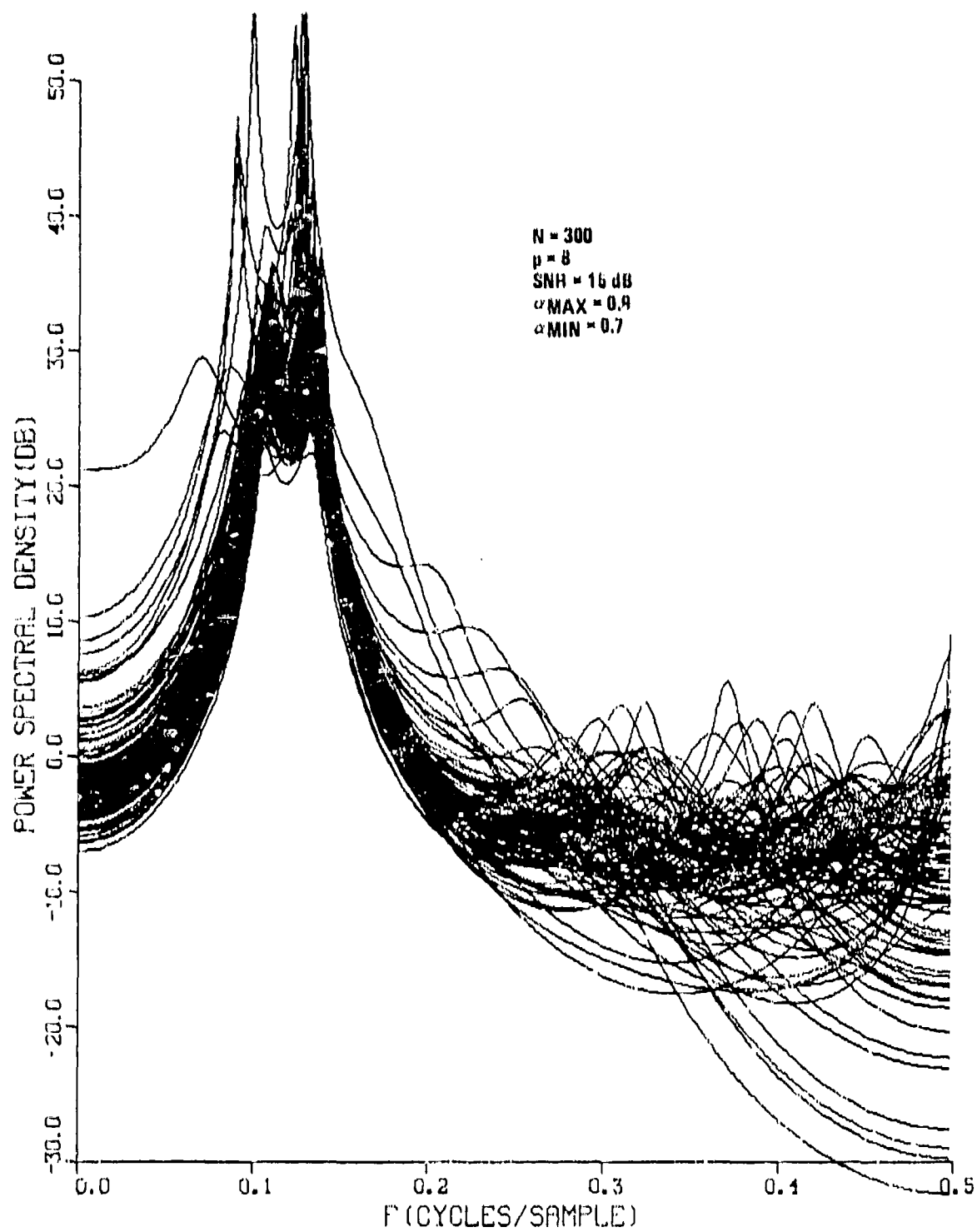


FIGURE 3. Noise Compensated Spectral Estimate of
Noise Corrupted AR(4) Process

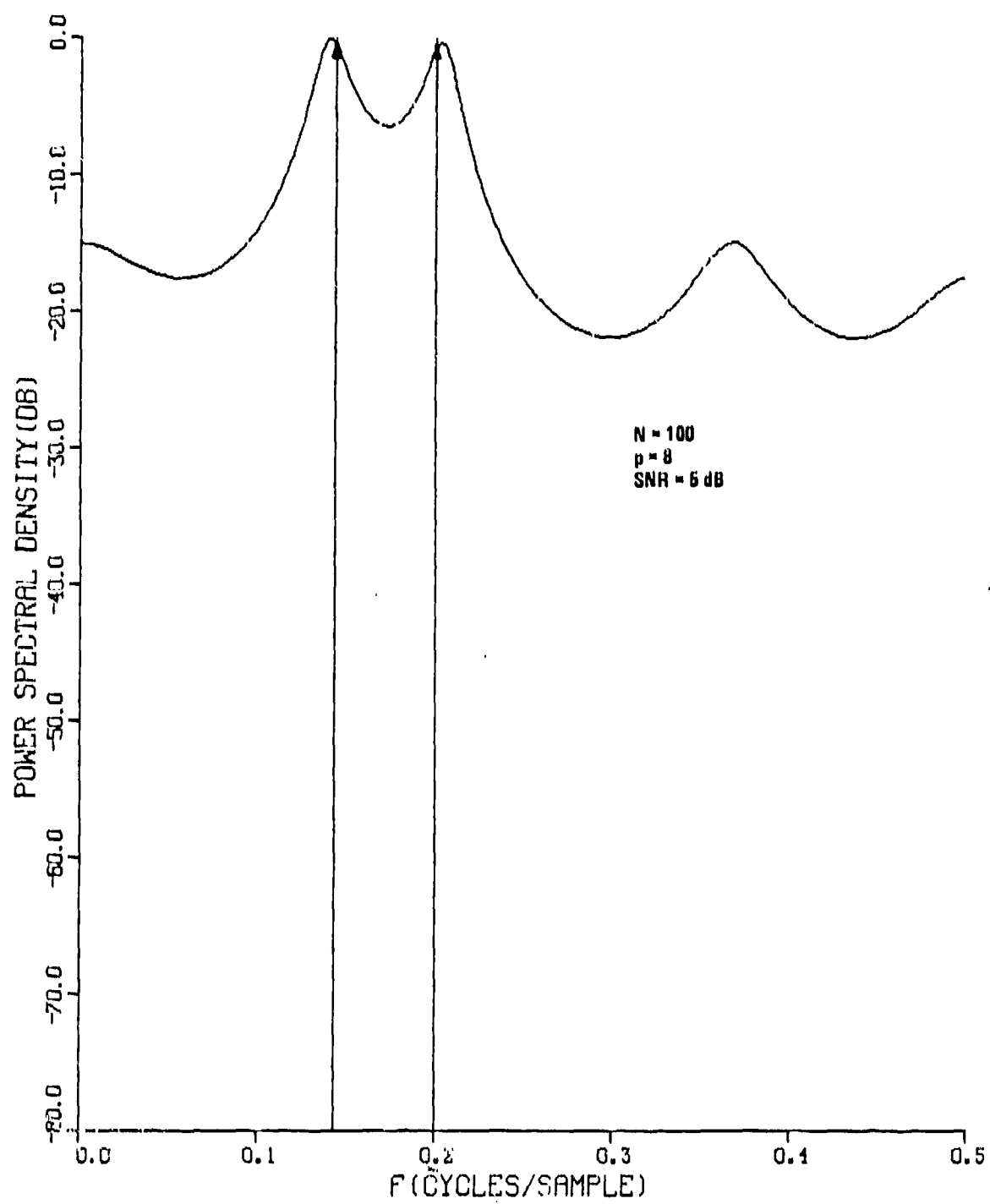


FIGURE 4. Burg Spectral Estimate of Two Sinusoids in White Noise

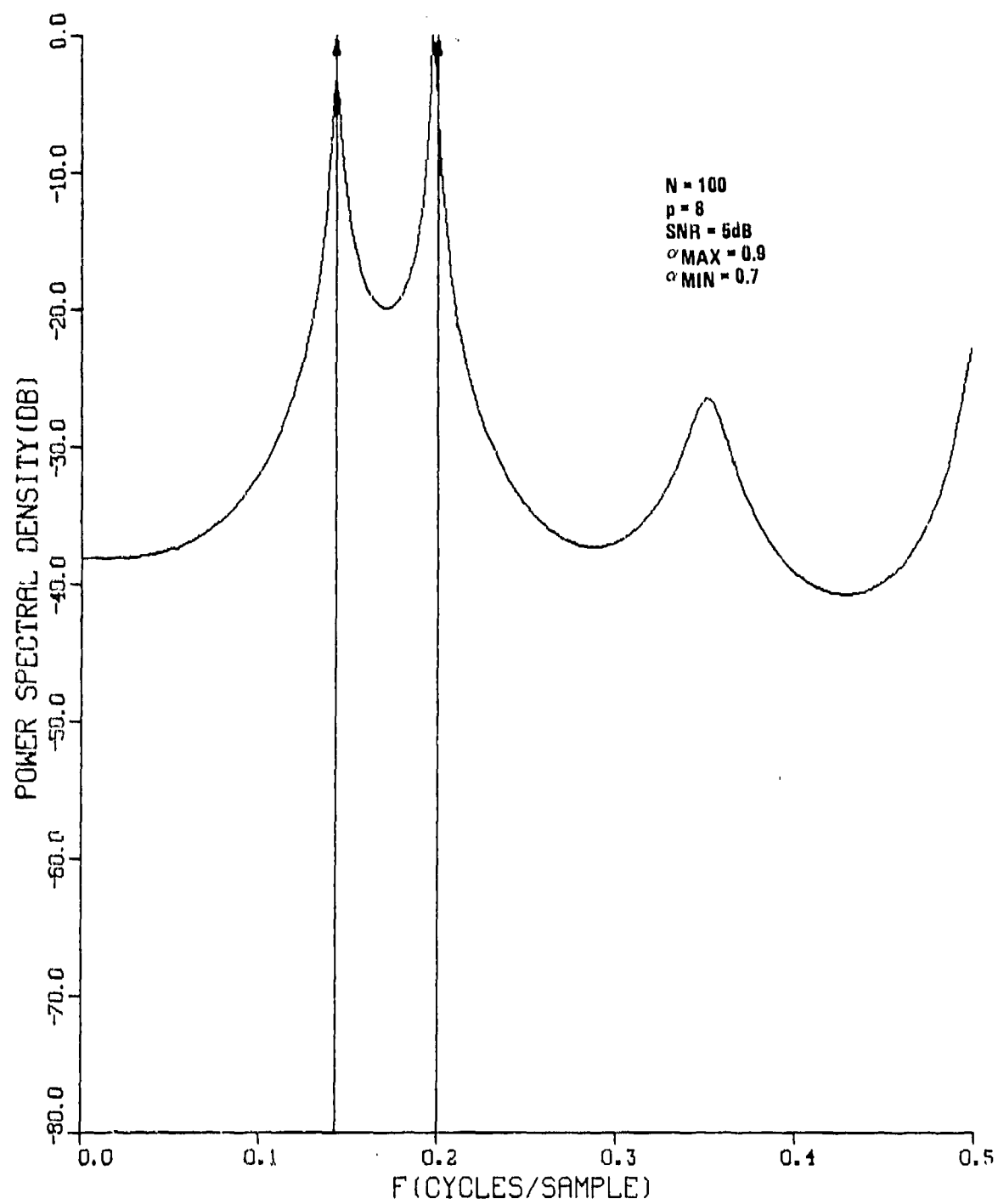


FIGURE 5. Noise Compensated Spectral Estimate of Two Sinusoids in White Noise

ORDER DETERMINATION FOR AUTOREGRESSIVE SPECTRAL ESTIMATION

M. KAVEH AND S. P. BRUZZONE

Department of Electrical Engineering
University of Minnesota
Minneapolis, MN 55455

Abstract

A method related to Akaike's Information Criterion for determining the order of an autoregressive (AR) spectral estimator is discussed. Examples are shown that compare the performance of this method with that of the well-known AIC.

Introduction

The trade-off between the bias (or resolution) and variance of a spectral estimator is the central issue in spectral estimation by any method. For the traditional (Blackman and Tukey type) spectral estimators, this trade-off is reflected in the choice of the spectral window type and the maximum lag of autocorrelation function used. This subject, referred to as window carpentering, is discussed in detail by Jenkins and Watts [1], and is straightforward because resolution is well-defined in terms of the spectral window bandwidth.

With the popularity of data adaptive (notably the autoregressive (AR)) spectral estimation methods, similar resolution-variance trade-offs are in order. Specifically, well-defined methods are needed to determine the order of the (AR) spectral estimator for a given data sample. Furthermore for practical applications, these methods need to be on-line and as much as possible objective in nature. This problem is complicated, however, due to the data dependent nature of the resolution of the AR spectral estimator (e.g., no well-defined window bandwidth). Therefore, the question of order determination for the spectral estimator seems to be best posed as a procedure for obtaining a compromise between the AR model fit and the variance of the estimated AR parameters as a function of the model order.

Akaike [2, 3] and Parzen [5] have recently introduced some methods for automatic determination of orders of autoregressive processes. One method, based on Akaike's Information Criterion (AIC), has gained special popularity. In this paper, we follow the derivations on which AIC is based, introduce appropriate modifications to account for practical estimation procedures and

derive a modified information criterion designated the Conditional AIC (CAIC). Finally we present the results of a number of numerical simulations that compare the performances of AIC and CAIC for spectral estimation.

Akaike's Information Criterion

Akaike derived his information criterion, AIC, as an estimate of the asymptotic relative goodness of fit of the model to the observation. Although his derivations were based on information theoretic arguments, the resulting parameters were the same as the maximum likelihood estimates. In this section we review the steps involved in obtaining AIC [3] as they pertain to the derivation of the new criterion. We assume the time series to be described by

$$x_t = \sum_{i=1}^L a_i x_{t-i} + u_t, \quad t=0, \dots, N \quad (1)$$

$$x_{-L}, \dots, x_0 = 0$$

where u_t is zero-mean white and Gaussian and L is to be determined. Through asymptotic arguments, Akaike defines an information criterion, related to the maximum likelihood of the estimates of a_i , $\hat{a}_i$, as:

$$\begin{aligned} \text{AIC}(\hat{\underline{A}}) &= (-2) \ln (\text{maximum likelihood}) \\ &+ E_{\infty} N \cdot ||\hat{\underline{A}} - \underline{A}||^2 \end{aligned} \quad (2)$$

where E_{∞} denotes asymptotic expectation, $\hat{\underline{A}}$ and $\underline{A}$ are $L \times 1$ vectors of the coefficients a_i and their estimates $\hat{a}_i$. The practical AIC which is related to the full-information likelihood function of a Gaussian process is then given by

$$\text{AIC}(L) = N \ln (\text{MLE of innovation variance}) + 2L \quad (3)$$

and the order L is chosen that minimizes $\text{AIC}(L)$.

The New Criterion

Since the exact maximum likelihood (full information maximum-likelihood) estimates are generally not available, the conditional MLE one based on Yule-Walker equations or Burg's algorithm, of the innovation variance are normally used in (3). We propose using the conditional maximum likelihood (CML) function in (2). This function is based exactly on the available data and we believe is a more sensitive indicator of the behavior of the estimates used in practice. Thus, in the following the CML estimate of $\underline{A}$ and its covariance

function are considered, in order to obtain tractable expressions for (2).

The conditional (partial information) likelihood function for the time series in (1) is given by:

$$L(\underline{A}, \sigma_u^2, x_1, \dots, x_L) = \frac{1}{(2\pi\sigma_u^2)^{\frac{N-L}{2}}} \exp \left[\frac{T}{2\sigma_u^2} \right] \quad (4)$$

where σ_u^2 is the variance of the innovation sequence u_t ,

$$\underline{C}^T = [1, -a_1, -a_2, \dots, -a_L] \text{ and} \quad (5)$$

$$D_{ij} = D_{ji} = \sum_{k=L+1}^N x_{k-i+1} x_{k-j+1}.$$

Furthermore, the CML estimation of σ_u^2 is given by:

$$\hat{\sigma}_u^2 = \hat{\underline{C}}^T \hat{\underline{D}} \hat{\underline{C}} / (N-L) \quad (6)$$

and a lower bound for the variance of the estimates of a_i follows from the Fisher's information matrix to be [6]

$$\text{var}[\hat{a}_i] \geq \frac{1}{N-L} \sigma_u^2 \underline{\Lambda}^{ii} \quad (7)$$

where $\underline{\Lambda}^{ii}$ is the diagonal element of the inverse of the $(L+1)$ -sample covariance matrix of x_t . It can also be shown [6] that for an AR model

$$\sigma_u^2 \underline{\Lambda}^{ii} = 1, \text{ thus} \quad (8)$$

$$\text{var}[\hat{a}_i] \geq \frac{1}{N-L}$$

We now proceed to define an expression for (2) based on the CML estimates of $\hat{A}$ and $\hat{\sigma}_u^2$. The expression for the CML given in (4) is now substituted in (2) for the maximum likelihood and using (8) for the second term in (2) and $(N - L)$ for N we have:

$$CAIC(L) = (N - L)\ln(2\pi\hat{\sigma}_u^2) + (\alpha - 1)L \quad (9)$$

The factor $\alpha \geq 1$ is included to account for the asymptotic nature of the criterion and the fact that (8) is a lower bound for the variance of a_i . A similar parameter was also suggested for AIC [7] and in [3] Akaike discusses a possible approach for choosing α . Since $CAIC(L)$ as given by (9) is dependent on the variance of x_t , the test is standardized by introducing a normalized innovation variance so that

$$CAIC(L) = (N - L)\ln[\hat{\sigma}_u^2/(\text{var } x_t)] + (\alpha - 1)L \quad (10)$$

Thus $CAIC(0) = 0$. The factor α is chosen to give more or less weight to the error in the estimation of the parameters. In other words, resolution can be increased at the expense of the variance of the estimates by decreasing α . We have found, empirically, values of 3.5-4 to give the most stable and reasonable indication of the order.

Simulation Results

We have tested the performance of CAIC relative to AIC on a number of time series models reported previously. The data included normal as well as uniform distributions. The estimates were based on CML (least-square) and Yule-Walker methods. In the great majority of cases, CAIC performed as well or superior to AIC. Examples of these can be found in [8]. Some estimated spectra based on orders determined by AIC and CAIC are also shown in Figures 1-3. Yule-Walker equations with autocorrelation function estimates given by

$$r_i = \frac{1}{N} \sum_{j=1}^{N-i} x_j x_{j+i}$$

were used. The example shown in Figure 1 indicates the relative stability of CAIC. Figure 2 shows that the model order chosen by AIC results in spurious peaks, while giving higher peak resolution than the CAIC based one. Figure 3 shows that an increase in white noise level increased the AIC order to the point that spurious spectral peaks became pronounced while CAIC remained nearly the same, showing the relative stability of CAIC.

Conclusions

Order determination for the AR spectral estimator was discussed. A new order indicator was introduced that is closely related to the AIC method of order determination for AR processes. The new criterion, CAIC, fits the practical estimation modes more closely and was found to be a relatively stable indicator of the order which trades off resolution and variance of the estimates.

References

1. Jenkins, G. M., Watts, D. G., 1968, "Spectral Analysis and its Applications", San Francisco, Holden-Day.
2. Akaike, H., 1970, "Statistical Predictor Identification", Annals of Inst. Stat. Math, Vol. 22, No. 2.
3. Akaike, H., 1974, "A New Look at Statistical Model Identification", IEEE Trans. on Automatic Control, Vol. AC-19, No. 6, December.
4. Akaike, H., 1977, "A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting", Research Memo No. 126, The Institute of Statistical Mathematics, November.
5. Parzen, E., 1975, "Multiple Time Series: Determining the Order of Approximating Autoregressive Schemes", Tech. Report No. 23, July, SUNY Buffalo Dept. Computer Science.
6. Kaveh, M., to be published, "Order Determination for Least-Squares Predictor Identification".
7. Bhansali, R. J., Downham, D. Y., 1977, "Some Properties of the Order of an Autoregressive Model Selected by a Generalization of Akaike's FPE Criterion", Biometrika, Vol. 64.
8. Kaveh, M., 1979, "A Modified Akaike Information Criterion", Proceedings of the 17th CDC, January.

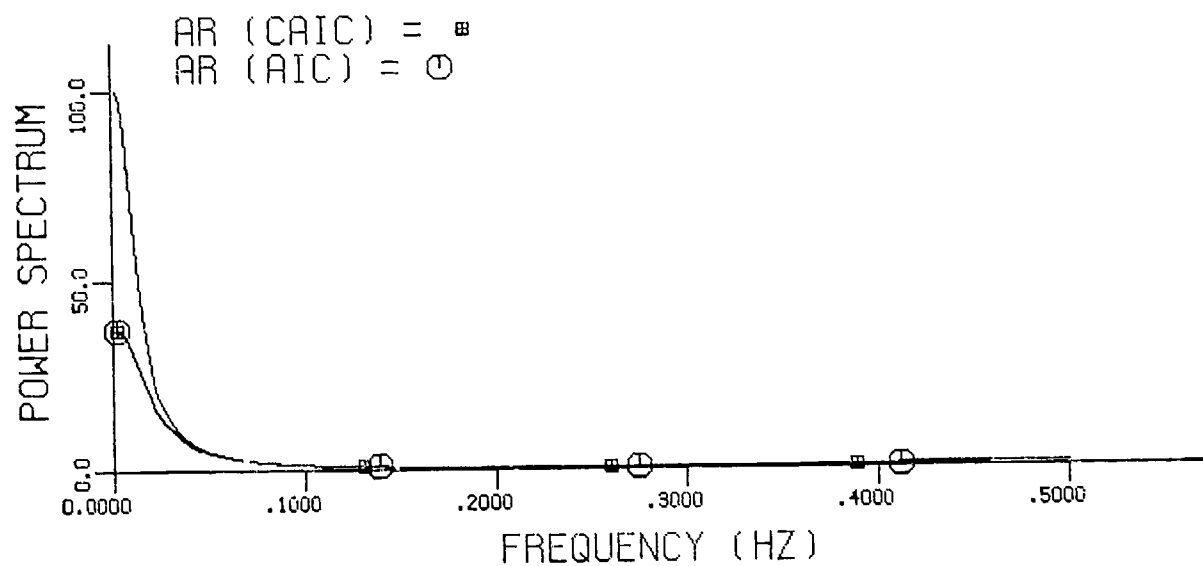


Fig. 1(a)

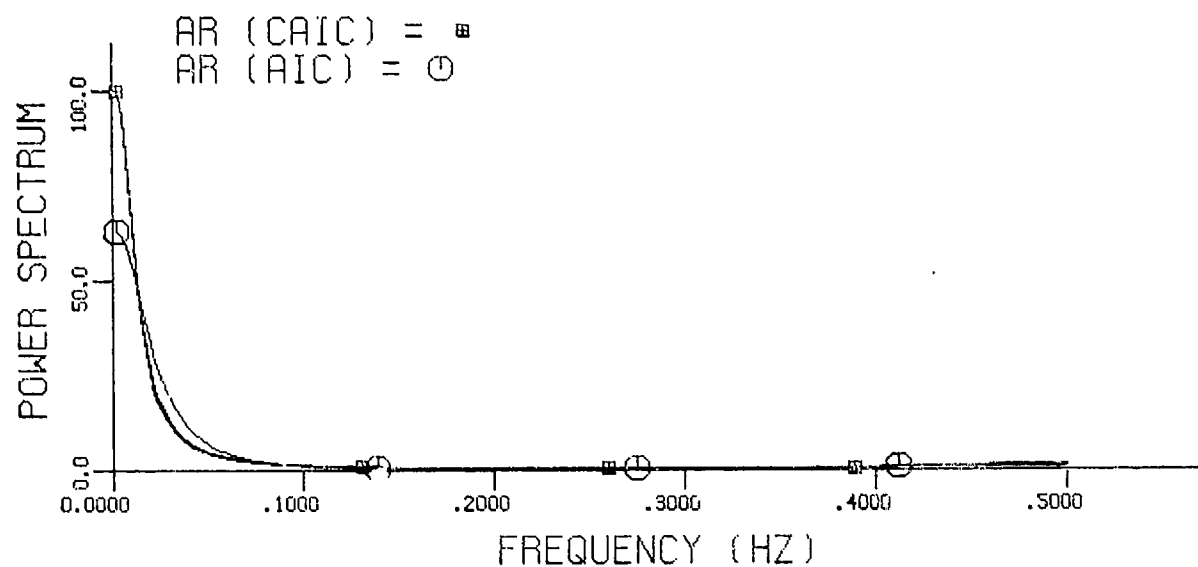


Fig. 1(b)

FIGURE 1. Power spectrum of $x_t = .4x_{t-1} + .5x_{t-2} + u_t$.

a) $N=100$, CAIC=2, AIC=2.

b) $N=500$, CAIC=2, AIC=8.

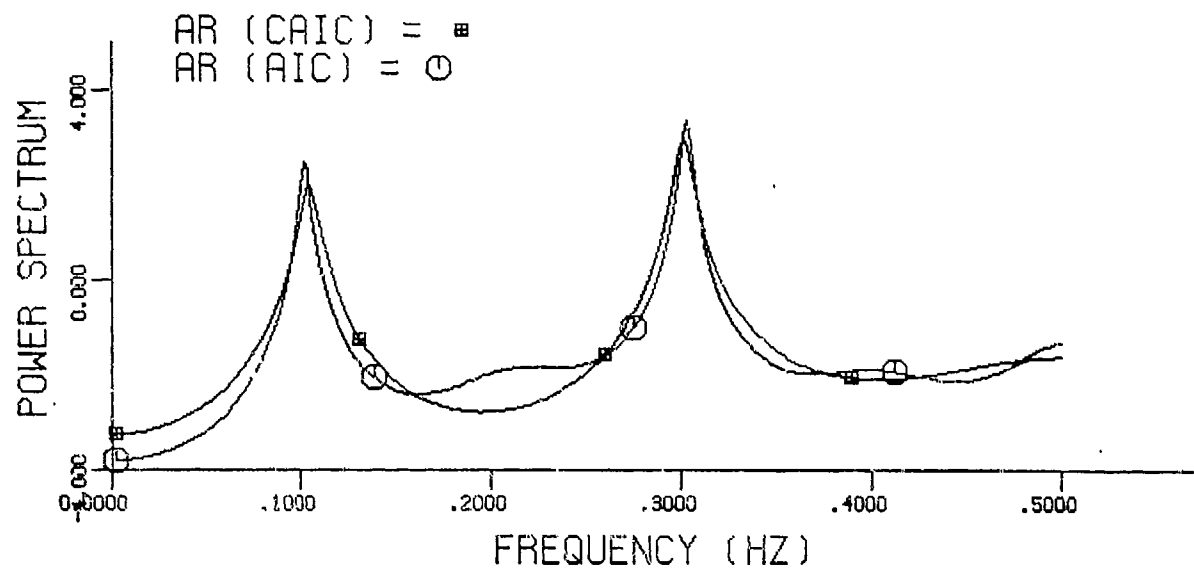


FIGURE 2. Log power spectrum of $x_t = \sin(.6\pi t + 30^\circ) + 0.707\sin(.2\pi t + 60^\circ) + n_t$, $\sigma_n^2 = .15$; $N=100$; CAIC=6, AIC=10.

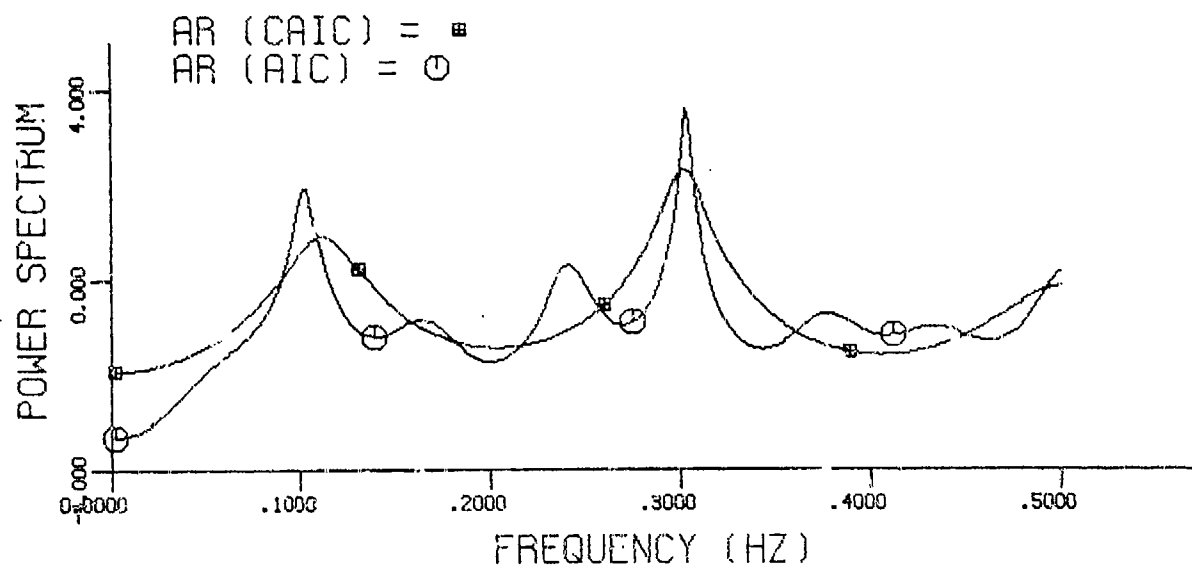


FIGURE 3. Log Power spectrum of $x_t = \sin(.6\pi t + 30^\circ) + 0.707\sin(.2\pi t + 60^\circ) + n_t$; $\sigma_n^2 = .316$, $N=100$; CAIC=5, AIC=15.

146-BL ARK

DIFFICULTIES PRESENT IN ALGORITHMS FOR DETERMINING THE RANK
AND PROPER POLES WITH PRONY'S METHOD

MICHAEL L. VAN BLARICUM

Effects Technology, Inc.
5383 Hollister Avenue
Santa Barbara, California 93111

Introduction

This author presented a paper [1] at the 1978 RADC Spectral Estimation Workshop which reviewed the algorithm of Prony. This Prony algorithm is used to extract a system's natural resonances and associated residues from transient response data. This procedure has great potential use in the analysis of transient electromagnetic response data such as those from EMC and EMP testing and from transient radar scattering.

Last year's paper addressed the algorithm and discussed the problem of noise and its effect on Prony's method with several solution methods being suggested. In addition a few examples were given. This paper will focus on the specific problem of the determination of the order of the system being analyzed. That is, it will address the question: How do you a priori determine how many poles are contained in the response data?

The process of extracting the natural resonances and their associated residues from a transient signal has four main steps as shown in Figure 1. The details of these steps were discussed in reference [1]. A brief review is given here.

The first step is the determination of the order of the system. At this step one decides how many poles the system response function has so that the proper model order can be obtained. It has been found, through trial and error, that if the order of the system is underestimated then the extracted poles will deviate substantially from the true poles. Similarly if the order of the system is overdetermined the algorithm produces extraneous poles. The presence of the extraneous poles causes the residues of the true poles to be inaccurate and also results in unnecessary computation time. In addition, as will be shown, attempting to solve an overdetermined system will result in an ill conditioned matrix and present numerical problems. The presence of noise in data makes the determination of the system order a very complex problem.

Once the order of the system has been determined the coefficients of Prony's difference equation must be solved. This step basically involves the solution of a linear matrix equation. The degree of difficulty of this step depends on the noise level in the data and on the proper determination of the order of the system. Generally this step is solved using a least squares method. However other solutions, such as recursive techniques, may be more applicable depending on the condition of the data.

Once the difference equation coefficients are obtained the roots of an N th order polynomial, N being the system order, must be found. Many root finding routines exist but Muller's method [2] appears to be the optimal method. While this is a key step in the procedure it is totally dependent on the accuracy of the coefficients which were obtained in the previous step.

The final step is the solution of the residues which are associated with the system poles or singularities. These residues are obtained by solving a simple linear matrix equation. Generally a least squared error (L_2 norm) solution is used. However, Schaubert [3] has shown that if the uniform norm (L_∞ norm) solution is found the accuracy of the residues is much better. The L_∞ norm solution does require the solution of a nonlinear problem and hence requires more computation time. In many problems, such as target identification, the residues are not even required and hence this is certainly not a critical step.

It is this author's opinion that the first step, the determination of the system order, is the key step in the total procedure. Up to this time many methods have been used to determine the order but they either break down when noise is present or they are dependent on trial and error or the intervention of the user. For analysis of massive amounts of data, as in the case of EMP data, or for radar target identification, a totally automated method is a must. The remainder of this paper will discuss the problem of order determination. Some specific techniques will be discussed and numerical examples will be presented. Finally specific recommendations as to the direction of future research in this area will be made.

Mathematical Preliminaries

In order to discuss the different methods for determining the order of a system response before attempting to extract the resonances, it is necessary to review some of the details of the Prony algorithm. The details of Prony's method can be found in [1]. The notation used in [1] will be followed here for easy reference.

The premise used in applying Prony's method is that the system to be modeled can be represented by a finite sum of complex exponentials as

$$R(t_n) = R_n = \sum_{i=1}^N A_i e^{S_i n \Delta t}, \quad n=0,1,\dots,M-1 \quad (1)$$

where $R(t_n)$ is the system response, the S_i are the complex poles and the A_i are the corresponding residues. Expression (1) is written in discrete data form where it has been assumed that the M samples are taken at equal time increments Δt . Equation (1) is M nonlinear equations in $2N$ unknown.

The solution of the S_i and the A_i in (1) is based on the fact that the R_n must satisfy a difference equation of order N which may be written as

$$\sum_{p=0}^N \alpha_p R_{p+k} = 0, \quad k = 0, 1, \dots, \gamma-1, \quad (2)$$

where γ is the value of $M-N$. This difference equation is referred to as Prony's difference equation. Equation (2) is usually rewritten by defining α_N equal to 1 so that the equation has the form

$$\sum_{p=0}^N \alpha_p R_{p+k} = -R_{N-k}. \quad (3)$$

If $2N$ data samples are used, then equation (3) can be solved exactly for the α 's. If more than $2N$ samples are desired, then one can use a least squared error fit to (3). It is the solution of this equation which is the second step of Figure 1 and which yields the coefficients, α_p , for the polynomial of the third step. For convenience equations (2) and (3) can be rewritten in matrix notation as

$$A x = b. \quad (4)$$

The matrix A is filled with the discrete response samples R_{p+k} and is either of dimension γ by $N+1$ for equation (2) or γ by N for equation (3). For equation (2) the vector x is of length $N+1$ and contains the unknown coefficients α and the vector b is equal to zero. For equation (3) the vector x is of length N and the vector b is of length γ .

In order to obtain a proper solution to either equation (2) or (3) it is necessary to know the value of N . This is equivalent to saying that for the matrix problem defined in (4) we need to know the rank of the matrix A . If N is picked too small then the matrix equation is underdetermined and will give wrong answers. If N is picked too large then the matrix A is rank

deficient and will likely cause the matrix to be ill conditioned.

The Eigenvalue Method

This author [4] has presented a method for determining the rank of the matrix A based on its eigenvalues. The details can be found in reference [4]. The basic theory is that the response, R_n , of a system which contains exactly N resonances will satisfy the difference equation (2) exactly. Another way of looking at the problem is that there are exactly N mode vectors for an N^{th} order system where the mode vector X_j is defined as

$$X_j = \begin{bmatrix} z_j^0 \\ z_j^1 \\ z_j^2 \\ \vdots \\ z_j^Y \end{bmatrix} \quad (5)$$

where $z_j = e^{s_j \Delta t}$. It can be shown that the matrix A is made up of a linear combination of these mode vectors. For the least squares or pseudo-inverse solution a square, $(N+1)$ by $(N+1)$ for equation (2), matrix Φ is defined as

$$\Phi = A^T A$$

This square matrix Φ will have $N+1$ eigenvectors and associated eigenvalues. The N mode vectors defined by (5) will be linearly independent and will have projections on N of the eigenvectors of the system. There will be one eigenvector which is orthogonal to the N mode vectors and its eigenvalue will be zero. Hence, the process for determining the order of the system is to fill matrix Φ to some dimension M by M . The corresponding M eigenvalues of the system are found and checked to see if one or more are equal to zero. If there are L eigenvalues equal to zero, then N would be equal to $M-L$. If L is not equal to one, then the matrix Φ is recomputed to order $N+1$ by $N+1$, and the eigenvalues are regenerated. This final step is done because the eigenvector corresponding to the zero eigenvalue is the vector of coefficients of the difference equation.

This procedure works fine as long as the system response data does not contain noise. With noisy data the theory starts to fail because the system is no longer exactly the sum of N exponentials. Reference [4] shows that for small levels of noise with variance σ^2 the $N+1$ eigenvalue for the Φ matrix should be equal to $\gamma\sigma^2$ instead of zero.

Eigenvalue Examples and Difficulties

Figure 2 shows an example of the thirteen eigenvalues for a twelve pole system. Note that for the no noise case, the thirteenth eigenvalue is less than 10^{-10} or practically zero. For the noise case the thirteenth eigenvalue is equal to about 3×10^{-3} . The value of γ_0^2 for this case is 2.5×10^{-3} . From the example of Figure 2 it appears that there should be no difficulty in deciding what the proper order of the system is even in noisy data. This is theoretically the case. However the following example shows some inherent difficulties.

Table 1 presents the results of twenty Monte Carlo trials performed for seven different sets of noisy data. The signal to noise ratio ranged from 3.3dB to 35.4dB. Note that while the predicted value of the $N+1$ eigenvalue was very close to the mean of the actual numerically determined value, the standard deviation of the poles is very large when the noise level is high. Hence, while the proper order was determined, the poles resulting from this model order were in error. Experience has shown that picking a higher order model will give better accuracy to the resulting true poles. However there will be extraneous poles also present which, if nothing is known about the system, will potentially be hard to distinguish from the true poles. Hence while the proper rank of the Φ matrix has been found the resulting answer will not necessarily be that which is sought. In addition further examples have shown that as the noise level increases it is difficult to determine what is the exact cutoff point. That is, the difference between the N and the $N+1$ eigenvalue is so small that a true cutoff point is difficult to determine. This problem is shown graphically in Figure 3.

Preliminary efforts have shown that the use of recursive techniques to solve the matrix equation once the proper rank has been determined reduces the large errors described above and shown in the example of Table 1. The determination of the proper rank is still a problem when faced with the difficulty illuminated in Figure 3.

The HFTI Method

Another approach to the determination of the rank of the system of equation (3), where the A matrix is γ by N , is given by the HFTI algorithm [5]. The HFTI algorithm, described in detail in [5], is designed to specifically to give a least squares solution to a rank-deficient problem by using Householder transformations. The method transforms the matrix-vector combination $[A:b]$ of (4) to a matrix-vector combination $[R:c]$ using premultiplying Householder transformations with column interchanges. All subdiagonal elements in the matrix R are zero and its diagonal elements satisfy $|r_{ii}| \geq |r_{i+1,i}|$, $i = 1, \dots, \ell - 1$, where $\ell = \min[\gamma, N]$. The proper rank of the system can be found by comparing the diagonal values, $|r_{ii}|$, and looking for a cutoff as was done for the eigenvalues of the previous section. The

subroutine HFTI will accept an input parameter τ which it compares with the diagonal elements to determine a new pseudorank k for the system. The pseudorank k is defined as:

$$k = i, \text{ when } |r_{ii}| > \tau \text{ and } |r_{i+1,i+1}| < \tau .$$

The method then calculates a minimal length solution vector x for the problem defined by the first k rows of $[R:c]$.

Figure 4 shows an example of the resulting diagonal elements $|r_{ii}|$. HFTI was used on a fourth order system presented by a rank deficient matrix of dimensions 20 by 6. The figure shows the results for both clean and noisy response data. The problem here is the same as in the eigenvalue method. That is, it is difficult to know what value of τ should be specified so that the proper rank of the system can be determined.

Summary and Conclusions

The two previous sections briefly discussed two methods for determining the order or rank of a system so that the proper poles can be extracted. While the methods are basically straight forward the actual application of the method is limited if noise is present in the data. The problem which now must be addressed is how does one determine the optimum cutoff point for the eigenvalue method or the τ parameter for the HFTI method. It appears that numerical parameter studies need to be run to get a better feel for this procedure. Preliminary results bode optimistically on obtaining an optimal cutoff procedure so that the entire resonance extraction procedure can be automated.

References

1. Michael L. Van Blaricum, "A Review of Prony's Method Techniques for Parameter Estimation," 1978 RADC Spectral Estimation Workshop.
2. J.D. Lawrence, "Polynomial Root Finder," Lawrence Livermore Laboratory, Livermore, CA, CIC Note C212-010, January 1966.
3. D.H. Schaubert, "Application of Prony's Method to Time Domain Reflectometer Data and Equivalent Circuit Synthesis," HDL-TR-1857, Harry Diamond Labs, June 1978.
4. M.L. Van Blaricum and R. Mittra, "Problems and Solutions Associated with Prony's Method for Processing Transient Data," IEEE Trans. on Ant. and Prop., January 1978, pp. 174-182.

5. C.L. Lawson and R.J. Hanson, Solving Least Squares Problems, Prentice Hall.

TABLE 1. Results for Example 2:
 $R(t) = e^{-2t} \sin(\pi t)$, $\gamma = 50$, $\Delta t = 0.1$ S

Standard Deviation of Noise σ	Standard Deviation of Pole		Theoretical N + 1 Eigenvalue	Mean Value of N + 1 Eigenvalue	Signal to Noise Level dB
	Real Part	Imaginary Part			
0.04	0.800	0.566	0.08	0.0836	3.3
0.03	0.566	0.311	0.045	0.047	5.8
0.02	0.371	0.179	0.02	0.021	9.4
0.01	0.187	0.085	0.005	0.0052	15.4
0.009	0.167	0.077	0.00405	0.00423	16.3
0.005	0.094	0.043	0.00125	0.00131	21.4
0.001	0.017	0.009	0.5×10^{-4}	0.52×10^{-4}	35.4

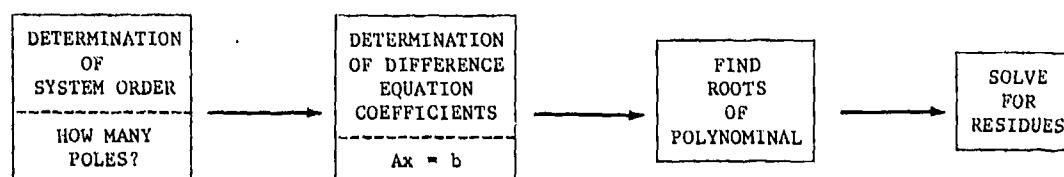


FIGURE 1. The Elements of the Extraction of Natural Resonances and Associated Residues from a Transient Signal.

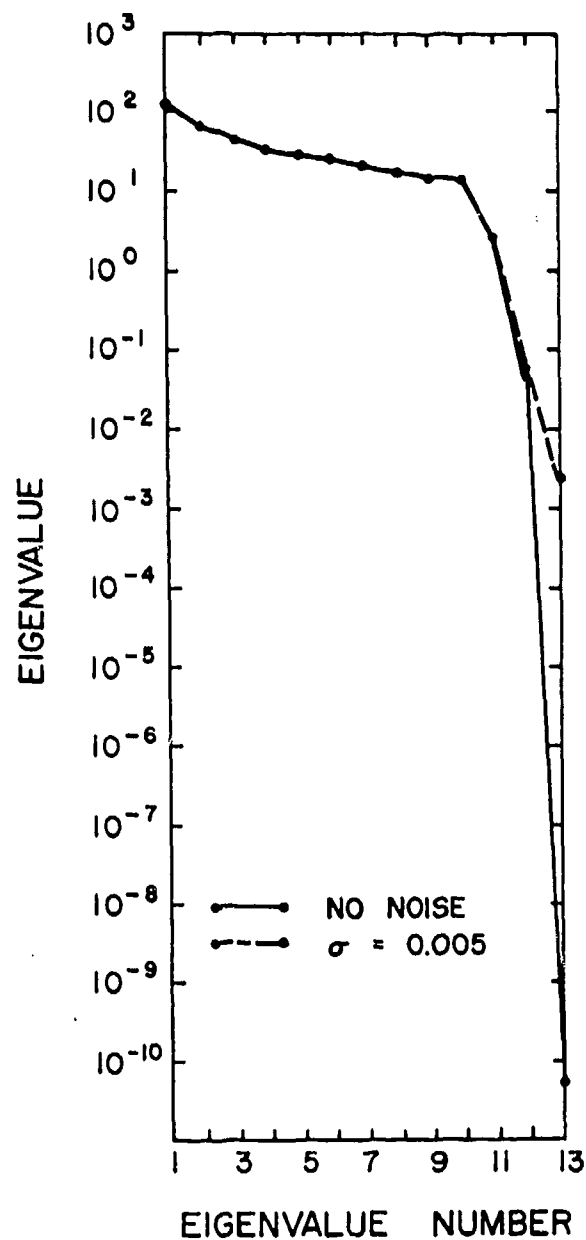


FIGURE 2. Resulting Eigenvalue from Response Generated Using 12 Poles for Non-Noise Case and Noise with $\sigma=0.005$, $\gamma=100$.

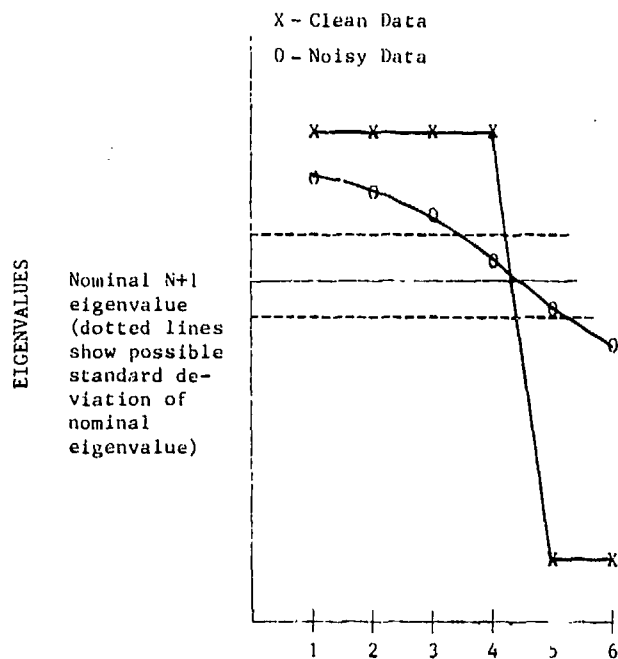


FIGURE 3. Example Showing the Lack of a Clean Break-point or Knee for a Fourth Order Noisy Data Response as Compared to the Clean Data Response.

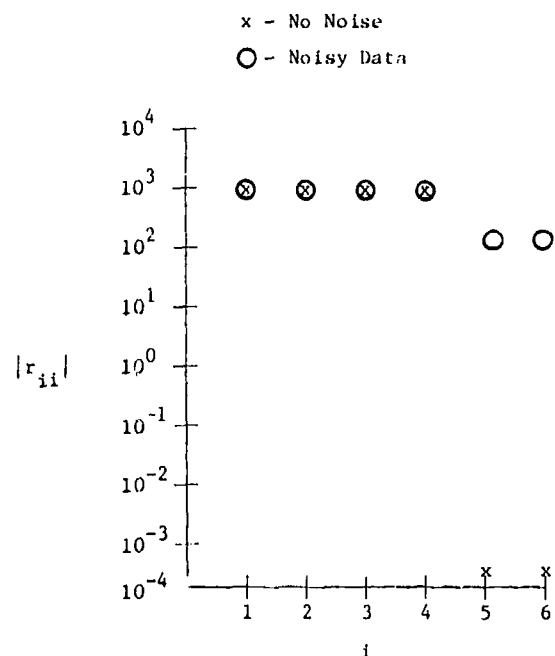


FIGURE 4. Results of $|r_{ii}|$ from HFTI using a Four Pole System with and without Noise. The Matrix was Rank Deficient at Dimensions 20 by 6.

156 - Blank

A UNIFYING MODEL FOR SPECTRAL ESTIMATION

CHARLES BYRNE

The Catholic University of America
Washington, D.C. 20064

and

RAYMOND FITZGERALD

Naval Research Laboratory
Washington, D.C. 20375

Abstract

The Fourier transform of a band-limited time function cannot be determined from finitely many observations. For this reason spectral estimation necessarily involves the substitution, for the original Fourier transform, of a function that is so determined. Some approaches make explicit the nature of the substitution, by assuming that the transform has a fairly simple form (a rational function, for example), whose parameters can be computed from the data. With other methods, such as those that rely on time-domain extrapolation or on iterative approximations, it is not always clear what is the substitution that the method is introducing.

In this article we discuss an explicit model, or substitute, for the Fourier transform, based on over-sampled data. Our optimal windowing or modified DFT model is seen to coincide with the minimum energy estimate of de Figueiredo. Upon examining the methods of Cadzow and of Kolba and Parks, we find this same model implicit in both techniques.

Introduction

According to the sampling theorem, a band-limited signal of finite energy can be reconstructed from uniformly spaced samples, provided the interval between sample times is small enough. Indeed, if $x(t)$ is a complex function of a real variable t , of finite energy, and its Fourier transform,

$$X(w) = \int_{-\infty}^{\infty} x(t) e^{iwt} dt, \quad (1)$$

is zero for w not in the band $W = [-\sigma, \sigma]$, then we have that

$$X(w) = \Delta \sum_{-\infty}^{\infty} x(n\Delta) e^{in\Delta w}, \quad |w| \leq \pi/\Delta, \quad (2)$$

for any sampling rate $0 < \Delta \leq \pi/\sigma$. Taking the inverse Fourier transform, we get, for all t ,

$$x(t) = \sum_{-\infty}^{\infty} x(n\Delta) [\sin(n\Delta\sigma - t\sigma)] / (n\pi - t\pi/\Delta) \quad (3)$$

In practice, because our observations are necessarily limited to some finite interval of time, say $[-d, d]$, only finitely many of the Fourier coefficients, $x(n\Delta)$, will be known. As has been noted by Levi [4], an arbitrarily band-limited signal can be made to pass through any finite set of points. Consequently, any procedure that estimates $X(w)$ from finitely many observations introduces a substitute, or model, for $X(w)$ that is determined by the data, or what is the same thing, makes additional assumptions about $X(w)$ which make it possible to reconstruct $X(w)$ from our limited knowledge. These models for the Fourier transform to be estimated are, at times, explicitly described, as in [2], where de Figueiredo uses the criterion of minimum energy to derive a model using splines. Other explicit models appear in the host of articles on rational approximation and ARMA schemes. With other approaches, such as those that rely on time-domain extrapolation or iterative approximation, the exact nature of the model being introduced is not always clear. Nevertheless, the operative model is the essence of the procedure and must be clearly understood before that procedure can be adequately compared with other methods.

In this article we propose an explicit model for the estimation of the Fourier transform from over-sampled data. We derive it as a best approximation to the unknown Fourier transform, and then note that it can also be derived as a minimum energy estimate, as was done in [2]. One can also view this model as a windowing procedure, that uses coefficients that are optimal in the sense described below, and are dependent upon the data. We then turn to a study of implicit models, dealing specifically with the recent work of Cadzow [1] and of Kolba and Parks[3]. As we discover, the extrapolation suggested by both of these methods can be by-passed, and if this is done, the operative model becomes the same optimal windowing presented earlier. It is in this sense that our model is described as "unifying".

An Optimal Windowing Model

Suppose that our data consists of the values $x(-M\Delta), \dots, x(M\Delta)$, where $0 < \Delta \leq \pi/\sigma$. The truncated DFT model for $X(w)$ is

$$X(w) = \Delta \sum_{-M}^M x(j\Delta) e^{ij\Delta w}, \quad |w| \leq \sigma. \quad (4)$$

This model is relatively easy to compute as an FFT and is optimal, in the sense that it is the polynomial of the form

$$\Delta \sum_{-M}^M a(j\Delta) e^{ij\Delta w} \quad (5)$$

that minimizes the mean square error

$$\int_{-\pi/\Delta}^{\pi/\Delta} |X(w) - \Delta \sum_{-M}^M a(j\Delta) e^{ij\Delta w}|^2 dw. \quad (6)$$

However, this model can be improved, if we know that the data is over-sampled, and that $X(w)$ is zero off the band $W = [-\sigma, \sigma]$, where $\sigma < \pi/\Delta$. It is reasonable then to minimize not (6), but

$$\int_{-\sigma}^{\sigma} |X(w) - \Delta \sum_{-M}^M a(j\Delta) e^{ij\Delta w}|^2 dw. \quad (7)$$

According to the orthogonality principle (see, for example, [5], p. 197), we must then have

$$\int_{-\sigma}^{\sigma} (X(w) - \Delta \sum_{-M}^M a(j\Delta) e^{ij\Delta w}) (e^{-ik\Delta w}) dw = 0 \quad (8)$$

for $k = -M, \dots, M$. It follows that the $a(j\Delta)$ must satisfy the system of linear equations

$$x(k\Delta) = \Delta \sum_{-M}^M a(j\Delta) \sin((k-j)\Delta\sigma)/(k-j)\Delta\pi, \quad (9)$$

for $k = -M, \dots, M$. It is a happy result that these optimal coefficients are completely determined by the data. When the model for $X(w)$ is (5), with coefficients obtained from (9), we shall say that the model is the modified DFT (MDFT). This MDFT model can be viewed as a data-dependent windowing, with optimal coefficients.

In the recent article [2] de Figueiredo presents a spline approximation technique for spectral estimation that employs a minimum energy criterion. It is easily verified that our MDFT model coincides with his, and so has minimum energy, subject to the data and the band W. The reader is also directed to the article [4], where the same model is discussed in a somewhat different context.

Two Extrapolation Methods and their Implicit Models

In [1] Cadzow suggests that we by-pass his sometimes slowly converging iterative procedure for estimating the Fourier transform and adopt a one-step extrapolation method. Involved in this method is the assumption that there is a function $z(t)$, of finite energy, satisfying the integral equation

$$x(t) = \int_{-d}^d z(s) \sin((t-s)\sigma) / (t-s)\pi \, ds, \quad (10)$$

whose Fourier transform agrees with $X(w)$ on the band W. Knowing only $x(-M\Delta), \dots, x(M\Delta)$, we approximate the integral equation by a system of linear equations

$$x(k\Delta) = \Delta \sum_{-M}^M z(j\Delta) \sin((k-j)\Delta\sigma) / (k-j)\Delta\pi, \quad (11)$$

for $k = -M, \dots, M$. Once we have the values of $z(j\Delta)$, we can extrapolate $x(q\Delta)$, $|q| > M$, using (11), with $q\Delta$ replacing $k\Delta$. The Fourier transform is estimated from a sufficiently expanded set of samples and extrapolations. However, we may proceed somewhat differently. Having computed the solution of (11), we are able to write the $X(w)$ in closed form; using (2) and (11) and rearranging terms we obtain, with $h(t) = (\sin(\sigma t)) / \pi t$,

$$X(w) = \Delta \sum_{-M}^M z(j\Delta) \left[\Delta \sum_{-\infty}^{\infty} h(n\Delta - j\Delta) e^{i(n-j)\Delta w} \right] e^{ij\Delta w}, \quad (12)$$

where the term in square brackets is the Fourier expansion of the function that is 1 on W and 0 off of W. So the operative model is the MDFT. Of course, if the extrapolation is used, the MDFT model is only approximated, due to the truncation that necessarily results.

Let us turn to the extrapolation method of Kolba and Parks [3]. Consider equation (3), with $t = k\Delta$. We then have, for each

fixed value of k ,

$$x(k\Delta) = \sum_{n=-\infty}^{\infty} x(n\Delta) \sin((k\Delta - n\Delta)\sigma) / (k - n)\pi \quad (13)$$

Viewing (13) as the inner product of two sequence, $\{x(n\Delta)\}$ representing $X(w)$, and $\{\sin((k\Delta - n\Delta)\sigma) / (k - n)\pi\}$, we see that the second sequence represents the linear functional that extracts the $k\Delta$ -th coefficient. Knowing only $x(k\Delta)$, $k = -M, \dots, M$ suggests that we approximate f_m , $|m| > M$, by a linear combination of the f_k , $|k| \leq M$, labeled f'_m . So

$$f'_m = \sum_{j=-M}^M b_{mj} f_j \quad (14)$$

We seek optimal coefficients, so applying the orthogonality principle, we get

$$\sin((m\Delta - k\Delta)\sigma) / (m\Delta - k\Delta)\pi = \sum_{j=-M}^M b_{mj} \sin((j\Delta - k\Delta)\sigma) / (j\Delta - k\Delta)\pi \quad (15)$$

for each $k = -M, \dots, M$. If we let b_m be the resulting solution vector, the best extrapolation of our data is then given by

$$x(m\Delta) = b_m \cdot y, \quad |m| > M \quad (16)$$

where y is the vector containing the data, $x(k\Delta)$, $|k| \leq M$. This is the extrapolation method due to Kolba and Parks [3]. It is an easy exercise to show that it is the same extrapolation obtainable from the MDFT model. The extrapolation can, as before, be by-passed, and $X(w)$ be obtained directly. The model is once again the MDFT.

Conclusion

The MDFT model presented here is an optimal data-dependent windowing, as well as the minimum energy estimate. It is also the implicit model being used in the one-step method of Cadzow, and in the best linear extrapolator procedure of Kolba and Parks. By uncovering the implicit model we are able to provide further justification for the use of each of these methods, as well as to make unnecessary the comparisons of these methods that focus needlessly on superficial differences in development, while

ignoring the common model.

It is also possible to derive the MDFT model using a maximum entropy criterion, or as a moving average. The details will be presented elsewhere.

References

1. Cadzow, J.A., "An Extrapolation Procedure for Band-limited Signals," IEEE Transactions ASSP-27, 4-12 (1979).
2. de Figueiredo, R. "Optimal Estimation of Essentially and Strictly Band-limited Signals and their Spectrum by Generalized Splines," IEEE ICASSP, April 1979, p. 194-199.
3. Kolba, D.P., and Parks, T.W., "Extrapolation and Spectral Estimation for Band-limited Signals," IEEE ICASSP, April 1978 p. 372-374.
4. Levi, L., "Fitting a Band-limited Signal to Given Points," IEEE Transactions on Information Theory, Volume IT-11, July 1965, p. 372-376.
5. Papoulis, A., Signal Analysis, McGraw-Hill, N.Y., 1977.

A COMPARISON OF THE BURG AND THE KNOWN-AUTOCORRELATION
AUTOREGRESSIVE SPECTRAL ANALYSIS OF COMPLEX SINUSOIDAL
SIGNALS IN ADDITIVE WHITE NOISE

Robert W. Herring

Communications Research Centre
Shirley Bay
P.O. Box 11490, Station "H"
Ottawa, Ontario K2H 8S2
CANADA

Abstract

Burg's algorithm for Maximum Entropy autoregressive spectral estimation is analyzed for the cases of one and two complex sinusoidal signals in additive white noise. For the latter case are found two biases which can account for the line splitting and line shifting that occur in simulation studies when the SNR is very high. These biases vanish completely if the two complex sinusoids are in phase quadrature at the middle of the data record; if there is an integral number of half-cycles of difference frequency contained in the data record, then the spectral estimate will be biased although the effects believed to cause line-splitting will be eliminated. Results of simulation studies to support these conjectures are presented.

1. Introduction

The Burg algorithm for the autoregressive spectral analysis of time-series data [1,2,3], sometimes referred to as the maximum-entropy method (MEM), is known to be inappropriate for the case of sinusoidal signals in additive white noise. This inappropriateness had been demonstrated both theoretically [4-6] and in practice [7-9]. A theoretically correct model [4-6] for the generating process for an N-pole complex sinusoidal signal in additive white noise is an N-pole, N-zero network with identical gain weights in its feedback (pole) and feedforward (zero) parts, being excited with a white-noise input (see Fig. 1). Autoregressive analysis models the generating process for the data as an all-pole network excited by white noise. Since a zero in the generating process network can be simulated exactly only by an infinite number of poles, it is clear that when autoregressive analysis is used, in principle an infinite set of either autocorrelation or time series data must be used in order to achieve correct results.

Fougere [9] has stated that, in the high signal-to-noise ratio (SNR) case, the Burg algorithm is overconstrained. In the case of simulation studies, this overconstraint causes errors in the estimated frequencies of spectral lines and the false splitting of spectral lines known to have been generated by a single pole (or a pair of poles, in the case of real signals). Fougere has developed an algorithm which avoids this phenomenon, but the algorithm is based on a gradient-search technique which lacks the intrinsic efficiency of the unmodified Burg algorithm.

In spite of its known limitations, the Burg algorithm is often used because of its computational efficiency. This report explores analytically the expected response of the Burg algorithm to time-series data comprising one or two complex sinusoids, with and without the presence of additive white noise. It is shown that only in very special cases does the Burg algorithm lead to the same results as are achieved when the true autocorrelation functions of the signals are known.

2. Review of Autoregressive Spectral Analysis

Autoregressive spectral analysis is based on the idea that, if it is somehow possible to design a feedforward (all-zero) filter which has as its input the data to be analyzed, and has as its output random white noise, then the power spectrum of the input data is given by the reciprocal of the power transfer function of the filter. Since this filter accounts for all the predictability inherent in the input signal and has as its output only unpredictable random white noise, it is often referred to as a prediction-error filter (PEF).

There are several well-known techniques for estimating the PEF corresponding to a given set of data. When only amplitude time-series data are available, most of these depend upon estimates of the autocorrelation function derived from the time-series data. The Burg algorithm, however, attempts to avoid possible biases or inconsistencies in such estimates of the autocorrelation function by deriving an estimate of the PEF coefficients directly from the data.

In this report the algorithm for generating the PEF when the true autocorrelation function is known is reviewed in Section 2.1 and the Burg algorithm is reviewed in Section 2.2. The results from these sections are then used to generate the sets of PEFs corresponding to the cases of one and two complex sinusoids in additive white noise, and the properties of these sets of PEFs are compared and contrasted.

2.1 The Known-Autocorrelation (KA) Algorithm [10,11]

Let it be assumed that N equispaced samples $R(n)$ of the complex autocorrelation function have been given, for $n=0,1,2,\dots,N-1$, where N may be finite or infinite. It is assumed that the Nyquist sampling criterion has been met. Then the system of equations to be solved is

$$-\sum_{m=0}^M R(k-m)\alpha(m,M) = P(M)\delta(k) \quad \begin{matrix} 0 \leq k \leq M \\ 0 \leq M \leq N-1 \end{matrix} \quad (1)$$

where the $\alpha(m,M)$ s for $m=0,1,2,\dots,M$ are sets of PEF coefficients, each value for M denoting a different set; the $P(M)$ s are called the output error powers and are real; and $\delta(k)=1$ if $k=0$, $\delta(k)=0$ otherwise, is the Kronecker delta function. In order to maintain proper scaling, the leading terms of the PEFs, $\alpha(0,M)$, $M=0,1,\dots,N-1$, are set equal to -1 by definition.

Since negative indices for $R(k-m)$ occur in the set of equations (1), it is necessary to note that $R(-n)=R^*(n)$, where the asterisk (*) denotes complex conjugation. This last fact allows (1) to be re-written in the alternative form:

$$-\sum_{m=0}^M R(k'+m)\alpha^*(m,M) = P(M)\delta(k') \quad \begin{matrix} -M \leq k' \leq 0 \\ 0 \leq M \leq N-1 \end{matrix} \quad (2)$$

Equations (2) imply that the same result is obtained if the complex conjugate of a PEF is applied to the time-reversed autocorrelation data. This "reverse-conjugate" symmetry is used in the derivation of the Burg algorithm.

If the set of linear equations (1) for a particular value of M is written in matrix form, it can be seen that the matrix of autocorrelation samples $[R(k,m)]$, where $R(k,m)=R(k-m)$, is an $M \times M$ Toeplitz matrix. Therefore (1) can be solved by applying the algorithm developed by Levinson, Robinson and Durbin, otherwise known as the Levinson recursion. Following [11] the recursive solution to (1) can be written as:

$$P(0) = R(0) \quad (3a)$$

$$\alpha(M,M) = -\sum_{m=0}^{M-1} \alpha(m,M-1)R(M-m)/P(M-1) \quad (3b)$$

$$\alpha(m,M) = \alpha(m,M-1) - \alpha(M,M)\alpha^*(M-m,M-1) \quad (3c)$$

$$P(M) = (1 - |\alpha(M,M)|^2) P(M-1) \quad \begin{matrix} m = 1, 2, \dots, M-1 \\ M = 1, 2, \dots \end{matrix} \quad (3d)$$

Note that at each successive stage of the recursion, the introduction of one new autocorrelation sample, $R(M)$, generates but one independent value $\alpha(M,M)$ for the M th-order PEF; all other coefficients of the M th-order PEF are determined from linear combinations of the coefficients of the $(M-1)$ th-order PEF and their complex conjugates, using $\alpha(M,M)$ as indicated by (3c).

The $\alpha(M,M)$ s are sometimes referred to as the reflection coefficients, because of the analogy of their appearance in (3d) with a similar equation which occurs in the theory of a signal propagating through a layered medium and being partially reflected at each layer interface.

In executing the recursion of equations (3a-d) it can occur (at least in theory) that, for some particular value of M , say M_0 , $P(M_0)=0$. This implies that $|\alpha(M_0,M_0)|=1$. This condition can arise only in the case where the signal being analyzed can be modelled as M_0 complex sinusoids with no additive noise (see Sections 3.1 and 3.2); in general M_0 is not finite. In particular, M_0 cannot be finite when additive white noise is present [4].

For each order M of PEF, an estimate of the power spectrum based on $(M+1)$ values of the autocorrelation function is given by

$$X_{KA}(\omega, M) = \frac{P(M)}{\left| \sum_{m=0}^M \alpha(m,M) \exp(-jm\omega) \right|^2} \quad (4)$$

where ω is the normalized angular frequency in radians with $-\pi \leq \omega \leq \pi$, and the subscript KA refers to the known-autocorrelation case. As discussed in Section 2.0 and as examination of (1) will indicate, the PEFs are "spiking" or whitening filters, since all but one of their output values are zero. The power spectrum of such an output signal is independent of frequency, or "white". The denominator in the right-hand term of (4) is the power transfer function of the PEF, which if multiplied by $X_{KA}(\omega, M)$, the estimated spectrum of the signal, yields the constant, white-noise spectrum $P(M)$. Thus, since both $P(M)$ and the power spectrum of the PEF can be calculated, the signal power spectrum $X_{KA}(\omega, M)$ can be estimated from (4) for successively higher orders of PEF.

2.2 The Burg Algorithm

The Burg algorithm is a procedure for estimating the reflection coefficients directly from a set of time-series amplitude data. It avoids the biases introduced into the spectral estimate when the autocorrelation is estimated from the data and the known-autocorrelation algorithm is then applied; however, as is shown in Sections 4.1 and 4.2, the Burg algorithm introduces biases of its own sort.

Let it be assumed that a set of N time-series amplitude data $x(n)$, $n=0,1,2,\dots,N-1$ have been given, and that $x(n) \neq 0$, where the brackets $\cdot$ denote expected value. The PEFs are derived sequentially. Each successively higher order PEF is applied to the data in both directions simultaneously, and the average of its forward and backward output error powers is minimized by adjusting only its reflection coefficient. The remaining coefficients of each PEF depend on the sequence of reflection coefficients through the functional relationship defined by (3c). The motivation for this procedure is its analogy with that defined by (1), (2), and (3a-d).

Following e.g., [3,11-14] and taking proper note of the occurrences of complex conjugation in the complex data case, the Burg algorithm can be written in a lattice-filter formulation:

$$f_M(n) = f_{M-1}(n) - \Gamma(M,M)b_{M-1}(n-1) \quad (5a)$$

$$b_M(n) = b_{M-1}(n-1) - \Gamma^*(M,M)f_{M-1}(n) \quad \begin{matrix} n = M, M+1, \dots, N-1 \\ M = 1, 2, \dots, N-1 \end{matrix} \quad (5b)$$

and

$$f_0(n) = x(n) = b_0(n) \quad n = 0, 1, 2, \dots, N-1 \quad (5c)$$

(See Fig. 2, where z^{-1} denotes the unit time-delay operator.) The series $f_M(n)$ is the output from the M th-order PEF applied to the input signal $x(n)$ in the forward direction, and is expressed in terms of the output series from the $(M-1)$ th-stage of the lattice. The series $b_M(n)$ is the output from the M th-order PEF, conjugated and applied to the input data in the reverse or backward direction, and again is expressed in terms of the output series from the $(M-1)$ th-stage of the lattice. The $\Gamma(M,M)$ s are the reflection coefficients.

The sum of the forward and backward output error energies at each stage of the lattice is given by

$$E(M) = \sum_{n=M}^{N-1} (|f_M(n)|^2 + |b_M(n)|^2) \quad M = 0, 1, 2, \dots, N-1 \quad (6)$$

The formula for computing $\Gamma(M,M)$, the Burg estimate of the M th-order reflection coefficient, is derived by substituting eqns. (5a) and (5b) into (6), setting $[dE(M)/d\Gamma(M,M)] = 0$ and solving for $\Gamma(M,M)$.

$$\Gamma(M,M) = \frac{\sum_{n=M}^{N-1} b_{M-1}^*(n-1)f_{M-1}(n)}{\sum_{n=M}^{N-1} (|b_{M-1}(n-1)|^2 + |f_{M-1}(n)|^2)} \quad (7)$$

Notice the similarity of (7) to a single-lag unwindowed cross-correlation of the forward and backward output series.

In order to obtain spectral estimates, it is usual to let the output error powers π be defined as

$$\pi(0) = E(0)/(2N) \quad (8a)$$

and

$$\pi(M) = (1 - |\Gamma(M,M)|^2)\pi(M-1) \quad M = 1, 2, \dots, N-1 \quad (8b)$$

by analogy with (3a) and (3d) respectively. Then, letting the $\rho(M,M)$ s be defined by

$$\rho(m,M) = \Gamma(m,M-1) - \Gamma^*(M,M)\Gamma^*(M-m,M-1) \quad m = 1, 2, \dots, M-1 \quad (9)$$

by analogy with (3c), and $\rho(0,M) = 1$ by definition, the M th-order Burg power spectrum estimate $X_B(\omega, M)$ is given by

$$X_B(\omega, M) = \frac{\pi(M)}{\left| \sum_{m=0}^M \rho(m,M) \exp(-jm\omega) \right|^2} \quad (10)$$

by analogy with (4).

2.3 The Zeroes of the Prediction-Error Filters (PEFs)

By employing standard z-transform techniques, the z-transforms of the PEFs for the KA case and the Burg case can be written as polynomials in the complex variable z. For the KA case, this polynomial is $F_{KA}(z, M)$ where

$$F_{KA}(z, M) = 1 - \sum_{m=1}^M \alpha(m, M) z^{-m} \quad (11)$$

Then (4) can be rewritten as

$$X_{KA}(\omega, M) = P(M) / |F_{KA}[\exp(j\omega), M]|^2 \quad (12)$$

where the denominator is the squared magnitude of $F_{KA}(z, M)$ evaluated around the unit circle ($|z| = 1$). Similarly, for the Burg case, the polynomial is $F_B(z, M)$, where

$$F_B(z, M) = 1 - \sum_{m=1}^M \beta(m, M) z^{-m} \quad (13)$$

and (10) can be rewritten as

$$X_B(\omega, M) = \eta(M) / |F_B[\exp(j\omega), M]|^2 \quad (14)$$

It is apparent from (12) and (14) that if any zeroes of $F_{KA}(z, M)$ or $F_B(z, M)$ lie near or on the unit circle, then the magnitude of the spectral estimators $X_{KA}(\omega, M)$ or $X_B(\omega, M)$ will be large at locations on the unit circle in the vicinity of such zeroes. Conversely, zeroes lying close to the origin of the complex z-plane will have little influence on the peaks of the spectrum, but will affect its magnitude away from the peaks. Thus some insight into the character of an autoregressive spectral estimate can be gained by studying the locations of the zeroes of its associated PEF.

3. The One-Pole Complex Sinusoid Case

The formula for a complex signal $x_1(n)$ consisting of a single complex sinusoid in the presence of additive white noise is given by

$$x_1(n) = A_1 \exp(jn\omega_1) + \epsilon(n) \quad (15)$$

where n is any positive or negative integer, or zero; A_1 is the complex amplitude of the complex sinusoid; ω_1 is the angular frequency of the complex sinusoid, normalized so that $-\pi < \omega_1 \leq \pi$; and $\epsilon(n)$ is additive white noise having the property

$$\langle \epsilon^*(n) \epsilon(n+k) \rangle = |\epsilon|^2 \delta(k) \quad (16)$$

where $|\epsilon|^2$ is used to denote the variance of $\epsilon(n)$ and $\delta(k)$ is again the Kronecker delta function.

3.1 The KA Estimate of the PEF

The autocorrelation function $R_1(k)$ of the signal defined by (15) is given by

$$R_1(k) = |A_1|^2 \exp(jk\omega_1) + |\epsilon|^2 \delta(k) \quad (17)$$

where k is any positive or negative integer, or zero, and in general $R(k)$ is defined by

$$R(k) = x^*(n) x(n+k) \quad (18)$$

Substitution of (17) into (3a) gives the result

$$P_1(0) = |\epsilon|^2 (\sigma_1^2 + 1) \quad (19)$$

where $\sigma_1^2 = |A_1|^2 / |\epsilon|^2$ is the signal-to-noise ratio (SNR). If there is no noise, i.e. $|\epsilon|^2 = 0$, then $P_1(0) = |A_1|^2$.

The following general result can be derived from (3b), (3c) and (3d) when $|\epsilon|^2 \neq 0$ and $|A_1|^2 \neq 0$:

$$\alpha_1(m, M) = [\sigma_1^2 / (M\sigma_1^2 + 1)] \exp(jm\omega_1) \quad (20a)$$

$$= [1 / (M + \sigma_1^{-2})] \exp(jm\omega_1) \quad (20b)$$

and

$$P_1(M) = |\epsilon|^2 \left\{ [(M+1)\sigma_1^2 + 1] / [M\sigma_1^2 + 1] \right\} \quad \begin{matrix} M = 1, 2, 3, \dots \\ m = 1, 2, 3, \dots, M \end{matrix} \quad (21)$$

In the noise-free case ($|\epsilon|^2 = 0$ and $\sigma_1^2 \rightarrow \infty$), (20a) or (20b) imply that $\alpha_1(1, 1) = \exp(j\omega_1)$ so that $|\alpha_1(1, 1)|^2 = 1$. Then (3d) implies that $P_1(1) = 0$. Thus, for this special case, the sequences defined by (20) and (21) terminate at $M=1$. Otherwise for this signal model the sequences are infinite, with $P_1(M) \rightarrow |\epsilon|^2 (1 + M^{-1})$ as $M \rightarrow \infty$.

The z-transform of the KA PEF (cf. (11)) is

$$F_{KA}^{(1)}(z, M) = 1 - [\sigma_1^2 / (M\sigma_1^2 + 1)] \sum_{m=1}^M z^{-m} \exp(jm\omega_1) \quad (22)$$

where the superscript 1 denotes a one-pole complex sinusoidal signal. For $M=1$, $F_{KA}^{(1)}(z, 1)$ has a zero at $z_0 = [\sigma_1^2 / (\sigma_1^2 + 1)] \exp(j\omega_1)$. For $M=2$, $F_{KA}^{(1)}(z, 2)$ has zeroes at

$$z_0 = \left\{ \frac{-2\sigma_1^2}{\sigma_1^2 + \sigma_1 \sqrt{9\sigma_1^2 + 4}} \right\} \exp(j\omega_1) \quad (23)$$

so that for high SNR ($1 \ll \sigma_1^2 < \infty$), the zeroes occur at approximately $1.0 \exp(j\omega_1)$ and $-0.5 \exp(j\omega_1)$, and for low SNR ($\sigma_1^2 \ll 1$), the zeroes occur at approximately $\pm \sigma_1 \exp(j\omega_1)$.

Some algebraic manipulation shows that, for the product $M\sigma_1^2$ sufficiently greater than 1, an approximate root z_0 of (22) is given by

$$z_0 \sim \left[1 - \frac{2}{(M-1) M\sigma_1^2} \right] \exp(j\omega_1) \quad (24)$$

This means that, as $M\sigma_1^2 \rightarrow \infty$, the estimated location of the pole corresponding to the single complex sinusoid asymptotically approaches the correct location $\exp(j\omega_1)$ on the complex z-plane along a radius oriented at an angle corresponding to the true frequency of the signal. This is true even when $\sigma_1^2 \ll 1$.

Numerical solutions of (24) show that the other zeroes tend to distribute themselves with approximately uniform angular separation and approximately constant radius inside the unit circle so as to account for the uniform spectrum of the additive white noise. The radius at which the zeroes occur varies inversely with the SNR.

3.2 The Burg Estimate of the PEF

Let it be assumed that N samples of the signal defined by (15) have been given:

$$x_1(n) = A_1 \exp(jn\omega_1) + e(n) \quad n = 0, 1, 2, \dots, N-1 \quad (25)$$

Substituting (25) into (6) by using (5c) and then taking the expected value of $E_1(0)$ yields

$$\langle E_1(0) \rangle = 2N|e|^2(\sigma_1^2 + 1) \quad (26)$$

It is not so easy to calculate $\langle B_1(1,1) \rangle$, since examination of (7) shows that it is necessary to derive the expected value of a quotient of correlated random variables. In general, this requires that the statistics of the random variables be specified and the problem be solved numerically. This will not be done here.

However, for sufficiently high SNR (e.g., $\sigma_1^2 > 100$ or 20 dB) the approximation $(1+x)^{-1} \approx 1-x$, where $x = e(n)/A_1$, can be used to approximate the denominator of (7). Then the following approximate result, which is independent of the statistical distribution of the white noise $e(n)$, is obtained:

$$\langle B_1(1,1) \rangle = \left\{ 1 / [1 + B_1(1,N)\sigma_1^{-2}] \right\} \exp(j\omega_1) \quad (27)$$

where

$$B_1(1,N) = (N-2)/(N-1)^2 \quad (28)$$

and $2/B_1(1,N) \approx 1$ for $2 \ll N \ll \infty$. Comparison of (27) with (20b) for $M=1$ shows that for high SNR $B_1(1,1)$ is a biased estimate of $\alpha_1(1,1)$. However, $B_1(1,1)$ correctly estimates the angular frequency ω_1 of the single complex sinusoid, and the bias term $B_1(1,N)$ monotonically approaches unity as N becomes large.

From (26), (27) and (8a) and (8b) it can be shown that

$$\langle \pi_1(0) \rangle = |e|^2(\sigma_1^2 + 1) \quad (29)$$

and, to the same degree of approximation as was used to obtain (27)

$$\langle \pi_1(1) \rangle = 2B_1(1,N)|e|^2 \quad (30)$$

Comparison of (30) with (21) shows that $\Pi_1(1)$ is a biased estimator of $P_1(1)$, since for σ_1^2 large, $P_1(1) = 2|e|^2$.

The result (30) implies that the lattice-filter outputs $f_1(n)$ and $b_1(n)$ as defined by (5a) and (5b) have low SNR, since their expected power is at most only a factor of four greater than the additive white noise power $|e|^2$. Therefore it would again be necessary to specify the detailed statistics of the additive noise before the higher order reflection coefficients could be estimated. This will not be done here.

3.3 Discussion of the One-Pole Complex Sinusoid Case

The results of Sections 3.1 and 3.2 show that, for high SNR ($\sigma_1^2 \gg 100$) the first order ($M=1$) PEF generated by the Burg algorithm is biased as compared to the PEF generated by the KA technique. This bias, however, monotonically decreases as the number of data N is increased, and both the KA and the Burg algorithms correctly estimate the frequency of the complex sinusoid.

It is impossible to investigate the properties of the Burg PEFs for the low SNR case, or for orders higher than $M=1$ for the high SNR case, without specifying the statistical distribution of the additive white noise. This problem has not been considered here.

4. The Two-Pole Complex Sinusoid Case

The sampled signal $x_2(n)$ consisting of two complex sinusoids in the presence of additive white noise is given by

$$x_2(n) = A_1 \exp(jn\omega_1) + A_2 \exp(jn\omega_2) + \epsilon(n) \quad (31)$$

where $A_k = |A_k| \exp(j\phi_k)$ is the complex amplitude of the k th sinusoid, ϕ_k is its arbitrary initial phase at $n=0$, and ω_k is its angular frequency, normalized so that $-\pi < \omega_k \leq \pi$, for $k=1,2$. $\epsilon(n)$ is additive white noise, as in (15). It is apparent that if $A_2 = A_1^*$ and $\omega_2 = -\omega_1$ then $x_2(n)$ is a sampled real sinusoid in additive white (complex) noise.

Equation (31) can be written in a form which will be subsequently more tractable:

$$x_2(n) = A_0 \exp[j(n\omega_0 + \phi_0)] \times \{r \exp[j(n\Delta\omega + \Delta\phi)] + r^{-1} \exp[-j(n\Delta\omega + \Delta\phi)]\} + \epsilon(n) \quad (32)$$

where $A_0 = |A_1 A_2|^{1/2}$ is the geometric mean of the magnitudes of the two amplitudes; $r = [|A_1|/|A_2|]^{1/2}$ is the square root of the ratio of the magnitudes of the amplitudes; $\phi_0 = (\phi_1 + \phi_2)/2$ is the mean initial phase; $\Delta\phi = (\phi_1 - \phi_2)/2$ is one-half the difference between the two phases; $\omega_0 = (\omega_1 + \omega_2)/2$ is the mean angular frequency; and $\Delta\omega = (\omega_1 - \omega_2)/2$ is one-half the difference between the two frequencies.

4.1 The KA Estimate of the PEF

The autocorrelation function $R_2(k)$ of the signal defined by (32) is given by

$$R_2(k) = A_0^2 (r^2 + r^{-2}) \exp(jk\omega_0) [\cos k\Delta\omega + j\rho(r) \sin k\Delta\omega] + |\epsilon|^2 \delta(k) \quad (33)$$

where

$$\rho(r) = (r^2 - r^{-2}) / (r^2 + r^{-2}) \quad (34)$$

and k is any positive or negative integer, or zero. Substitution of (33) into (3a) gives the result

$$P_2(0) = |\epsilon|^2 (\sigma_2^2 + 1) \quad (35)$$

where $\sigma_2^2 = A_0^2 (r^2 + r^{-2}) / |\epsilon|^2$ is the SNR. If there is no noise, then $P_2(0) = A_0^2 (r^2 + r^{-2})$ is the signal power.

From (33), (35), (3b) and (3d), the following general results for the reflection coefficient and the output error power for $M=1$ can be derived when $|\epsilon|^2 \neq 0$ and $A_k \neq 0$:

$$\gamma_2(1,1) = \exp(j\omega_0) \left\{ [\cos \Delta\omega + j\rho(r) \sin \Delta\omega] / [1 + \sigma_2^{-2}] \right\} \quad (36)$$

and

$$P_2(1) = A_0^2 (r^2 + r^{-2}) \left\{ [1 - \rho^2(r)] \sin^2 \Delta\omega + \sigma_2^{-2} (2 + \sigma_2^{-2}) \right\} / \{1 + \sigma_2^{-2}\} \quad (37)$$

For high SNR ($\sigma_2^2 \gg 1$), the first reflection coefficient is given approximately by

$$\gamma_2(1,1) = \exp(j\omega_0) [\cos \Delta\omega + j\rho(r) \sin \Delta\omega] \quad (38)$$

For $r=1$, $\rho(r)=0$ and the zero of $F_{KA}^{(2)}(z,1)$ (see Section 2.3) is located at $z_0 = \cos \Delta\omega \exp(j\omega_0)$, which lies on a radius oriented at the mean angular frequency ω_0 . This zero moves towards the limiting values of $\exp(j\omega_1)$ as $r \rightarrow \infty$ and $\rho(r) \rightarrow 1$, or $\exp(j\omega_2)$ as $r \rightarrow 0$ and $\rho(r) \rightarrow -1$, and the single complex sinusoid case is approached in each case.

The output error power is, using (38), (35) and (8b),

$$P_2(1) = (2A_0^2 \sin^2 \Delta\omega) / (r^2 + r^{-2}) \quad (39)$$

which is essentially the signal power attenuated by the factor $2 \sin^2 \Delta\omega / (r^2 + r^{-2})$. This factor is unity when $r=1$ and $\Delta\omega = \pi/2$, and decreases as r or $\Delta\omega$ deviate from these values.

For low SNR ($\sigma_2^2 \ll 1$), the reflection coefficient and the output error power are given by

$$u_2(1,1) = \exp(j\omega_0) \alpha_2^2 [\cos\Delta\omega + j\rho(r)\sin\Delta\omega] \quad (40)$$

and

$$P_2(1) = |u_2(1,1)|^2 + A_0(r^2 + r^{-2}) \quad (41)$$

Thus the frequency estimated from the location of the zero of the $M=1$ order PEF lies in the range bounded by $\omega_0 \pm \Delta\omega$, and the output error power is essentially the unattenuated signal plus noise power.

For $M=2$, (33), (36) and (3b) can be combined to yield

$$u_2(z,2) = -\exp(j2\omega_0) \left\{ \frac{[1-\rho^2(r)]\sin^2\Delta\omega - \alpha_2^{-2} [\cos 2\Delta\omega + j\rho(r)\sin 2\Delta\omega]}{[1-\rho^2(r)]\sin^2\Delta\omega + \alpha_2^{-2} [2+\alpha_2^{-2}]} \right\} \quad (42)$$

and (37), (42), and (3d) can be combined to yield

$$P_2(2) = |u_2(2,2)|^2 = \frac{2[1-\rho^2(r)]\sin^2\Delta\omega(2 + \cos 2\Delta\omega) + \alpha_2^{-2}(3+\alpha_2^{-2})}{[1-\rho^2(r)]\sin^2\Delta\omega + \alpha_2^{-2}(2+\alpha_2^{-2})} \quad (43)$$

For high SNR ($\alpha_2 \rightarrow 1$) and

$$\alpha_2^2 [1-\rho^2(r)]\sin^2\Delta\omega \gg |\cos 2\Delta\omega + j\rho(r)\sin 2\Delta\omega| \quad (44)$$

(42) reduces to

$$u_2(2,2) = -\exp(j2\omega_0) \quad (45)$$

It can be shown from (45), (36) and (3c) that

$$u_2(1,2) = \exp(j\omega_0) 2 \cos\Delta\omega \quad (46)$$

so that, in the limit as the left-hand side of (44) approaches infinity, the zeroes of $F_{KA}^{(2)}(z,2)$ approach $z_0 = \exp[j(\omega_0 \pm \Delta\omega)]$, the true locations of the poles of the complex sinusoids. In this same limit, $|u_2(2,2)|^2 = 1$, so that $P_2(2)=0$ and the recursion terminates.

For the intermediate case, where $\alpha_2^2 \gg 1$ but (44) is not satisfied, i.e.,

$$\alpha_2^2 [1-\rho^2(r)]\sin^2\Delta\omega \approx |\cos 2\Delta\omega + j\rho(r)\sin 2\Delta\omega| \quad (47)$$

(42) can be reduced to

$$u_2(2,2) = \exp(j2\omega_0) \cos\Delta\omega [0.5 \cos\Delta\omega + j\rho(r)\sin\Delta\omega] \quad (48)$$

and it can be shown from (48), (38) and (3c) that

$$u_2(1,2) = 0.5 \exp(j\omega_0) [\cos\Delta\omega + j\rho(r)\sin\Delta\omega] \quad (49)$$

so that the zeroes of $F_{KA}^{(2)}(z,2)$ occur at $[\cos\Delta\omega + j\rho(r)\sin\Delta\omega]\exp(j\omega_0)$ and $-0.5[\cos\Delta\omega + j\rho(r)\sin\Delta\omega]\exp(j\omega_0)$. Comparison of these results with (23), which gives the zero locations of the PEF for a one-pole signal is very similar to that for the one-pole signal at high SNR. Comparison with (38) shows that one of the zeroes of $F_{KA}^{(2)}(z,2)$ remains the same as that of $F_{KA}^{(2)}(z,1)$ at high SNR; i.e., the two poles are estimated as one by the $M=2$ PEF if (44) is not satisfied.

For the low SNR case ($\alpha_2 \rightarrow 1$), (42) reduces to

$$u_2(2,2) = \exp(j2\omega_0) \left\{ \alpha_2^2 [\cos 2\Delta\omega + j\rho(r)\sin 2\Delta\omega] \right\} \quad (50)$$

and $P_2(2)$ is the same as $P_2(1)$ as given by (41). Since for this case both $|u_2(2,2)|$ and $|u_2(1,1)|$ are proportional to α_2^2 , then from (3c) it is clear that, to first order in α_2^2 ,

$$u_2(1,2) = u_2(1,1) \quad (51)$$

where $u_2(1,1)$ is given by (40). The zeroes of $F_{KA}^{(2)}(z,2)$ in this case occur at approximately $z = \alpha_2^2 [\cos 2\Delta\omega + j\rho(r)\sin 2\Delta\omega] + \alpha_2^2 [\cos\Delta\omega + j\rho(r)\sin\Delta\omega]/2 \exp(j\omega_0)$. For $r=1$ and thus $\rho(r)=0$, the zeroes occur at $z = \alpha_2^2 [\cos 2\Delta\omega] + \alpha_2^2 [\cos\Delta\omega]/2 \exp(j\omega_0)$, which lie close to the origin of the complex z -plane, on a diameter of the unit circle passing through the point $z = \exp(j\omega_0)$. For $r < 1$ ($\rho(r) < 1$) or $r > 1$ ($\rho(r) > 1$) the zeroes tend to the locations $z = \alpha_2^2 [1 + \alpha_2/2] \exp(j\omega_1)$ or $z = \alpha_2^2 [1 + \alpha_2/2] \exp(j\omega_2)$ respectively. Thus it is apparent that for low SNR, the spectral estimate corresponding to $M=2$ is incapable of resolving the spectral peaks corresponding to the poles of the two complex sinusoidal signals.

There appears to be no straightforward recursion formula for the KA PEF coefficients, as in the case of the single sinusoid example of Section 3.1. Therefore, following this approach, it is not easy to determine the behavior of the KA PEFs in the case of large M and, in particular, whether resolution of the two sinusoids is to be expected for the product $M\alpha_2^2$ sufficiently large, independent of the value of α_2 . This problem, however, has been solved using powerful matrix techniques, by Marple [6].

4.2 The Burg Estimate of the PEF

Let it again be assumed (cf. Section 3.2) that N samples of the signal defined by (32) have been given:

$$x_2(n) = A_0 \exp[j(n\omega_0 + \phi_0)] \left\{ r \exp[j(n\Delta\omega + \Delta\phi)] + r^{-1} \exp[-j(n\Delta\omega + \Delta\phi)] \right\} + v(n) \quad (52)$$

n = 0, 1, 2, ..., N-1

Substituting (52) into (6) and (7), using (5c), and then applying (8a) and taking expected values with respect to the additive white noise only yields

$$\langle |I_2(0)|^2 \rangle = A_0^2 (r^2 + r^{-2}) [1 + 2 \cos \Delta\phi_{\text{mid}} G(N, \Delta\omega) / (r^2 + r^{-2})] + |v|^2 \quad (53)$$

for the expected signal-plus-noise power. Here

$$G(N, \Delta\omega) = \frac{\sin(N\Delta\omega)}{N \sin \Delta\omega} \quad (54)$$

is the common grating-function frequency response of a normalized, uniformly weighted discrete Fourier transform of N data, and

$$\Delta\phi_{\text{mid}} = (N-1)\Delta\omega + 2\Delta\phi \quad (55)$$

is the phase difference between the two complex sinusoidal components, reckoned at the middle of the data set. Note that there may or may not be a datum at the middle of the data set, according to whether N is an odd or even integer, respectively. Also note that $\Delta\phi$ has not been averaged, but rather is assumed to be a fixed parameter of the particular set of data being analyzed. This assumption corresponds to the usual practical case, where only one set of data is available.

Again, it is not easy to calculate the expected values of the reflection coefficients unless the assumption of high SNR is made. In that case the same approximations can be made as in the derivation of (26) to get

$$\langle R_2(1,1) \rangle = \exp(j\omega_0) \left\{ \frac{\cos \Delta\omega + j r(r) \sin \Delta\omega + 2 \cos \Delta\phi_{\text{mid}} G(N-1, \Delta\omega) / (r^2 + r^{-2})}{1 + 2 \cos \Delta\phi_{\text{mid}} \cos \Delta\omega G(N-1, \Delta\omega) / (r^2 + r^{-2}) + v_2^{-2} [1 + B_2(1, N)]} \right\} \quad (56)$$

where the bias term $B_2(1, N)$ is of order $(N-1)^{-1}$ and is given by eqn. (A1) of Appendix A.

Comparison of (56) and (36) shows that for high SNR $R_2(1,1)$ is a biased estimate of $\alpha_2(1,1)$ unless $N \rightarrow \infty$. For infinite SNR and finite N, however, $R_2(1,1)$ becomes an unbiased estimate of $\alpha_2(1,1)$ if

$$\cos \Delta\phi_{\text{mid}} = 0 \quad (57)$$

or

$$G(N-1, \Delta\omega) = 0 \quad (58)$$

Similarly, comparison of (53) and (35) shows that $|I_2(0)|^2$ is a biased estimate of $P_2(0)$ unless either (57) is satisfied or

$$G(N, \Delta\omega) = 0 \quad (59)$$

It is impossible to satisfy (58) and (59) simultaneously for N finite, but when (57) is satisfied the Burg spectral estimate (14) is unbiased for M=1 and infinite SNR.

Progressing now to the second stage (M=2) of the Burg algorithm, it is found that the algebra becomes all but intractable unless the condition of infinite SNR is assumed. For this special case, (56), (52), (5a-c) and (7) can be used to derive the following expression for $R_2(2,2)$:

$$\begin{aligned} R_2(2,2) = & -\exp(j2\omega_0) \times \left\{ (1 - 2 \cos \Delta\phi_{\text{mid}} G(N-2, \Delta\omega) / (r^2 + r^{-2})) \right. \\ & - 2 \cos \Delta\phi_{\text{mid}} G(N-1, \Delta\omega) \times [\cos \Delta\phi_{\text{mid}} G(N-2, \Delta\omega) [\cos \Delta\omega + j r(r) \sin \Delta\omega] \\ & \quad - 2 \cos \Delta\omega / (r^2 + r^{-2})] \\ & + \cos^2 \Delta\phi_{\text{mid}} G^2(N-1, \Delta\omega) \times \{ [\cos 2\Delta\omega + j r(r) \sin 2\Delta\omega] \\ & \quad - 2 \cos \Delta\phi_{\text{mid}} G(N-2, \Delta\omega) / (r^2 + r^{-2}) \} \\ & \left. / \left\{ (1 - 2 \cos \Delta\phi_{\text{mid}} G(N-2, \Delta\omega) / (r^2 + r^{-2})) \right\} \right\} \end{aligned}$$

$$\begin{aligned}
& - 2 \cos \Delta \phi_{\text{mid}} \cos \Delta \omega G(N-1, \Delta \omega) \times \left\{ \cos \Delta \phi_{\text{mid}} G(N-2, \Delta \omega) - 2/(r^2 + r^{-2}) \right\} \\
& + \cos^2 \Delta \phi_{\text{mid}} G^2(N-1, \Delta \omega) \times \left\{ 1 - 2 \cos \Delta \phi_{\text{mid}} \cos 2 \Delta \omega G(N-2, \Delta \omega) / (r^2 + r^{-2}) \right\} \quad (60)
\end{aligned}$$

Examination of (60) shows that, even for infinite SNR, the "correct" value of $-\exp(j2\omega_0)$ for the reflection coefficient $\Gamma_2(2,2)$ is not realized for N finite unless either (57) or (58) is satisfied. Realization of either of these conditions for infinite SNR will, as examination of (60) show, cause the magnitude of the reflection coefficient to be unity. Thus it can be inferred that, for sufficiently high SNR, the two crucial factors affecting the reflection coefficient computed using the Burg algorithm are the phase difference between the two complex sinusoids at the middle of the data record, $\Delta \phi_{\text{mid}}$, and the number of cycles of difference frequency contained in the finite-length data record.

4.3 Discussion of the Two-Pole Complex Sinusoid Case

The results of Section 4.1 and 4.2 show that even for infinite SNR, the first ($M=1$) and second ($M=2$) order PEFs generated by the Burg algorithm are biased as compared to the PEFs generated by the KA technique. It is clear from the appearance of the grating function (54) in the equations (56) and (60) for the first- and second-order reflection coefficients that the magnitude of such biases will have inverse dependence on N , the length of the data record.

The effect of this bias is to reduce the magnitude of the reflection coefficient and thus to allow significant levels of uncanceled signal energy to propagate beyond the stage $M=2$ in the Burg algorithm. Then PEFs of successively higher order can be based on this coherent "leakage" signal. However, whenever one of the criteria described by (57) or (58) is satisfied, no significant coherent leakage signal is propagated beyond the stage $M=2$. It is conjectured that it is the presence or absence of this coherent leakage signal beyond the stage $M=2$ that determine whether or not line splitting will occur for PEFs of some higher order. Results both of previously published [7,8] and new simulation studies support this conjecture, as indicated in Section 5 below.

5. Results of Some Simulation Studies

In this section are presented the results of some studies of the performance of the complex Burg algorithm for the analysis of signals known to be comprised of two complex sinusoids in the presence of very weak additive complex white noise ($\sigma_n^2 = 77$ dB). These studies parallel and extend a set of studies performed by Fougere et al [8] using the real-arithmetic Burg algorithm to estimate spectra of a single real sine-wave signal in the presence of very weak additive real white noise.

It will be necessary to make comparisons between complex data of the form (52) and real data $x_s(n)$ of the form

$$x_s(n) = \sin(n\omega_s + \phi_s) + i_s(n) \quad n = 0, 1, \dots, N-1 \quad (61)$$

which is comprised of N samples of a real sine wave with initial phase ϕ_s plus additive uncorrelated noise samples $i_s(n)$. The angular frequency ω_s is given by

$$\omega_s = 2\pi f \Delta t \quad (62)$$

where Δt is the sampling interval (sec) and f is the signal frequency (Hz). Equation (61) can easily be rewritten in the form of (52) by letting $A_0 = 0.5$, $r = 1$, $\phi_0 = 0$, $\Delta \phi = \phi_s - \pi/2$, $\omega_0 = 0$ and $\Delta \omega = \omega_s$. Then the phase difference between the two complex components of the sine wave reckoned at the middle of the data record is, according to (55), given by

$$\Delta \phi_{\text{mid}} = (N-1)\omega_s + 2\phi_s - \pi \quad (63)$$

The results of Fougere et al have been extended by allowing the value of r , the ratio of the positive frequency to negative frequency signal amplitude, to range from 1 to ∞ in a series of six steps. These steps are denoted by the letters A-F, and the relevant signal parameters are summarized in Table 1. For all steps the total signal power $A_0^2(r^2 + r^{-2})$ was maintained constant and equal to 0.5, the power of a real unit-amplitude sine wave. Also, the values of $\phi_0 = 0$ and $\omega_0 = 0$ were maintained for all the trials. These restrictions do not limit the generality of the results obtained. It is clear that an arbitrary phase rotation of the entire data-set will have no effect on its power spectrum. It is also clear that since terms of the form $\exp(jm\omega_0)$ can be factored out of the $u(m, M)$ s and $R(m, M)$ s, the effect of non-zero ω_0 is simply to shift the estimated spectrum along the frequency axis (in a circular or end-around fashion) by the amount ω_0 . Finally, it should be noted that for each set of cases examined, the same set of noise-data samples was used with all sets of signal data.

5.1 Case 1

The signal data for Case 1 consisted of 21 samples of two complex sinusoids with angular frequencies $\omega_{1,2} = 2\pi/20$, so that $\Delta \omega = \omega_1$. $\Delta \phi_{\text{mid}}$ was stepped from $-\pi$ to $+\pi$ in increments of $2\pi/9$ radians, so that in all instances $\cos \Delta \phi_{\text{mid}} \neq 0$. All spectra were estimated using (14) and a length 20 ($M=19$) Burg PEF.

$$\langle e_2(1,1) \rangle = \frac{\{j\rho(r) + 0.4 \cos \Delta\phi_{mid} / (r^2 + r^{-2})\}}{1 + 10^{-7.7} [1 + B_2(1.6)]} \quad (64)$$

and (60) reduces to

$$B_2(2,2) = -(1 - 0.04 \cos^2 \Delta\phi_{mid}) / (1 + 0.04 \cos^2 \Delta\phi_{mid}) \quad (65)$$

Consideration of (64) and (65) shows that here the Burg PEF should have greatest bias for $r=1$, when $(r^2 + r^{-2})$ has its minimum value of 2, and that the bias should vanish for $r=\infty$, when $\rho(r)=1$ and $\langle B_2(1,1) \rangle = 1 - 4 \times 10^{-8}$. For the case $r=\infty$, (60) and hence (65) are not valid.

Examination of Figs. 9-14 supports all these conjectures. These figures show orthographic projections of the 91 spectra, with ω ranging from $-\pi$ to $+\pi$ across the page, and $\Delta\phi_{mid}$ increasing "into" the page. The labelled vertical bar to the left of the spectra indicates a variation of 20 dB in power spectral density. It is clear that there is no line-splitting for $\Delta\phi_{mid} = -5\pi/2, -3\pi/2$ and $-\pi/2$, and that the splitting shows a quasi-cosinusoidal dependence on $\Delta\phi_{mid}$ as suggested by the form of (64). The splitting became less severe as $r^2 \rightarrow \infty$, again as might be inferred from (64), until for Case 2E the weaker signal pole was correctly estimated as a single pole. Examination of the zeroes of the PEFs and the residue powers showed measurable splitting of the stronger signal poles even for this case. Case 2F of course showed no line splitting, since there was then only one signal pole.

The "banding" effect visible most clearly in Fig. 9 (Case 2A) and to a decreasing extent in subsequent figures can be explained on the basis that when $\cos \Delta\phi_{mid} = 0$, $|B_2(2,2)| = 1$ so that the output error power was greatly reduced in those cases. This caused a shift in the level of the spectrum, as can be seen from the dependence through (8b) of the numerator of (14) on this quantity. This also explains the obvious drop in spectral level in Fig. 14, where $|B_2(1,1)| = 1$ for all values of $\Delta\phi_{mid}$.

Examination of the residue powers showed that, when line-splitting occurred, a significant portion of the signal power was accounted for by each pole of a split pair; in fact for Case 2A and $|\cos \Delta\phi_{mid}| < 1$, 70% of the signal powers appeared at the more severely deviated poles, and only 30% at less severely deviated poles. That the greater portion of the signal power was associated with the more deviated pole appeared for these data to be true in general. The present theory makes no prediction as to why this should be the case, although in principle the theory for the infinite SNR model could be extended to do so.

5.3 Case 3

The signal data for Case 3 consisted of 25 sets of 101 samples of two complex sinusoids with $\Delta\phi_{mid} = 2\pi$ in all cases. Again $\omega_0 = 0$ was chosen, and $\Delta\omega$ was stepped from $2\pi \times 0.0125$ to $2\pi \times 0.4925$ inclusive in increments of $2\pi \times 0.02$. Thus Case 3A parallels Case 4 of [8], where 101 samples were taken at intervals $\Delta t = 0.01$ s of real unit-amplitude sine waves with $\phi_s = \pi/4$ and f_s stepped from 1.25 Hz to 49.25 Hz inclusive in steps of 2 Hz. In all cases the spectra were estimated using (14) and a length 25 ($M=24$) Burg PEF.

Figures 15-20 (Cases 3A-F) show the spectral estimates obtained from the data. These figures are again orthographic projections with $\Delta\omega$ increasing "into" the page. Any comments on the detailed structure of the line-splitting shown would necessarily be speculative, but some general observations can be made.

The first remark is that line-splitting appears to become less severe as r was increased in value, as examinations of (56) and (60) might infer, and in fact line-splitting vanished for $r=\infty$ as discussed in Section 5.2.

The second remark concerns the possible dependence of the spectral level on the values of $G(N-1, \Delta\omega)$ and $G(N-2, \Delta\omega)$. These values are given for the plotted spectra in Table 2. It is interesting to note that the minimum spectral level occurred at the minimum values for $|G(N-1, \Delta\omega)|$, $|G(N-2, \Delta\omega)|$ and $|\cos \Delta\omega|$, and the higher spectral levels were observed when $G(N-2, \Delta\omega) < 0$, or $\Delta\omega > 2\pi \times 0.25$. Examination of (60) shows that as $\Delta\omega$ exceeds the value $\pi/2$ certain terms change sign in such a manner as to decrease the magnitude of $B_2(2,2)$ and thus perhaps to increase the magnitude of the numerator of (14) through the relation (8b).

Finally, for Case 3F it is again observed that the spectral level drops as r and the bias term in (56) vanishes, similar to Case 2F. For Case 3F only, the estimated poles accurately reflected the true signal pole locations, were unsplit and had correct residue powers. For all other cases, examination of the locations of the poles of the Burg spectral estimate showed the existence of multiple poles with significant residue power in the vicinity of the true locations of the two signal poles.

TABLE 2
Values of $(\Delta\omega/2\pi) \times 100$, $G(N-1, \Delta\omega)$ and $G(N-2, \Delta\omega)$ for data from Case 3

$(\Delta\omega/2\pi) \times 100$	$G(N-1, \Delta\omega)$	$G(N-2, \Delta\omega)$	$\cos \omega$
1.25	0.1275	0.1823	0.9969
3.25	0.0493	0.0488	0.9792
5.25	0.0309	0.0295	0.9461
7.25	0.0227	0.0206	0.8980
9.25	0.0182	0.0125	0.8358
11.25	0.0154	0.0118	0.7604
13.25	0.0135	0.0092	0.6730
15.25	0.0122	0.0071	0.5750
17.25	0.0113	0.0058	0.4679
19.25	0.0107	0.0033	0.3535
21.25	0.0103	0.0024	0.2334
23.25	0.0101	0.0011	0.1097
25.25	0.0100	-0.0002	-0.0157
27.25	0.0101	-0.0014	-0.1409
29.25	0.0104	-0.0028	-0.2639
31.25	0.0108	-0.0042	-0.3827
33.25	0.0115	-0.0058	-0.4955
35.25	0.0125	-0.0076	-0.6004
37.25	0.0135	-0.0098	-0.6959
39.25	0.0160	-0.0126	-0.7804
41.25	0.0191	-0.0165	-0.8526
43.25	0.0243	-0.0224	-0.9114
45.25	0.0340	-0.0328	-0.9558
47.25	0.0582	-0.0579	-0.9551
49.25	0.2123	-0.2142	-0.9889

6. Summary and Conclusions

The theoretical properties of autoregressive spectral analysis schemes have been analyzed when the signal under investigation is known to be comprised of one or two complex sinusoids in additive white noise. This latter case includes as a special case data comprising a single real sine wave in additive white noise. It has been shown that when the autocorrelation of the signal is known, the frequency of a single complex sinusoid can always be extracted, independent of the signal-to-noise ratio (SNR), provided enough samples of the autocorrelation are available. It was also shown that the Burg algorithm correctly extracts the frequency of a single sinusoid in additive white noise from the complex amplitude time-series data, provided the SNR is sufficiently high.

The situation when two complex sinusoids are present is much more complicated. It was found to be fairly difficult to derive general equations describing the spectrum for the known autocorrelation (KA) case, even for a two-pole autoregressive model. Nevertheless these equations served as a useful touchstone for the extremely complicated Burg equations for the analysis of time-series amplitude data. Detailed theoretical analysis showed that, unlike the KA case, the Burg spectral estimate is expected to be sensitive both to the number of cycles of the difference frequency between the two components contained in the finite-length data record, and in particular to the relative phase difference between the two complex sinusoidal components at the middle of the data record. Finally, simulation results were shown to be fully compatible with the conjectured basis of line-splitting presented here.

7. Acknowledgement

This work is supported by the Department of National Defence under Research and Development Branch Project No. 33C69.

Appendix A

The bias term $B_2(1, N)$ found in the expression for $\langle B_2(1, 1) \rangle$ is:

$$B_2(1, N) = (N-1)^{-1} \left\{ 1 + [(N-2)/(N-1)] \times \right. \\ \left. [\cos \Delta\omega + j_1(r) \sin \Delta\omega + \cos \Delta\omega_{mid} \cos \Delta\omega G(N-2, \Delta\omega)/(r^2 + r^{-2})] \right. \\ \left. [\cos \Delta\omega + j_1(r) \sin \Delta\omega + 2 \cos \Delta\omega_{mid} \cos \Delta\omega G(N-1, \Delta\omega)/(r^2 + r^{-2})] \right\} \quad (A1)$$

References

1. Burg, J.P., "Maximum Entropy Spectral Analysis", presented at the 37th Meeting of the Society of Exploration Geophysicists, Oklahoma City, 31 October 1967.
2. Burg, J.P., "A New Analysis Technique for Time Series Data". Presented at the NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics, Enschede, Netherlands, 1968.
3. Burg, J.P., "Maximum Entropy Spectral Analysis". Ph.D. Thesis Stanford University, Stanford, California, May 1975.
4. Ulrych, T.J. and R.W. Clayton, "Time series modelling and maximum entropy". Phys. Earth Planet. Inter., 12, 1976, pp. 188-200.
5. Frost, O.L., "Power Spectrum Estimation", in "Aspects of Signal Processing, Part 1", D. Reidel Publishing Company, Dordrecht-Holland, 1977, pp. 125-162.
6. Marple, L.A., "Conventional Fourier, Autoregressive, and Special ARMA Methods of Spectrum Analysis". Engineer's Degree Thesis, Stanford University, Stanford, California, December 1976.
7. Chen, W.Y. and G.R. Stegen, "Experiments with maximum entropy power spectra of sinusoids", J. Geophys. Res., 79, No. 20, 10 July 1974, pp. 3019-3022.
8. Fougere, P.F., E.J. Zawalick and H.R. Radoski, "Spontaneous line splitting in maximum entropy power spectrum analysis", Phys. Earth Planet. Inter. 12, 1976, pp. 201-207.
9. Fougere, P.F., "A solution to the problem of spontaneous line splitting in maximum entropy power spectrum analysis", J. Geophys. Res., 82, No. 7, 1 March 1977, pp. 1051-1054.
10. Makhoul, J., "Linear prediction: A tutorial review", Proc. IEEE, 63, No. 4, April 1975, pp. 561-580.
11. Makhoul, J., "Lattice methods in spectral estimation", Proceedings of the RADC Spectrum Estimation Workshop, 24, 25 and 26 May 1978, pp. 159-173, AD-A054650.
12. Smylie, D.E., G.K.C. Clarke and T.J. Ulrych, "Analysis of Irregularities in the Earth's Rotation", Methods in Computational Physics, 13, Academic Press, New York, 1973, pp. 391-430.
13. Andersen, N., "On the calculation of filter coefficients for maximum entropy spectral analysis", Geophysics 39, No. 1, February 1974, pp. 69-72.
14. Herring, R.W., "A Review of Maximum Entropy Spectral Analysis", CRC Technical Note No. 685, June 1977.
15. Johnsen, S.J. and N. Andersen, "On power estimation in maximum entropy spectral analysis", Geophysics 43, No. 4, June 1978, pp. 681-690.

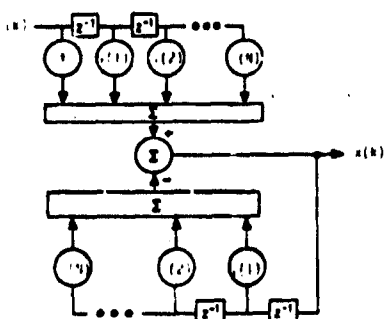


Fig. 1 Network for generating N complex sine waves in additive white noise from complex white noise input (after Fig. 4.1 of [6]).

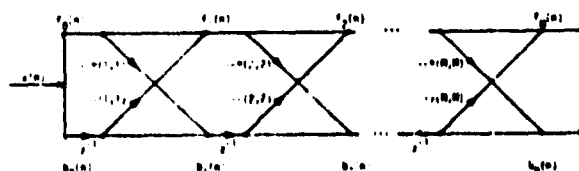


Fig. 2 Basic all-zero lattice network (after [11]).

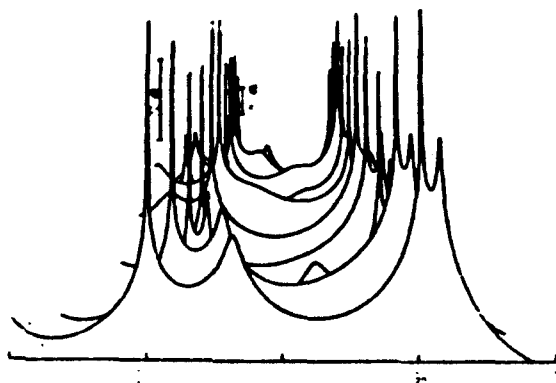


Fig. 3 Case 1A. $r = 1$

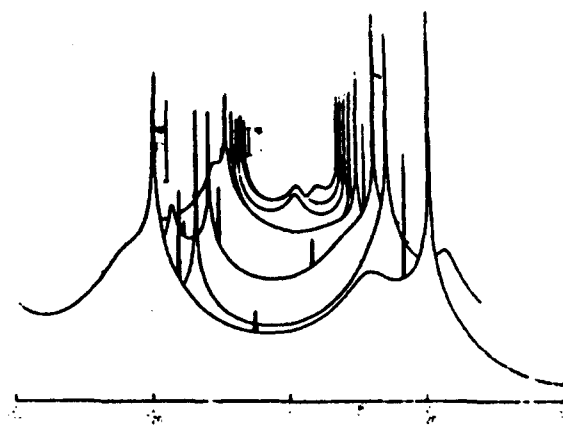


Fig. 6 Case 1D. $r = \sqrt{100}$

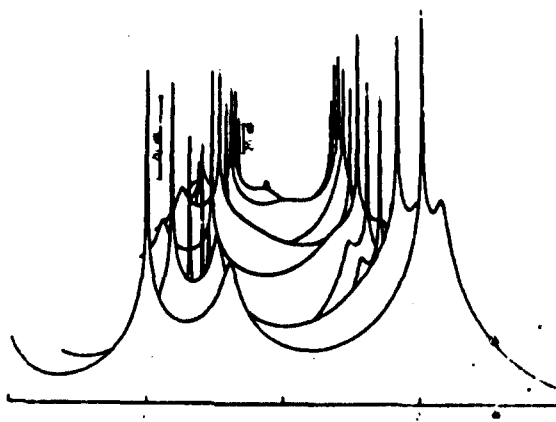


Fig. 4 Case 1B. $r = \sqrt{2}$

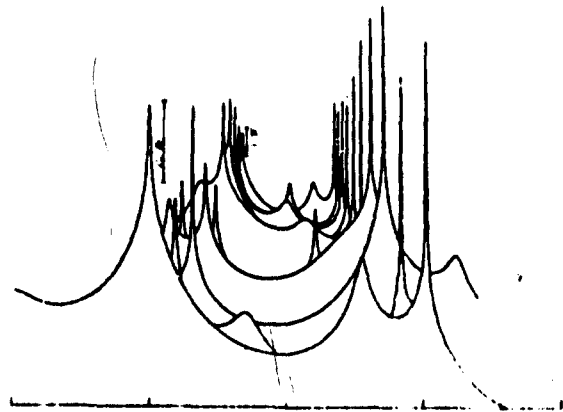


Fig. 7 Case 1E. $r = \sqrt{1000}$

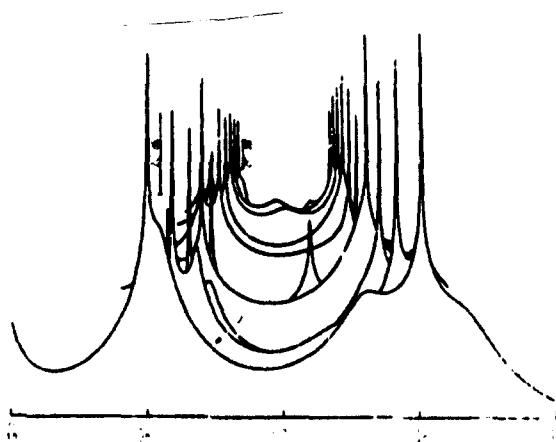


Fig. 5 Case 1C. $r = \sqrt{10}$

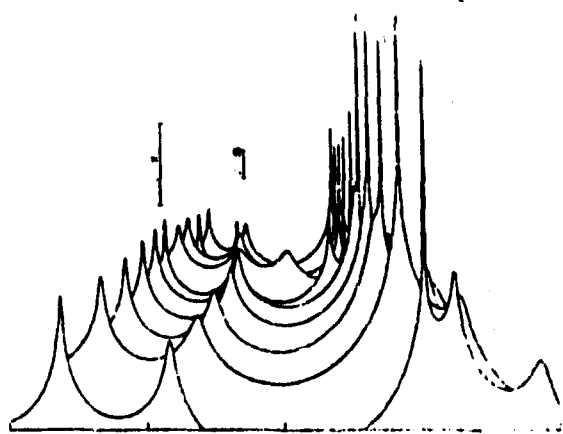


Fig. 8 Case 1F. $r = \infty$

Figs. -3-8. Cases 1A-F. Estimated spectral power vs. angular frequency and $\Delta\phi_{mid}$. $\omega = \omega_0 = 2\pi/20$. $N=21$. $M=19$. $\Delta\phi_{mid}$ stepped from $-\pi$ to π in increments of $2\pi/9$ radians. (Perspective projections.)

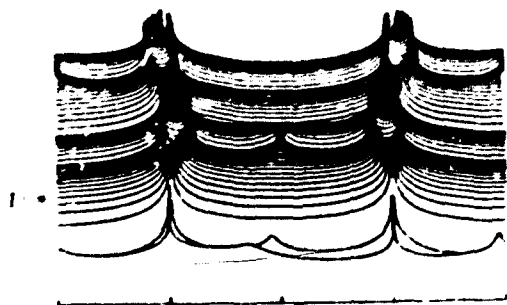


Fig. 9 Case 2A. $r=1$

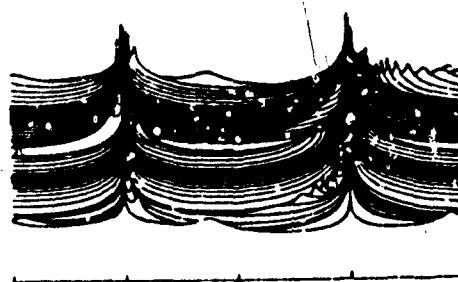


Fig. 12 Case 2D. $r = \sqrt{100}$

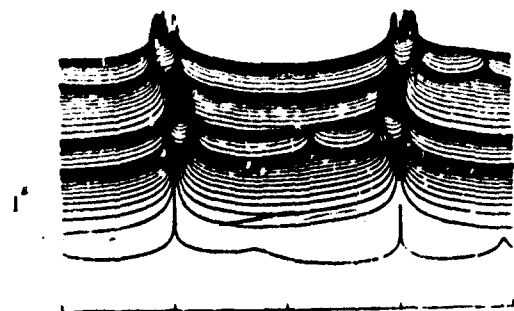


Fig. 10 Case 2B. $r = \sqrt{2}$

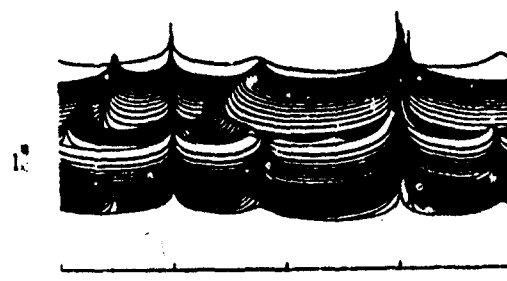


Fig. 13 Case 2D. $r = \sqrt{1000}$

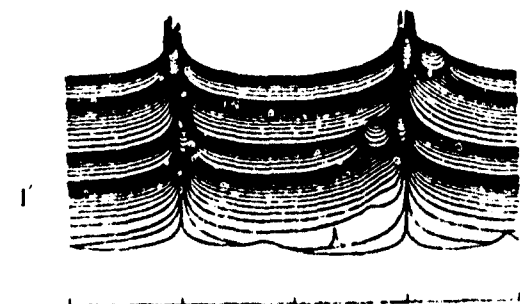


Fig. 11 Case 2C. $r = \sqrt{10}$

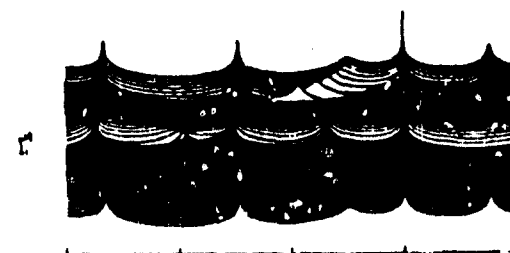


Fig. 14 Case 2F. $r = \infty$

Figs. 9-14. Cases 2A-F. Estimated spectral power vs. angular frequency and $\Delta\phi_{mid}$. $\omega_1 = -\omega_2 = \pi/2$. $N=6$. $M=5$. $\Delta\phi_{mid}$ stepped from $-5\pi/2$ to $\pi/2$ in increments of $2\pi/90$ radians. (Orthographic projections.)

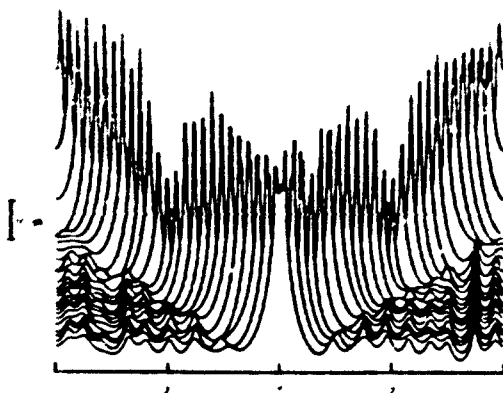


Fig. 15 Case 3A. $r = 1$

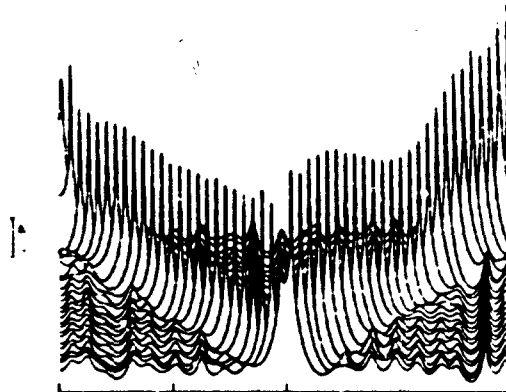


Fig. 18 Case 3D. $r = \sqrt{100}$

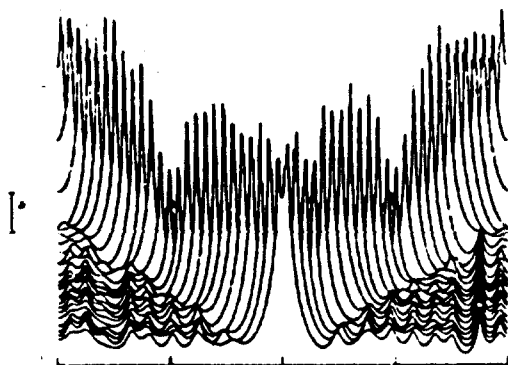


Fig. 16 Case 3B. $r = \sqrt{2}$

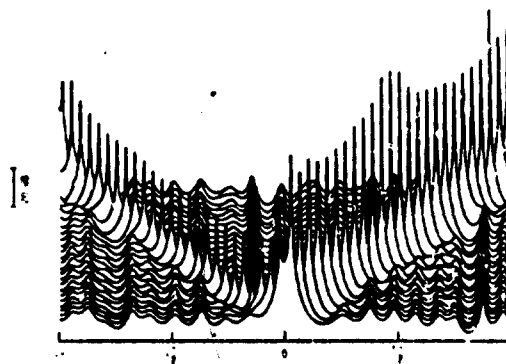


Fig. 19 Case 3E. $r = \sqrt{1000}$

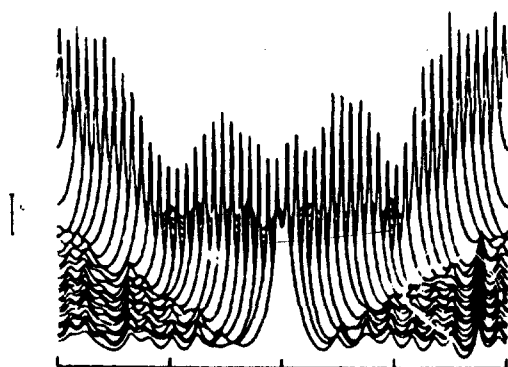


Fig. 17 Case 3C. $r = \sqrt{10}$

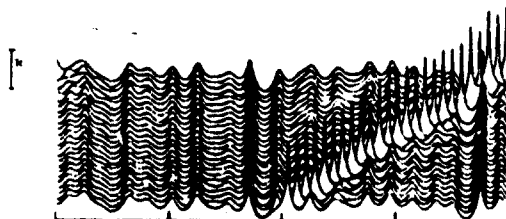


Fig. 20 Case 3F. $r = \dots$

Figs. 15-20. Cases 3A-F. Estimated spectral power vs. angular frequency. $\Delta\omega_{mid} = 2\pi$. $N=101$. $M = 24$. $\omega_1 = -\omega_2$ stepped from $2\pi \times 0.0125$ to $2\pi \times 0.4925$ in increments of $2\pi \times 0.02$. (Orthographic projections.)

A TWO-DIMENSIONAL MAXIMUM ENTROPY SPECTRAL ESTIMATOR

Salim Roucos
Dept. of Electrical Engineering
University of Florida
Gainesville, FL 32611

D.G. Childers
Dept. of Electrical Engineering
University of Florida
Gainesville, FL 32611

ABSTRACT

Using the ideas from one-dimensional (1-D) maximum entropy spectral estimation, we derive a 2-D spectral estimator by extrapolating the 2-D sampled autocorrelation (covariance) function. The maximum entropy method used here maximizes the entropy of a set of random variables. The extrapolation (prediction) process under this maximum entropy condition is shown to correspond to the most random extension or equivalently to the maximization of the mean square prediction error when the optimum predictor is used. The 2-D extrapolation must be terminated by the investigator. The Fourier transform of the extrapolated autocorrelation function is our 2-D spectral estimator. Using this method, one can apply windowing prior to calculating the spectral estimate.

A specific algorithm for estimating the 2-D spectrum is presented and its computational complexity is estimated. The algorithm has been programmed and computer examples are presented.

I. INTRODUCTION

For time series or one-dimensional (1-D) data, one may consider the maximum entropy (ME) formulation [1] as a procedure for deriving a spectral estimator such that the entropy of the signal is maximized subject to the constraint that the spectral estimate is consistent with the known autocorrelation values. This spectral estimator is the same as that derived by autoregressive or linear predictive methods [2]. Some authors have considered this criterion as a smoothing or a whitening process [3-8], an interpretation which has been advanced for both the ME and linear predictive (LP) methods.

For 1-D data the solution to the ME spectral estimation problem is achieved via a polynomial spectral factorization. However, for the 2-D case such an approach is not fruitful since 2-D polynomials cannot in general be factored, i.e., the Fundamental Theorem of Algebra does not hold. For this reason there has been some concern in the literature about the existence of a

2-D ME spectral estimator. This is to say, while it seemed quite natural to extend Shannon's 1-D entropy ideas to 2-D, there was no assurance that a 2-D ME spectral estimator existed and if it did, was it unique?

Barnard and Burg [10] originally hypothesized such a 2-D ME spectral estimator, giving the expression for the estimator and suggesting that it could be derived via a Lagrangian multiplier approach. Ables [3] agreed with this approach and suggested a modified constraint on the 2-D ME function which would account for noisy data. Ponsonby [11] attempted to derive the 2-D ME spectral estimator following the suggestions in [3,10] but was unable to determine a closed form analytical solution due to the nonlinear integral equations. An iterative numerical solution was developed by Wernecke and D'Addario [5] for application to radio astronomic data. Wernecke [4] justifies the ME entropy reconstruction model in terms of its smoothing properties. From the viewpoint of image processing an alternate model has been used [12-15].

The 2-D ME spectral estimator is not the result of a simple extension of the 1-D solution. In fact, Woods [16] has provided a constructive proof of the existence and uniqueness of a 2-D discrete Markov random field which agrees with known correlation values in a nearest neighbor array, thus placing the 2-D spectral estimator on a firm theoretical footing. As might be anticipated the derivation and the algorithm are considerably more complicated than that for the 1-D case. The corresponding spectrum is the 2-D ME spectrum [10]. Wood's algorithm is in some ways an improvement on Ong's [17] ME algorithm. However, the computational load is still quite extensive for even small (3 X 3) data arrays (a 5 X 5 correlation array). The computation time is dependent upon the degree of approximation used in calculating the spectrum; typical values for Wood's algorithm are 5 minutes and 20 minutes for an IBM 360/44 for two different approximations to the maximum entropy spectrum for a 5 X 5 correlation array.

In our search to find a more efficient 2-D spectral estimator we decided to extrapolate the sampled autocorrelation function, considered as that of a set of random variables, under the maximum entropy condition and then Fourier transform the extrapolated autocorrelation function to obtain the estimate of the spectral density of the random field.

Some of the properties of our estimator are not investigated but its existence and uniqueness under certain conditions are proven. The argument that this estimator will yield a high resolution spectral estimator is based on the analogy with the one-dimensional case, where a maximum entropy extension of the autocorrelation function resulted in a better spectral estimation than an extension of the autocorrelation function such as that obtained by appending zeros to artificially extend the duration of the autocorrelation function.

We illustrate our results with computed examples and show that our algorithm appears to be more efficient, computationally, than that of Woods [16].

II. OVERVIEW OF OUR APPROACH

The 1-D algorithm does not require that the actual extrapolation be performed, rather the spectral estimate is obtained directly from a system of linear equations in the known correlation values (Yule-Walker equations).

Our approach for the 2-D case is to formulate the problem such that we actually perform an extrapolation of the known 2-D autocorrelation values. At each step of the extrapolation, we hypothesize that an additional sample of the random data field is available. However, its probability structure as a random variable with respect to the previous set of samples is not completely known. The 2-D ME extension used in the paper is that the entropy of the new set of random variables is maximized subject to the constraint that these random variables are samples from a stationary random field. By successively choosing the location of the hypothetical samples, we are defining a one to one map between the set of points in the (x,t) plane to the set of integers. This defines an order for the random variables. (In particular this order is implicit in the autocorrelation matrix of the random variables.) This map will be called the spatio-temporal "1-D" extension path. The uniqueness of the spectral estimation with respect to the choice of this path is unsolved.

The autocorrelation extension process is reduced to a linear prediction model, in its general sense, where the new random variable in the "1-D" extension path is predicted from the existing (known) random variables. The corresponding prediction error of the model is maximized so as to produce the most random extension of the autocorrelation matrix.

In the 1-D case, once the order of the LP or AR model is selected, the solution to the linear difference equations is determined by the initial conditions. If one attempts to derive a higher order model from the extended autocorrelation matrix, the coefficients of the new model reduce identically to the coefficients of the original lower order model.

For the 2-D case the order of the model is dictated by the number of available samples of the random data field. Therefore, the model order increases with each extension step, yielding new prediction filter coefficients at each iteration. In contrast with the 1-D maximum entropy extension method, the extension for the 2-D case must be terminated by the investigator.

Note specifically that the extension process described here deals with the autocorrelation values where as the 1-D Burg technique [10] is applied to the actual data.

A. Formulation of the Extension

Let the real random data field be denoted as $f(x,t)$ such that the process is Gaussian with zero mean and stationary in time and space. The autocorrelation function is given by

$$R(\rho,\tau) = E [f(x,t) f(x+\rho, t+\tau)] \quad (1)$$

where E denotes the ensemble average. This function is also the autocovariance since we have assumed a zero mean process. This function is symmetric and positive semidefinite. We may, of course, generalize to complex random data fields.

We denote our matrices as

$$\Lambda_N = E \begin{bmatrix} f(x_1, t_1) \\ f(x_2, t_2) \\ \vdots \\ f(x_N, t_N) \end{bmatrix} \cdot [f(x_1, t_1) \ f(x_2, t_2) \dots f(x_N, t_N)] \quad (2)$$

where the ordering of Λ_N is arbitrary in contrast with the uniform sampling 1-D case. Because of this the matrices are not necessarily Toeplitz even though the process is stationary. However, $R_{ij} = R_{ji}$, thus, Λ_N is symmetric and positive semidefinite.

The power spectrum is defined as [18,19]

$$P(f, k) = \iint R(\rho, \tau) e^{-2\pi i(ft - kx)} dt dx \quad (3)$$

We assume the 2-D ($N \times N$) covariance (autocorrelation) matrix Λ_N , is known and obtained by sampling the spatio-temporal field at $[(x_1, t_1), (x_2, t_2), \dots, (x_N, t_N)]$. We ask, what should the extended covariance matrix be if an additional sample of the random field at location (x_{N+1}, t_{N+1}) is available? The stationarity of the random field is maintained and the entropy of the new set of $(N+1)$ random variables is maximized with respect to the unknown autocovariance values.

We illustrate the extrapolation by an example.

Example

Imagine the random field is sampled at four locations $[(0,0), (0,1), (1,1), (1,0)]$ as shown in Figure 1, then

$$\Lambda_4 = \begin{bmatrix} R_{11} & R_{12} & R_{13} & R_{14} \\ R_{21} & R_{22} & R_{23} & R_{24} \\ R_{31} & R_{32} & R_{33} & R_{34} \\ R_{41} & R_{42} & R_{43} & R_{44} \end{bmatrix} = \begin{bmatrix} R(0,0) & R(0,1) & R(1,1) & R(1,0) \\ R(0,1) & R(0,0) & R(1,0) & R(1,-1) \\ R(1,1) & R(1,0) & R(0,0) & R(0,1) \\ R(1,0) & R(1,-1) & R(0,1) & R(0,0) \end{bmatrix}$$

We assume that Λ_4 is known.

Note that an element R_{ij} of the above matrix corresponds to sampling the autocorrelation function $R(m, n)$ at a point dictated by the ordering of the

random variable via the "1-D" extension path. These samples are shown in Figure 2.

Now suppose an additional hypothetical sample at (0,2) in the (x,t) plane is available (as shown by the open circle in Figure 1), then

$$\Lambda_5 = \begin{bmatrix} & & & & R_{15} \\ & & & & R_{25} \\ & & & & R_{35} \\ & & & & R_{45} \\ R_{51} & R_{52} & R_{53} & R_{54} & R_{55} \end{bmatrix} = \begin{bmatrix} & & & & R(0,2) \\ & & & & R(0,1) \\ & & & & R(-1,1) \\ & & & & R(-1,2) \\ x & x & x & x & R(0,0) \end{bmatrix} \quad (5)$$

The elements in the last row are determined by symmetry. Only $R(0,2)$ and $R(-1,2)$ are undetermined.

The entropy of this set of random variables is $\log |\Lambda_5|$, hence we want to maximize $|\Lambda_5|$. Letting $\underline{R}_e = (R_{51}, R_{52}, R_{53}, R_{54})$, we have

$$|\Lambda_5| = \begin{vmatrix} \Lambda_4 & \underline{R}_e \\ \underline{R}_e^T & R(0,0) \end{vmatrix} = |\Lambda_4| \left| -\underline{R}_e^T \Lambda_4^{-1} \underline{R}_e + R(0,0) \right| \quad (6)$$

Since Λ_4 is fixed, we need to maximize $R(0,0) - \underline{R}_e^T \Lambda_4^{-1} \underline{R}_e$ with respect to $R(0,2)$ and $R(-1,2)$. This will be shown to correspond to maximizing the prediction error of an optimum linear predictor in the next section.

III. THE "MOST RANDOM" EXTENSION

We may illustrate our idea by reconsidering the previous example, but we number the random variables in the (x,t) plane as $[v_i]$, $i = 1, 2, 3, 4$ where the correspondence between v_i and its location on the (x,t) plane is arbitrary. We let v_5 denote the open circle random variable at (0,2). The linear prediction estimate of v_5 in terms of the other variables is

$$\hat{v}_5 = \sum_{i=1}^4 a_i v_i \quad (7)$$

The prediction error is

$$e_5 = v_5 - \hat{v}_5 = \sum_{i=1}^5 -a_i v_i \text{ where } a_5 = -1 \quad (8)$$

We determine the coefficients, a_i , such that $E[e_5^2] = \overline{e_5^2}$ is minimized.

By orthogonality

$$\begin{aligned} \overline{e_5 v_i} &= 0 \\ &= -\sum_{j=1}^5 a_j \overline{v_j v_i} = -\sum_{j=1}^5 a_j R_{ij}, \quad i = 1, 2, 3, 4 \end{aligned} \quad (9)$$

In matrix notation the solution becomes

$$\underline{A} = \underline{\Lambda}_4^{-1} \underline{R}_e \quad (10)$$

where $\underline{A}$ is the column vector (a_1, a_2, a_3, a_4) , $\underline{\Lambda}_4$ is assumed nonsingular and $\underline{R}_e$ is the column vector $(R_{15}, R_{25}, R_{35}, R_{45})$.

The mean square prediction error when using the optimum predictor $\underline{A} = \underline{\Lambda}_4^{-1} \underline{R}_e$ is given by

$$e_0^2 = R_{55} - \underline{R}_e^T \underline{A} = R(0,0) - \underline{R}_e^T \underline{\Lambda}_4^{-1} \underline{R}_e \quad (11)$$

which gives the variance of the error as a function of the autocorrelation function when an optimum predictor is used. Note that $\underline{R}_e$ is not completely known and the optimum predictor depends upon $\underline{R}_e$. Hence if we choose $\underline{R}_e$ such that the corresponding optimum predictor has the largest prediction error of all other choices of $\underline{R}_e$ consistent with the stationarity constraint, then this extension is exactly the maximum entropy extension as described in the previous section.

IV. EXISTENCE AND UNIQUENESS OF THE SOLUTION

Our method is basically one of extrapolation which we have shown reduces to the maximization of the determinant,

$$\left| \underline{\Lambda}_{N+1} \right| = \begin{vmatrix} \underline{\Lambda}_N & \underline{R}_e \\ \underline{R}_e^T & R(0,0) \end{vmatrix} \quad (12)$$

subject to the constraints that some components of $\underline{R}_e$ are known, say m of them. Rearrange $\underline{R}_e$ such that the unknown components occur at the top of $\underline{R}_e$ ($\underline{R}_e$ is a column vector) to obtain,

$$\begin{vmatrix} \underline{\Lambda}_N^r & \underline{R}_e^u \\ & \underline{R}_e^k \\ \underline{R}_e^{uT} & \underline{R}_e^{kT} & R(0,0) \end{vmatrix} \quad (13)$$

where the superscripts r,u,k stand for rearranged, unknown, and known, respectively. This determinant is equal to,

$$\left| \Lambda_N^r \right| \left| R(0,0) - \begin{bmatrix} \underline{R}_e^u & \underline{R}_e^k \end{bmatrix} \Lambda_N^{r-1} \begin{bmatrix} \underline{R}_e^u \\ \underline{R}_e^k \end{bmatrix} \right| \quad (14)$$

which we wish to maximize with respect to the components of $\underline{R}_e^u$; since $e^2 = R(0,0) - \underline{R}_e^T \Lambda_N^{-1} \underline{R}_e \geq 0$ is required for an autocorrelation matrix extension, and $R(0,0)$ is fixed, we have to minimize

$$\begin{bmatrix} \underline{R}_e^u & \underline{R}_e^k \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \underline{R}_e^u \\ \underline{R}_e^k \end{bmatrix} \quad (15)$$

where we properly partitioned Λ_N^{r-1} ; expanding, we minimize

$$\left[\underline{R}_e^u \quad \underline{R}_e^k \right] \begin{bmatrix} A \underline{R}_e^u + B \underline{R}_e^k \\ C \underline{R}_e^u + D \underline{R}_e^k \end{bmatrix} \quad (16)$$

with respect to the elements of $\underline{R}_e^u$. This gives

$$\underline{R}_e^u = -A^{-1} B \underline{R}_e^k \quad (17)$$

Hence $\underline{R}_e^u$ exists and is unique as long as we can assume that Λ_N is positive definite. But at any step of the extrapolation we have,

$$\left| \Lambda_{N+1} \right| = \left| \Lambda_N \right| \left| R(0,0) - \underline{R}_e^T \Lambda_N^{-1} \underline{R}_e \right| \quad (18)$$

Thus Λ_{N+1} will be positive definite if the prediction error variance $R(0,0) - \underline{R}_e^T \Lambda_N^{-1} \underline{R}_e$ is positive. Therefore if the covariance matrix at step N is positive definite the extrapolation will yield a positive definite extension, provided the error variance is not null.

V. THE MAXIMUM ENTROPY TWO-DIMENSIONAL SPECTRAL ESTIMATOR

An efficient computational algorithm follows noting that Λ_{M+1}^{-1} can be recursively computed from Λ_M^{-1} .

We assume Λ_N to be known and positive definite.

1. Compute Λ_N^{-1} , let $M = N$.
2. Generate the position of a hypothetical new sample (assuming an extrapolation path).

3. Find which components of $\underline{R}_e$ are known by stationarity.
4. Rearrange Λ_M^{-1} to obtain $(\Lambda_M^r)^{-1}$.
5. Compute $\underline{R}_e^u = -A^{-1} [B \underline{R}_e^k]$ where A and B are submatrices of $(\Lambda_M^r)^{-1}$.
6. Compute $\Lambda_M^{-1} \underline{R}_e$ and the prediction error $R(0,0) = \underline{R}_e^T \Lambda_M^{-1} \underline{R}_e$. If the prediction error is zero, stop.
7. Compute Λ_{M+1}^{-1} according to the recursion relations*.
8. Let $M = M+1$.
9. Is this extension adequate (i.e., is M large enough)?
If not, go to step 2.
10. Construct from the MXM autocorrelation matrix found in step 5 an LXL autocorrelation array. This LXL array is the extended autocorrelation function with maximum lag of $\frac{L-1}{2}$ in each dimension.
11. Take a two-dimensional FFT of the LXL array constructed in step 10.

Steps 9, 10, and 11 above need some further explanation. Suppose one decides that the final extended autocorrelation (covariance) array is to be of order LXL; this is the LXL array that is transformed in step 11. However, M in step 9 must be much larger than L. The reason for this is that we must select the autocorrelation values with the appropriate lags from the MXM matrix to construct the LXL array. This is best illustrated by examining Equation (4) again. Here Λ_4 is a 4X4 matrix with the various R_{ij} entries. However, note that the maximum autocorrelation lag for this matrix is only unity in any one direction. If a matrix with autocorrelation values for larger lags is desired then the extension process must be continued. In general one can convince oneself that $M = (\frac{L-1}{2} + 1)^2$ and that L is odd because of the symmetry of the autocorrelation function.

The computational complexity of this algorithm may be estimated as follows:

- i) less than $3M^2$ operations are required at step 5
- ii) less than $2M^2$ operations are required at step 6
- iii) less than $6M^2$ operations are required at step 7

This gives less than $11M^2$ operations per iteration. In the case where we start with $N=5$ and extend Λ_N to a 121×121 matrix, then less than 19×10^6 operations are needed not including the FFT nor the necessary row and column interchange operations. This corresponds to an LXL array of 21×21 . For the case of a 3×3 data (auto correlation) array (a 5×5 covariance matrix) the computation time on an Amdhal 470 was 10 sec for extrapolation to a 21×21 (LXL) covariance array and 54 sec for extrapolation to a 29×29 (LXL) covariance array including the appropriate FFT calculations. The storage required is 150 Kbytes (single precision) which exceeds Wood's 56 Kbytes but our program appears to execute faster and does require additional storage for the FFT.

We illustrate this algorithm in Figures 3, 4, and 5 where the spectra of the actual 29×29 (LXL) covariance data, the 5×5 covariance data (extended with zeros only), and the algorithmically extended data are shown. Only one quadrant of the spectrum is plotted for convenience. The original data field are samples from two sinewaves of frequencies

$$f_t = 0.207 \text{ Hz}, f_x = 0.318 \text{ (cm)}^{-1} \text{ and} \\ f_t = 0.41 \text{ Hz}, f_x = 0.11 \text{ (cm)}^{-1} \text{ respectively}$$

The noise level is $\sigma^2 = 0.8$ and the signal power is 1 for each sinewave.

These figures show three contour levels at 97%, 50%, and 10% of the normalized peak value in each power spectrum. Note that the maximum entropy extended spectrum has well defined peaks at the 50% level, i.e., the peaks are easily resolved. However, the peaks in the 5×5 covariance array (extended with zeros to 29×29) are not resolved at the 50% (1/2 power) level.

VI. CONCLUSION

The proposed extension of the autocorrelation (covariance) function under the maximum entropy condition is such that at the M^{th} step $K+M$ samples (at the locations specified by the initial available autocorrelation values and the particular choice of the extrapolation path) of the random field satisfy a system of partial difference equations, i.e., linear prediction equations. However, the prediction error is only orthogonal to the samples involved in the prediction and is not necessarily white. Whereas in the one-dimensional case the maximum entropy extension of the autocorrelation function results in a predictor whose prediction error is orthogonal to all past samples of the random process, i.e., the error is white.

One interesting uniqueness problem remains unsolved, namely, the dependence of the extended autocorrelation function on the extrapolation path. If the arbitrary choice of an extrapolation path does not yield a unique maximum entropy spectrum, then it is possible to determine constraints on the selection of an extrapolation path which may yield a unique spectrum.

REFERENCES

- [1] J.P. Burg, 1967, "Maximum entropy spectral analysis," presented at the 37th meeting Soc. of Exploration Geophysicists, Oklahoma City, OK.
- [2] A. van der Bos, 1971, "Alternative interpretation of maximum entropy spectral analysis," IEEE Trans. Inform. Theory, Vol. 17, pp. 493-494.
- [3] J.G. Ables, 1974, "Maximum entropy spectral analysis," Astron. Astrophys. Suppl., vol. 15, pp. 353-393.
- [4] S.J. Wernecke, Sept./Oct. 1977, "Two-dimensional maximum entropy reconstruction of radio brightness," Radio Sc. Vol. 12, No. 5, pp. 831-844.
- [5] S.J. Wernecke and L.R. D'Addario, April 1977, "Maximum entropy image reconstruction," IEEE Trans. Computers, Vol. C-26, pp. 351-364.
- [6] J. Makhoul, 1975, "Linear prediction: a tutorial review," Proc. IEEE, Vol. 63, pp. 561-580.
- [7] J.D. Markel and A.H. Gray, Jr., 1976, "Linear Prediction of Speech," New York: Springer-Verlag.
- [8] C.E. Shannon and W. Weaver, 1964, "The Mathematical Theory of Communication," Urbana: The University of Illinois Press, pp. 93-95 (paper back edition).
- [9] S. Treitel, P.R. Gutowski, and E.A. Robinson, 1977, "Empirical spectral analysis revisited," to appear in J.H.J. Miller (Ed.), Topics in Numerical Analysis, Vol. 3, New York: Academic Press.
- [10] T. Barnard and J.P. Burg, May 1969, "Analytical studies of techniques for the computation of high-resolution Wavenumber spectra," Advanced Array Research Report #9, Texas Instruments, Inc., (AD 855345).
- [11] J.E.B. Ponsonby, 1973, "An entropy measure for partially polarized radiation and its application to estimating radio sky polarization distributions from incomplete 'Aperture Synthesis' data by the maximum entropy method," Mon. Not. R. Astr. Soc., Vol. 163, pp. 369-380.
- [12] B.R. Frieden, April 1972, "Restoring with maximum likelihood and maximum entropy," J. Opt. Soc. Am., Vol. 62, pp. 511-518
- [13] B.R. Frieden, 1975, "Image enhancement and restoration," in T.S. Huang (Ed.), Topics in Applied Physics, Vol. 6, New York, Springer-Verlag, pp. 177-248.
- [14] B.R. Frieden and D.C. Wells, Jan. 1978, "Restoring with maximum entropy, III. Poisson sources and backgrounds," J. Opt. Soc. Am., Vol. 68, pp. 93-103.

- [15] R. Kikuchi and B.H. Soffer, Dec. 1977, "Maximum entropy image restoration. I. The entropy expression," J. Opt. Soc. Am., Vol. 67, pp. 1656-1665.
- [16] J.W. Woods, Sept. 1976, "Two-dimensional Markov spectral estimation," IEEE Trans. Infor. Theory, Vol. 1T-22, pp. 552-559.
- [17] C. Ong, April 1971, "An investigation of two new high-resolution two dimensional spectral estimate techniques," Long Period Array Processing Report #1, Texas Instruments, Inc.
- [18] J. Capon, August 1960, "High-resolution frequency-wavenumber spectrum analysis," Proc. IEEE, Vol. 57, pp. 1408-1418.
- [19] C.S. Halpeny and D.G. Childers, June 1975, "Composite wavefront decomposition via multidimensional digital filtering of array data," IEEE Trans. on Circuits and Systems, Vol. CAS-22, pp. 552-562.
- [20] D.E. Symlie, G.K.C. Clarke, and T.J. Ulrych, 1973, "Analysis of irregularities in the earth's rotation," in B. Adler, S. Fernback, and M. Rotenberg (Eds.), Methods in Computational Physics, Vol. 13, B.A. Bolt (Ed.) New York: Academic Press, Inc., pp. 391-430.
- [21] T.J. Ulrych and T.N. Bishop, February 1975, "Maximum entropy spectral analysis and autoregressive decomposition," Rev. Geophys. and Space Phys., Vol. 13, pp. 183-200

ACKNOWLEDGEMENT

This work was supported in part by the Naval Coastal Systems Laboratory, Panama City, Florida.

*The resursion relation is

$$\Lambda_{M+1}^{-1} = \begin{bmatrix} \Lambda_M^{-1} + \frac{(\Lambda_M^{-1} \underline{R}_e)(\Lambda_M^{-1} \underline{R}_e)^T}{R(0,0) - \underline{R}_e^T \Lambda_M^{-1} \underline{R}_e} & \frac{-\Lambda_M^{-1} \underline{R}_e}{R(0,0) - \underline{R}_e^T \Lambda_M^{-1} \underline{R}_e} \\ \frac{-(\Lambda_M^{-1} \underline{R}_e)^T}{R(0,0) - \underline{R}_e^T \Lambda_M^{-1} \underline{R}_e} & \frac{1}{R(0,0) - \underline{R}_e^T \Lambda_M^{-1} \underline{R}_e} \end{bmatrix}$$

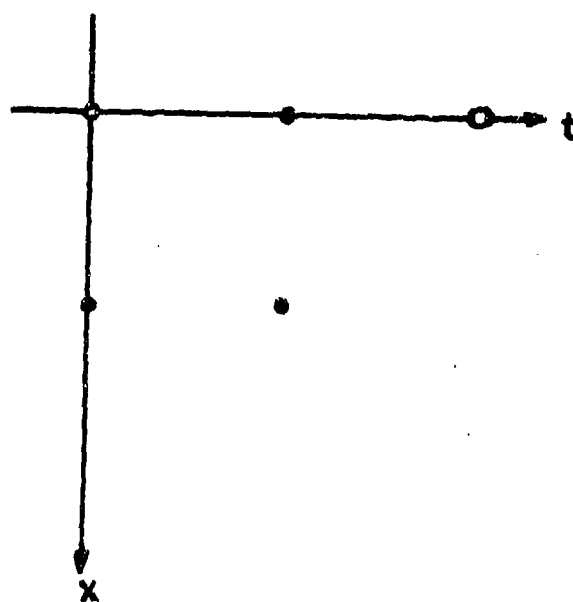


Figure 1. Spatio-temporal location of random data field samples.

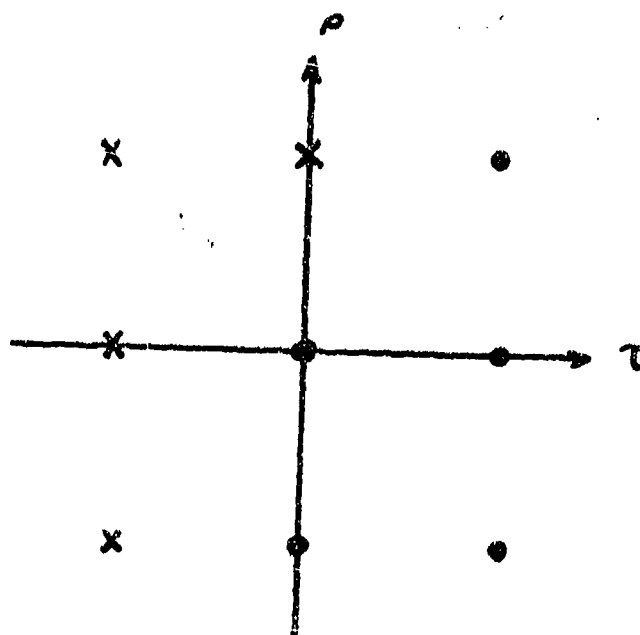


Figure 2. Spatio-temporal location of known autocorrelation (covariance) values.

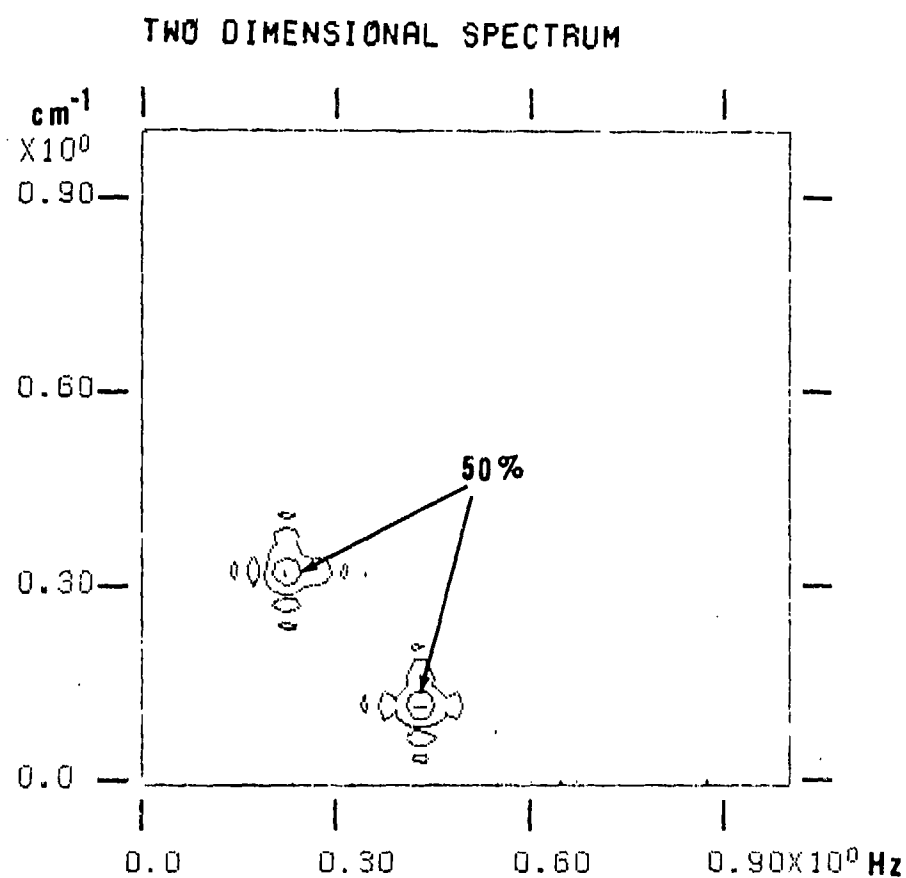


Figure 3. Spectrum of two sinusoids calculated by transforming the truncated (29X29) known spatio-temporal autocorrelation matrix. Contours are at 97%, 50%, and 10% of the maximum peak spectral value.

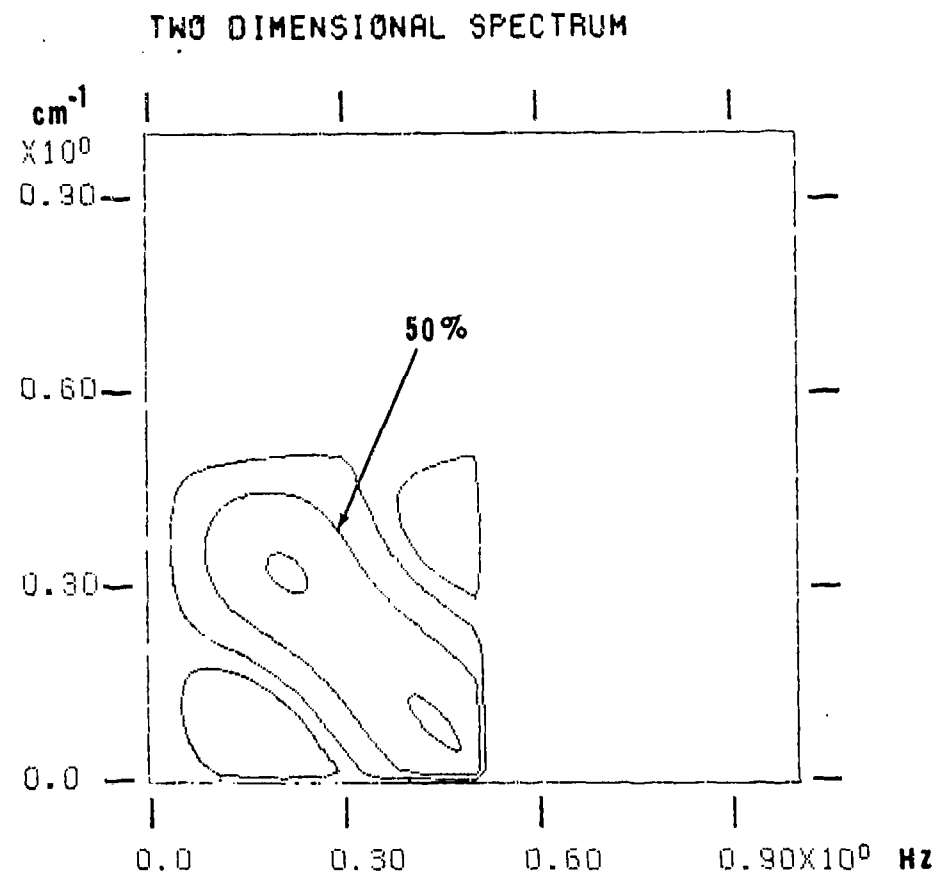


Figure 4. Spectrum of two sinusoids calculated by transforming the truncated (5X5) known spatio-temporal autocorrelation matrix with zeros appended to extend the autocorrelation matrix to 29X29. Contours are the same as for Figure 3.

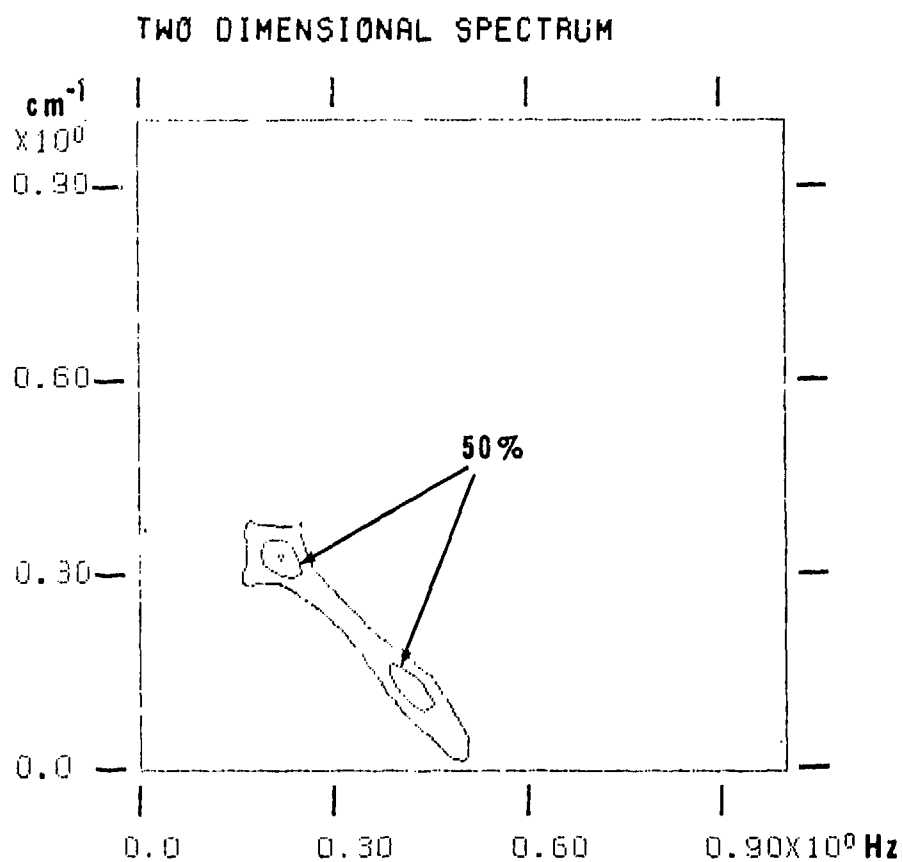


Figure 5. Spectrum of two sinusoids calculated via the algorithm described in the text, by extending the 5X5 autocorrelation matrix to 29X29. Contours are the same as for Figure 3.

194 - BLANK

Spectral Estimation and Signal Extrapolation in One and Two Dimensions

Anil K. Jain*

Signal and Image Processing Laboratory
Department of Electrical and
Computer Engineering
University of California
Davis, California 95616

ABSTRACT

In this paper, we consider extrapolation and spectral estimation of discrete time (or space) signals in one (or two) dimensions. The paper is divided into two parts. In the first part, we present some recent results [6] for extrapolation of bandlimited discrete time signals. These results show the relationship between several recently reported extrapolation algorithms by Papoulis, [1], Sabri and Steenaart [2], Cadzow [3] and others [4,5] and also yield some new algorithms. In the second part, first we show how the one dimensional algorithms could be extended to two dimensions. Then we introduce a two-dimensional semi-causal prediction algorithm for spectral estimation of discrete random fields. This algorithm requires solution of linear equations and realizes a particular minimum variance ARMA model for the spectral estimate. It gives superior resolution compared to two-dimensional (FFT based) periodograms or the two-dimensional and autoregressive (AR) spectral estimates.

PART ONE: EXTRAPOLATION OF DISCRETE TIME BANDLIMITED SIGNALS

1.1 Problem Definition

A discrete signal $y(k)$, $k = 0, \pm 1, \pm 2, \dots$, is called bandlimited if its Fourier transform

$$Y(f) = \sum_{k=-\infty}^{\infty} y(k) \exp(-j2\pi kf), \quad -\frac{1}{2} \leq f \leq \frac{1}{2} \quad (1)$$

satisfies the relation

*Research supported in part by the Army Research Office, Durham N.C. under grant No. DAAG29-78G0206 and in part by RADC under a Multi-effort Post Doctoral program. Paper presented at RADC Spectrum Estimation Workshop, Rome, N.Y., Oct. 1979.

$$Y(f) = 0, \frac{1}{2} > |f| > \sigma \quad (2)$$

This implies $y(k)$ comes from bandlimited continuous signal which is over-sampled with respect to its Nyquist rate. This occurs quite often when a system observes signals over a wide bandwidth. We are given a set of time limited, noise free observations

$$z(k) = \begin{cases} y(k) & , -M \leq k \leq M \\ 0 & , \text{otherwise,} \end{cases} \quad (3)$$

Given $\{z(k)\}$, the problem is to find an estimate of $y(k)$ outside the interval $[-M, M]$. We define the infinite vector

$$y = [\dots y(-k) \dots y(-1), y(0), y(1), \dots, y(k) \dots]^T \quad (4)$$

a bandlimiting operator L , and a time-limiting operator W , as infinite matrices,

$$L = \{\ell_{i,j}\} , \ell_{i,j} = \frac{\sin 2\pi(i-j)\sigma}{\pi(i-j)} , i, j = 0, \pm 1, \pm 2 \dots \quad (5)$$

$$W = \{w_{i,j}\} , w_{i,j} = \begin{cases} 1, & i=j, -M \leq i, j \leq M \\ 0 & , \text{otherwise} \end{cases} \quad (6)$$

1.2 Properties of L

1. Let S be a $(2M+1) \times \infty$ matrix operator whose elements are

$$s(i,j) = \begin{cases} 1, & j = \pm 1, \pm 2, \dots, \pm M \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

Basically, S maps $(2M+1)$ elements from an infinite vector into a finite vector. Consider the $(2M+1) \times (2M+1)$ matrix $\hat{L} \triangleq SLS^T$ and the matrix $S^T S \triangleq W$. The operator W replaces the elements outside $[-M, M]$ of an infinite vector by zeros. S^T extrapolates the elements of a $(2M+1) \times 1$ vector by zeros.

2. The operator L is idempotent, i.e., $L^2 = L$. (8)

3. For every $M < \infty$, $\hat{L}$ is positive definite. Moreover, all the eigenvalues of $\hat{L}$ lie in the interval $(0,1)$ i.e., $0 < \lambda(\hat{L}) < 1$, $M < \infty$.

For proofs of these and other properties see Jain and Ranganath [6].

1.3 Matrix Formulation of the Extrapolation Problem

Let $y(k)$, $k=0, +1, \dots$ be a discrete-time, bandlimited signal as defined in (1) and (2). If z denotes a $(2M+1) \times 1$ vector of the observations, then $z = Sy$. Since y is bandlimited, it must satisfy $Ly = y$, so that we can write

$$z = SLy. \quad (9)$$

Now, if we define an $\infty \times (2M+1)$ matrix

$$H = SL \quad (10)$$

then (9) becomes

$$z = Hy \quad (11)$$

where z is a $(2M+1) \times 1$ vector and y is an infinite vector. The extrapolation problem is now, simply, to find an estimate of y given z . Equation (11) as such does not have a unique solution because H is a rectangular matrix. In other words, for discrete bandlimited signals given over a finite interval, it is not possible to extrapolate uniquely. One way to make the solution of (11) unique is to look for minimum norm least squares (MNLS) solution defined as

$$y^+ = \min \{ \|y\|^2; H^T H y = H^T z \}, \quad \|x\|^2 = x^T x$$

Note that a solution of $H^T H y = H^T z$ minimizes the least squares error $\|z - Hy\|^2$. y^+ is that least squares solution which has the minimum norm. Conceptually, the problem is quite straightforward now and a large number of algorithms are available to find y^+ . Using the definition of H [see (10)] and the properties of L , these algorithms assume simple form in many cases and will be briefly stated.

We note that a continuous bandlimited signal given over any finite interval is analytic and can, therefore, be extrapolated uniquely to its original values outside the given interval.

1.4 Iterative Extrapolation Algorithms: Let y_n represent an estimate of y at n th iteration. Following [6], we can write down the gradient and the conjugate gradient algorithms as follows.

1.4.1 The Gradient Method

$$y_{n+1} = y_n + \epsilon H^T (z - Hy_n), \quad y_0 = 0, \quad 0 < \epsilon < 2/\lambda_{\max}(H^T H) \quad (12)$$

Using (10) and the properties of L , this equation simplifies to give

$$y_{n+1} = \epsilon f_1 + (I - \epsilon LW)y_n, y_0 = 0; f_1 \triangleq H^T z = LS^T z \quad (13)$$

Convergence is achieved for $0 < \epsilon \leq 2/\lambda_{\max}(\hat{L})$. Now it can be shown by induction that since f_1 is bandlimited i.e., $Lf_1 = f_1$, each y_n is also bandlimited. Hence, (13) can be written as

$$y_{n+1} = \epsilon f_1 + L(I - \epsilon W)y_n \quad (14)$$

For $\epsilon=1$, this becomes the discrete version of Papoulis iterative algorithm [1] reported in [2]. However, the convergence would be best (for constant ϵ) if we let

$$\epsilon = \epsilon_{\text{opt}} = 2/((\lambda_{\min}(H^T H) + \lambda_{\max}(H^T H)) = 2/\lambda_{\max}(\hat{L}) \geq 2 \quad (15)$$

Further improvement in convergence is obtained if we go to the steepest descent [6] or the conjugate gradient algorithm. Thus, it is seen that Papoulis' method, based on a successive energy reduction method (see Gerchberg [4]), is a special case in the class of one step gradient methods. It is easily shown that y_n converges to y^+ [6].

1.4.2 Conjugate Gradient Method

Using the property $L^2 = L$, the gradient vectors $\{g_k\}$ and the conjugate direction vectors $\{d_k\}$ can be shown to be bandlimited and this algorithm becomes

$$\left. \begin{aligned} y_{k+1} &= y_k + \alpha_k d_k, & y_0 &= 0 \\ d_{k+1} &= -g_{k+1} + \beta_k d_k, & d_0 &= -g_0 \\ g_k &= LWy_k - LS^T z = g_{k-1} + \alpha_{k-1} LWd_{k-1} \\ \beta_k &= \frac{\sum_{n=-M}^M g_{k+1}(n) d_k(n)}{\sum_{n=-M}^M (d_k(n))^2} \\ \alpha_k &= \frac{\sum_{n=-M}^M d_k(n) (y_k(n) - z(n))}{\sum_{n=-M}^M (d_k(n))^2} \end{aligned} \right\} \quad (16)$$

This is a two step gradient algorithm and iteration by iteration, has better convergence than the ordinary gradient method discussed earlier.

1.5 Generalized Inverse Extrapolation Filter

The generalized inverse of H is given by

$$H^+ = H^T(HH^T)^{-1} = LS^T(SLS^T)^{-1} = LS^T\hat{L}^{-1} \quad (17)$$

which exists since $\hat{L}$ is nonsingular. Hence we can directly evaluate the MNLS estimate of y as

$$y^+ = H^+z \quad (18)$$

In practice, $\hat{L}$ could be quite ill-conditioned depending on σ and M and has to be stabilized [6]. This can be done by either using a singular value expansion in which terms corresponding to small eigenvalues of $\hat{L}$ are discarded or by adding a small positive quantity to the diagonal terms of $\hat{L}$ in (17). The generalized inverse of (17) is called the Extrapolation Matrix in the context of signal extrapolation. It has appeared in the context of image restoration, see e.g., Helstrom, Rino, Jain and others [7-9]. Recently, it has also been derived by Cadzow by a different procedure for signal extrapolation. Another extrapolation matrix (whose size is infinite, if we extrapolate the signal to infinity) has been suggested by Sabri et al [2]. In [6] we show that this matrix does not exist and its finite approximation (used in [2]) is ill-conditioned.

1.6 Discrete Prolate Spheroidal Wave Functions and Singular Value Expansion

It is known that a continuous band-limited signal can be extrapolated outside its observation interval, exactly, via the PSWF expansion [10]. In the case of discrete signals, a similar expansion is possible when we consider the minimum norm least squares extrapolated estimate via the singular value expansion of the matrix H [6].

Papoulis and Bertram [11], Slepian [5] Jain and Ranganath [6], Algazi [12], and others have studied the properties and applications of these functions in digital signal processing. For singular value decomposition of $H = SL$, we consider the eigenvalue problems associated with H^TH and HH^T , i.e.,

$$LWL\phi_k = \lambda_k\phi_k, \quad (19)$$

$$SLS^T\psi_k = \lambda_k\psi_k, \quad -M \leq k \leq M \quad (20)$$

where $\lambda_k > 0$, and $\{\phi_k\}$ are $\infty \times 1$ and $\{\psi_k\}$ are $(2M+1) \times 1$ orthonormal vectors. It can be shown that ϕ_k must be bandlimited vectors, (i.e., $L\phi_k = \phi_k$), and are related to ψ_k via the relations

$$\psi_k = \frac{1}{\sqrt{\lambda_k}} S\phi_k \quad (21)$$

$$\phi_k = \frac{1}{\sqrt{\lambda_k}} L S^T \psi_k \quad (22)$$

Equation (21) states that $(2M+1) \times 1$ vector ψ_k is simply obtained by selecting the $(2M+1)$ elements $\{\phi_k(m), -M \leq m \leq M\}$ of ϕ_k and scaling them by $\lambda_k^{-1/2}$. Equation (22) is remarkable in that the $\infty \times 1$ vector ϕ_k is obtained by simply low pass filtering the sequence $\{\psi_k(m)\}$ and scaling the result by $\lambda_k^{-1/2}$. This means ϕ_k is the extrapolation of ψ_k , obtained by simple low pass filtering and scaling. Also noteworthy is the fact that the sequence $\{\phi_k(m), -\infty \leq m \leq \infty\}$ is orthogonal over the interval $-M \leq m \leq M$ as well as over the infinite interval. This property is similar to that of the continuous PSWFs. The extrapolated signal is obtained by writing the singular value expansion

$$H^+ = \sum_{k=-M}^M \frac{1}{\sqrt{\lambda_k}} \phi_k \psi_k^T \quad (23)$$

which gives $y^+ = H^+ z$, as

$$y^+(m) = \sum_{k=-M}^M \frac{a_k}{\sqrt{\lambda_k}} \phi_k(m), \quad a_k = \psi_k^T z = \sum_{m=-M}^M \psi_k(m) y(m) \quad (24)$$

It is easy to check that $y^+(m) = y(m)$ for $m \in [-M, M]$.

1.7 Mean Square Extrapolation Filter

In the presence of additive noise, uncorrelated with y , (15) is modified to give

$$z = S L y + n \triangleq H y + n \quad (25)$$

where n is the $(2M+1) \times 1$ noise vector. Now we look for the best linear mean square extrapolation of z and is given by the Wiener filter estimate

$$y = R L^T S^T (S L R L^T S^T + R_n)^{-1} z \quad (26)$$

where R and R_n are the autocorrelation matrices of y and n respectively. Since $\{y(k)\}$ is a bandlimited signal, $L R L^T = L [E y y^T] L^T = E[(L y)(L y)^T] = E y y^T = R$. Hence, the above equation becomes

$$\hat{y} = R L S^T (S R S^T + R_n)^{-1} z$$

In the worst case when we do not know R , one may set $R = L$. Now, if $R_n \rightarrow 0$, $\hat{y} \rightarrow y^+$, the MNLS extrapolated estimate.

1.8 Recursive Extrapolation

Now we present a recursive least squares algorithm based on Kalman filtering techniques where the extrapolated signal estimate is updated recursively as a new observation sample arrives. From (25), the k th observation $z(k)$ can be written as

$$z(k) = h_k^T y + n_k \quad k = 0, 1, \dots \quad (27)$$

where h_k^T is the k^{th} row of L and n_k is zero mean white Gaussian noise.

The state equation for the unknown extrapolated vector y can be written as

$$y_{k+1} = y_k, \quad y_0 = y, \quad \text{cov}(y_0) \triangleq P_0 = L \triangleq \{\alpha_{k-\ell}\} \quad (28)$$

The Kalman filter associated with equations (27)-(28) is called the recursive least squares filter and is given by

$$\hat{y}_{k+1} = \hat{y}_k + g_k(z(k) - h_k^T \hat{y}_k), \quad y_0 = 0 \quad (29)$$

where $\hat{y}_k$ is the k^{th} estimate of y and g_k is the Kalman filter gain. Using the properties of L , the Riccati equation which is associated with the Kalman filter can be simplified considerably. Details are given in [6].

Examples:

Let the observations model be

$$z(k) = \sin(.099\pi k) + \sin(.085\pi k) + n_k, \quad -8 \leq k \leq 8$$

It is known that the spectrum of the signal lies in the interval $[-.1, .1]$ i.e., $\sigma=0.1$. Figure 1 and 2 show the original signal $y(k)$ and the observations when there is no noise. Figures 3, 6, and 4 respectively show the extrapolated estimates obtained by the Papoulis (30 iterations), the conjugate gradient (10 iterations), the generalized inverse algorithms. In theory, these algorithms are equivalent. However, due to differences in their numerical properties, the results are different. Faster convergence of the conjugate gradient method is evident from Figs. 3 and 6. Although the gen-

eralized inverse of (17) always exists, it exhibits instability (Fig. 4) due to ill conditioning of $\tilde{L}$. However, the stabilized inverse of (24) improves the extrapolated estimate (Fig. 5) greatly. When the observations contain small noise (SNR=21.6dB) performance of these algorithms is considerably degraded. For example, Fig. 7 shows the result of the conjugate gradient method. This is not unexpected because the noise was not accounted for in the extrapolation algorithm. The mean square extrapolation filter (Fig. 8) greatly improves the result. Application of these algorithms in spectral estimation radar signal processing are considered in [6].

PART TWO: TWO DIMENSIONAL EXTRAPOLATION AND SPECTRAL ESTIMATION

2.1 Extrapolation of Bandlimited Sequences

The results of part one can be easily extended to a two dimensional bandlimited sequence $y(m,n)$ which is known over a finite observation window say, $[-M,M] \times [-M,M]$. Then using Kronecker products and mapping $y(m,n)$ into a lexicographically ordered vector $y(k)$, we can write an equation analogous to (9) as

$$Z = SLy, \quad S = S \otimes S, \quad L = L \otimes L \quad (30)$$

The properties of the two dimensional low pass operator L can be determined and the two dimensional version of the various algorithms can be derived. For example, the gradient algorithm corresponding to (13) becomes (after using properties of Kronecker products of matrices)

$$Y_{n+1} = eF_1 + Y_n - eLWY_nWL, \quad F_1 \triangleq LS^TZSL, \quad Y_0 = 0 \quad (31)$$

where Z is the matrix of observations on $[-M,M] \times [-M,M]$, and Y_n is the $(\infty \times \infty)$ extrapolated MNLS estimate of Y at iteration n . This algorithm is simple in that it requires separable row by row and column by column operations. Details of this and other two dimensional version of the foregoing algorithms will appear elsewhere.

2.2 Two Dimensional Maximum Entropy Spectral Estimation

For a bandlimited sequence, a spectral estimate can be obtained from the Fourier transform of the extrapolated signal. The foregoing extrapolation methods become inapplicable when sampling is done at the Nyquist rate. For one dimensional signals, autoregressive (AR) or equivalently the maximum entropy (ME) method have been found useful in obtaining high resolution spectral estimates. In this case, the solution of the nonlinear ME spectral estimate equations reduces to an equivalent set of linear Toeplitz, AR equations which are easily solvable. This equivalence between the ME and AR methods is due to the existence of spectral factorization theorem in one dimension. In two dimensions, if we are given, say, the autocorrelations $\{r_{m,n}\}$ on a rectangular window $W^* = [-p,p] \times [-q,q]$, the ME estimate of the spectral density function (SDF) must be the form

$$S(z_1, z_2) = \beta^2 / \left[\sum_{m,n \in W^*} a_{m,n} z_1^m z_2^n \right]; |z_1| = 1, |z_2| = 1 \quad (32)$$

where $\{a_{m,n}\}$ must be determined from $\{r_{m,n}\}$ given on W^* . Now, if $\{r_{m,n}\}$ is a positive definite sequence on W^* , it is not sufficient to guarantee a positive ME spectral estimate. In other words, the array $\{r_{m,n}\}$ defined on W^* need not have any positive definite extension[†] [14,15]. Besides existence difficulties, the nonlinear problem of determining $a_{m,n}$ can no longer be reduced to a linear problem (as in 1-D case) of autoregression. This is because factorization of a two dimensional rational SDF as a product of two complex conjugate rational functions is not always possible. Hence, if the ME solution existed, one would generally resort to iterative methods e.g., see [13]. Unfortunately, there are convergence difficulties and the results do not seem to be very attractive.

2.3 Discrete Random Fields and their SDFs

An obvious extension of one dimensional AR spectral estimation method is to assume that the SDF to be estimated has a stationary random field realization

$$u_{i,j} = \sum_{m,n \in S} \alpha_{m,n} u_{i-m,j-n} + e_{i,j}; E u_{i,j} u_{i+k,j+l} \triangleq r(k,l) \quad (33)$$

where $\{u_{i,j}\}$ is a two dimensional discrete random field and S is a set of suitably chosen index pairs (m,n) . We define

$$\sum_{m,n \in S} \alpha_{m,n} u_{i-m,j-n} \triangleq \hat{u}_{i,j} \quad (34)$$

as prediction estimate of $u_{i,j}$ and $e_{i,j}$ is the prediction error. We consider three types of predictors characterized by S as (see Fig. 9)

$$S = \begin{cases} \{n \geq 1, \forall m\} U \{n=0, m \geq 1\} & : \text{Causal model} \end{cases} \quad (35a)$$

$$S = \begin{cases} \{n \geq 1, \forall m\} U \{n=0, \forall m \neq 0\} & ; \text{Semicausal model} \end{cases} \quad (35b)$$

$$S = \begin{cases} \{\forall (m,n) \neq (0,0)\} & ; \text{Noncausal model} \end{cases} \quad (35c)$$

The "causal model" (35a) defines causality in the sense of raster scanning the field $\{u_{i,j}\}$ column by column. Often one is interested only in representations where $\alpha_{m,n}$ are non-zero only over a finite window W , called the prediction window, which is a subset of S . In that event (33) is a stochastic

[†]This was pointed out to the author by B. Dickinson [15].

difference equation which realizes the rational SDF

$$S_u(z_1, z_2) = S_e(z_1, z_2) / [1 - \sum_{m,n \in W} \alpha_{m,n} z_1^{-m} z_2^{-n}] [1 - \sum_{m,n \in W} \alpha_{m,n} z_1^m z_2^n] \quad (36)$$

where S_e is the SDF of $\{e_{i,j}\}$. In general, $\{e_{i,j}\}$ could be a moving average field.

2.4 Minimum Variance Representations (MVRs): If, for a given W the random field of (33) is such (i.e., $\{\alpha_{m,n}\}$) that the prediction error has the minimum variance, then it is called an MVR. The orthogonality condition associated with such a representation requires that $E e_{i,j} u_{i-m,j-n} = 0$, whenever $m,n \in S$ which gives

$$r(k,l) - \sum_{m,n \in W} \alpha_{m,n} r(k-m, l-n) = \beta^2 \delta_{k,0} \delta_{l,0} ; k,l \in S_0 \quad (37)$$

where $S_0 = SU[0,0]$. Defining $\alpha_{0,0} = -1$ and $W_0 = WU[0,0]$, and mapping the arrays $\{\alpha_{m,n}\}$, $\{u_{m,n}\}$, defined on W_0 into vectors $\underline{\alpha}$ and $\underline{u}$ respectively (37) reduces to the equation

$$R \underline{\alpha} = -\beta^2 \underline{1} \quad \text{or} \quad \underline{\alpha} = -\beta^2 R^{-1} \underline{1} = -\beta^2 b_{i_0} \quad (38)$$

where R is the covariance matrix of the vector u . The vector $\underline{1}$ takes a value 1 at a location, say i_0 , which corresponds to the (0,0) location in the window W_0 ; and b_{i_0} is the i_0 th column of R^{-1} . Since $\alpha_{0,0} = -1$, one obtains from (38)

$$\beta^2 = 1/b_{i_0}(i_0) \quad (39)$$

Once β^2 is obtained, $\underline{\alpha}$ is directly computed from (38). Thus, we need only the (block Toeplitz) matrix R to solve for β^2 and $\underline{\alpha}$.

2.5 Causal Models and AR Spectral Estimation:

A common example of causal prediction is to consider $W_0 = [0,p] \times [0,q]$ which gives rise to a (single quadrant) predictor, causal in each dimension. The orthogonality condition for these models requires $e_{i,j}$ to be a white noise field so that $S_e = \beta^2$ and the spectral density function is

$$S_u = \beta^2 / |1 - \sum_{m,n \in W} \alpha_{m,n} z_1^{-m} z_2^{-n}|^2 ; |z_1| = 1, |z_2| = 1 \quad (40)$$

where $\{\alpha_{m,n}\}$ are obtained via (38) and (39) with $i_0 = 1$ and R is a $(p+1) \times (p+1)$ block Toeplitz matrix of basic dimension $(q+1) \times (q+1)$ i.e., $R = \{R_{i-j}\}$,

$0 \leq i, j \leq p$, where $R_k = \{r(k, i-j)\}$, $0 \leq i, j \leq q$. Comparison with (32) shows that the SDF estimated via (40) is of the form of (32) since we can write the denominator polynomial in (40) as in (32) via the relation

$$a_{m,n} = a_{-m,-n} = \sum_{i=0}^p \sum_{j=0}^q \alpha_{i,j} \alpha_{i+m,j+n}; m,n \in W^* \quad (41)$$

where $\alpha_{i,j} = 0$ if $(i,j) \notin W_0$ and W^* is the window $[-p,p] \times [-q,q]$. The element of the block Toeplitz matrix R are defined on this window. However, in general, the coefficients $\{a_{m,n}\}$, (and the SDF) obtained via (38)-(41) are not the quantities that one would obtain by solving the ME nonlinear equations. Alternatively, the covariances realized by the random field of (33) whose $\{\alpha_{i,j}\}$ are determined via (38)&(39) (from a given positive definite R obtained from given autocorrelations $r(k,l)$ on W^*), need not match exactly the given values $r(k,l)$ on the window W^* . If the two sets autocorrelation matched, then (36) would also be the ME spectrum.

Example.

As an example, consider the autocorrelation model.

$$r(m,n) = \cos 2\pi(0.05m + 0.2n) + 0.5 \cos 2\pi(0.2m + 0.05n) + 0.25 \delta_{m,0} \delta_{n,0} \quad (42)$$

The SDF has two delta functions at $(.05, .2)$ and $(.2, 0.05)$ in the positive quadrant of the frequency plane. Assuming $r(m,n)$ are available on a 5×5 window $W^* = [-2,2] \times [2,2]$ ($p=q=2$), the spectrum estimated according to (40) is shown as a contour plot in Fig.10. The two original peaks have merged into a single peak at roughly half way between those peaks. Increasing the size of the observation window to 7×7 or higher did not improve the resolution. This is so because many covariance functions may not be realizable even by an infinite order ($p=q=\infty$) single quadrant causal models. Other causal structures such as non-symmetric half plane models may improve results, but the order of the model required to achieve desired resolution may be prohibitively high.

2.6 Spectral Estimation via Semicausal Models

In the case of semicausal models, the prediction window selects samples which are in the past in one of the directions and in the past as well as future in the other direction. Correspondingly, the model allows prediction on a symmetric half plane.

As an example, let the prediction window W_0 be $[-p,p] \times [0,q]$. The maximum variance condition applied at $n=0$ requires $E e_{i,j} u_{i+m,j} = 0, \forall m \neq 0$, which yields the condition

$$r_e(m,n) = E e_{i,j} e_{i+m,j+n} = \beta^2 \left[1 - \sum_{m=1}^p \alpha_{m,0} (\delta_{i-m,0} + \delta_{i+m,0}) \right] \quad (43)$$

Hence the SDF of $e_{i,j}$ is

$$S_e = \beta^2 \left[1 - \sum_{m=1}^p \alpha_{m,0} (z_1^{-m} + z_1^m) \right] \quad (44)$$

This implies $\{\epsilon_{i,j}\}$ is a moving average field (in the 'i' variable) and is a white noise field in the 'j' variable. The SDF realized by this model is of the form

$$S_u(z_1, z_2) = S_e / \left| 1 - \sum_{m,n \in W} \alpha_{m,n} z_1^{-m} z_2^{-n} \right|^2; |z_1|=1, |z_2|=1 \quad (45)$$

which in view of (43), is not necessarily an "all-pole" model (unless S_e is a factor in the denominator polynomial). Corresponding to (38) and (39) here $i_0 = p+1$, R is a $(2p+1) \times (2p+1)$ block Toeplitz matrix of basic dimension $(q+1) \times (q+1)$ i.e., $R = \{R_{i-j}\}$, $-p \leq i, j \leq p$; $R_k = \{r(k, i-j)\}$, $0 \leq i, j \leq q$, and α and 1 are vectors of size $(2p+1)(q+1)$. Thus the elements of R are defined over a window $W^* = [-2p, 2p] \times [-q, q]$.

Some interesting and important facts about the foregoing semicausal model are in order (1) Comparison with (32) shows that (45), in general, is not a maximum entropy spectrum, since it is not an "all-pole" model. (2) The semicausal representation of (33) where $S=W$ and $\{\epsilon_{i,j}\}$ is defined in (43), is a special autoregressive moving average (ARMA) model of a two dimensional random field. (3) The main equation, (38), for obtaining the spectral estimate is linear.

Examples: Consider the autocorrelation function

$$r(m,n) = \sin 2\pi(m+n)/8 + \sin 2\pi(m+n)/8.05 + 7\% \text{ white noise} \quad (46)$$

and the window $W_0 = [-2, 2] \times [0, 2]$. Thus $r(m,n)$ are available over $W^* = [-4, 4] \times [-2, 2]$. Figure 11 shows that the FFT based periodogram is unable to resolve the two closely spaced peaks which should occur at $[0.25, 0.25]$ and $[0.2445, 0.2445]$ in the first quadrant. However, the semicausal model spectrum (Fig. 12) easily resolves these peaks.

Remarks

1. Semicausal models are not only useful in high resolution spectral estimation, but also in finding random field realizations of known spectra (i.e., spectral factorization). For example, consider the irrational covariance function $r(m,n) = \exp(-.05\sqrt{k^2 + l^2})$. Table 1 compares the mismatch of the model covariances with the actual covariances on a 5×5 window, for causal and semicausal realizations of equal order. Clearly, the semicausal model provides a much better fit.

2. Minimum variance semicausal models were first introduced by Jain in [16] for semirecursive filtering of images. Such models lead to algorithms which are recursive in one of the (causal) dimensions and are nonrecursive in the other (noncausal) dimension. Often, the nonrecursive part of the algorithm can be implemented via a fast unitary transform yielding an efficient overall algorithm. Such models also arise when one considers finite difference approximations of parabolic partial differential equations [17,18]. Applications of such models in image restoration and data compression have been studied in [18-21].

3. It can be shown that any minimum variance semicausal model of finite order (i.e., the transfer function is rational) can be factorized to yield a minimum variance causal model (i.e., an AR model). In general, this causal model would be defined on a non-symmetric half plane (NSHP) and would be of infinite order. In practice, one may obtain an approximate rational, causal NSHP realization of semicausal model (or equivalently of its SDF) by a suitable truncation of this factorization (consistent with stability requirements). The foregoing procedure is therefore applicable for design of semicausal and/or causal digital filters whose magnitude of the frequency response is specified.

4. For separable spectra (i.e., $S(z_1, z_2) = S_1(z_1)S_2(z_2)$), spectral factorization is possible and therefore, the two dimensional causal, semicausal and noncausal MVRs as well as the ME method, all yield the same estimates.

2.7 Noncausal Models

The equations for noncausal models can also be derived by specifying W_0 and using (38) and (39). The minimum variance condition for such models requires $\epsilon_{i,j}$ to be a moving average field with SDF

$$S_c = \beta^2 \left[1 - \sum_{m,n \in W} \alpha_{m,n} z_1^{-m} z_2^{-n} \right] \quad (47)$$

where W is a noncausal window (e.g., $W_0 = [-p,p] \times [-q,q]$, $W_0 = WU[0,0]$).

Using (47) in (36), it is seen that the SDF of $u_{i,j}$ is an all pole model and is therefore of the form of ME spectrum. However, due to existence difficulties explained earlier, an admissible covariance matrix R would not guarantee a non-negative spectrum. Hence the usefulness of these models in spectral estimation is severely limited.

3. CONCLUSIONS

In summary, we have considered extrapolation and spectral estimation algorithms for discrete signals in one and two dimensions. For bandlimited (oversampled) signals observed over a region of finite (and small) support, we recommend the signal be extrapolated first followed by a suitable spectral estimator (e.g., ME or smoothed periodogram etc.). Several existing extrapolation algorithms were shown to be unified under the minimum norm least

squares criterion of extrapolation, and was shown to yield new improved extrapolation algorithms via the conjugate gradient, least squares and recursive methods. For other two dimensional sequences, the minimum variance semicausal models seem to yield high resolution spectra compared to other methods. It requires solution of linear equations but yields an ARMA spectral estimate. Not described in this paper, are the implementation and other obvious practical considerations such as i) efficient solution of the block Toeplitz equations (38),(39),ii) use of two dimensional data (rather than autocorrelations $r(k,l)$). iii) determination of the order of the model etc. These and other related considerations as well as details of semicausal modeling are reported in [22].

ACKNOWLEDGMENT

The author is grateful to S. Ranganath for computer implementing several of the algorithms reported here and also to Phil Jackson of the Environmental Research Institute of Michigan for verifying the semicausal model spectral estimation algorithm and Figs. 11 and 12.

References

1. A. Papoulis, "A New Algorithm in Spectral Analysis and Band-limited Extrapolation," IEEE Trans. Circuits Sys., Vol. CAS-22, pp735-742, Sept. 1975.
2. M.S. Sabri and W. Steenaart, "An Approach to Band-limited Signal Extrapolation: The Extrapolation Matrix," IEEE Trans. Circuits Sys., Vol. CAS-25, pp. 74-78, Feb. 1978.
3. J. A. Cadzow, "Improved Spectral Estimation from Incomplete Sampled Data Observations," Proc. RADC Spectrum Estimation Workshop, pp. 109-123, May 1978.
4. R. W. Gerchberg, "Super Resolution Through Error Energy Reduction," Optica Acta. Vol.21, pp. 709, 1974.
5. D. Slepian, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - V: The Discrete Case," The Bell Syst. Tech. J., Vol. 57, No. 5, pp. 1371-1430, May-June, 1978.
6. A. K. Jain and S. Ranganath, Extrapolation and Spectral Estimation Techniques for Discrete Time Signals, Technical Phase Report, RADC-TR-79-124, Rome Air Development Center, Griffiss Air Force Base, N.Y. 13441, May 1979.
7. C. W. Helstrom, "Image Restoration by the Method of Least Squares," J. Opt. Soc. Am. Vol. 57, pp. 297-303, March 1967.
8. C. L. Rino, "Bandlimited Image Restoration by Linear Mean Square Estimation," J. Opt. Soc. Am., Vol. 59, pp. 547-558, May 1969.

9. A. K. Jain, "An Operator Factorization Method for Restoration of Blurred Images," IEEE Trans. Computers, Vol. C-26, pp. 1061-1071, Nov. 1977.
10. D. Slepian, et.al., "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty Principle," Bell Syst. Tech. J., Vol. 40, No. 1, pp. 43-84, 1961.
11. A. Papoulis and M. S. Bertram, "Digital Filtering and Prolate Functions," IEEE Trans. Circuit Theory, Vol. CT-19, No. 6, pp. 674-681, Nov. 1972.
12. V. R. Algazi and M. Suk, "On the Frequency Weighted Least-Square Design of Finite Duration Filters," IEEE Trans Circuit Sys., Vol. CAS-22, No. 12, pp. 943-953, 1975.
13. A. K. Jain and S. Ranganath, "Two Dimensional Spectral Estimation," Proc. RADC Workshop on Spectral Estimation, Rome N.Y., pp. 151-157, May 1976.
14. W. Rudin, "The Extension Problem for Positive Definite Functions," Ill. J. Math., Vol. 7, pp. 532-539, 1963.
15. B. Dickinson, "Two-Dimensional Markov Spectrum Estimates Need Not Exist," (to appear)
16. A. K. Jain, "A Semicausal Model for Recursive Filtering of Two Dimensional Images," IEEE Trans. Computers, Vol. C-26, pp. 345-350, April 1977.
17. A. K. Jain, "Partial Differential Equations and Finite Difference Methods in Image Processing," Part I: Image Representation," J. Optimiz. Th. Appl. Vol. 23, pp. 817-834, Sept. 1977.
18. A. K. Jain and J. R. Jain, Part II (of above): Image Restoration, IEEE Trans. Aut. Contr., Vol. AC-23, pp. 817-834, Oct. 1978.
19. E. Angel and A. K. Jain, "Frame to Frame Restoration of Diffusion Images," IEEE Trans. Aut. Contr., Vol. AC-23, pp. 850-855, Oct. 1978.
20. A. K. Jain and S. H. Wang, "Stochastic Image Models and Hybrid Coding," Final Report NOSC Contract, NOSC953-77-C-003 MJE, Dept. Elect. Engr. SUNY Buffalo, New York, Oct. 1977.
21. S. H. Wang, "Applications of Stochastic Models in Image Data Compression," Ph.D. Thesis, Dept. Elect. Engr. SUNY, Buffalo, 1979.
22. A. K. Jain, Final Report, ARO grant DAAG29-78G0206, Signal and Image Processing Laboratory, Dept. Elec. Engr., U.C. Davis, CA 95616 (to appear)

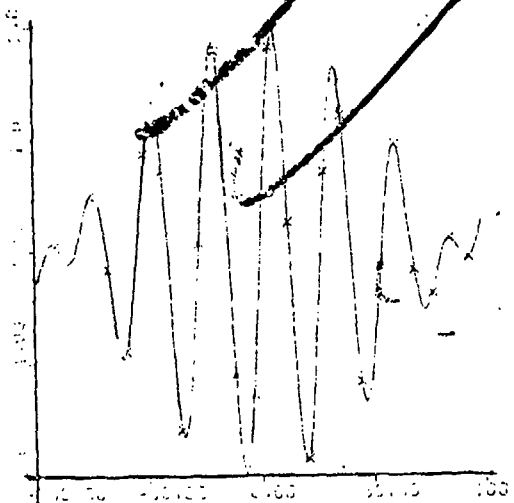


Figure 1: Original Signal

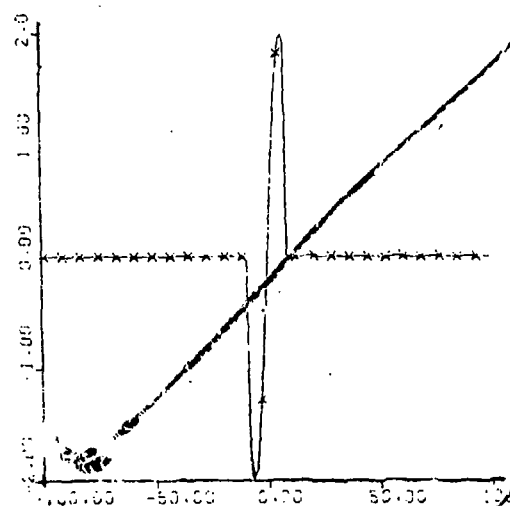


Figure 2: Given Observations
(17 samples)

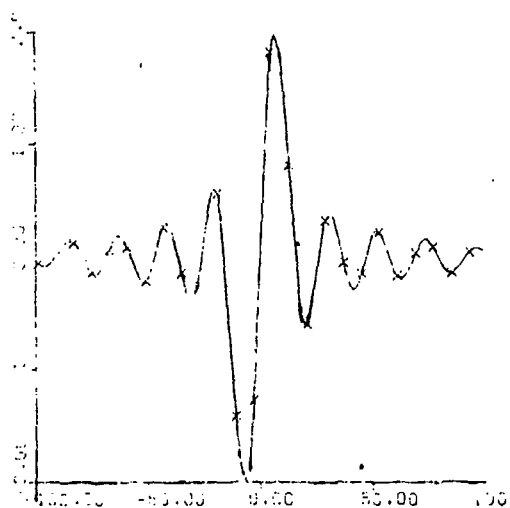


Figure 3: Extrapolation by
Papoulis' (Gradient) Algorithm
(30 iterations)

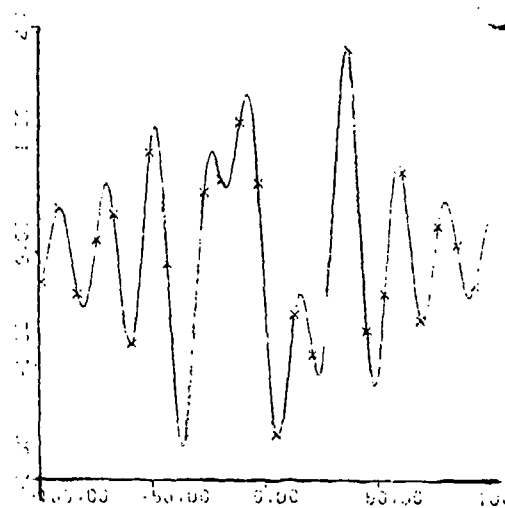


Figure 4: Extrapolation Estimate by
Generalized Inverse Extrapolation
Matrix H^+ .

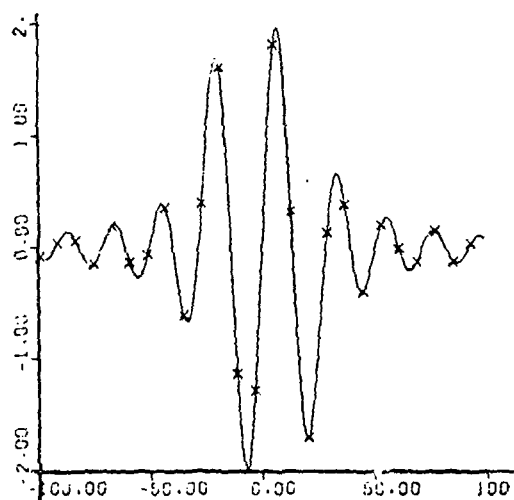


Figure 5: Extrapolation by
Stabilized Inverse

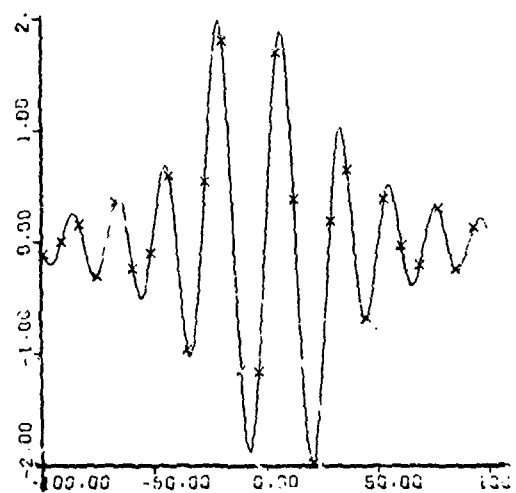


Figure 6: Extrapolation by
Conjugate Gradient Al-
gorithm (10 iterations)

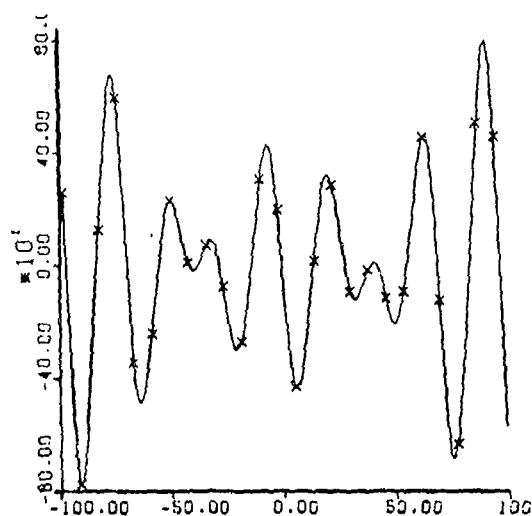


Figure 7: Extrapolation of Noisy
Data by Conjugate Gradient
Algorithm

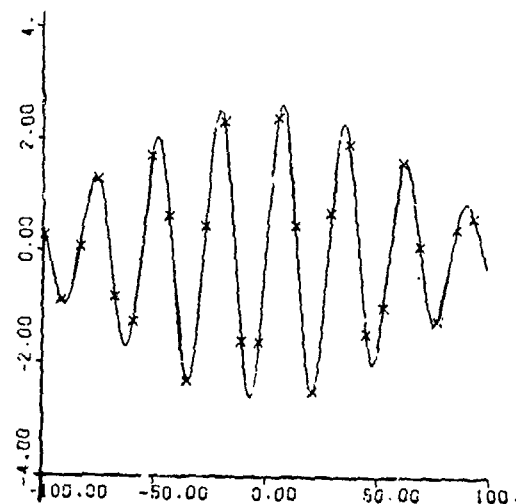
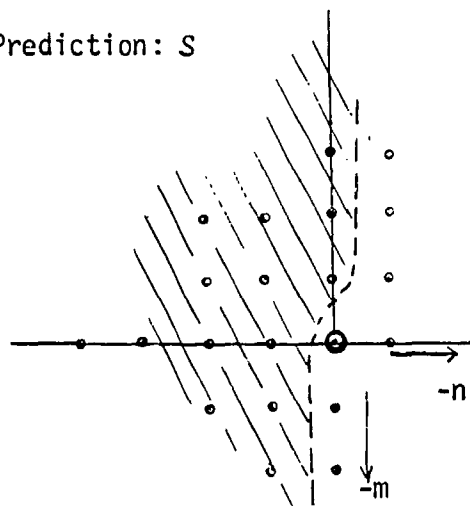
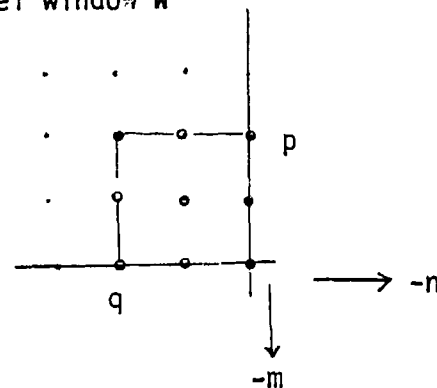


Figure 8: Extrapolation of Noisy
Data by the Mean Square
Extrapolation Filter.

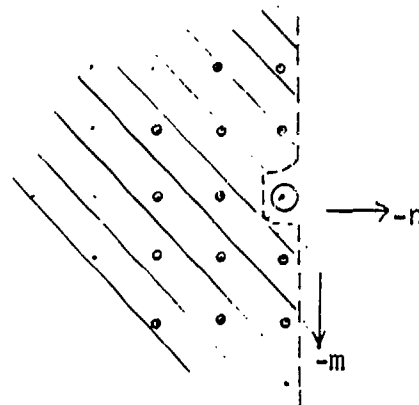
Causal Prediction: S



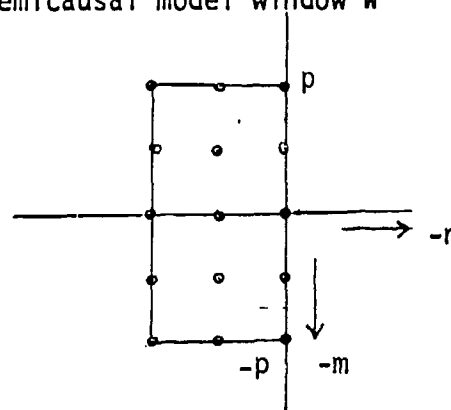
Example of a single quadrant causal model window W



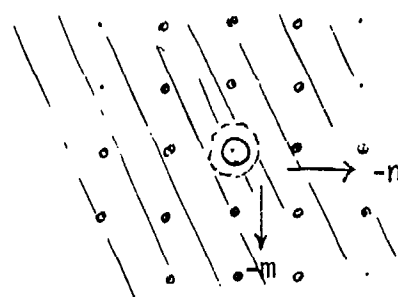
Semicausal Prediction: S



Semicausal model window W



Noncausal Prediction: S



Noncausal model window W

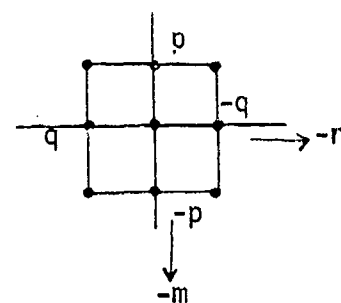


Figure 9: Three Types of Two Dimensional Models

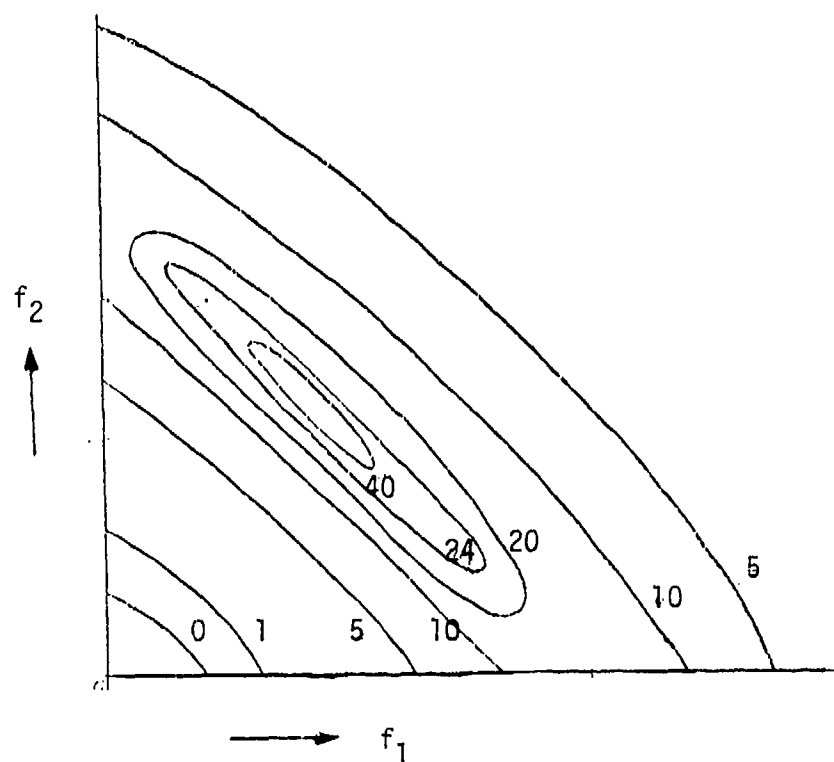


Figure 10: Causal Model Spectrum Contours

Table 1: Comparison between Causal and Semicausal Random Field Realizations

$n \uparrow$	.905	.894	.868	$n \uparrow$	.208	.315	.389	$n \uparrow$	.039	.041	.047
	.951	.932	.894		.111	.231	.315		.019	.025	.036
	1.00	.951	.905		0.	.111	.208		0.	.017	.031
	$\rightarrow m$				$\rightarrow m$				$\rightarrow m$		
Actual covariances $r(m,n)$ $r(-m,n)=r(m,-n)=r(-m,-n)$				Causal Model Covariance Mismatch				Semicausal Model Covariance Mismatch			

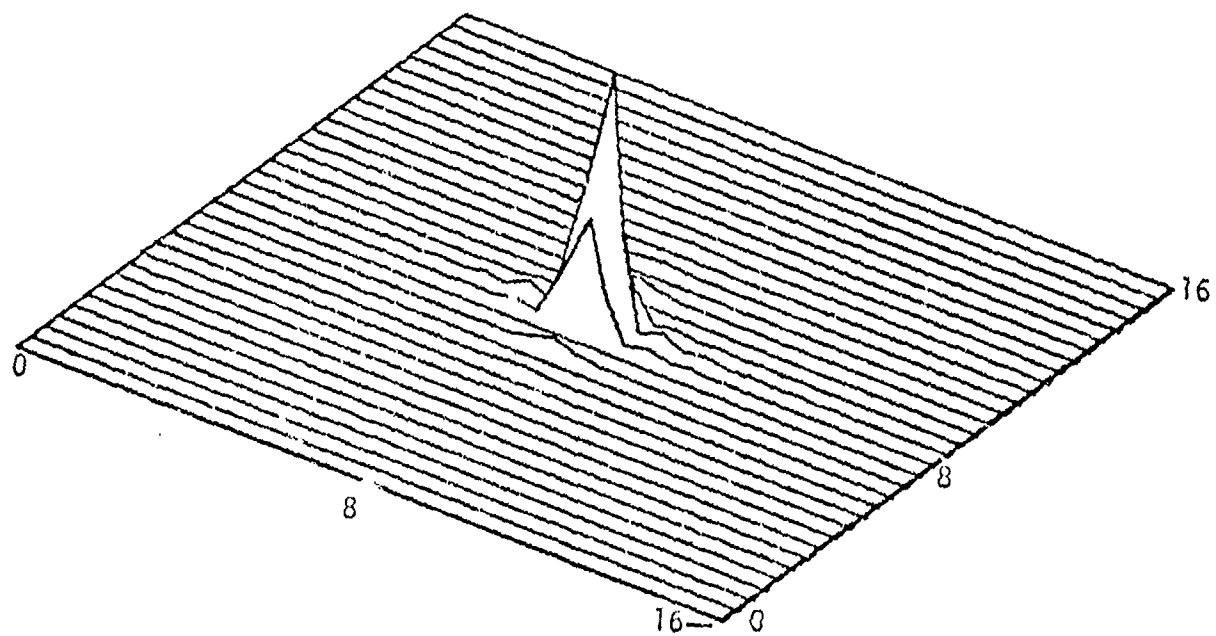


Figure 11: FFT Spectrum (peaks are unresolved)

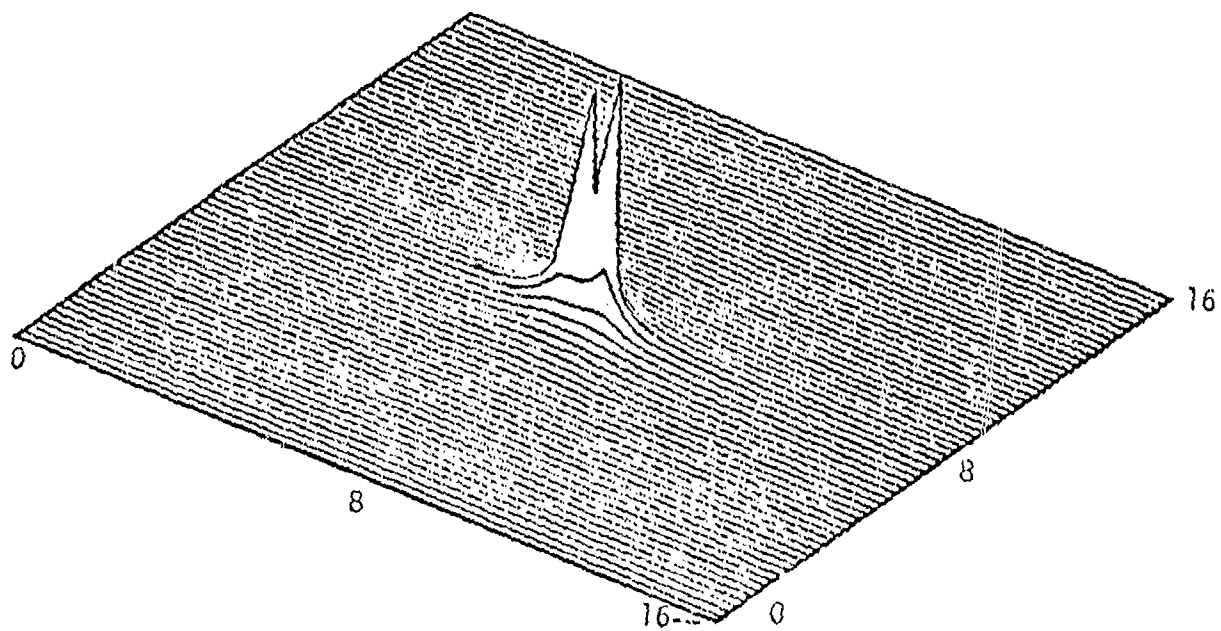


Figure 12: Semicausal Model Spectrum (peaks are resolved)

ANTENNA SPACIAL PATTERN VIEWPOINT OF MEM, MLM, AND ADAPTIVE ARRAY RESOLUTION

WILLIAM F. GABRIEL

Radar Division
Naval Research Lab
Washington, DC 20375

Abstract

The Burg maximum entropy method (MEM) and the maximum likelihood method (MLM) nonlinear spectral estimation techniques are compared with their similar adaptive array antenna counterparts. The comparison permits an examination of their principles of operation from the antenna array spacial pattern viewpoint, and qualifies their superresolution performance behavior. Also, the real-time adaptive resolution of two incoherent sources located within a beamwidth was simulated, and results are presented over an array output SNR range of 0 to 40 DB.

Introduction

Adaptive array processing techniques are being investigated to determine their applicability to high-resolution location of sources/targets. The work was motivated by high-resolution performance reported in the field of spectral analysis in recent years, particularly from the two nonlinear techniques generally identified as the maximum entropy method (MEM) [1-3] and the maximum likelihood method (MLM) [4-6]. MEM and MLM bear a very close relationship to nonlinear adaptive array processing techniques. It is the purpose of this paper to point out a few of these similarities, examine their principles of operation from the "spacial filter" pattern viewpoint of adaptive array antennas, and to discuss some of the limitations to be expected in their superresolution behavior.

MEM and the Adaptive Sidelobe Canceller

The Burg MEM has been shown to be equivalent to least mean square (LMS) error linear prediction [7-9], where an optimum K point prediction filter predicts the nth value of a sequence from K past values.

$$\hat{y}_n = \sum_{k=1}^K a_k y_{n-k} \quad (1)$$

where $\hat{y}_n$ is the predicted sample, the a_k are optimum weighting coefficients,

and the K past samples of y_{n-k} are presumed known. Define the difference between this predicted value $\hat{y}_n$ and the true value of y_n as the error, ϵ_n , which is to be LMS minimized over a larger data sequence of N samples, $N > K$.

$$\epsilon_n = (y_n - \hat{y}_n) \quad (2)$$

The z-transforms associated with this discrete convolution may be written,

$$\epsilon(z) = \left[1 - \sum_{k=1}^K a_k z^{-k} \right] Y(z) \quad (3)$$

where the expression within the brackets may be defined as the filter transform function, $H(z)$, consisting of a polynomial with K roots or zero factors. If we optimize the weights a_k in such a manner that the spectrum of ϵ approaches white noise, then the unknown spectrum of the input is approximated by,

$$|Y(\omega)|^2 = \left| \frac{\epsilon(\omega)}{H(\omega)} \right|^2 \approx \frac{(\text{CONSTANT})}{\left| \sum_{k=1}^K a_k e^{-j\omega k} \right|^2} \quad (4)$$

Conversion of the above linear prediction filter to a weighted linear array of spacial sensors is straightforward [10, 11], with the simplest configuration illustrated in Fig. 1. We assume that our sensor elements are equally spaced, and that narrowband filtering precedes our spacial domain processing. The n th "snapshot" signal sample at the k th element will consist of independent gaussian receiver noise, η_{kn} , plus L incoherent source voltages,

$$E_{kn} = \eta_{kn} + \sum_{i=1}^L J_i e^{j(ku_i + \phi_{in})} \quad 1 \leq k \leq K \quad (5)$$

$$\text{where } u_i = 2\pi \left(\frac{d}{\lambda} \right) \sin \theta_i$$

d = element spacing, assumed near $\lambda/2$

λ = wavelength

θ_i = spacial location angle of i^{th} source

J_i = amplitude of i^{th} source

ϕ_{in} = random phase of i^{th} source, n^{th} sample

k = element index

n = snapshot sample index

A "snapshot" is defined as one simultaneous sampling of the aperture signals at all array elements, and we assume that N snapshots of data are available.

A brief examination of Fig. 1 from the standpoint of adaptive arrays leads to the conclusion that it is identical in configuration to a special subclass commonly referred to in the literature as a "sidelobe canceller" [12, 13]. A typical sidelobe canceller configuration from Applebaum [13] is illustrated in Fig. 2. For the benefit of those who may not be familiar with them, it should be noted that the unweighted mainbeam "element" is usually different and of much higher gain than the others, and the elements may or may not be equally spaced. They are designed to be operated on the basis of many successive snapshots (assuming digital operation) because their environment generally involves weak desired signals and an abundance of interference source data. They are a prediction filter in the sense that, after convergence, they are predicting the signal at the phase center of the mainbeam element.

The pertinence of the adaptive sidelobe canceller to our linear prediction filter is that their spacial filter pattern analysis is well-developed and can be applied directly to achieve a better understanding of the super-resolution performance behavior. A further point is that real-time operation is readily achieved via most of the current adaptive algorithms, provided that the number of snapshots is sufficient to reach convergence in whitening G. Convergence may require as little as 2 snapshots or as many as several thousand, depending upon the particular algorithm and the parameters of the source distribution.

Spacial Filter Patterns

The spacial filter function for the array of Fig. 1 is simply the adapted pattern after convergence, which is commonly referred to as the steady-state adapted pattern and may readily be computed from the inverse of the sample covariance matrix [13],

$$\underline{W}_0 = \underline{\mu} \underline{M}^{-1} \underline{S}^* \quad (6)$$

$$\underline{f}^{*t} = [0, 0, 0, 0, 0, 0, 0, 1] \quad (7)$$

$$\underline{M} = \frac{1}{N} \sum_{n=1}^N \underline{M}_n \quad (8)$$

$$\underline{M}_n = \begin{bmatrix} \underline{E}_n^* & \underline{E}_n^t \end{bmatrix} \quad (9)$$

where $\underline{E}_n$ is the n^{th} "snapshot" signal sample vector whose element components are given by equation (5), $\underline{M}_n$ is the n^{th} snapshot contribution to the covariance matrix, $\underline{M}$ is the sample covariance matrix averaged over N snapshots, $\underline{S}^*$ is the quiescent weight steering vector, μ is a scalar quantity, and $\underline{W}_0$ is the optimum weight vector. Note that the steering vector $\underline{S}^*$ injects a zero weight on every element except for the end element, thus causing the quiescent pattern of the array to be that of the single end element. Fig. 3 illustrates a typical quiescent (single element) pattern and an adapted pattern obtained from an 8 element linear array with two far-field, incoherent, 30 dB sources located at 18 and 22 degrees. The adapted pattern weights were computed per equation (6) from the inverse of the covariance matrix averaged over 1024 simulated snapshots. Note that the two pattern nulls (zeros) align perfectly with the locations of the two sources. Of course, the array signals in this simulation were corrupted only by receiver noise (no element errors are included) and an average over 1024 snapshots is indeed steady-state. Another important point to note is that nulls in such an adapted pattern may be located arbitrarily close together in terms of beamwidth, without violating any physical principle. Yet, because the nulls have served to locate two sources within a beamwidth, one may describe this as a "superresolution" pattern.

It is readily shown that this adapted pattern is obtained by subtracting the summed array output pattern from the element (mainbeam) pattern and, furthermore, that the summed array pattern consists of properly weighted "eigenvector beams" [14]. Written in terms of the eigenvector weights, we can express the optimum weights in the form,

$$\underline{W}_0 = \underline{S}^* - \sum_{i=1}^K \left(\frac{B_i - B_0}{B_i + B_0} \right) \hat{W}_{qi} \underline{e}_i \quad (10)$$

$$\hat{W}_{qi} = \left(\underline{e}_i^* \underline{S}^* \right)$$

where $\underline{e}_i$ is the i^{th} eigenvector of the covariance matrix, B_i is the i^{th} eigenvalue, and B_0 is the smallest eigenvalue corresponding to receiver noise power. Note that only the significant eigenvectors corresponding to $B_i > B_0$ need be considered here. An adaptive array forms one such eigenvector beam for each degree of freedom consumed in nulling out the spacial source distribution. Fig. 4 illustrates the two eigenvector beams required for this two-source example. It should be emphasized here that the true resolution and signal gain of the array is reflected in these eigenvector beams. They demonstrate the importance of having as wide an aperture as possible, because the superresolution capability in the adapted pattern is a percentage of the true resolution of these beams. Also, since the superresolution nulls are formed via the subtraction of these beams of conventional width, it follows that the nulls will be rather delicate and very sensitive

to system imperfections and signal fluctuations.

The desired "spacial spectrum pattern" is then obtained from equation (4) as simply the inverse of the adapted pattern. Fig. 5 illustrates this inverse for the two-source example, in comparison with the output of a conventional beam scanned through the two sources. Several comments are in order concerning such inverse patterns:

a. They are not true antenna patterns, because there is no combination of the element weights that could produce such a peaked spacial pattern. They are simply a function computed from the reciprocal of a true antenna pattern.

b. Linear superposition does not hold in either the inverse or the original adapted pattern, because of the nonlinear processing involved.

c. The heights of the peaks do not correspond with the relative strengths of the sources, because the depths of the adapted pattern nulls do not. In general, the adaptive null depth will be proportional to the square of the SNR of a source [14], but even this relationship fails when there are multiple sources closely spaced.

d. There is no real-signal output port associated with such a pattern, because it is not a true antenna pattern. An output could be simulated, of course, by implementing the equivalent all-pole filter and driving it with white noise.

e. They do emphasize the locations of the zeros (nulls) of the adaptive array filter polynomial.

f. They are inherently capable of superresolution.

g. They achieve good "contrast" with the quiescent pattern background ripple (equivalent of "sidelobes") because of the aforementioned proportionality to the square of source strengths.

h. Spacial information is gained beyond that obtained from a conventional array beam which is scanned through the sources, because the array degrees of freedom are utilized in a more effective, data adaptive manner.

To get a feel for real-time operation performance with realistic weight update averaging, simulations were run in which an eight element array had its weights computed from the Howells-Applebaum recursive algorithm [15]. Weight update averaging was performed via a dynamic time constant in accordance with the reciprocal of the closed-loop bandwidth, α ,

$$\left(\frac{1}{\alpha}\right) = \frac{\tau_0 + T P_r}{1 + P_r} \quad (11)$$

where τ_0 = quiescent conditions slow time constant

T = high-power fast time constant

P_r = snapshot power ratio $\left(\frac{|S + N|^2}{|\bar{N}|^2}\right)$

where we approach the value, $\tau_0/2$, under quiescent conditions when $P_r \approx 1$, and we approach the value of T when $P_r \gg 1$. This formulation permits us to satisfy the 10 percent bandwidth criterion at high power levels to avoid noisy weights [14] by choosing the value of $T = 3.2$, and yet the quiescent condition time constant need be no worse than $\tau_0 = 200$. The larger value for τ_0 is necessary in order to have a relatively stable quiescent pattern.

Fig. 7 illustrates typical snapshot spectrum plots, after convergence, for our two-source case at two different SNR levels. Note the considerable fluctuations which occur in these plots near the peaks, which merely reflects the null fluctuations in the adapted pattern. The deteriorating conditions exhibited in Fig. 7b are indicative of the resolution capability nearing its limit, i.e., if the source power levels are reduced further, then the adaptive array cannot resolve them at that particular spacing.

A summary of the approximate resolution capability limit for the adaptive array spacial filter operating against two incoherent sources is illustrated in Fig. 8. This performance curve is universal in nature because the abscissa is source separation in beamwidths, and the ordinate is source SNR measured at the array output, i.e., element SNR multiplied by the number of elements in the array. Thus, the curve can be utilized for any number of array elements in a linear array configuration. Note that at low ordinate SNR values, we actually have negative SNR at the elements. The curve tells us that we can separate two sources at arbitrarily small spacings, provided we have sufficient SNR and, also, provided that our element data samples are sufficiently accurate. Recall that the simulations involved here did not include any element errors.

If there are more than two sources within a beamwidth or if coherence exists among the sources, then difficulties mount rapidly and the filter null points may not accurately represent the spacial locations of the sources.

MLM and Adaptive Directional Constraints

The maximum likelihood spectral estimate is defined as a filter designed to pass the power in a narrow band about the signal frequency of interest, and to minimize or reject all other frequency components in an optimal manner [4, 5]. This is identical to the use of a zero-order mainbeam directional gain constraint in adaptive arrays [16, 17], where the "spacial spectrum" would be estimated by the output residual power, P_o , from the optimized adapted array weights,

$$P_o = \underline{W}_o^{*t} \underline{M} \underline{W}_o \quad (12)$$

where $\underline{W}_o = \mu \underline{M}^{-1} \underline{S}^*$ (optimized weights)

$\underline{M}$ = covariance matrix estimate

$\underline{S}^*$ = mainbeam direction steering vector

μ = scalar quantity

Under the zero-order gain constraint, we require $\underline{S}^t \underline{W}_o = 1$, whereupon μ becomes

$$\mu = (\underline{S}^t \underline{M}^{-1} \underline{S}^*)^{-1} \quad (13)$$

Substituting μ and $\underline{W}_o$ into equation (12) then results in,

$$P_o = \frac{1}{\underline{S}^t \underline{M}^{-1} \underline{S}^*} \quad (14)$$

Upon sweeping the steering vector, $\underline{S}^*$, for a given covariance matrix inverse, P_o will estimate the spacial spectrum. Interestingly, this result is identical (within a constant) to the spectrum obtained from the inverse of the output residual power from an unconstrained optimized adapted array, and the principle of operation is the output from a continuously adapting pattern formed by subtracting eigenvector beams from the quiescent uniform illumination steering vector "mainbeam" as it scans.

Fig. 6 illustrates the output spectrum plotted from P_o for the two-source case utilized for Figs. 3, 4, and 5. Note that in comparison with Fig. 5, this MLM spectrum has peaks which are about 18 dB lower and thus of less resolution capability. However, the two peaks have located the sources correctly and, in addition, the peak values reflect the true power levels of the sources. This is in agreement with the observations of Lacoss [5] and others. Although this technique has less resolution than the previous one and requires more computation in plotting the output spectrum, it does

offer several rather significant advantages:

- a. The output power is directly referenced to receiver noise power, thus permitting calibration and measurement of relative source strength.
- b. If the sources can be resolved, then a psuedo-linear-superposition holds at the peaks, and they should reflect the true relative source strengths.
- c. The output of this filter is a real signal, and if the filter pass-band is steered to a particular source, one can monitor that source at full array gain while rejecting all other sources.
- d. The residual background spacial ripple (the equivalent of pattern "sidelobes") is very low and well behaved.
- e. It is not necessary to have the elements equally spaced. Thus, one should take advantage of this property to spread them out for a wider aperture and substantially increase the resolution for a given number of elements. This is done in the field of Geophysics [4]. By so doing, it is very likely that this method could equal the resolution of the previous technique.

References

1. D. G. Childers, 1978, "Modern Spectrum Analysis", IEEE Press (Note: This book contains complete copies of the following references Nos. 2, 3, 4, 5, 6, 7, 8, 10).
2. J. P. Burg, 1967, "Maximum Entropy Spectral Analysis", Proc. of the 37th Meeting of the Society of Exploration Geophysicists.
3. J. P. Burg, 1968, "A New Analysis Technique for Time Series Data", presented at the NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics, Enschede, Netherlands.
4. J. Capon, August 1969, "High-Resolution Frequency-Wavenumber Spectrum Analysis," Proc. IEEE, Vol. 57, pp. 1408-1418.
5. R. T. Lacoss, August 1971, "Data Adaptive Spectral Analysis Methods", Geophysics, Vol. 36, pp. 661-675.
6. J. P. Burg, April 1972, "The Relationship Between Maximum Entropy Spectra and Maximum Likelihood Spectra", Geophysics, Vol. 37, pp. 375-376.
7. A. Van den Bos, July 1971, "Alternative Interpretation of Maximum Entropy Spectral Analysis", IEEE Trans. on Information Theory, IT-17, pp. 493-494.

8. L. J. Griffiths, April 1975, "Rapid Measurement of Digital Instantaneous Frequency", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 23, pp. 207-222.
9. D. R. Morgan and S. E. Craig, December 1976, "Real-Time Adaptive Linear Prediction Using the Least Mean Square Gradient Algorithm", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 24, pp. 494-507.
10. R. N. McDonough, December 1974, "Maximum-Entropy Spatial Processing of Array Data", Geophysics, Vol. 39, pp. 843-851.
11. W. R. King, March 1979, "Maximum Entropy Spectral Analysis in the Spatial Domain", NRL Report 8298.
12. P. W. Howells, September 1976, "Explorations in Fixed and Adaptive Resolution at GE and SURC", IEEE Trans. on Antennas and Propagation, Vol. 24, pp. 575-584.
13. S. P. Applebaum, "Adaptive Arrays", IEEE Trans. on Antennas and Propagation, Vol. 24, pp. 585-598, September 1976.
14. W. F. Gabriel, February 1976, "Adaptive Arrays - An Introduction", Proc. of IEEE, Vol. 64, pp. 239-272.
15. I. S. Reed, J. O. Mallett, and L. E. Brennan, November 1974, "Rapid Convergence Rate in Adaptive Arrays", IEEE Trans. on Aerospace and Electronic Systems, Vol. 10, pp. 853-863.
16. O. L. Frost, August 1972, "An Algorithm for Linearly Constrained Adaptive Array Processing", Proc. of IEEE, Vol. 60, pp. 926-935.
17. S. P. Applebaum and D. J. Chapman, September 1976, "Adaptive Arrays with Mainbeam Constraints", IEEE Trans. on Antennas and Propagation, Vol. 24, pp. 650-662.

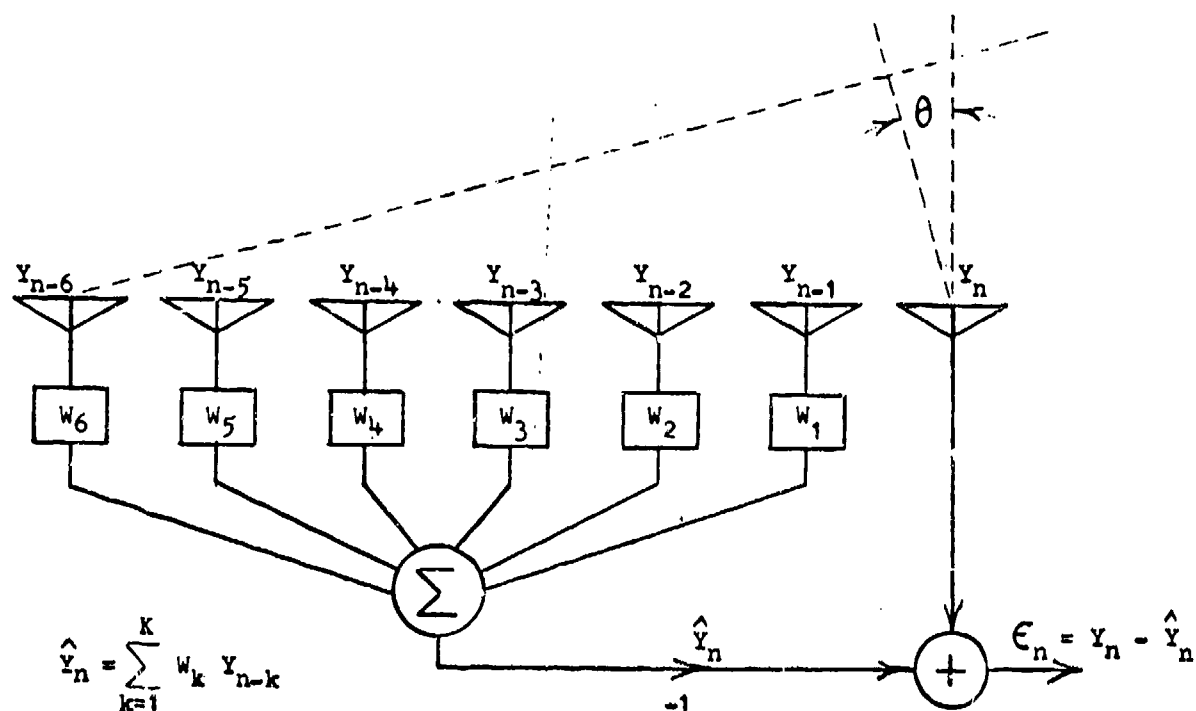


FIGURE 1. Array Aperture Linear Prediction Spatial Filter Model

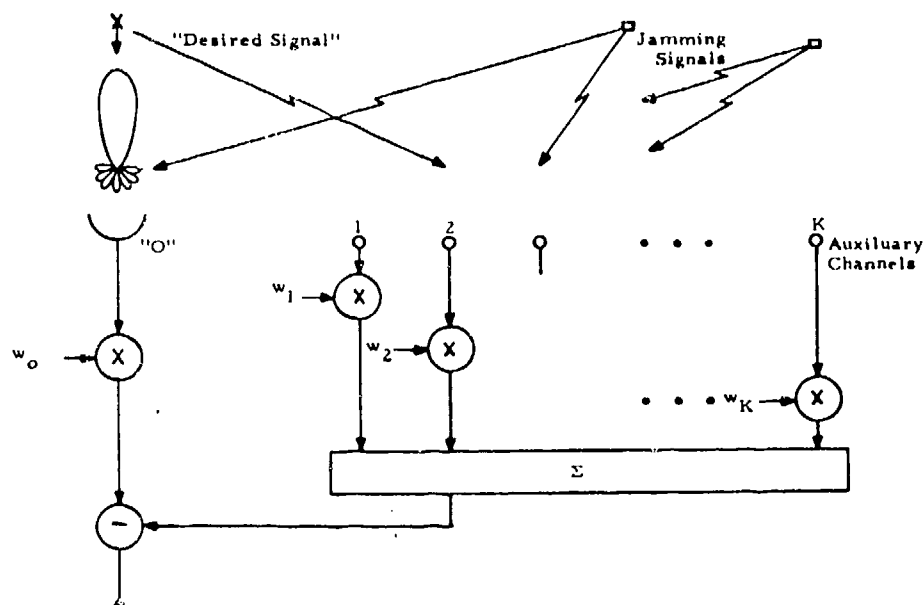


FIGURE 2. Typical Adaptive Array Sidelobe Canceller Configuration

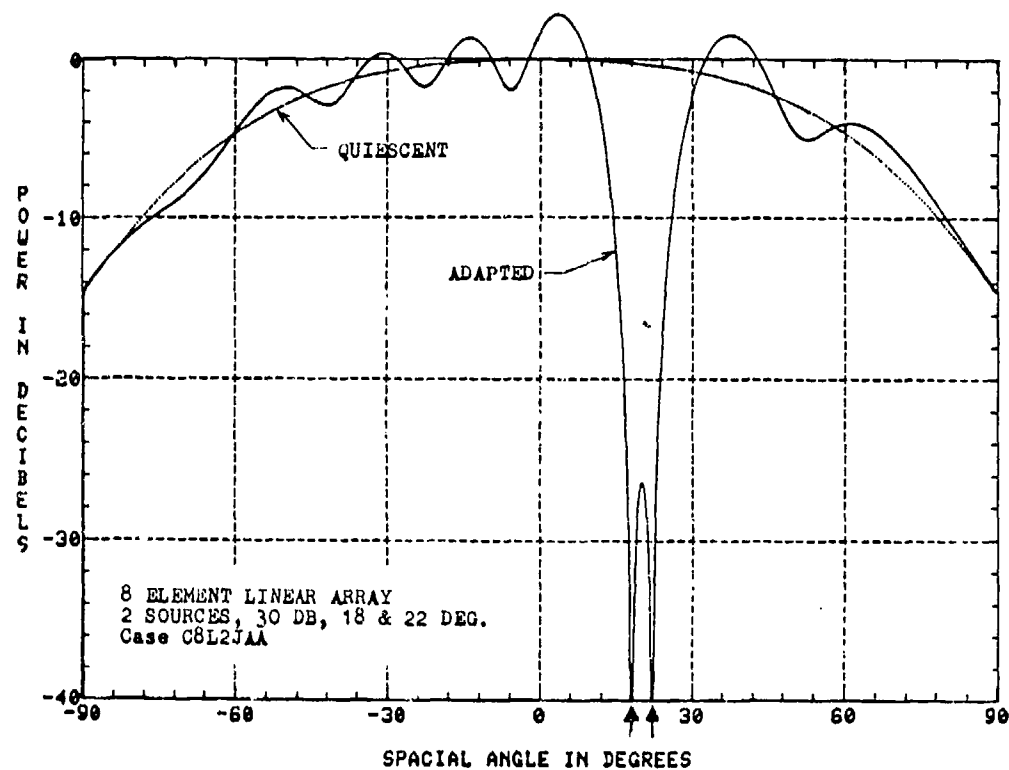


FIGURE 3. Quiescent (Single Element) and Adapted Patterns for Two-Source Case, Covariance Matrix Inverse Algorithm, 1024 Datasnaps

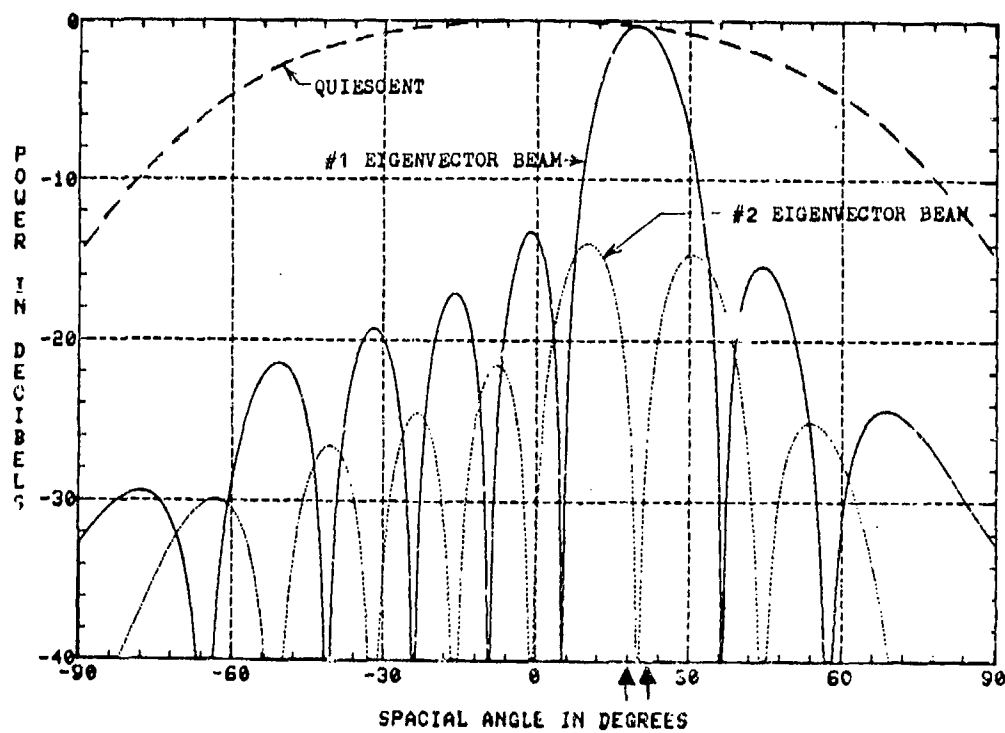


FIGURE 4. Eigenvector Component Beam Patterns for the Two-Source Case of Figure 3

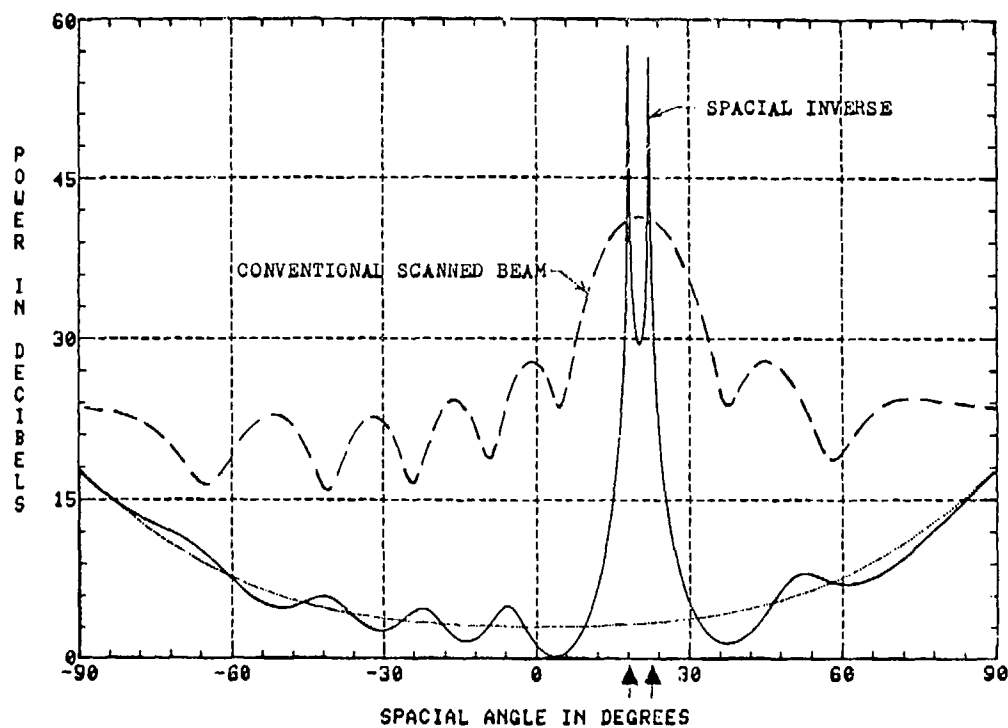


FIGURE 5. Spacial Spectrum Inverse Pattern for the Two-Source Case of Figure 3, and Comparison with Output of Conventional Scanned Beam

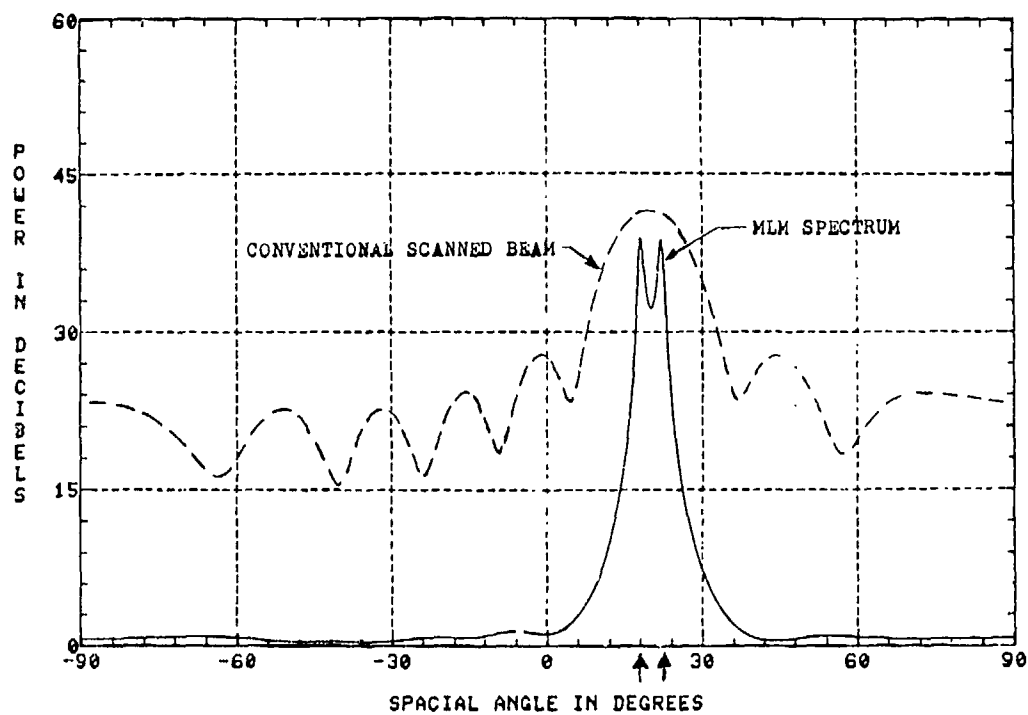


FIGURE 6. MLM Spacial Spectrum Plotted from Residual Power of Adaptive Zero-Order Mainbeam Constraint for the Two-Source Case of Fig. 3

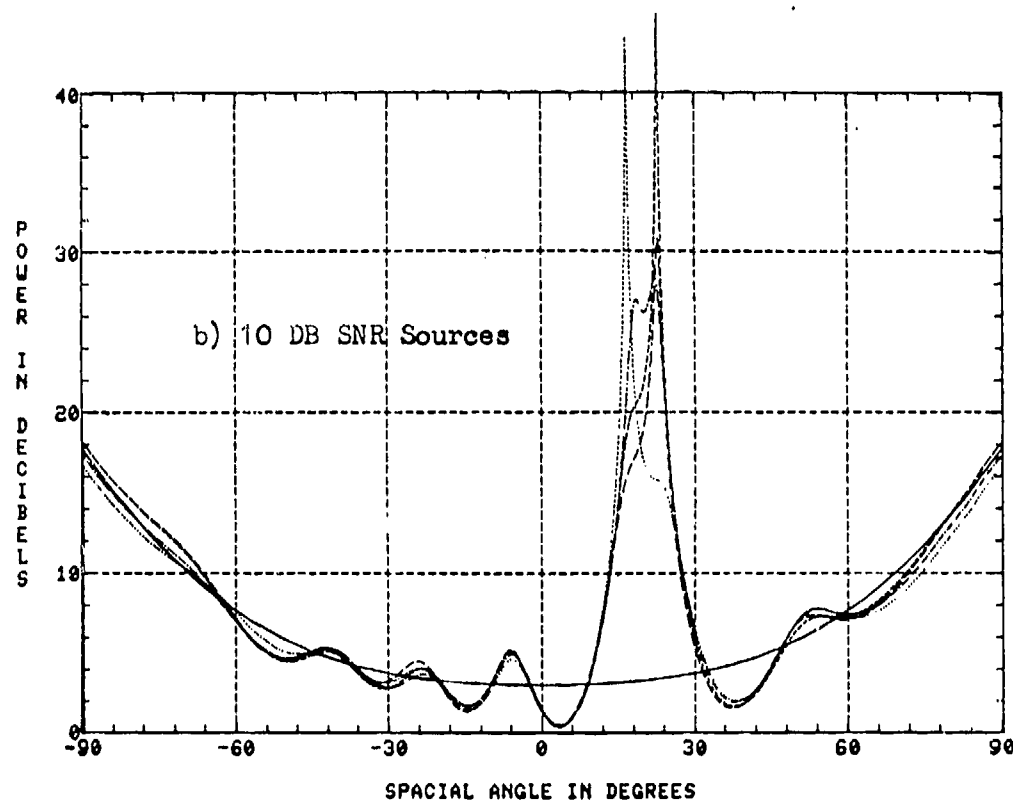
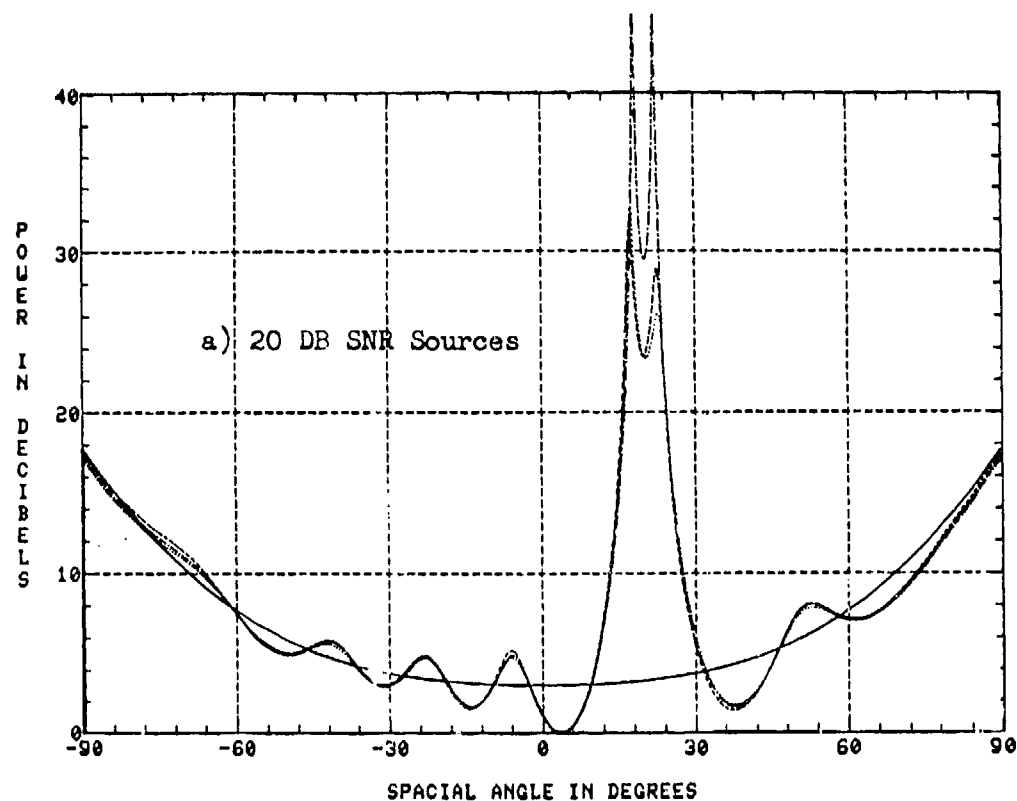


FIGURE 7. Typical Spacial Spectrum Snapshot Plots after Convergence. Two Sources Located at 18 and 22 Degrees

(NRL)

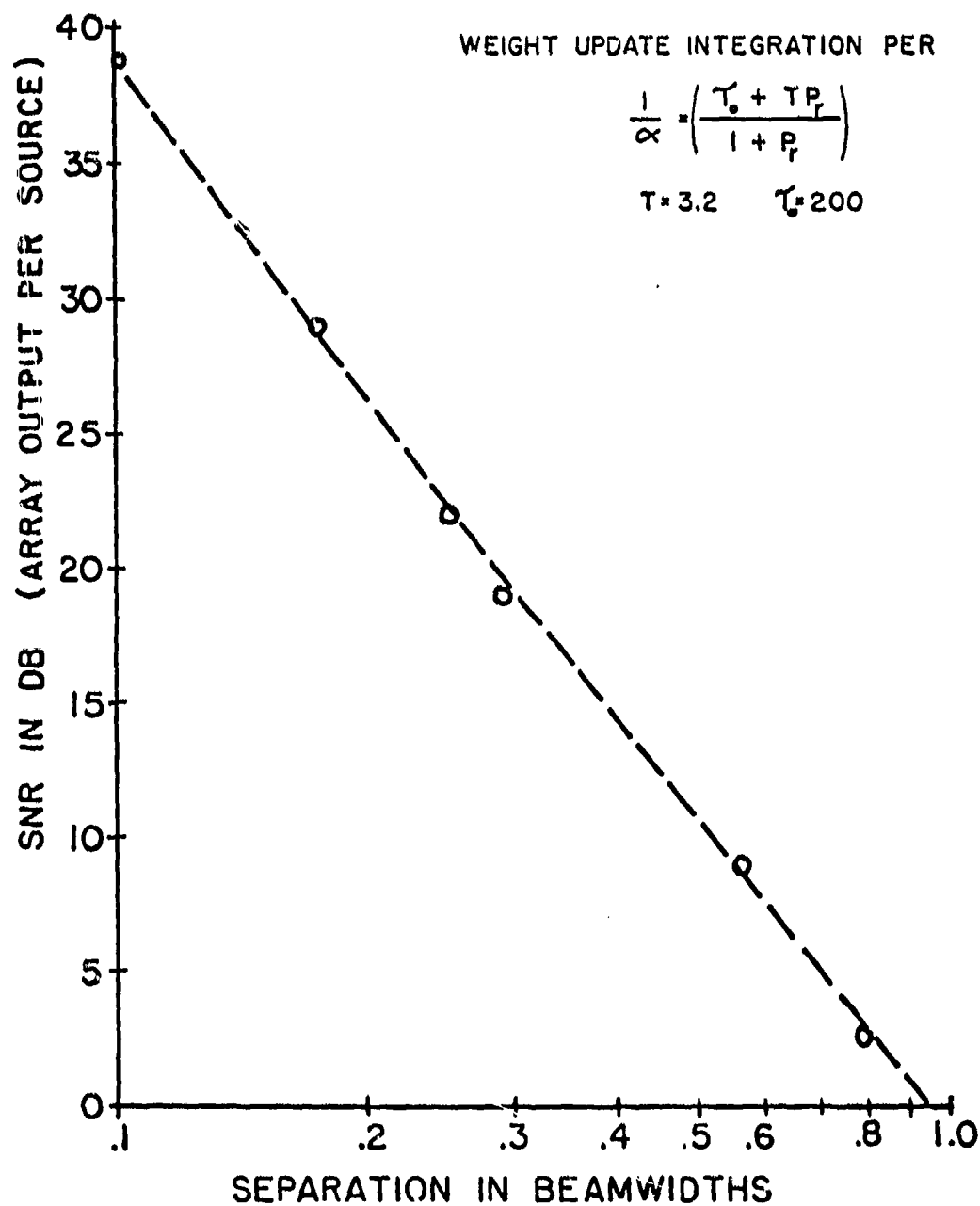


FIGURE 8. Universal Approximate Resolution Limit for Two Incoherent Sources, Simulation Conditions: Narrowband, No Array Errors, $\lambda/2$ Element Spacing, Linear Array, Gaussian Receiver Noise

APERTURE SAMPLING PROCESSING FOR
GROUND REFLECTION ELEVATION MULTIPATH CHARACTERIZATION *

JAMES E. EVANS and DAVID F. SUN

M.I.T. Lincoln Laboratory
Lexington, Mass. 02173

Abstract

The angular resolution and tracking of closely spaced targets is a classical radar problem which is receiving increased attention, and terrain multipath (e.g., reflections) has long been recognized to be a principal limitation on the achievable accuracy of radar elevation trackers at low elevation angles. This paper discusses the use of aperture sampling processing to improve the angular resolution/tracking and to characterize the multipath environment. The received signal is measured along the antenna aperture and the modern "high resolution" spectral estimation techniques (e.g., maximum likelihood and maximum entropy method) are applied to the spatial sample data. Experimental results of applying these techniques to field data from an L-band elevation array are presented. It is shown that maximum entropy processing offers improved performance in resolving multipath features and low angle tracking.

1. Introduction

This paper presents the results of an experimental program to obtain a better quantitative understanding of low angle microwave propagation phenomena and to assess the potential for improved elevation tracking performance by aperture sampling processing. It has long been recognized that terrain multipath (e.g., reflections and/or shadowing) are a principal limitation on the achievable accuracy of radar elevation trackers at low angles [1,2]. Figure 1-1a illustrates the propagation phenomena of interest. Since elevation tracker antennas generally have quite directional patterns in the elevation plane, a critical factor in refining and predicting the performance of an elevation tracker is the distribution of the received signal power as a function of elevation angle (i.e., the so-called angular power spectrum) [1]. Figure 1-1b illustrates the angular power spectrum that might arise with the multipath environment shown in Figure 1-1a. With our approach, we treat the problem of multipath environment characterization and target elevation angle estimation as one of estimating the angular power spectrum of the received signal.

* This work was sponsored by the Federal Aviation Administration. "The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the official policies, either expressed or implicit, of the United States Government".

The approach taken here is based on the adaptive processing of the received aperture information. Our goals are (1) to obtain higher resolution angular power spectrum than would be obtainable with the "conventional" beam sum spectrum for better characterization of the multipath environment and (2) to achieve better estimation of the target elevation angle than would be achievable with the "standard" tracking methods (e.g., monopulse) by utilizing this knowledge of the multipath characteristics. The starting point in our approach is measuring the (complex) received waveform (i.e., the amplitude and phase) at various points along the receiving antenna aperture. Next, we apply several high resolution spectral analysis techniques to the spatial sample data. Here, specifically, we consider the use of the maximum likelihood (ML) and the maximum entropy (ME) spectral estimation methods which have been applied in time series analysis and seismic/sonar array processing [3-6] for resolving the closely spaced spectral lines.

Although there is a simple duality between space and time (see Fig. 1-2) which permits one to apply time series analysis techniques, several features of our problem differ significantly from the usual time series application:

- (1) the data samples are complex (thus alleviating the "peak splitting" phenomena encountered with real data samples)
- (2) the number of data samples is generally small (e.g., 5-50)

and

- (3) in many cases the signals received from different directions are highly correlated in time, such that the sample spatial covariance is quite nonstationary. The consequence of this is that the results for certain phase relationships can differ substantially from the results for the ensemble covariance.

At various points in the subsequent discussion, we will illustrate the impact of these various features on the overall performance.

The remainder of the paper is organized as follows. Section II briefly describes the high resolution algorithms employed in our work and shows some examples of applying them to synthetic data. These algorithms are then used in analyzing field data from an L-band terrain reflection measurement program. These experimental results are presented in Section III, followed by the summary of the results in the last section.

II. HIGH RESOLUTION SPECTRUM ESTIMATION PROCEDURES

There has been much discussion of the maximum likelihood (ML) and the maximum entropy (ME) techniques recently in the geophysics and time series analysis literature. Thus, we will only present the key ideas together with pertinent references. The presentation of the various techniques can be

facilitated by use of vector notation. The notational convention we will use is that column vectors or matrices are represented by underlined lower-case letters. The asterisk denotes conjugate transposition. Underlined uppercase letters represent Hermetian matrices. Thus vector $\underline{s}$ represents the complex sensor outputs of the line array while the covariance matrix $\underline{R}$ has as its i, j th entry

$$R_{ij} = \overline{s_i^* s_j}$$

The ML estimation had its genesis in seismic array beamforming under conditions of directional interference [3] and adaptive array nulling of incoherent interfering sources [7]. The problem is formulated as determining the minimum variance unbiased estimate of the power from a given angle subject to the interference (complex) covariance matrix. If the interference were Gaussian with a known covariance matrix $\underline{Q}$ (e.g., via measurements in the absence of the desired signal), the maximum likelihood estimate of the power in a plane wave from angle θ would be given by

$$p_{ML}(\theta) = \underline{e}^* \underline{Q}^{-1} \hat{\underline{R}} \underline{Q}^{-1} \underline{e} / (\underline{e}^* \underline{Q}^{-1} \underline{e})^2 \quad (1)$$

where $\underline{e} = \exp(j2\pi x_k \sin \theta)$ is the received signal vector corresponding to a unit plane wave from angle θ and $\hat{\underline{R}} = \underline{s} \underline{s}^*$ is the sample covariance matrix.

Unfortunately, when the interfering signals are multipath and/or coherent jammers, $\underline{Q}$ cannot be measured independently of the desired signal. Capon [3] suggests using the sample covariance matrix $\hat{\underline{R}}$ as an estimate of $\underline{Q}$, so that the angle power spectrum estimate is then given by

$$p_{ML}(\theta) = (\underline{e}^* \hat{\underline{R}}^{-1} \underline{e})^{-1} \quad (2)$$

The estimate (2) may be contrasted to the standard "beam sum" (BS) angle power spectrum estimate of

$$\begin{aligned} p_{BS}(\theta) &= |\underline{e}^* \underline{s}|^2 \\ &= \left| \sum s(x_k) e^{-j2\pi x_k \sin \theta} \right|^2 \\ &= \underline{e}^* \hat{\underline{R}} \underline{e} \end{aligned} \quad (3)$$

This particular estimate gives rise to a $\sin K\theta/K\theta$ beam pattern which has a high sidelobe level (-13 dB). By weighting the data samples, lower sidelobes are obtained at the "cost" of wider beamwidths (i.e., poorer resolution) [8].

The use of maximum entropy method for high resolution spectrum estimation has been justified by a variety of arguments [4-6]. The physically most meaningful argument for radar applications is that the received angular spectrum can be represented by a finite number of poles in the complex plane, i.e.,

$$P(0) = \left| \sum_{i=1}^N (e^{j2\pi\delta\sin\theta} - z_i) \right|^{-2} \quad (4)$$

where z_i lies on or within the unit circle and the spatial samples are taken at points $x_k = k\delta$. The case of z_i on the unit circle could correspond to discrete plane waves while z_i inside the unit circle might correspond to an extended target (e.g., diffuse reflections). Time samples with the spectrum of (4) may be generated by passing a white noise process through an all pole filter of order N , which is a standard model in autoregression time series analysis. For the bulk of the data described here, only a short nonstationary set of spatial samples were available. Therefore, the Burg algorithm (modified for complex data values) has been used to determine the values of [9].

Much of the interest in the ME and ML algorithms has been generated by experiments in applying these methods to synthetic data. Fig. 2-1 and 2-2 show examples of applying the various algorithms to synthetic data consisting of 1 and 2 plane waves, respectively, with independent noise added to each sensor sample.* The actual angular power spectrum in each case consists of impulse functions at the plane wave angles. In the case of a single plane wave, all three estimates give the same peak location; however, the high resolution techniques more closely approximate the actual spectrum shape.

The example with two plane wave components is a case where the components are too close (0.84 standard beamwidths) to resolve by classical means. In this particular case, only the ME method gives an estimate close to the actual spectrum. The failure of the ML technique in this case is noteworthy also in view of its success in resolving plane wave signals when given the ensemble covariance as illustrated in Fig. 2-3. Although the ME technique was successful in the case shown in Fig. 2-2 at smaller separation angles (e.g., 0.25 standard beamwidths) and unfavorable phase relationships (e.g., 0° at the array center), it too is unsuccessful even at high signal to noise ratio.

This significant discrepancy between resolution performance for the ensemble and certain sample functions provides impetus for studies of alternative estimators. In this context, mention should be made of theoretical bounds on the performance of optimal two plane wave (sinusoid) parameter estimators which suggest that significant improvements in resolution performance at 0° or 180° phases may require very high signal to noise ratios [10, 11].

A radar tracker can be viewed as attempting to determine the centroid of the angular power spectrum peak corresponding to the direct signal. For an elevation tracker, the direct signal generally corresponds to the peak with

*In these and the following figures of this type, each type of angular spectrum estimate has been individually normalized to yield 0 dB peak value.

the most positive elevation angle. Our initial effort has considered such a peak finding criteria for ME spectra. Conventional radar trackers typically approximate the beam sum centroid by determining the null of the ratio $\epsilon(\theta) = [\Delta(\theta)/\Sigma(\theta)]$ where the difference pattern $\Delta(\theta) \approx d\Sigma(\theta)/d\theta$. When only a direct signal is present at angle θ_d and θ is within 1 beamwidth of θ_d ,

$$\epsilon(\theta) \approx (N\delta) (\theta - \theta_d) \quad (5)$$

so that one can estimate θ without pointing the array at θ_d . This gives rise to an "off boresight" elevation tracker whereby θ is constrained to be $\geq 0.7/N\delta$ and (5) is used to estimate θ_d if the last estimate of θ_d is less than $0.7/N\delta$. This keeps the main lobes of $\Sigma(\theta)$ and $\Delta(\theta)$ pointed above the terrain and thus significantly reduces the errors due to multipath signals at elevation angles below $\theta_d[1]$.

III. Experimental Results

As mentioned earlier, the terrain multipath is a principal limitation on the achievable accuracy of radar elevation trackers at low elevation angles. Low angle tracker development and performance prediction has been inhibited by the lack of experimental data on the angular distribution of the scattered power [1,2]. The objective of the work reported here was to utilize the aperture sampling and high resolution spectral estimation methods discussed earlier to analyze the field measurement data of the L-band ground reflection signals for better characterization of the ground reflection elevation multipath and for improved estimation of the target elevation angle.

Figure 3-1 shows the aperture sampling equipment utilized in the field measurements. The sampled aperture consisted of a 5 element 6.5λ line array (for evaluation of a small aperture tracker performance) as well as a 9 element 26λ line array (for fine grain resolution of various multipath components). Sensors in both array configurations were uniformly spaced. The beamwidths of these two arrays were approximately 7° and 1.75° , respectively. The received signal consisted of 1090 MHz replies from a standard air traffic control radar beacon (ATCRB) on board an aircraft in response to the ground interrogations. The amplitude and phase* of the received signal at each of 11 L-band dipoles (used as sensor) was digitized and recorded on magnetic disks. Also recorded were the digitized elevation angle of the target aircraft obtained from a tracking theodolite and various relevant environmental data.

The received signal consists of a plane wave at positive elevation angle (corresponding to the direct signal coming from the aircraft ATCRB) and other plane waves generally at negative elevation angles (corresponding to various

* The RF phase was measured relative to a reference dipole while the amplitude was measured on calibrated log video receivers.

ground reflections from terrain features). Thus, as shown in Figure 1-1, we expect that the angular power spectrum of the received signal (i.e., the received signal power as a function of elevation angle) will consist of a narrow peak at the direct signal elevation angle, narrow peaks at the arrival angles of the major specular ground reflections and wider peaks in regions of diffuse scattering [1].

In the results presented below, the maximum entropy (ME) angular power spectrum was calculated using the Burg technique [9,14], and the filter length of the corresponding prediction error filter was determined using Akaike's final prediction error criterion [15].

Figure 3-2 shows the experimental angular power spectral estimates for a special measurement at MIT Lincoln Laboratory antenna test range where the elevation array was laid sideways horizontally on the ground so as to have only a single plane wave incident on the array. The results are seen to correspond closely to the synthetic data result of Figure 2-1, and are viewed as providing a degree of validation for our data recording and analysis procedure.

Ground reflection field measurements were made for various terrain conditions. Results for both near-flat terrain and rolling terrain are given below. For comparison purposes, both the experimental angular power spectral estimates from the field measured data and the corresponding simulated spectral estimates (using the multipath computer simulation program developed for the Microwave Landing System multipath simulations [13]) are shown in the same figure.

Figure 3-3 shows the angular power spectral estimates for a flight test in which the target helicopter was at an angle of 1.4° and at a range of 0.4 nmi. Figure 3-4 shows the terrain height profile and the corresponding ground model used to generate the simulated spectral estimates. The terrain in front of the receiving antenna array consisted of a fairly flat grass field adjacent to the main runway at Hanscom airport, Mass. Thus, it is expected that the ground reflected signal would be primarily a specular reflection from the fairly flat ground which had been attenuated by the grass cover. In both measured and simulated results, it can be seen that all three angular power spectral estimates suggest the presence of two signals (one direct signal and one ground reflected signal); however, the ME spectral estimate appears to offer higher resolution of the signals as well as lower background spectral level. It has been shown that the area under an ME spectral peak provides a good estimate of the component power [4]. Based on this estimate of the component power, the estimated specular reflected signal power relative to the direct signal in Figure 3-3a is -3 dB which compares reasonably well with -3.5 dB for the corresponding simulated result in Figure 3-3b. Also, the estimated arrival angles of the ground reflected signals agrees fairly well between the field measurement and the corresponding simulation results.

Figure 3-5 shows the spectral estimates for a flight test in which the target helicopter was at 4.2° and 0.6 nmi. This field test was taken at the golf course of Fort Devens, Mass. Figure 3-6 shows the terrain height profile and the corresponding ground model used to produce the simulated spectral estimates. Here, the terrain in front of the receiving antenna array has various downward and upward slopes within a roughly level horizon and the ground was covered very much uniformly by short grass. This type of rolling terrain can often give rise to the "focusing" terrain reflections, i.e., more than one specular reflection presenting at a given time. We see in Figure 3-5 that both the field measured result and the simulation result indicates the existence of two ground reflected signals, one at -6.0° and the other at -1.7° with the latter having lower multipath level. Again, the ME spectral estimate appears to yield better resolution of various arriving signals and to give lower background spectral level.

Figure 3-7 shows experimental results for target elevation angle estimation of a flight test at golf course of Fort Devens, Mass. The flight path of the target helicopter was vertical descent at a range of 0.6 nmi covering elevation angles from 7.5° to 1.5° . We see that the elevation angle estimator based on the ME spectral estimates generally yields smaller angular errors than the conventional monopulse, especially in the low elevation angle region.

IV. Summary

Our preliminary results from the analysis of the low angle terrain scattering field measurements by utilizing the high resolution spectral estimation techniques suggest that these modern spectral estimation methods, especially the ME method, can be effectively used for ground reflection elevation multipath characterization and for improved target elevation angle estimation. However, several problems associated with applying these promising techniques to such array data need additional study. These include (1) choice of the "correct" prediction error filter length in the ME method, (2) the proper estimation of the covariance matrix to be used in the ML method, and (3) alternative estimators which are less sensitive to the relative phase between the various received signals.

Acknowledgments

R. Sandholm and P. Lanzillotti designated and operated the equipment used in the experimental measurements. J. Reid developed the bulk of software and operated on-site computer in the field measurements. K. Roberts typed the manuscript and captioned the figures. I. Stiglitz provided encouragement and assistance in commencing and carrying out the studies reported here.

References

1. Barton, D., June 1974, "Low-Angle Radar Tracking", Proc. of IEEE, p. 687.
2. White, W.D., November 1974, "Low Angle Radar Tracking in the Presence of Multipath", IEEE Trans. on AES, Vol. AES-10, No. 6, p. 835.
3. Capon, J., August 1969, "High-Resolution Frequency-wave Number Spectrum Analysis", Proc. of IEEE, Vol. 57, p. 1408.
4. Lacoss, R.T., August 1971, "Data Adaptive Spectral Analysis Methods", Geophysics, Vol. 56, No. 6, p. 661.
5. Burg, J.P., October 1967, "Maximum Entropy Spectral Analysis", paper presented at 37th International SEC Meeting, Oklahoma City, Oklahoma.
6. Van Den Bos, A. 1971, "Alternative Interpretation of Maximum Entropy Spectral Analysis", IEEE Trans. on Inform. Thres., IT-17, p. 693.
7. Special Issue on Adaptive Antennas, IEEE Trans. on Antennas and Propagation, Vol. AP-24, No. 5, September 1976.
8. Harris, F., January 1978, "On the Use of Windows for Harmonic Analysis With the Discrete Fourier Transform", Proc. of IEEE, Vol. 68, No. 1, pp. 51-83.
9. Anderson, N., February 1974, "On the Calculation of Filter Coefficients for Maximum Entropy Spectral Analysis", Geophysics, Vol. 39, pp. 66-72.
10. Sklar, J. R., and Schweppe, F.C., 1964, "The Angular Resolution of Multiple Targets", M.I.T. Lincoln Laboratory, Rpt. 1964-2.
11. Pollon, G.E., 1967, "On the Angular Resolution of Multiple Targets", IEEE Trans. Aerosp. Electron. System (Corresp.), Vol. AES-3, pp. 145-148.
12. Peterson, A.M., et al, 1976, "Low Angle Radar Tracking", Standford Research Institute Report JSR74-7.
13. Capon, J., April 1976, "Multipath Parameter Computations for the MLS Simulation Computer Program", M.I.T. Lincoln Laboratory Project Report ATC-68, FAA-RD-76-55.
14. Bernard, T.E., "The Maximum Entropy Spectrum and the Burg Technique", Advanced Signal Processing Technical Report No. 1, Texas Instruments Inc., ALEX103-TR-75-01.
15. Akaike, H., 1970, "Statistical Predictor Identification", Ann. Inst. Statist. Math., Vol. 22, p. 205.
16. Cox, H., 1973, "Resolving Power and Sensitivity to Mismatch of Optimum Array Processors", Jour. Acoust. Soc. America, Vol. 54, pp. 771-785.

Figures

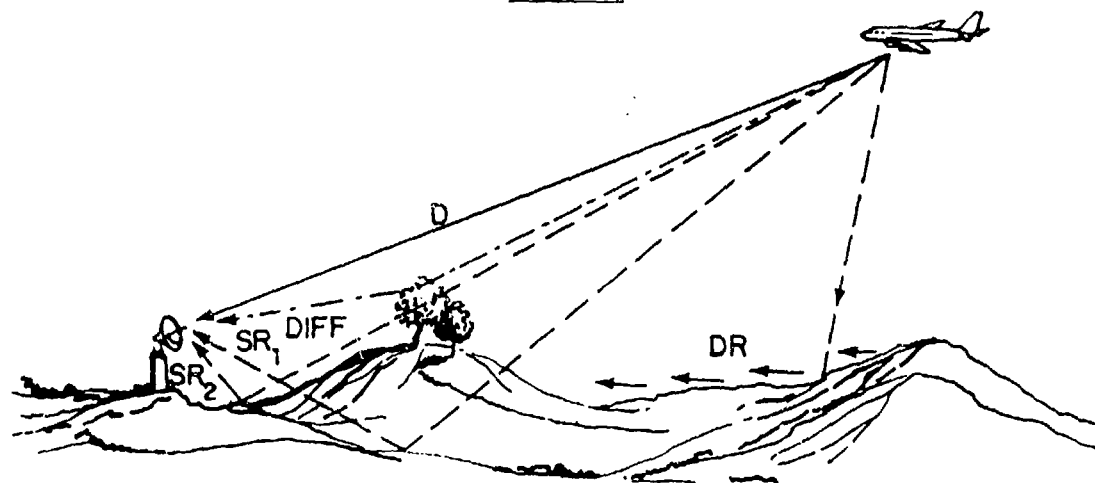


FIGURE 1-1a Multipath Propagation Phenomena of Interest.

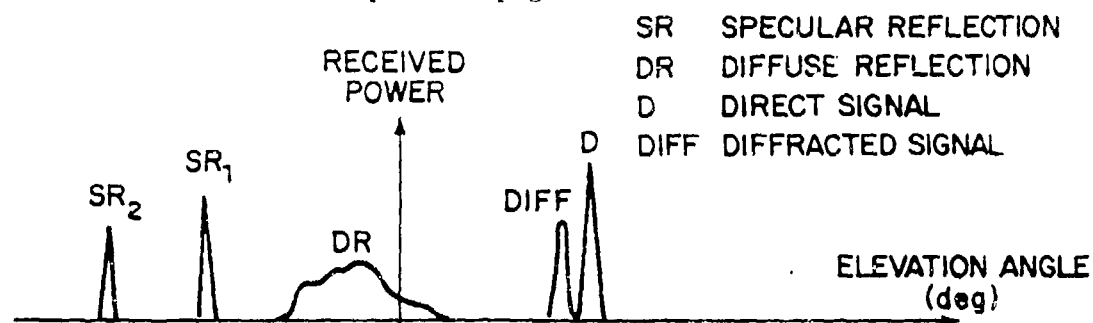


FIGURE 1-1b Relationship of Received Power to Low-Angle Multipath Environment.

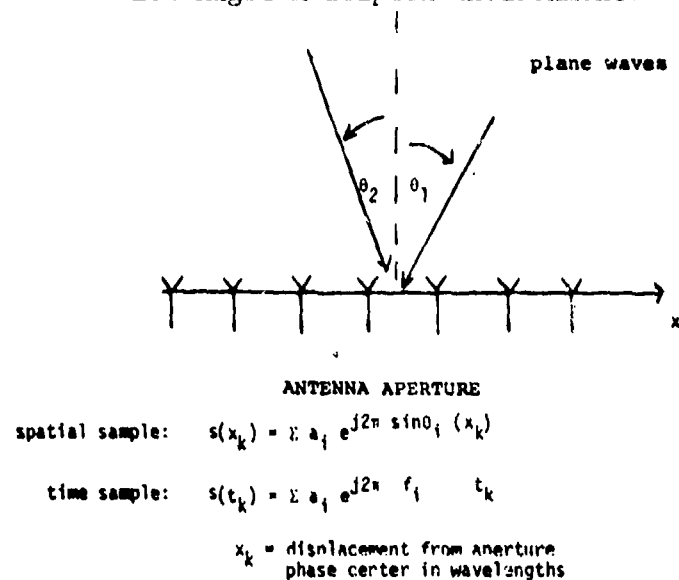


FIGURE 1-2 Duality Between Angular Estimation with Line Array and Frequency Estimation for Time Series.

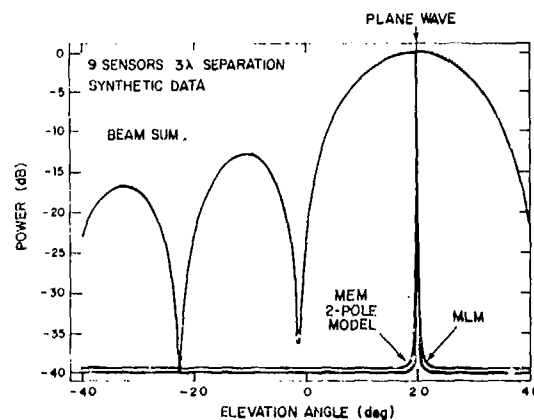


FIGURE 2-1 Application of High Resolution Techniques to Single Plane Wave Received Signal Model.

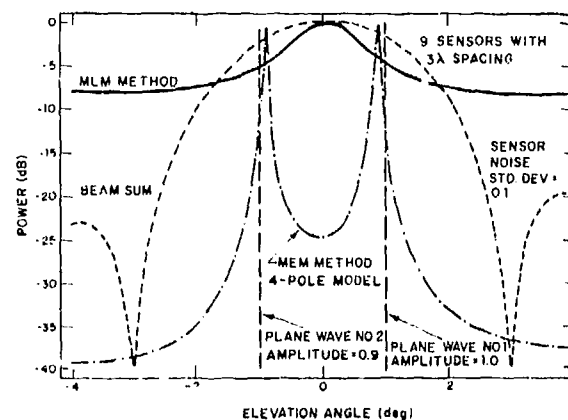


FIGURE 2-2 Application of High Resolution Techniques to Two Plane Wave Received Signal Model.

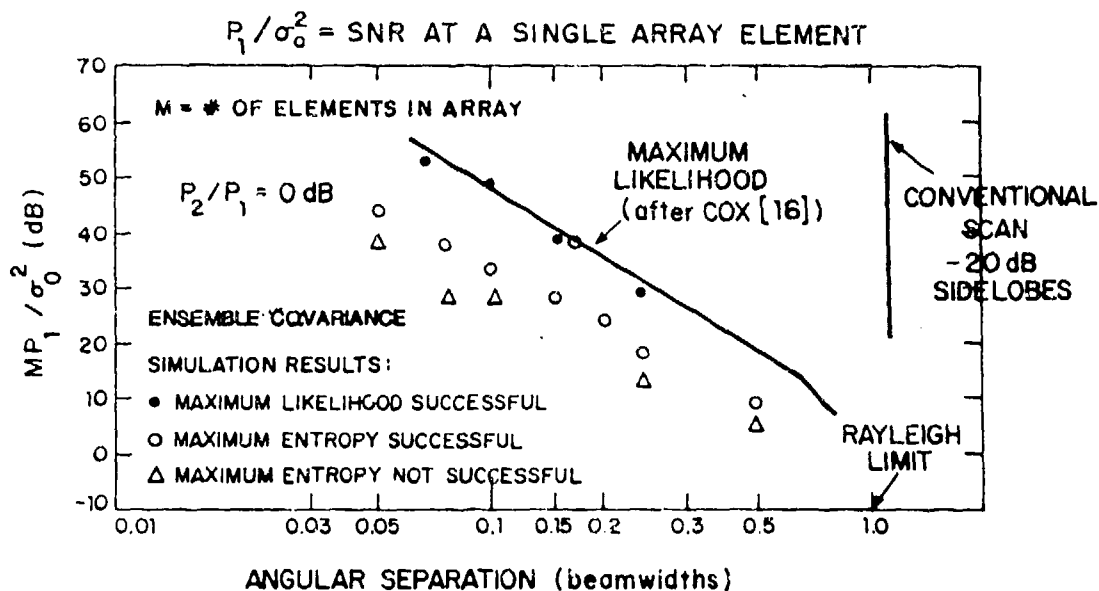


FIGURE 2-3 Comparison of Angular Spectrum Resolution for Equal Power Incoherent Sources.

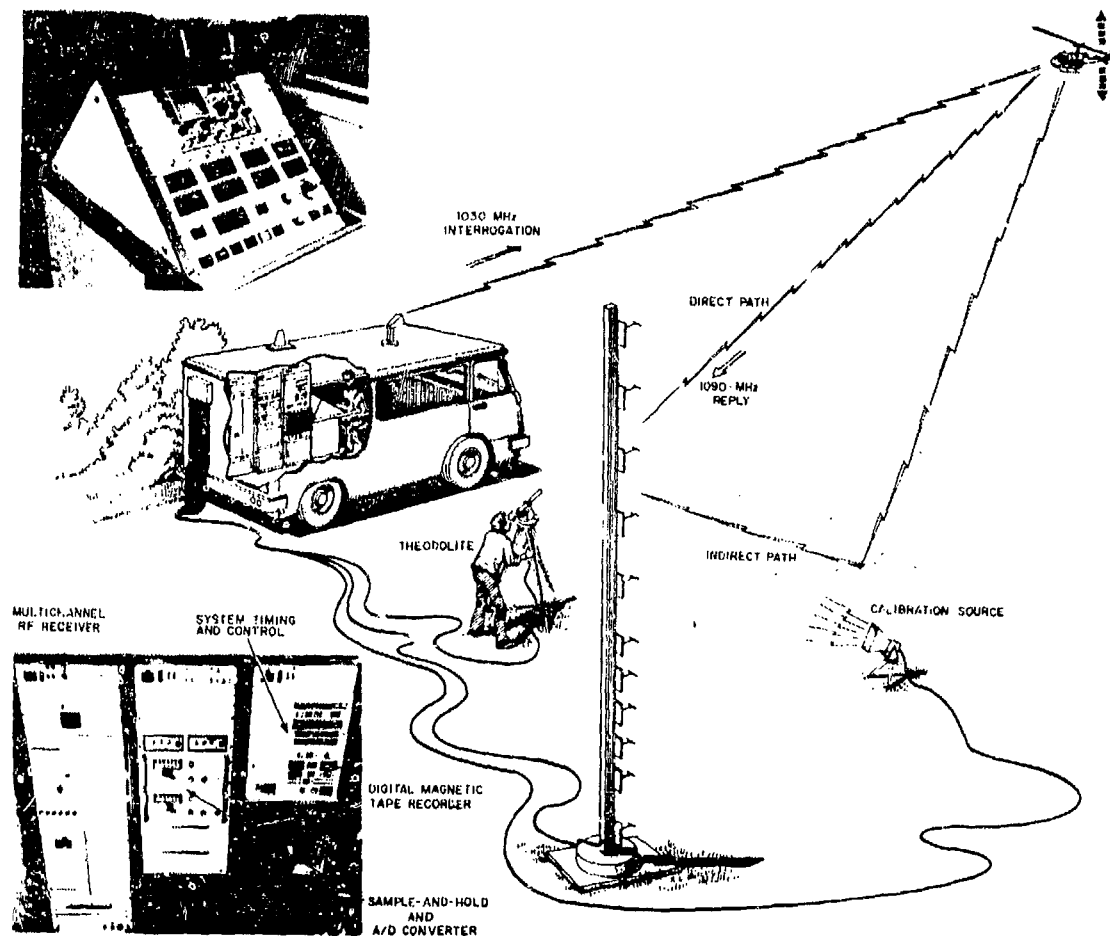


FIGURE 3-1 Terrain Multipath Experimental Configuration.

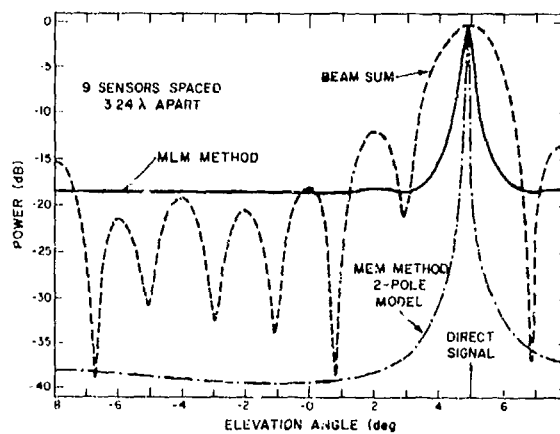
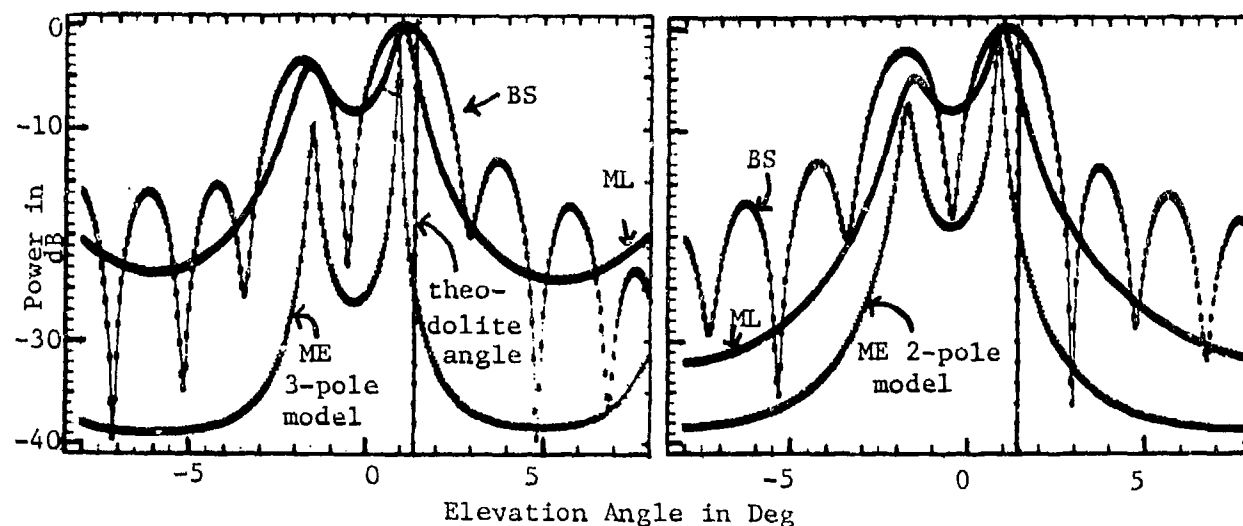


FIGURE 3-2 Angular Spectrum Estimates for Field Measurement with Only Direct Signal Present.



a) field measured results: 9 sensors spaced 3.24λ apart Hanscom field measurement 10/21/77

b) simulation results: 9 sensors spaced 3.24λ apart

FIGURE 3-3 Angular Spectrum Estimates for Field Measurements with Terrain Reflection: Near-Flat Terrain.

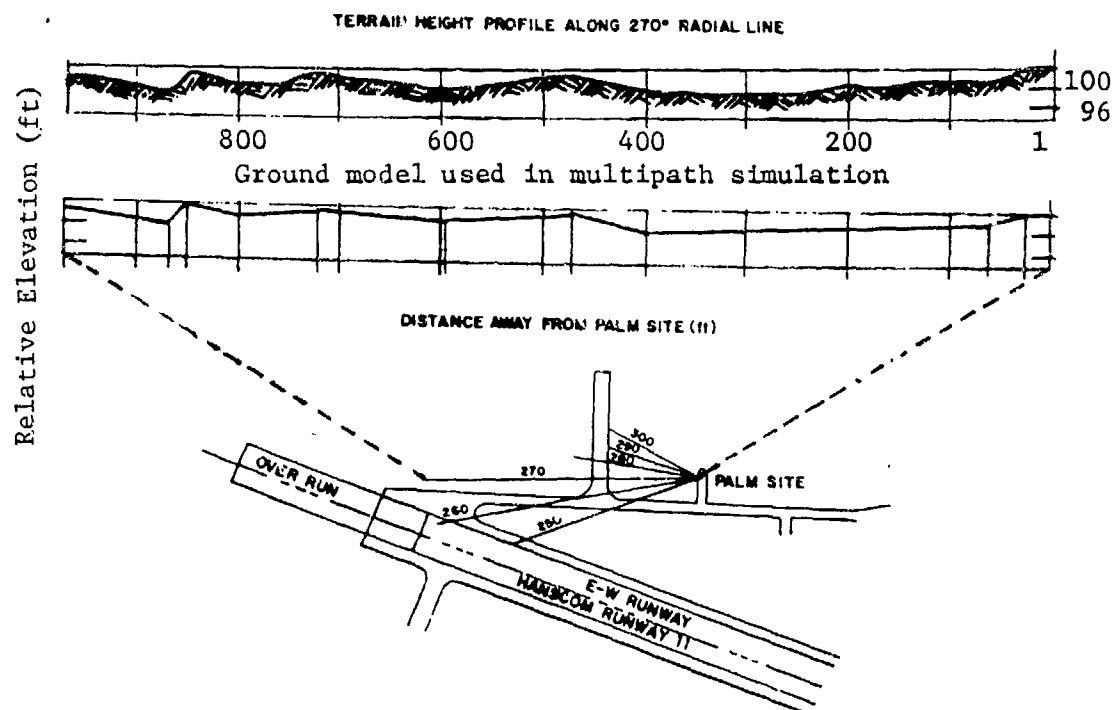
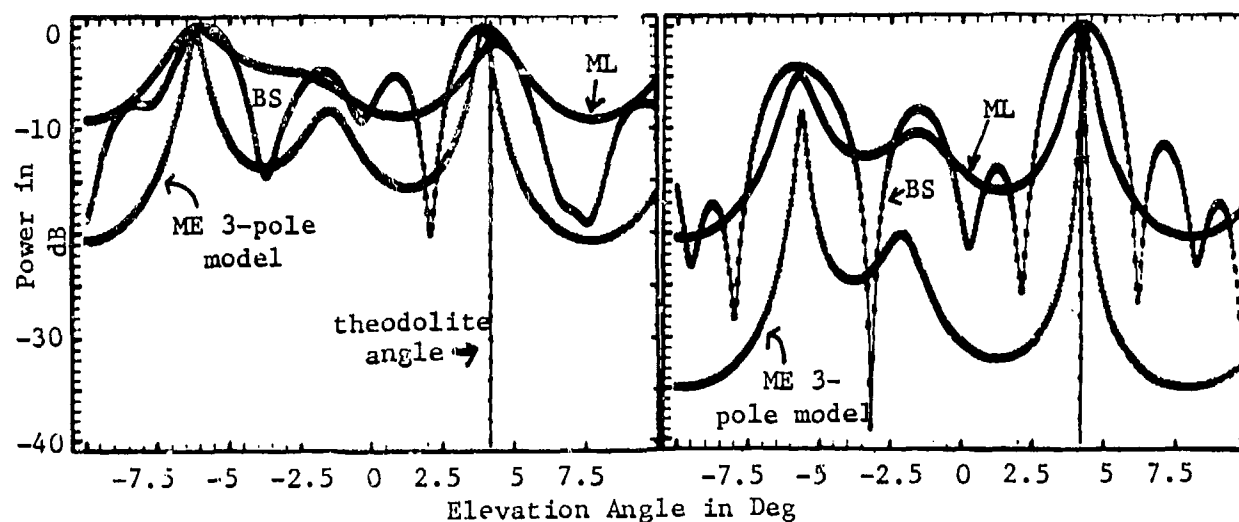


FIGURE 3-4 Terrain Height Profile at Hanscom AFB: Near-Flat Terrain.



- a) field measured results: 9 sensors 3.24 λ apart Fort Devens field measurements 5/1/78
- b) Simulation results: 9 sensors spaced 3.24 λ apart.

FIGURE 3-5 Angular Spectral Estimates for Field Measurements with Terrain Reflections: Rolling Terrain.

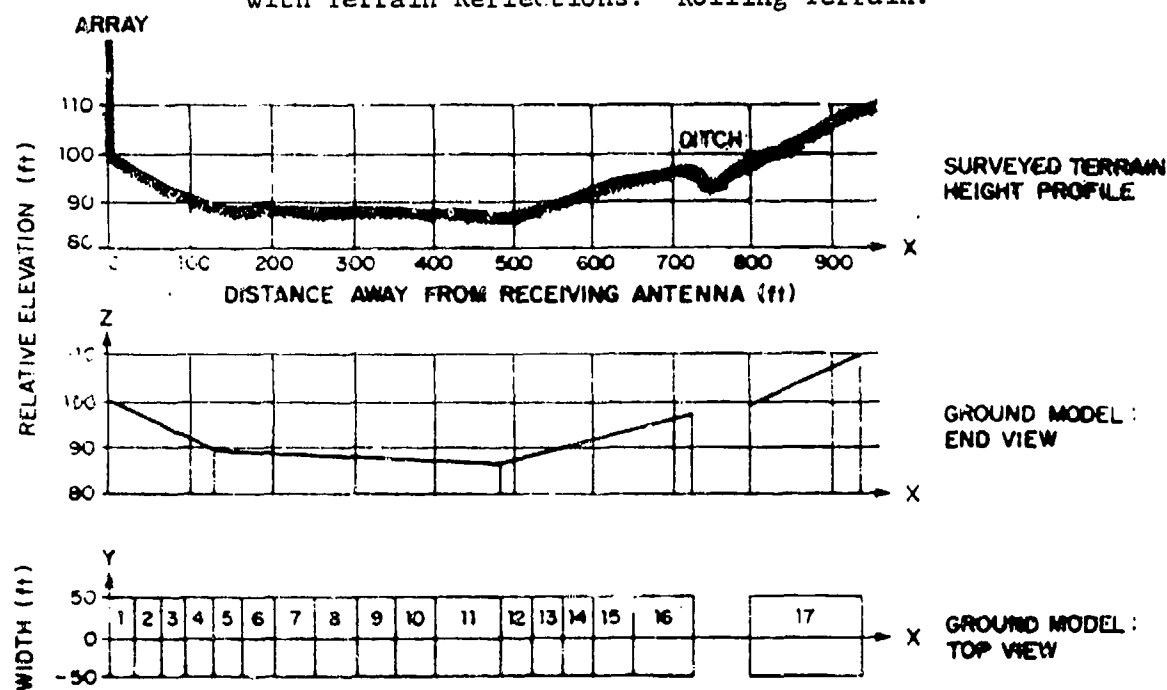


FIGURE 3-6 Terrain Height Profile at Fort Devens Golf Course: Rolling Terrain.

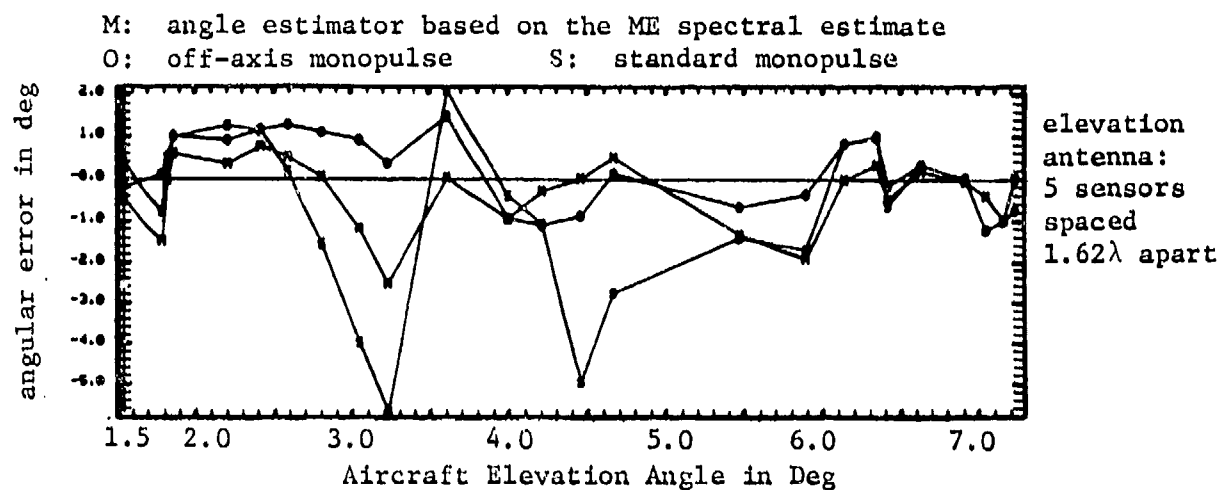


FIGURE 3-7 Angular Error in Target Elevation Angle Estimation for the Field Measurements at Fort Devens, Mass.

MULTIPLE EMITTER LOCATION AND SIGNAL PARAMETER ESTIMATION

RALPH SCHMIDT

ESL, Incorporated
495 Java Drive
Sunnyvale, CA 94086

Abstract

Processing the signals received on an array of sensors for the location of the emitter is of great enough interest to have been treated under many special case assumptions.

The general problem considers sensors with arbitrary locations and arbitrary directional characteristics (gain/phase/polarization) in a noise/interference environment of arbitrary covariance matrix.

This report is concerned first with the multiple emitter aspect of this problem and second with the generality of solution. A description is given of the Multiple Signal Classification (MUSIC) algorithm, which provides asymptotically unbiased estimates of

1. number of incident wavefronts present
2. directions-of-arrival (or emitter locations)
3. strengths and cross-correlations among the incident waveforms
4. noise/interference strength.

Examples and comparisons with methods based on Maximum Likelihood and Maximum Entropy, as well as conventional beam forming are included. An example of its use as a multiple frequency estimator operating on time series is included.

Introduction

The term multiple signal classification (MUSIC) is used to describe experimental and theoretical techniques involved in determining the parameters of multiple wavefronts arriving at an antenna array from measurements made on the signals received at the array elements.

The general problem considers antennas with arbitrary locations and arbitrary directional characteristics (gain/phase/polarization) in a noise/interference environment of arbitrary covariance matrix. The Multiple Signal Classification (MUSIC) approach is described; it can be implemented as an algorithm to provide asymptotically unbiased estimates of

1. Number of signals
2. Directions-of-arrival
3. Strengths and cross-correlations among the directional waveforms
4. Polarizations
5. Strength of noise/interference.

These techniques are very general and of wide application. Special cases of MUSIC are

1. Conventional Interferometry
2. Monopulse DF, i.e., using multiple colocated antennas
3. Multiple Frequency Estimation.

The Data Model

The waveforms received at the M array elements are linear combinations of the D incident wavefronts and noise. Thus, the multiple signal classification (MUSIC) approach begins with the following model for characterizing the received M vector X as in

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} = \begin{bmatrix} | & | & & | \\ a(\theta_1) & a(\theta_2) & \dots & a(\theta_D) \\ | & | & & | \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_D \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_M \end{bmatrix}$$

or

$$X = AF + W \quad (1)$$

The incident signals are represented in amplitude and phase at some arbitrary reference point (for instance the origin of the coordinate system) by the complex quantities $F_1, F_2, \dots, F_D$. The noise, whether "sensed" along with the signals or generated internal to the instrumentation, appear as the complex vector W .

The elements of X and A are also complex in general. The a_{ij} are known functions of the signal arrival angles and the array element locations. That is, a_{ij} depends on the i^{th} array element, its position relative to the origin of the coordinate system, and its response to a signal incident from the direction of the j^{th} signal. The j^{th} column of A is a "mode" vector $a(\theta_j)$ of responses to the direction-of-arrival θ_j of the j^{th} signal. Knowing the mode vector $a(\theta_1)$ is tantamount to knowing θ_1 (unless $a(\theta_1) = a(\theta_2)$ with $\theta_1 \neq \theta_2$, an unresolvable situation, a type I ambiguity).

In geometrical language, the measured X vector can be visualized as a vector in M dimensional space. The directional mode vectors $a(\theta_j) = a_{ij}$ for $i = 1, 2, \dots, M$, i.e., the columns of A , can also be so visualized. Equation (1) states that X is a particular linear combination of the mode vectors; the elements of F are the coefficients of the combination. Note that the X vector is confined to the range space of A . That is, if A has 2 columns, the range space is no more than a 2-dimensional subspace within the M space and X necessarily lies in the subspace. Also note that $a(\theta)$, the continuum of all possible mode vectors lies within the M space but is quite nonlinear. For help in visualizing this, see Figure 1. For example, in an azimuth-only DF system, θ will consist of a single parameter. In an azimuth/elevation/range system, θ will be replaced by θ, ϕ, r for example. In any case, $a(\theta)$ is a vector continuum such as a "snake" (azimuth only) or a "sheet" (AZ/EL) twisting and winding through the M space. (In practice, the procedure by which the $a(\theta)$ continuum is measured or otherwise established corresponds to calibrating the array.)

In these geometrical terms (see Figure 1), the problem of solving for the directions-of-arrival of multiple incident wavefronts consists of locating the intersections of the $a(\theta)$

continuum with the range space of A. The range space of A is, of course, obtained from the measured data. The means of obtaining the range space and, necessarily, its dimensionality (the number D of incident signals) follows.

The S Matrix

The MxM covariance matrix of the X vector is

$$S \triangleq \overline{XX^*} = A \overline{FF^*} A^* + \overline{WW^*}$$

or

$$S = APA^* + \lambda S_0 \quad (2)$$

under the basic assumption that the incident signals and the noise are uncorrelated. Note that the incident waveforms represented by the elements of F may be uncorrelated (the DxD matrix $P \triangleq \overline{FF^*}$ is diagonal) or may contain completely correlated pairs (P is singular). In general, P will be "merely" positive definite which reflects the arbitrary degrees of pairwise correlations occurring between the incident waveforms.

When the number of incident wavefronts D is less than the number of array elements M, then APA^* is singular; it has a rank less than M. Therefore

$$|APA^*| = |S - \lambda S_0| = 0 \quad (3)$$

This equation is only satisfied with λ equal to one of the eigenvalues of S in the metric of S_0 . But, for A full rank and P positive definite, APA^* must be nonnegative definite. Therefore λ can only be the minimum eigenvalue $\lambda_{\min}$.

Therefore, any measured $S = \overline{XX^*}$ matrix can be written

$$S = APA^* + \lambda_{\min} S_0, \quad \lambda_{\min} \geq 0 \quad (4)$$

where $\lambda_{\min}$ is the smallest solution to $|S - \lambda S_0| = 0$. Note the special case wherein the elements of the noise vector W are mean zero, variance σ^2 ; in which case, $\lambda_{\min} S_0 = \sigma^2 I$.

Calculating a Solution

The rank of APA^* is D and can be determined directly from the eigenvalues of S in the metric of S_0 . That is, in the complete set of eigenvalues of S in the metric of S_0 , $\lambda_{\min}$ will not always be simple. In fact, it occurs repeated $N = M - D$ times.

This is true because the eigenvalues of S and those of $S - \lambda_{\min} S_0 = APA^*$ differ by $\lambda_{\min}$ in all cases. Since the minimum eigenvalue of APA^* is zero (being singular), $\lambda_{\min}$ must occur repeated N times. Therefore, the number of incident signals estimator is

$$\hat{D} = M - \hat{N} \quad (5)$$

where $\hat{N}$ = the multiplicity of $\lambda_{\min}(S, S_0)$ and $\lambda_{\min}(S, S_0)$ is read " $\lambda_{\min}$ of S in the metric of S_0 ." (In practice, one can expect that the multiple $\lambda_{\min}$'s will occur in a cluster rather than all precisely equal. The "spread" on this cluster decreases as more data is processed.)

The Signal and Noise Subspaces

The M eigenvectors of S in the metric of S_0 must satisfy $Se_i = \lambda_i S_0 e_i$, $i = 1, 2, \dots, M$. Since $S = APA^* + \lambda_{\min} S_0$, we have $APA^* e_i = (\lambda_i - \lambda_{\min}) S_0 e_i$. Clearly, for each of the λ_i that is equal to $\lambda_{\min}$ - there are N - we must have $APA^* e_i = 0$ or $A^* e_i = 0$. That is, the eigenvectors associated with $\lambda_{\min}(S, S_0)$ are orthogonal to the space spanned by the columns of A ; the incident signal mode vectors!

Thus we may justifiably refer to the N dimensional subspace spanned by the N noise eigenvectors as the noise subspace and the D dimensional subspace spanned by the incident signal mode vectors as the signal subspace; they are disjoint.

The Algorithm

We now have the means to solve for the incident signal mode vectors. If E_N is defined to be the $M \times N$ matrix whose columns are the N noise eigenvectors, and the ordinary Euclidean distance (squared) from a vector Y to the signal subspace is $d^2 = Y^* E_N E_N^* Y$, we can plot $1/d^2$ for points along the $a(\theta)$ continuum as a function of θ . That is,

$$P_{MU}(\theta) = \frac{1}{a^*(\theta) E_N E_N^* a(\theta)} \quad (6)$$

(However, the $a(\theta)$ continuum may intersect the D dimensional signal subspace more than D times; another unresolvable situation occurring only for the case of multiple incident signals - a type II ambiguity.) It is clear from the expression that MUSIC is asymptotically unbiased even for multiple incident wavefronts because S is asymptotically perfectly measured so that E_N is also. $a(\theta)$ does not depend on the data.

Once the directions-of-arrival of the D incident signals have been found, the A matrix becomes available and may be used to compute the parameters of the incident signals. The solution for the P matrix is direct and can be expressed in terms of $(S - \lambda_{\min} S_0)$ and A. That is, since $APA^* = S - \lambda_{\min} S_0$,

$$P = (A^*A)^{-1}A^*(S - \lambda_{\min} S_0)A(A^*A)^{-1} \quad (7)$$

Including Polarization

Consider a signal arriving from a specific direction θ_0 . Assume that the array is not diverse in polarization; i.e., all elements are identically polarized, say, vertically. Certainly the DF system will be most sensitive to vertically polarized energy, completely insensitive to horizontal and partially sensitive to arbitrarily polarized energy. The array is only sensitive to the vertically polarized component of the arriving energy.

For a general or polarizationally diverse array, the mode vector corresponding to the direction θ_0 depends on the signal polarization. A vertically polarized signal will induce one mode vector and horizontal another, and right hand circular (RHC) still another.

Recall that signal polarization can be completely characterized by a single complex number q . We can "observe" how the mode vector changes as the polarization parameter q for the emitter changes at the specific direction θ_0 . It can be proven that as q changes through all possible polarizations, the mode vector sweeps out a two-dimensional "polarization subspace." Thus, only two independent mode vectors spanning the polarization subspace for the direction θ_0 are needed to represent any emitter polarization q at direction θ_0 . The practical embodiment of this is that only the mode vectors of two emitter polarizations need be calculated or kept in store for direction θ_0 in

order to solve for emitter polarizations where only one was needed to solve for DOA in a system with an array that was not polarizationally diverse.

These arguments lead to an equation similar to Equation (6) for $P(\theta)$ but including the effects of polarization diversity among the array elements.

$$P_{MU}(\theta) = \frac{1}{\lambda_{\min} \left(\begin{bmatrix} a_x^*(\theta) \\ a_y^*(\theta) \end{bmatrix} E_N E_N^* \begin{bmatrix} a_x(\theta) \\ a_y(\theta) \end{bmatrix} \right)} \quad (8)$$

where $a_x(\theta)$ and $a_y(\theta)$ are the two continua corresponding to, for example, separately taken x and y linear incident wavefront polarizations. The eigenvector corresponding to $\lambda_{\min}$ in Equation (8) provides the polarization parameter q since it is of the form $[1 \ q]^T$.

The Algorithm

In summary, the steps of the algorithm are

- Step 0: Collect data, form S
- Step 1: Calculate Eigenstructure of S in metric of S_0
- Step 2: Decide number of signals D ; Equation (5)
- Step 3: Evaluate $P_{MU}(\theta)$ vs. θ ; Equation (6) or (8)
- Step 4: Pick D peaks of $P_{MU}(\theta)$
- Step 5: Calculate remaining parameters; Equation (7).

The above steps have been implemented in several forms to verify and evaluate the principles and basic performance. Field tests have been conducted using actual receivers, arrays, and multiple transmitters. The results of these tests have demonstrated the potential of this approach for handling multiple signals in practical situations. Performance results are being prepared for presentation in another paper.

Comparison With Other Methods

In comparing MUSIC with ordinary beamforming (BF) Maximum Likelihood (ML) and Maximum Entropy (ME), the following expressions were used. See Figures 3 and 4.

$$P_{BF}(\theta) = a^*(\theta) S a(\theta)$$

$$P_{ML}(\theta) = \frac{1}{a^*(\theta) S^{-1} a(\theta)}$$

$$P_{ME}(\theta) = \frac{1}{a^*(\theta) c c^* a(\theta)}$$

where c is a column of S^{-1} . The beamformer expression calculates for plotting the power one would measure at the output of a beamformer (summing the array element signals after inserting delays appropriate to steer or look in a specific direction) as a function of the direction.

$P_{ML}(\theta)$ calculates the log likelihood function under the assumptions that X is a mean zero, multivariate Gaussian and that there is only a single incident wavefront present. For multiple incident wavefronts, $P_{ML}(\theta)$ becomes

$$P_{ML}(\theta) = \frac{1}{\lambda_{\min}(A_{\theta}^* S^{-1} A_{\theta})}$$

which implies a D dimensional search (and plot!)

$P_{ME}(\theta)$ is based on selecting 1 of the M array elements as a "reference" and attempting to find weights to be applied to the remaining $M-1$ received signals to permit their sum with a MMSE fit to the reference. Since there are M possible references, there are M generally different $P_{ME}(\theta)$'s obtained from the M possible column selections from S^{-1} . In the comparison plots, a particular reference was consistently selected.

An example of the completely general MUSIC algorithm applied to a problem of steering a multiple feed parabolic dish antenna is shown in Figure 5. $\frac{\sin x}{x}$ pencil beamshapes skewed slightly off boresight are assumed for the element patterns. Since the six antennas are essentially colocated, the DF capacity arises out of the antenna beam pattern diversity. The computer was used to simulate the "noisy" S matrix that would arise in practice for the conditions desired and then to subject

it to the MUSIC algorithm. Figure 5 shows how three directional signals are distinguished and their polarizations estimated even though two of the arriving signals are highly similar (90% correlated).

The application of MUSIC to the estimation of the frequencies of multiple sinusoids (arbitrary amplitudes and phases) for a very limited duration data sample is shown in Figure 6. The figure suggests that, even though there was no actual noise included, the rounding of the data samples to six decimal digits has already destroyed a significant portion of the information present in the data needed to resolve the several frequencies.

Summary and Conclusions

As this paper was being prepared, the works of Gething[1] and Davies[2] were discovered, offering a part of the solution discussed here but in terms of simultaneous equations and special linear relationships without recourse to eigenstructure. However, the geometric significance of a vector space setting and the interpretation of the S matrix eigenstructure was missed. More recent work by Reddi[3] is also along the lines of the work presented here though limited to uniform, collinear arrays of omnidirectional elements and also without clear utilization of the entire noise subspace. Ziegenbein[4] applied the same basic concept to time series spectral analysis referring to it as a Karhunen-Loeve Transform though treating aspects of it as "ad hoc". El-Behery and MacPhie[5] and Capon[6] treat the uniform collinear array of omnidirectional elements using the Maximum Likelihood method. Pisarenko[7] also treats time series and addresses only the case of a full complement of sinusoids; i.e., a 1 dimensional noise subspace.

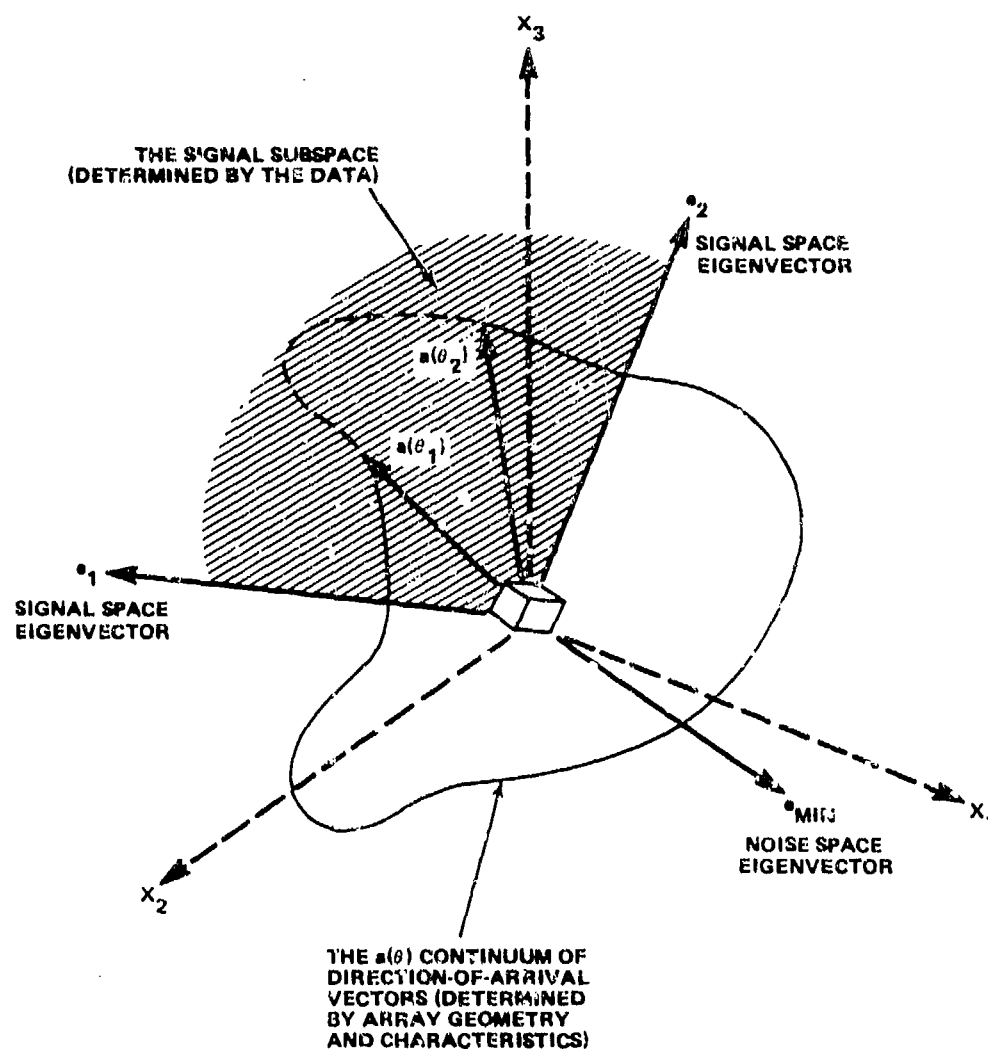
The approach presented here for Multiple Signal Classification (MUSIC) is very general and of wide application. The method is interpretable in terms of the geometry of complex M spaces wherein the eigenstructure of the measured S matrix plays the central role. MUSIC provides asymptotically unbiased estimates of a general set of signal parameters approaching the Cramer-Rao accuracy bound. MUSIC models the data as the sum of point source emissions and noise rather than the convolution of an all pole transfer function driven by a white noise (i.e., autoregressive modeling, Maximum Entropy) or maximizing a probability under the assumption that the X vector is zero mean, Gaussian (Maximum Likelihood for Gaussian data). In geometric

terms MUSIC minimizes the distance from the $a(\theta)$ continuum to the signal subspace whereas Maximum Likelihood minimizes a weighted combination all component distances.

No assumptions have been made about array geometry. The array elements may be arranged in a regular or irregular pattern and may differ or be identical in directional characteristics (amplitude/phase) provided their polarization characteristics are all identical. The extension to include general polarizationally diverse antenna arrays will be more completely described in a separate paper.

References

1. Gething, P.J.D., Oct. 1971, Analysis of Multicomponent Wavefields, Proc. IEE, Vol. 118, no 10.
2. Davies, D.E.N., March 1967, Independent Angular Steering of Each Zero of the Directional Pattern for a Linear Array, IEEE Trans. on Antennas and Propagation.
3. Reddi, S.S., Jan. 1979, Multiple Source Location - A Digital Approach, IEEE Trans. on Aerospace and Electronic Systems, Vol. AES-15, no. 1.
4. Ziegenbein, J., April 2-4, 1979, Spectral Analysis Using the Karhunen-Loeve Transform, 1979 IEEE Int'l. Conf. on ASSP, Washington, D.C., P. 182 - 185.
5. El-Behery, I.N., MacPhie, R.H., July 1977, Maximum Likelihood Estimation of Source Parameters from Time-Sampled Outputs of a Linear Array, J. Aconst. Soc. Am., Vol. 62, no. 1.
6. Capon, J., Aug. 1969, High Resolution Frequency-Wavenumber Spectrum Analysis, Proc. IEEE, Vol. 57, no. 8.
7. Pisarenko, V.F., 1973, The Retrieval of Harmonics from a Covariance Function, Geophy. J.R. Astr. Soc., no. 33, P. 374 - 386.



e_1, e_2, e_{MIN} ARE THE EIGENVECTORS OF S CORRESPONDING
 TO EIGENVALUES $\lambda_1 > \lambda_2 > \lambda_{\text{MIN}} > 0$
 e_1, e_2 SPAN THE SIGNAL SUBSPACE
 $a(\theta_1), a(\theta_2)$ ARE THE INCIDENT SIGNAL MODE VECTORS

FIGURE 1. Geometric Portrayal for Three-Antenna Case

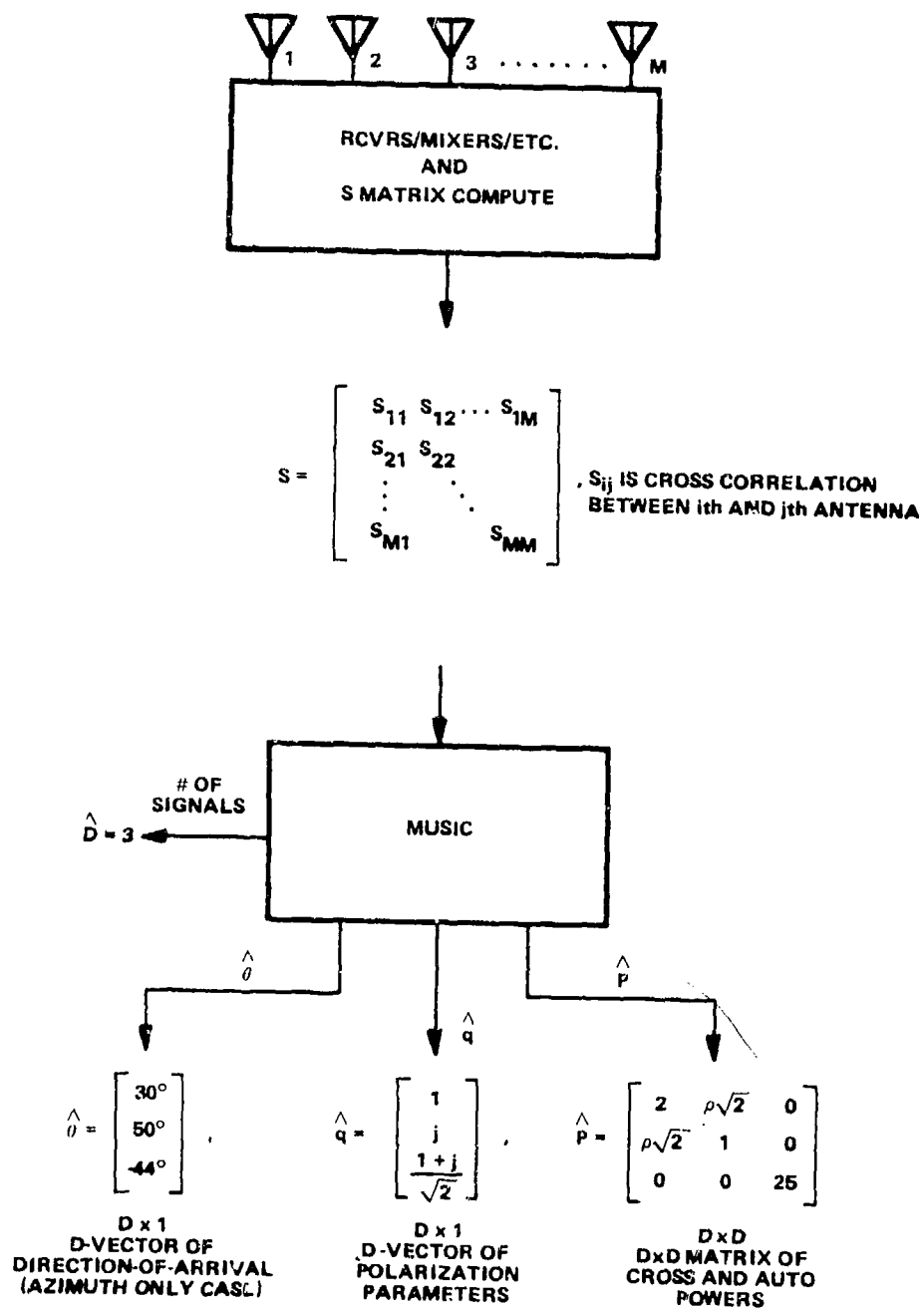


FIGURE 2. Block Diagram for Multiple Signal Classification

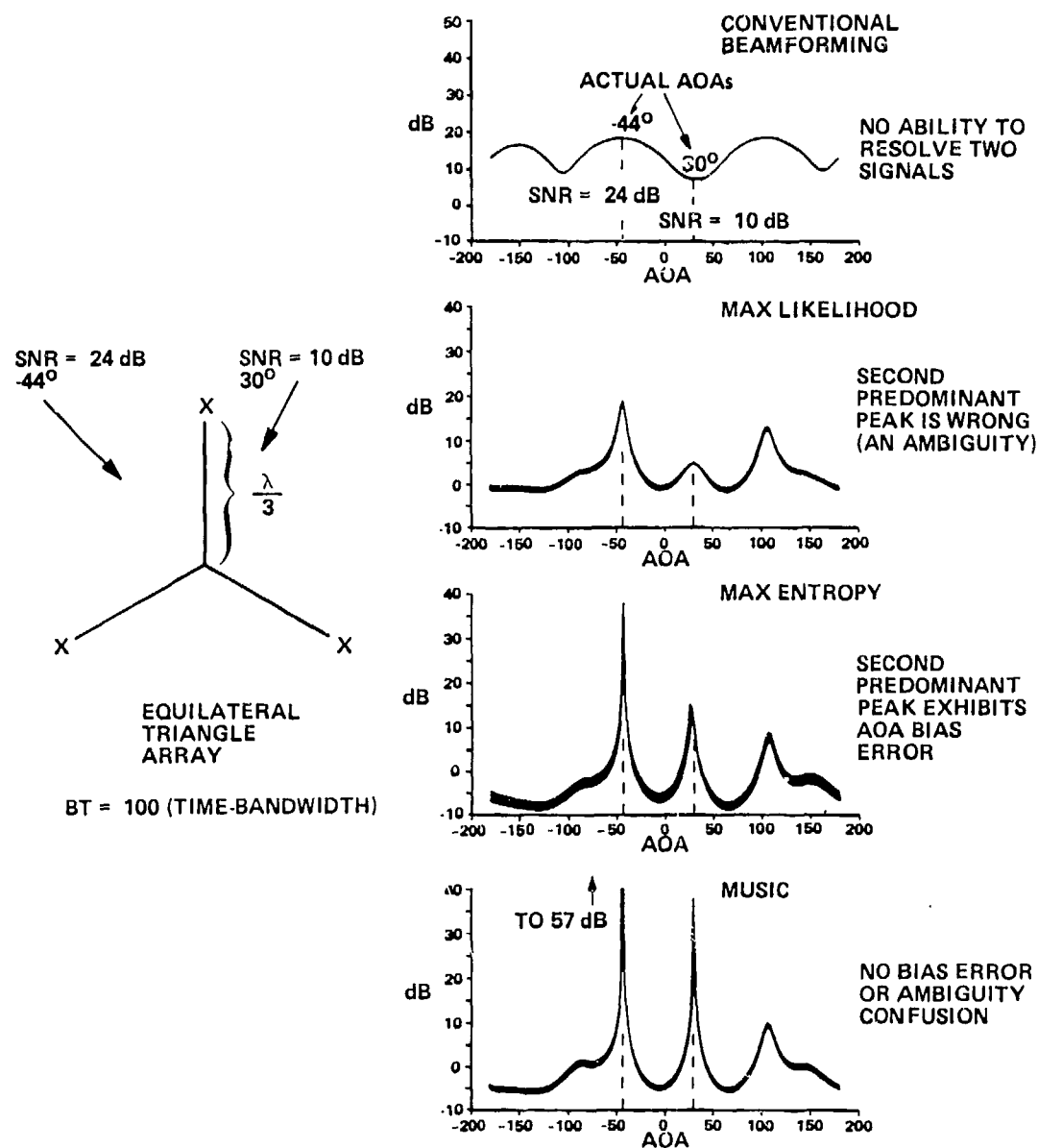


FIGURE 3. Example of Azimuth-Only DF Performance

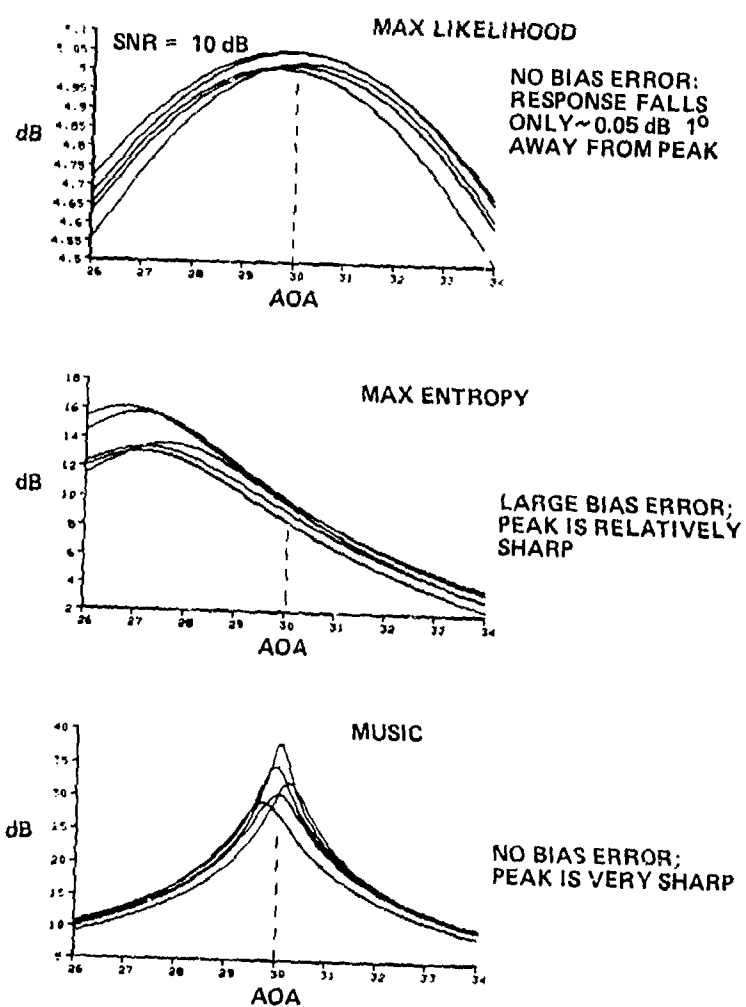


FIGURE 4. Example of Azimuth - Only DF Performance (Scale Expanded About Weaker Signal at 30°)

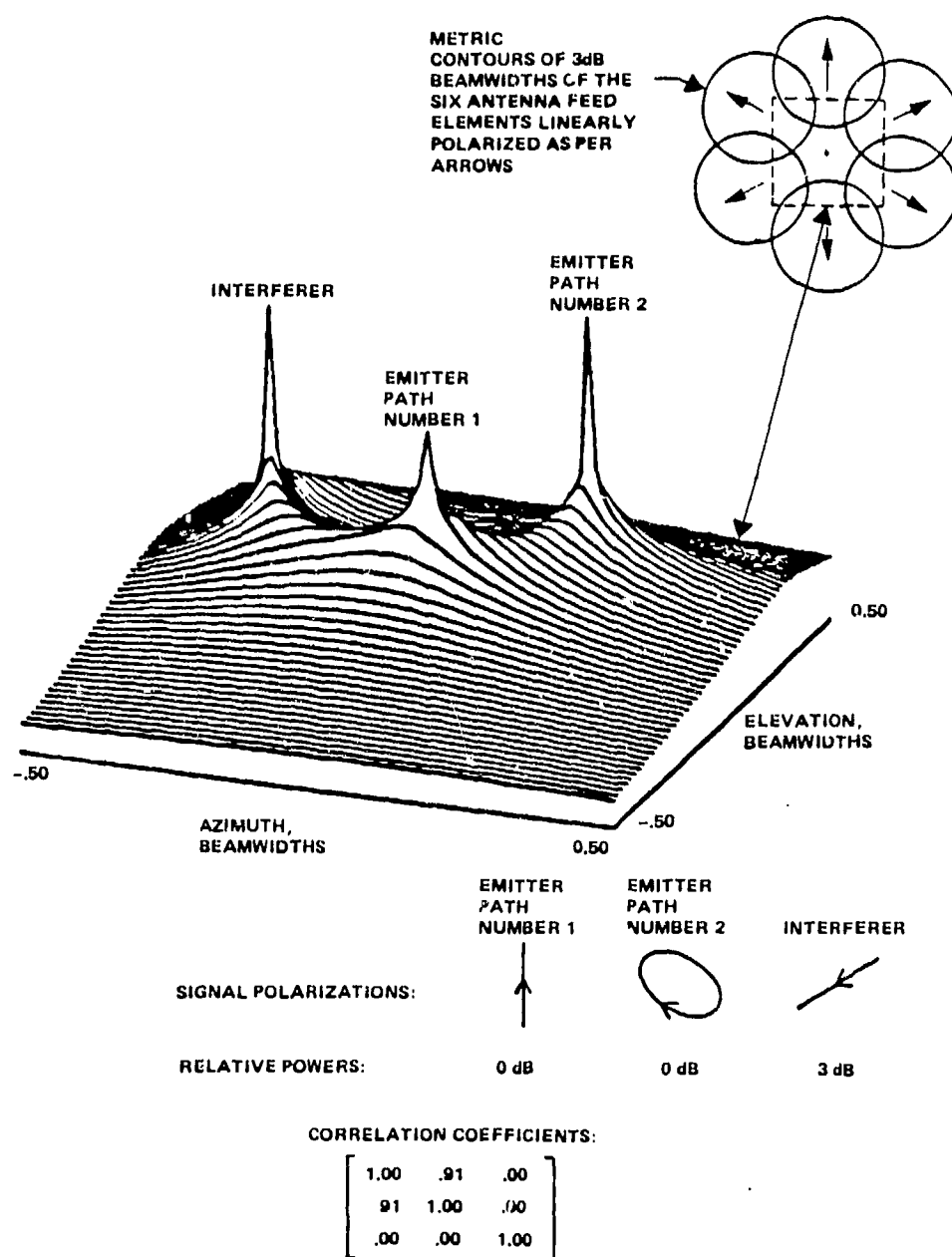


FIGURE 5. MUSIC Applied to a Multiple Feed, Parabolic Dish Antenna System

DATA: 16 COMPLEX TIME SAMPLES OF 6 CISOIDS

FREQ (Hz)	REL AMP (dB)
78.1	-62.33
134.1	-11.10
138.6	0.0
142.9	-11.10
152.9	-40.0
165.3	-26.02

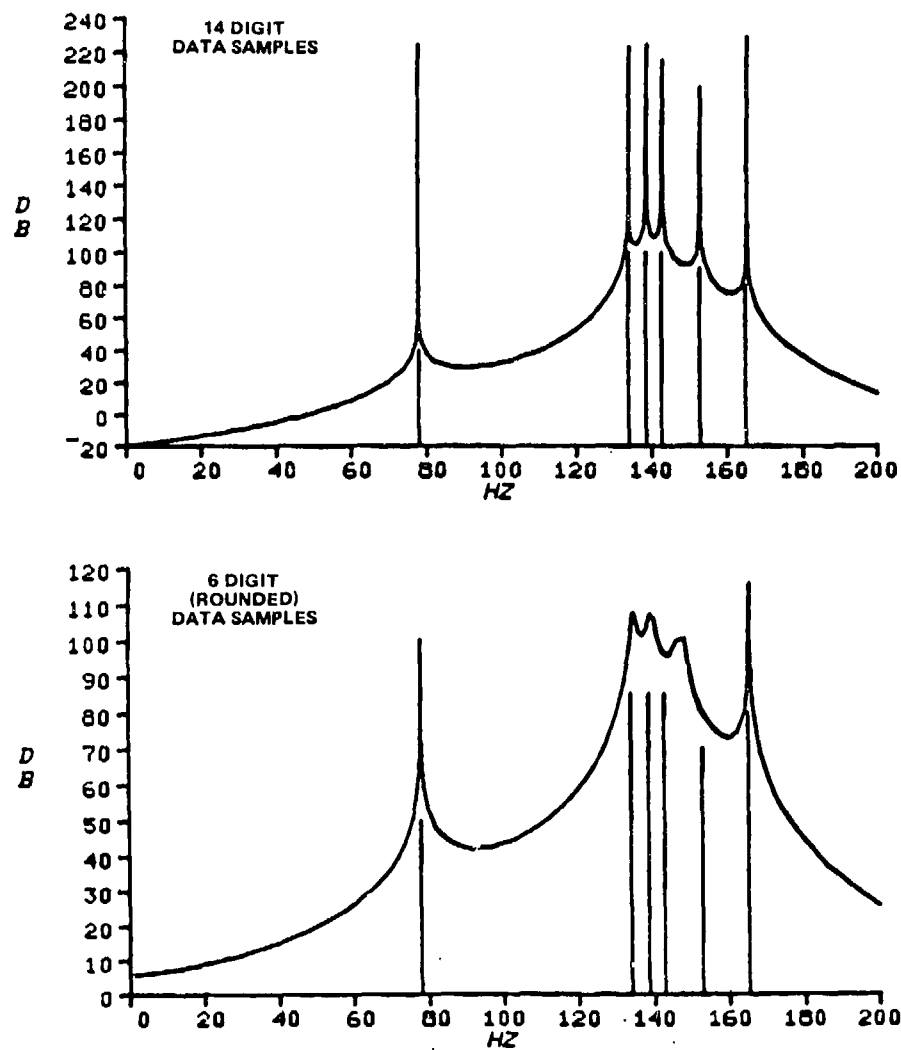


FIGURE 6. Example of MUSIC Used for Frequency Estimation

THE MAXIMUM ENTROPY SPECTRAL ESTIMATOR
USED AS A RADAR DOPPLER PROCESSOR†

SIMON HAYKIN and HING C. CHAN

Communications Research Laboratory
Faculty of Engineering
McMaster University
Hamilton, Ontario, Canada

Abstract

The paper describes the use of the maximum entropy method to estimate the Doppler shift of a moving target in the presence of additive white noise or colored noise consisting of clutter plus white noise. Computer simulation results are included, which show that in a background of additive white noise a Doppler processor based on the maximum entropy method is only slightly sub-optimal with respect to a conventional Doppler processor based on the discrete Fourier transform, whereas in the presence of additive clutter with a narrow spectral width (e.g., ground clutter) it is markedly superior in performance to the conventional processor for the case of low Doppler targets.

†This research was supported by the Department of Communications, Ottawa, Canada.

1. Introduction

In this paper we describe a novel application of the maximum entropy spectral estimator (MESE) as a radar Doppler processor in which the Doppler frequency shift produced by a moving target (e.g., an aircraft) as it moves radially with respect to the radar antenna, is used to detect the presence of the target in a background of stationary clutter and/or receiver noise. Two features make the MESE well-suited for this application: (1) its high frequency resolution capability, and (2) the fact that it can operate with a relatively small number of data samples, which is often the case in a pulsed radar environment.

In Section 2 we briefly review the relevant features of the MESE, and in Section 3 describe the use of this device for Doppler processing. In Section 4 we compare the performance of this processor with that of a Doppler processor which uses the combination of a double-delay line canceler and discrete Fourier transformer. The comparison is made for two different forms of interference at the receiver input: (1) the interference consists of pure white

Gaussian noise, and (2) the interference consists of white noise plus a clutter component.

2. The Maximum-Entropy Spectral Estimator (MESE)

Consider a complex-valued weakly stationary time series $\{x_n\}$, where $n = 1, 2, \dots, N$. The algorithm used to design a maximum-entropy spectral estimator (MESE) for such a time series involves two forms of prediction, namely, forward prediction and backward prediction, with the resulting prediction-errors denoted by $e_{f,n}^{(m)}$ and $e_{b,n}^{(m)}$ respectively, where $m = 0, 1, 2, \dots, M$ refers to the pertinent stage of computation. These two prediction errors may be computed using the equivalent lattice model of Fig. 1, where the number of stages in the lattice model is denoted by M . The set of numbers $\{\rho_m\}$, $m = 1, 2, \dots, M$, are called the reflection coefficients of the estimator. The input time series $\{x_n\}$ and the prediction-error time series $\{e_{f,n}^{(m)}\}$ are orthogonal. Furthermore, the successive stages of the equivalent lattice model are decoupled from each other, that is, the backward prediction-errors in the model are orthogonal to each other. Accordingly, we may state the following:

- (1) The reflection coefficient ρ_m at stage m of the model may be computed independently of the reflection coefficients of those following stage m . In fact, the global minimization of the prediction-error power with respect to ρ_m may be achieved as a sequence of local minimization problems, one at each stage. Specifically, we have

$$\rho_m = \frac{-2 \sum_{n=m+1}^N e_{f,n}^{(m-1)} e_{b,n-1}^{(m-1)*}}{\sum_{n=m+1}^N [|e_{f,n}^{(m-1)}|^2 + |e_{b,n-1}^{(m-1)}|^2]}, \quad m = 1, 2, \dots, M \quad (1)$$

$$\text{where } e_{f,n}^{(0)} = e_{b,n}^{(0)} = x_n \quad (2)$$

- (2) To compute the forward prediction error $e_{f,n}^{(m)}$ at stage m , it is sufficient to update the forward prediction error $e_{f,n}^{(m-1)}$ with a constant (namely, the reflection coefficient ρ_m) times the delayed backward prediction-error $e_{b,n-1}^{(m-1)}$, with both $e_{f,n}^{(m-1)}$ and $e_{b,n-1}^{(m-1)}$ referring to the preceding stage $m-1$. The orthogonality of the data and error sequences ensures that this new forward prediction-error $e_{f,n}^{(m)}$ is smallest in the least-squares sense and that previous coefficients need not be changed. Similar arguments hold for the forward prediction-error $e_{b,n}^{(m)}$ at stage m . We may thus write

$$e_{f,n}^{(m)} = e_{f,n}^{(m-1)} + \rho_m e_{b,n-1}^{(m-1)} \quad (3)$$

$$e_{b,n}^{(i)} = e_{b,n-1}^{(m-1)} + \rho_m^* e_{f,n}^{(m-1)} \quad (4)$$

where $m = 1, 2, \dots, M$.

- (3) The minimum prediction-error power P_m at stage m is defined in terms of the minimum prediction-error power P_{m-1} at stage $m-1$ and the reflection coefficient ρ_m at stage m as follows:

$$P_m = P_{m-1} (1 - |\rho_m|^2), \quad m = 1, 2, \dots, M \quad (5)$$

where

$$P_0 = \frac{1}{N} \sum_{n=1}^N x_n x_n^* \quad (6)$$

A filter having the time series $\{x_n\}$ as input and the prediction-error $\{e_{f,n}\}$ as output is known as the prediction-error filter. The order m of this filter defines the number of stages contained in the lattice model of Fig. 1. The coefficients of the prediction-error filter, denoted by $a_m^{(M)}$, may be computed by using the Levinson recursion:

$$a_m^{(M)} = a_m^{(n-1)} + \rho_M a_{M-m}^{(M-1)*}, \quad m = 0, 1, \dots, M \quad (7)$$

where the asterisk denotes complex conjugation. Note that

$$a_m^{(M)} = \begin{cases} 1, & \text{for } m = 0 \\ 0, & \text{for } m > M \end{cases} \quad (8)$$

Finally, having evaluated the pertinent set of filter coefficients, we may compute the maximum entropy spectral estimate of the given time series by using the formula

$$\hat{S}_x(f) = \frac{P_M}{2B \left| 1 + \sum_{m=1}^M a_m^{(M)} \exp(-jm2\pi f T_s) \right|^2} \quad (9)$$

where B is the bandwidth of the time series, and T_s is the sampling interval. With Nyquist rate sampling, we have $T_s = 1/2B$. Equation (9) clearly emphasizes the nonlinear nature of the maximum entropy spectral estimator. This formula is the basis of the Doppler processor to be described next.

3. A Doppler Processor Using the MESE

A radar Doppler processor utilizes the effect of Doppler shift on the echo reflected from a moving target in that the power spectrum of this echo is centered about a frequency which is shifted from the transmitted carrier frequency by an amount proportional to the radial velocity of the target. In addition to this target echo, the received signal contains a clutter component (produced by reflections from unwanted objects such as ground and weather disturbances) and a receiver noise component. Figure 2 shows the block

diagram of a Doppler processor using the maximum entropy spectral estimator, which is designed to compute the spectral estimate $\hat{S}_x(f)$ of the sampled form of the received signal at a pre-selected set of frequencies, uniformly spaced across the Doppler band of interest. This frequency spacing is determined by the resolution capability of the MESE. In the Doppler processor of Fig. 2, the logarithm of the MESE output, rather than the output itself, is compared with a threshold in order to determine whether a target is present or not. The reason for this operation is explained below.

In order to properly set the detection threshold level at the Doppler processor output, we need to know the statistical behavior of $\hat{S}_x(f)$. Owing to the nonlinear dependence of $\hat{S}_x(f)$ on the receiver input, we find that it is rather difficult to treat this problem analytically.

The results of an extensive computer simulation study [1] have shown that with white Gaussian noise as input, the statistics of the logarithm of the spectral estimate $\hat{S}_x(f)$, that is, the quantity defined by

$$\hat{Z}_x(f) = 10 \log_{10} \hat{S}_x(f) \quad (10)$$

may be closely modeled as Gaussian. This is illustrated in Fig. 3 where we have plotted the probability of false alarm P_{FA} versus the threshold level V_T . We see that the curve calculated by assuming a Gaussian model for $\hat{Z}_x(f)$ fits the experimental curve (obtained by Monte-Carlo simulation) almost perfectly. It is found that for a fixed value of data record length N , the mean value of $\hat{Z}_x(f)$ decreases with the order M of the prediction-error filter. On the other hand, for a fixed value of M , the mean value of $\hat{Z}_x(f)$ increases with N . With regard to the standard deviation of $\hat{Z}_x(f)$ it increases as the filter order M increases, whereas for a fixed value of M , it decreases as the record length N increases.

With clutter as input, the results of computer simulation show that:

- (a) If the spectrum around the frequency at which $\hat{Z}_x(f)$ is computed is symmetrical, the use of a Gaussian model for $\hat{Z}_x(f)$ is justified.
- (b) If the spectrum of the input around the frequency of interest is unsymmetrical, the statistics of $\hat{Z}_x(f)$ deviate markedly from a Gaussian one. The degree of deviation from a Gaussian model increases with the slope of the power spectrum of the clutter input at the frequency of interest. As this slope approaches zero, (corresponding to white noise input), the use of a Gaussian model provides an increasingly better fit for the statistics of $\hat{Z}_x(f)$.
- (c) Increasing the number of data samples N and the filter order M in a corresponding way tend to reduce the deviation from Gaussian statistics for $\hat{Z}_x(f)$.

4. Performance of the Doppler Processor

For the MESE to be useful as a means of measuring the unknown Doppler shift of a moving target, a threshold level must be set at the processor

output so as to realize prescribed values for the probability of false alarm P_{FA} and probability of detection P_D . In the case of thermal noise at the receiver input as the only source of interference, we simply have to maintain the mean and variance estimates for each Doppler cell of interest at prescribed values, since the logarithm of the Doppler processor output, namely, $\hat{Z}_x(f)$ is Gaussian at all frequencies across the Doppler band. In the case of a received signal containing a clutter component, however, the situation is more complex, because the statistics of $\hat{Z}_x(f)$ in the clutter-dominated region tend to deviate from a Gaussian behavior and the degree of deviation is dependent on the slope of the spectrum of the received signal at the frequency of interest. It appears, therefore, that in this case some form of correction in the threshold settings is required if a Gaussian model is assumed for the statistics of $\hat{Z}_x(f)$. This is usually true for parametric detectors operating with unknown input statistics.

In this section, the performance of a Doppler processor using the MESE is investigated and the results compared with that of one of two different configurations, depending on the input conditions:

- (1) A Doppler processor using the discrete-Fourier transformer (DFT) for the case of white Gaussian noise at the processor input; this processor is equivalent to a matched filter for non-fluctuating targets [2], [3].
 - (2) A Doppler processor using the combination of DFT and double-delay line canceler [2], [3] for the case of clutter plus white noise at the input.
- For this study, only small values of record length N are used, so that the results are compatible with surveillance radar requirements. Specifically, for the case of white Gaussian noise at the input, the performance of these processors are evaluated for $N = 8$ and $N = 16$, whereas for the case of white noise plus clutter, the evaluations are made for $N = 10$ and 18 . The two extra samples, in the latter case, are needed in order to initialize the double-delay line canceler (i.e., for transient effects to die out).

4.1 Input Containing White Gaussian Noise

Figure 4 shows plots of the probability of detection P_D versus the input signal-to-noise ratio (SNR) for the case of a non-fluctuating target at the normalized Doppler frequency $f = 0.25$ at a threshold level set to realize a probability of false alarm $P_{FA} = 10^{-6}$; the record length $N = 16$. Curve A applies to the DFT processor, whereas curves B, C, D, and E apply to the MESE processor with filter order $M = 1, 2, 3$, and 4 respectively. These results indicate that for $N = 16$, a MESE processor using a prediction-error filter of order $M = 2$ performs the best. Such a processor requires only 1.3 dB more SNR than the optimal 16-point DFT processor for $P_D = 0.9$ and $P_{FA} = 10^{-6}$.

For the case where $N = 8$, it is found that a MESE processor with $M = 1$ provides the best performance, and for this value of M , it requires 1.7 dB more SNR than the 8-point DFT processor for $P_D = 0.9$ and $P_{FA} = 10^{-6}$.

We conclude therefore that a properly designed Doppler processor using

the MESE performs only slightly sub-optimally compared to one using the DFT in the presence of additive white Gaussian noise at the input.

4.2 Input Containing Noise and Clutter

The evaluations in this case are made by using three sets of data labelled as CD-1, CD-2, and CD-3 whose spectra are shown plotted in Fig. 5. The conditions at the input are described by specifying two parameters, namely, the signal-to-noise ratio (SNR) and clutter-to-noise (CNR). For each target frequency of interest across the Doppler band the probability of detection P_D is computed with the threshold level set to realize a probability of false alarm $P_{FA} = 10^{-6}$. These computations are carried out for each of the data sets CD-1, CD-2, and CD-3. In the presence of clutter, it is found that the order M of the prediction-error filter in the MESE processor has to be equal to or greater than 2 so as to realize an acceptable target frequency selectivity.

For the input data set CD-1, Fig. 6 shows different plots of the required SNR versus the normalized Doppler frequency of the input for $N = 10$, $P_{FA} = 10^{-6}$, and $P_D = 0.9$, assuming a non-fluctuating target. Curve A refers to a Doppler processor using the combination of a double-delay line canceler and 8-point DFT. Curves B and C refer to a Doppler processor using the MESE with $M = 2$ and 3, respectively. We see that in this case, the use of the MESE outperforms the double delay-line canceler-DFT combination by a fairly large margin. For example, for the case when $P_D = 0.9$ AND $P_{FA} = 10^{-6}$, we find that, in the Doppler frequency range 0.03 to 0.12, a MESE processor with filter order $m = 3$ requires 3 to 10 db less SNR than the combination of a double delay-line canceler and DFT. However, for the case of data sets CD-2 and CD-3, the improvement resulting from the use of MESE is not as large, as may be seen by examining Figures 7 and 8 respectively.

Based on these results, we may make the following observations:

- 1) Compared to a Doppler processor using the combination of a double delay-line canceler and DFT, a processor using the MESE provides a substantial improvement in the detection of a slowly moving target in the presence of clutter with a very narrow spectral width (e.g., ground clutter). Therefore, by using the MESE, the frequency band of target visibility is extended by a sizeable margin.
- 2) In the case of clutter with a narrow spectral width, the prediction-error filter of the MESE processor should be as high as possible. In particular, the filter order M should be chosen so as to minimize the effect of clutter on neighboring Doppler components.

5. Conclusions

It has been shown, by means of computer simulation, that in the case of white noise as input, the statistics of the logarithm of the maximum entropy spectral estimate may be closely modeled as Gaussian. However, in the

presence of a clutter component, the statistics of the logarithm of this estimate deviate from a Gaussian model, with the deviation becoming more pronounced as the slope of the input clutter spectrum is increased.

A processor based on the maximum entropy estimator has been described for the measurement of the Doppler shift of a moving target. It has been shown that in the presence of additive white noise at the input, this processor is only slightly sub-optimal compared to a Doppler processor based on the discrete Fourier transform, which is optimum for the case of a non-fluctuating target. In the presence of a clutter component with narrow spectral width (e.g., ground clutter), however, we find that a Doppler processor using the MESE is markedly superior to the combination of a double delay-line canceler and discrete Fourier transformer, for low Doppler targets.

It should be emphasized that, although MESE-based processor has several useful features for the processing of radar signals, it does not completely eliminate the need for other conventional signal processing methods (e.g., the discrete Fourier transform). Rather, a MESE-based processor may be used as a way of extending the performance capability of conventional radar processors. Also, it should be emphasized that there is need for further investigations (both theoretical and experimental) concerning the statistical behavior of the MESE output and the full exploitation of the MESE for radar applications.

6. References

- (1) H.C. Chan and S. Haykin, March 1979, "Applications of the Maximum Entropy Method in Radar Signal Processing", Report CRL-62, Communications Research Laboratory, McMaster University.
- (2) R.J. McCaulay, February 22, 1972, "A Theory for Optimal MTI Digital Signal Processing, Part I, Receiver Synthesis", MIT Lincoln Laboratory, Technical Note 1972-14.
- (3) C.E. Muehe et al., June 1974, "New Techniques Applied to Air Traffic Control Radars", Proc. IEEE, Vol. 62, pp. 716 - 723.

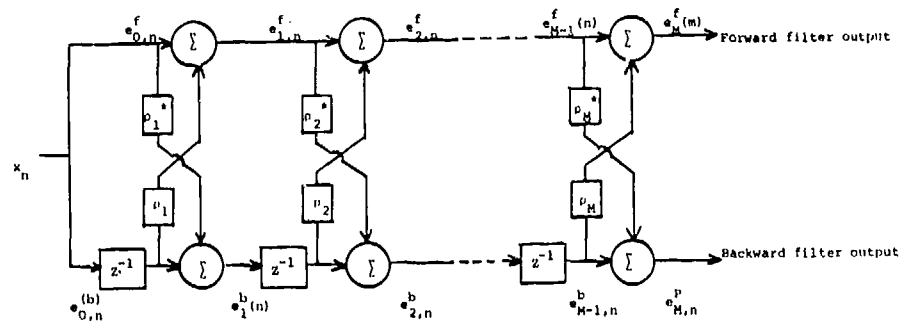


FIGURE 1. Equivalent lattice structure for the prediction-error filter.

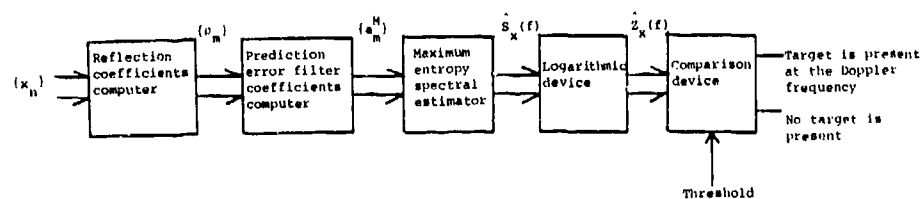


FIGURE 2. Block diagram of a Doppler processor using the MESE.

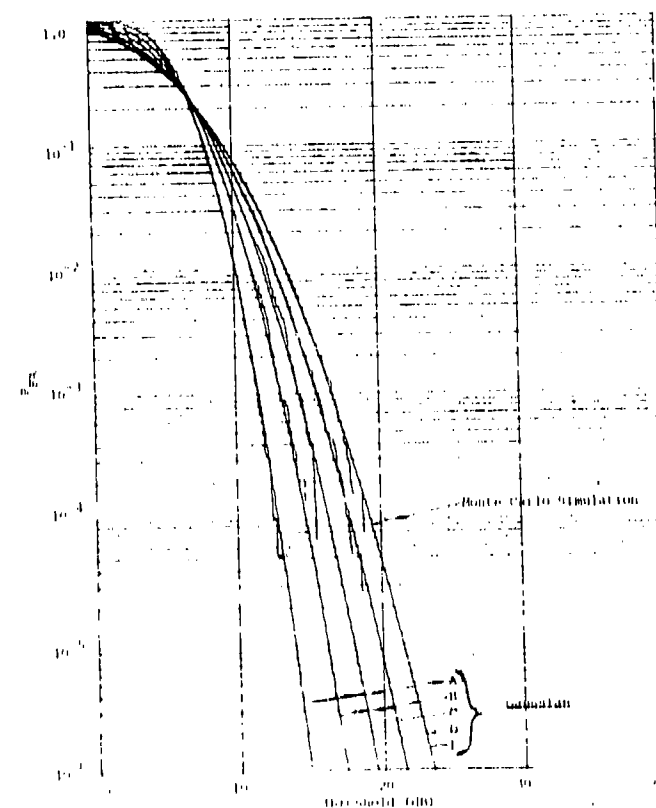


FIGURE 3. Curves of false alarm probability P_{FA} versus threshold for

MEEF processors of various order:

Curve A : $M = 1$

Curve B : $M = 2$

Curve C : $M = 3$

Curve D : $M = 4$

Curve E : $M = 5$

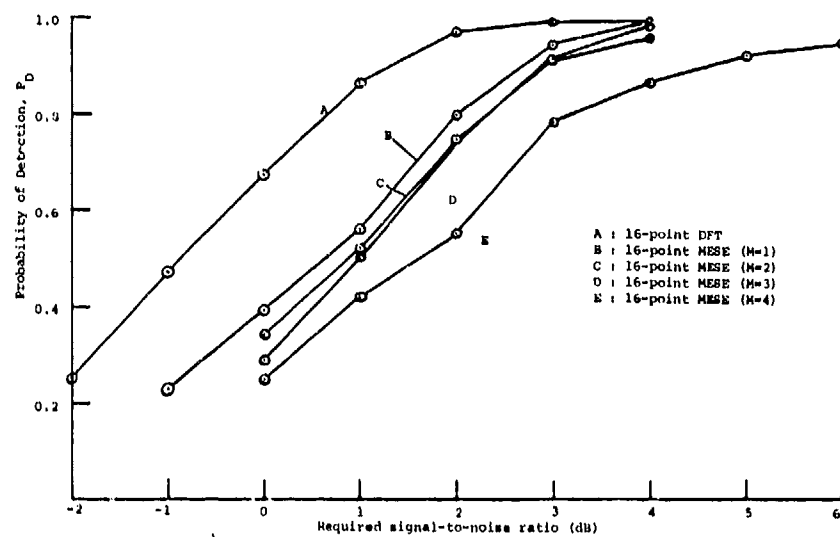


FIGURE 4. Curves of detection probability P_D versus signal-to-noise ratio for the DFT and MESE processors.

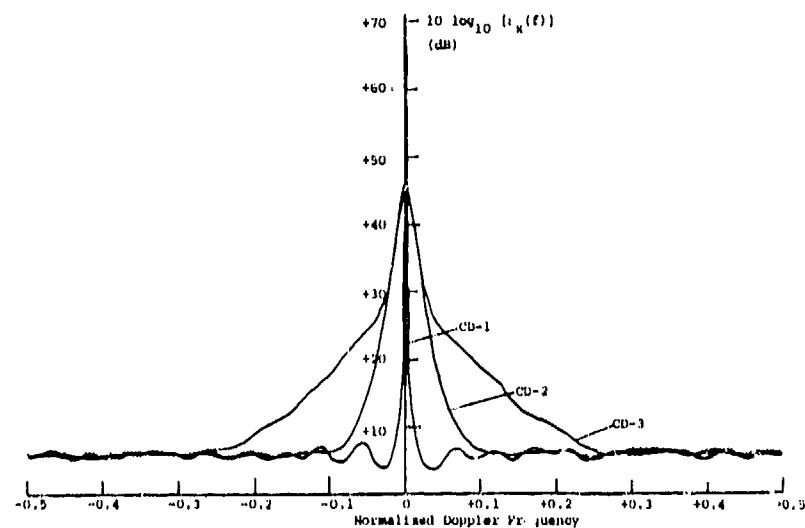


FIGURE 5. Spectral densities of three different data sets.

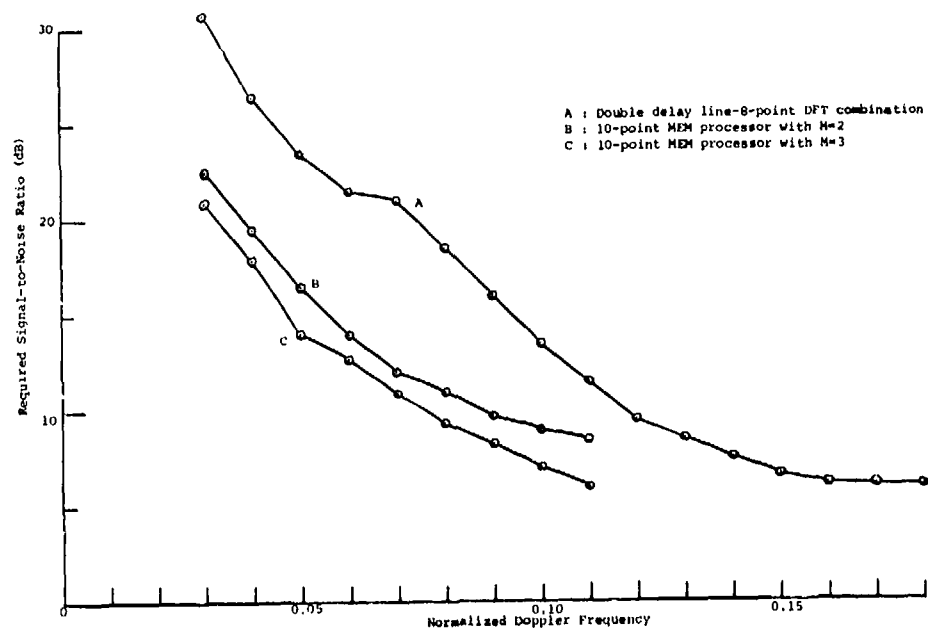


FIGURE 6. Curves of signal-to-noise ratio versus Doppler frequency for the conventional and MESE processors, based on data set CD1. Clutter-to-noise ratio = 22 db.

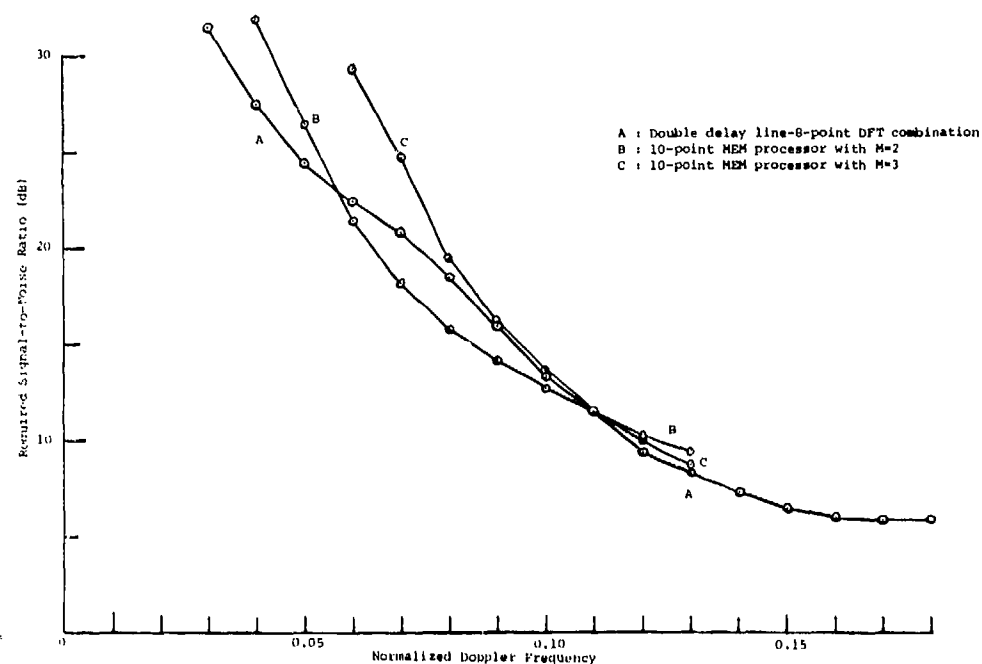


FIGURE 7. Curves of signal-to-noise ratio versus Doppler frequency for the conventional and MESE processors, based on data set CD2. Clutter-to-noise ratio = 22db.

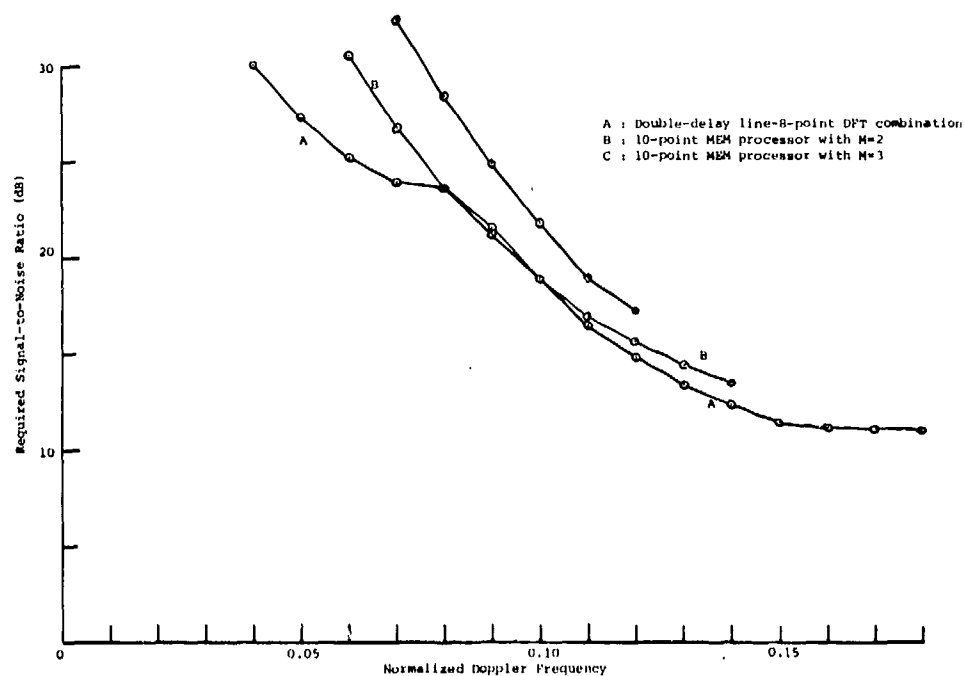


FIGURE 8. Curves of signal-to-noise ratio versus Doppler frequency for the conventional and MESE processors, based on data set CD3. Clutter-to-noise ratio = 22db.

272 - BLANK.

APPLICATIONS FOR MESA AND THE PREDICTION ERROR FILTER

WILLIAM R. KING

King Research
10209 Westford Dr.
Vienna, Virginia 22180

Abstract

In recognition of the whitening and resolution characteristics of MESA and the prediction error filter, it is demonstrated that these techniques may have several signal processing applications. Examples are provided to illustrate stable, high resolution power spectra, high resolution autocorrelation functions, and signal detection in strong interfering clutter.

Introduction

In 1967 Burg [1] utilized the prediction error filter and the maximum entropy method for estimating power spectra in the frequency domain. Since that time (12 years ago), the maximum entropy method has had little application for purposes other than estimation of power spectra in the frequency and wavenumber domains. Maximum entropy spectral analysis (MESA), which is a whitening filter, is recognized chiefly for its high resolution capability when applied to short data sets. However, the whitening characteristic of MESA may prove to be effective for reducing clutter in signal detection applications.

In a power spectral analysis MESA is often used in lieu of the Fourier transform. There are, of course, many applications for the Fourier transform, both narrowband and broadband, which may be considered as possible applications for MESA or the prediction error filter. In a demonstration of the versatility of MESA and the prediction error filter, these techniques are used for estimating power spectra, the autocorrelation function, and for detecting signals in typical radar interference clutter.

The MESA Technique

The MESA algorithm incorporates the Burg technique in all applications demonstrated in this paper. The MESA-Burg equations for complex data are given in slide 1.

The MESA power spectra is expressed by eqn. (1) for the space-wavenumber domain, where k is the wavenumber, N is the number of filter weights, Δx is the uniform antenna element spacing and γ_n is the n th filter weight. The total noise power P_N is given by eqn. (2), and the filter weights are evaluated with use of the iterative eqn. (4). However, the last filter weight is defined by eqn. (3). The terms of eqn. (3) contain the forward and backward prediction errors denoted by α and β that are defined by eqns. (5) and (6) as functions of the data samples.

Because MESA "snapshot" spectral patterns are inherently unstable and sometimes contain split spectral peaks, all MESA examples are computed by incorporating averaging techniques into the MESA algorithm. King [2] employed several averaging methods, and concluded that two such methods, averaged covariance matrix elements and averaged filter weights, are both useful methods for applying MESA to successive sets of data samples.

Antenna Patterns

Stable wavenumber power spectra (antenna patterns) are shown computed for an 8 element antenna using both averaging methods. In the first example a signal incident at 5 degrees from broad-side with a SNR of 10 dB. (per antenna element) is shown detected in slide 2 using averaged filter weights. And in slide 3 the same signal is detected by MESA using averaged covariance matrix elements. Side peak levels are substantially reduced and whitened with use of both averaging methods.

Two signals, separated 5 and 6 degrees, with a SNR of 10 dB each signal (each antenna element) are resolved using both averaging methods for the maximum number of filter weights permitted ($N=7$). The two signals, which are separated 5 degrees, are observed resolved in slide 4 using filter weights averaged 10 times. And in slide 5, two signals which are separated 6 degrees are shown resolved with use of covariance matrix averaged 10 times. Both averaging methods effectively stabilize even the highest order MESA patterns, while permitting excellent resolution and clutter suppression.

$$P(k) = \frac{P_N K}{\left| 1 + \sum_{n=1}^N \gamma_n^N e^{ikn(\Delta x)} \right|^2} \quad (1)$$

$$K = 2\pi/\lambda, \quad k = K \sin(\theta), \quad \Delta x = \lambda/2$$

WAVENUMBER POWER SPECTRA

$$P_1 = r_0^2$$

$$P_{N+1} = P_N \left[1 + (\gamma_{N+1}^{N+1})^2 \right] \quad (2)$$

ERROR POWER $\frac{P_N}{P_N}$

$$\gamma_1^N = 1$$

$$\gamma_{N+1}^{N+1} = \frac{\sum_{j=1}^{M-N+1} (\beta_j^N)^* \alpha_{j+N}^N}{\sum_{j=1}^{M-N-1} \left[(\beta_j^N)^2 + (\alpha_{j+N}^N)^2 \right]} \quad (3)$$

$$\gamma_n^N = \gamma_n^N + \gamma_{N+1}^{N+1} (\gamma_{N-n+2}^N)^* \quad (4)$$

PREDICTION ERROR COEFFICIENTS

$$\alpha_{j+1}^1 = x_{j+1}$$

$$\alpha_j^{N+1} = \gamma_{N+1}^{N+1} \beta_{j-N}^N + \alpha_j^N \quad (5)$$

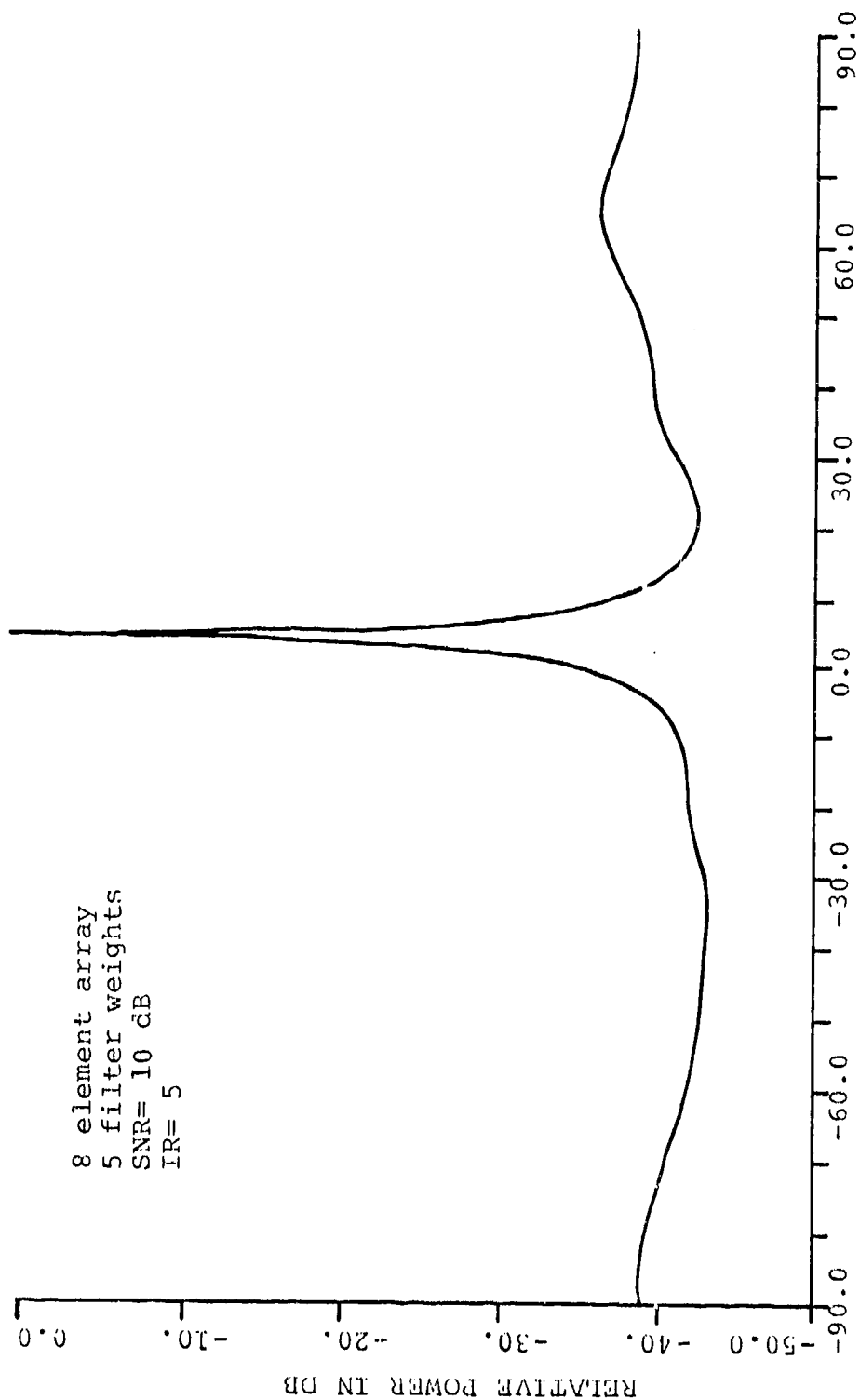
$$\beta_j^1 = x_j$$

$$\beta_j^{N+1} = (\gamma_{N+1}^{N+1})^* \alpha_{j+N}^N + \beta_j^N \quad (6)$$

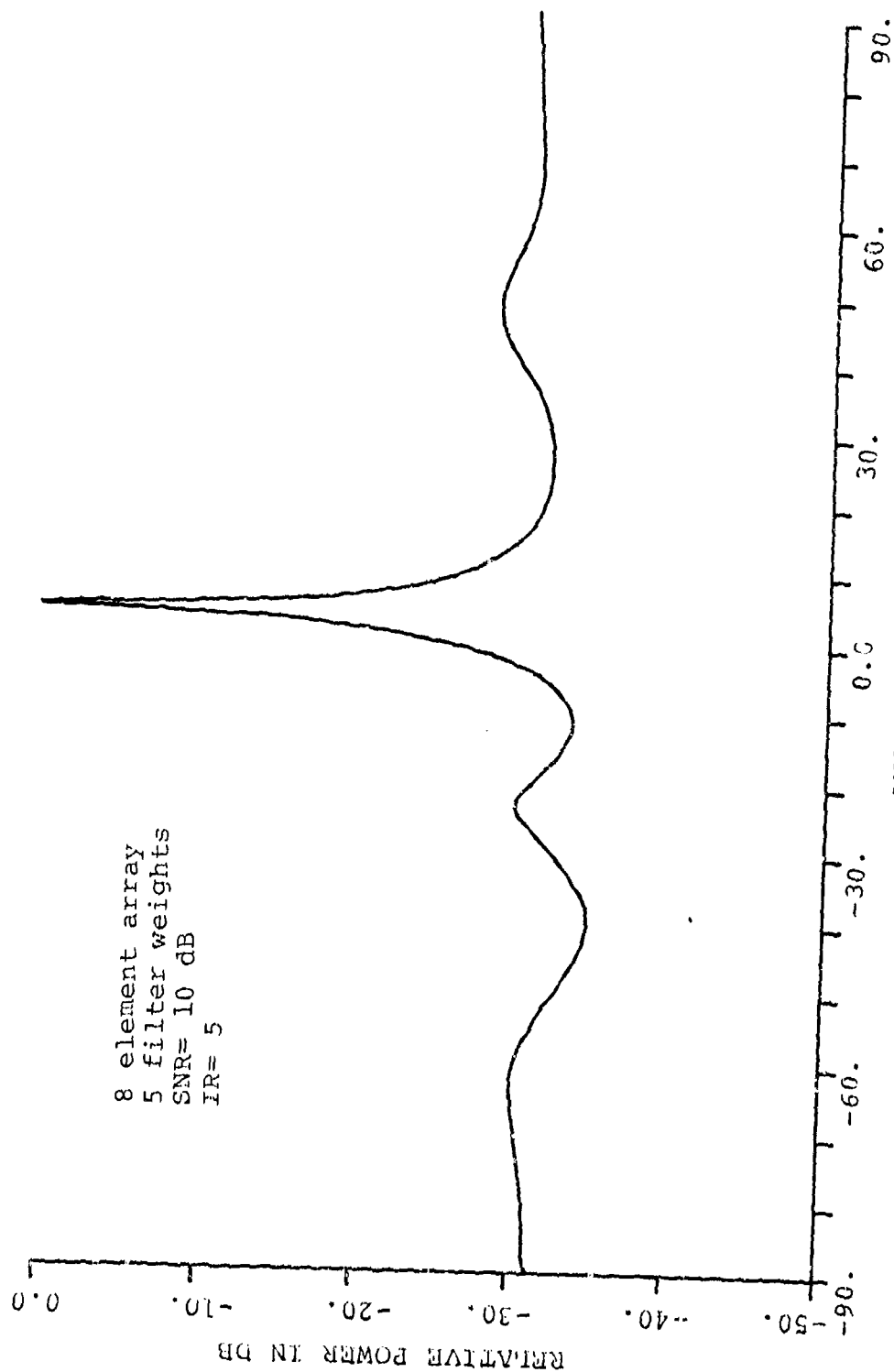
FORWARD AND BACKWARD PREDICTION ERRORS

SLIDE 1

MESA ANTENNA PATTERN



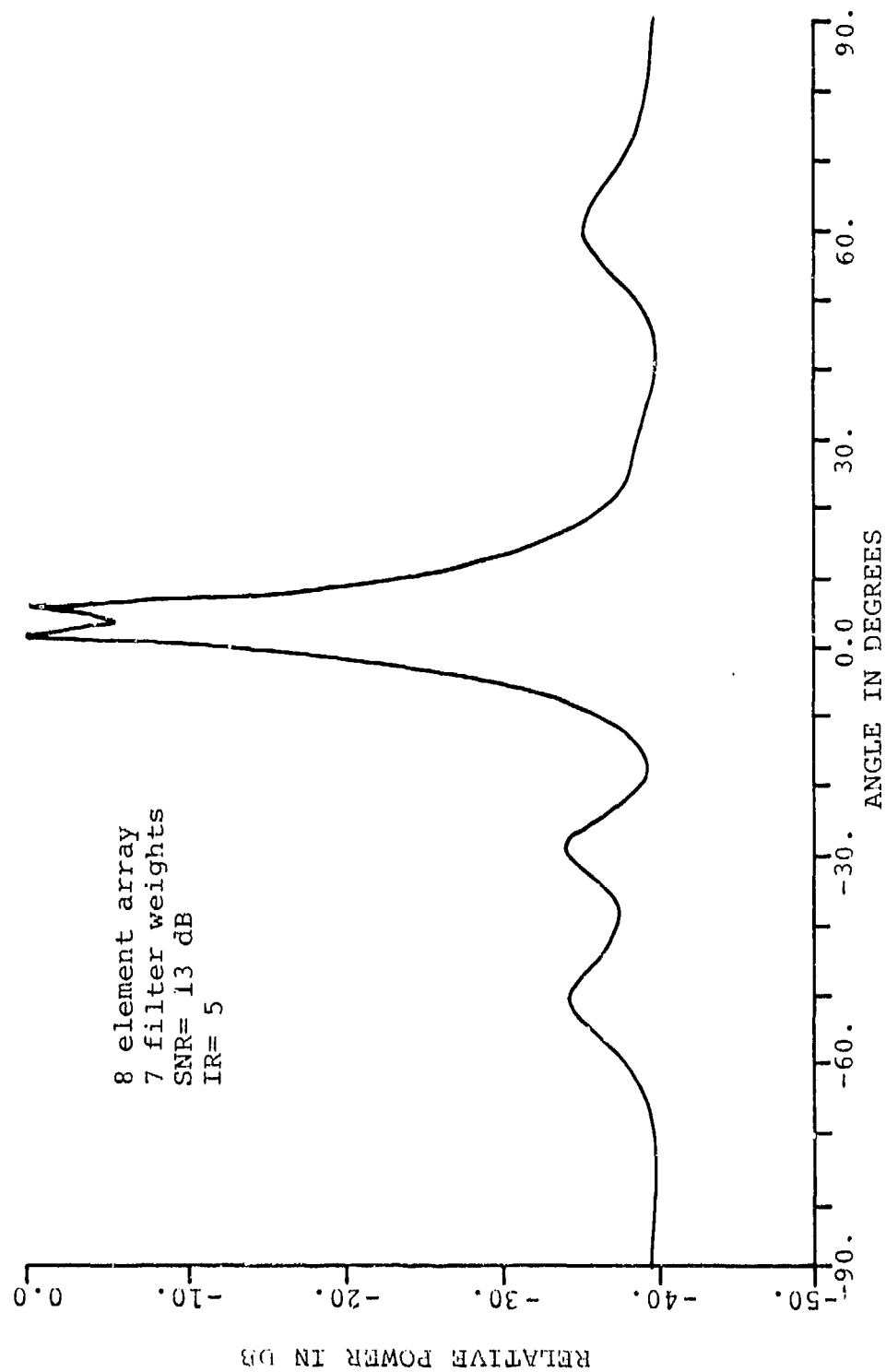
MESA ANTENNA PATTERN



ONE SIGNAL, AVERAGED COVARIANCE MATRIX, (L= 10)

SLIDE 3

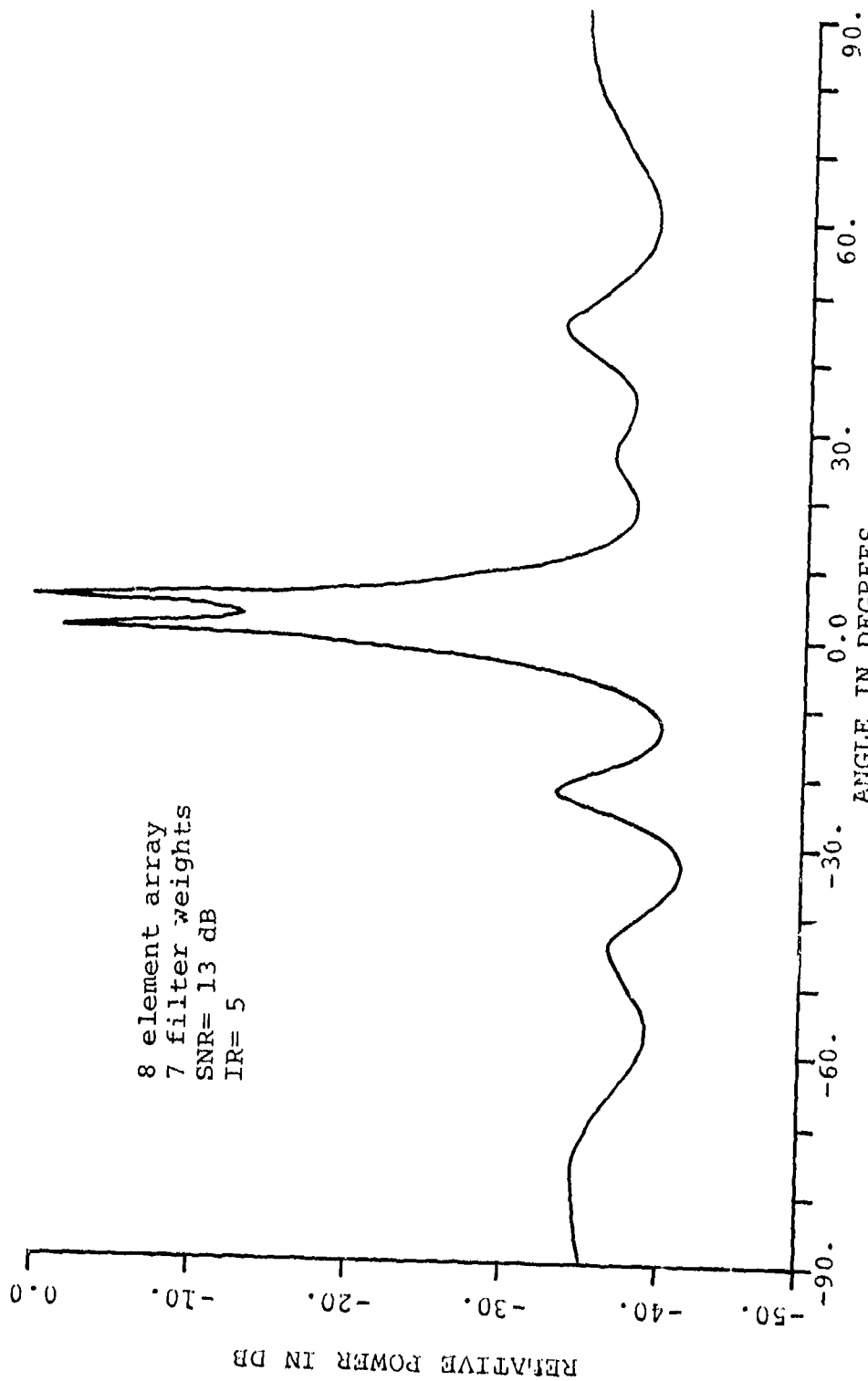
MESA ANTENNA PATTERN



TWO SIGNALS, AVERAGED FILTER WEIGHTS, (L= 20)

SLIDE 4

MESA ANTENNA PATTERN



TWO SIGNALS, AVERAGED COVARIANCE MATRIX, (L= 10)

SLIDE 5

Autocorrelation

The autocorrelation function $r(t)$, as noted in slide 6, may be defined by the Fourier transform of the power spectrum $P(\omega)$. And the power spectrum $P(\omega)$ may be derived from a time dependent function $x(t)$ using either the Fourier transform or MESA as indicated in the second equation of slide 6.

However, the autocorrelation function may also be evaluated using the prediction error transform by following the derivation outlined in slide 6. The prediction error is expressed as a linear function of the appropriately sampled power spectra density $P(\omega)$. The prediction error e_n is transformed to the time domain using the Fourier transform as indicated in slide 6. By transforming the prediction error to the time domain, the autocorrelation function $r(t)$ becomes inversely proportional to the prediction error transform as noted. For demonstration purposes the autocorrelation function is evaluated with both the FFT and the prediction error transform. The power spectral density $P(\omega)$, which is required by both methods, is evaluated only with the FFT.

As an example, a multipath type signal is displayed in slide 7. The signal consists of two gated sinusoids, one delayed by one pulse width and shifted in phase by 180 degrees. The transfer function, which is shown in the middle of slide 7, contains the anticipated modulation. In all computations the signal is treated as a complex function, although only the real part is displayed in slide 7. The autocorrelation function, shown in the lower section of slide 7, is evaluated with use of the FFT.

In slide 8 the autocorrelation function is computed for reduced bandwidths in order to illustrate the advantage of using the prediction error transform. The bandwidth is reduced by clipping the transfer function at both ends of the spectrum as desired. The upper plot of slide 8 depicts the conventional autocorrelation function (using the FFT) computed at only 10% of the original bandwidth. Peaks of the autocorrelation function are not resolved. However, in the middle plot the autocorrelation function computed using the prediction error transform (and only 10% bandwidth), contains the usual number of anticipated peaks. Even with the bandwidth reduced to only 5% of the total bandwidth, all 3 peaks of the autocorrelation function are present with use of the prediction error transform. Distortion effects have become noticeable at the 5% bandwidth level indicating the limitations of the prediction error transform.

$$r(t) = \int_{-\infty}^{\infty} P(\omega) e^{i\omega t} d\omega$$

AUTOCORRELATION FUNCTION
AS FOURIER TRANSFORM

$$P(\omega) = \frac{P_N(\Delta t)}{\left| \sum_{n=0}^N \gamma_n^N e^{i\omega n(\Delta t)} \right|^2}$$

MESA USING PREDICTION
ERROR TRANSFORM

$$r(t) \longleftrightarrow P(\omega)$$

TRANSFORM PAIRS

$$\hat{P}_\omega = \sum_{n=1}^N a_n P_{\omega-n}$$

PREDICTED POWER

$$e_\omega = P_\omega - \hat{P}_\omega$$

PREDICTION ERROR

$$e_\omega = \sum_{n=0}^N \gamma_n^N P_{\omega-n}$$

PREDICTION ERROR

$$E_N = r(t) \sum_{n=0}^N \gamma_n^N e^{-itn(\Delta\omega)}$$

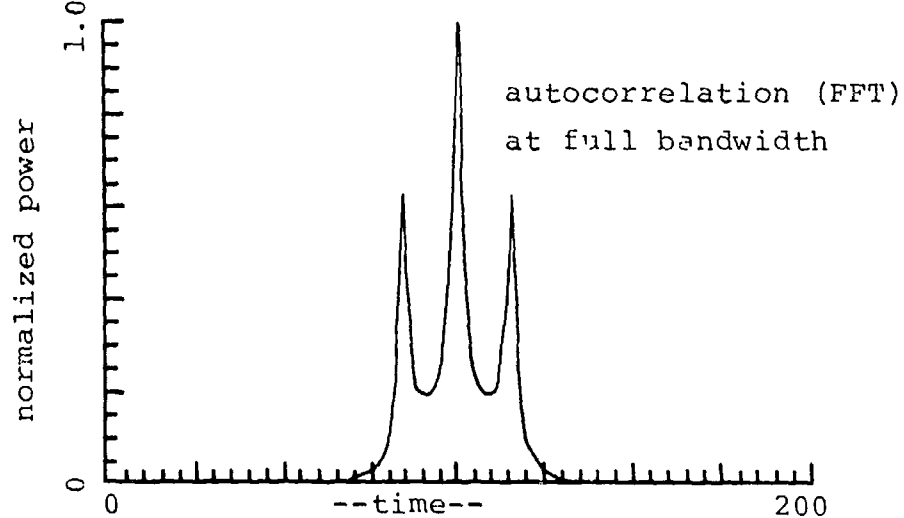
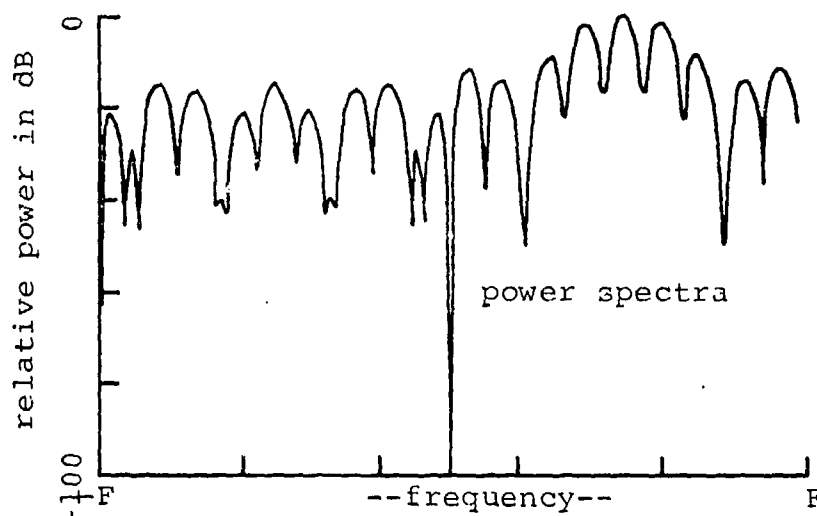
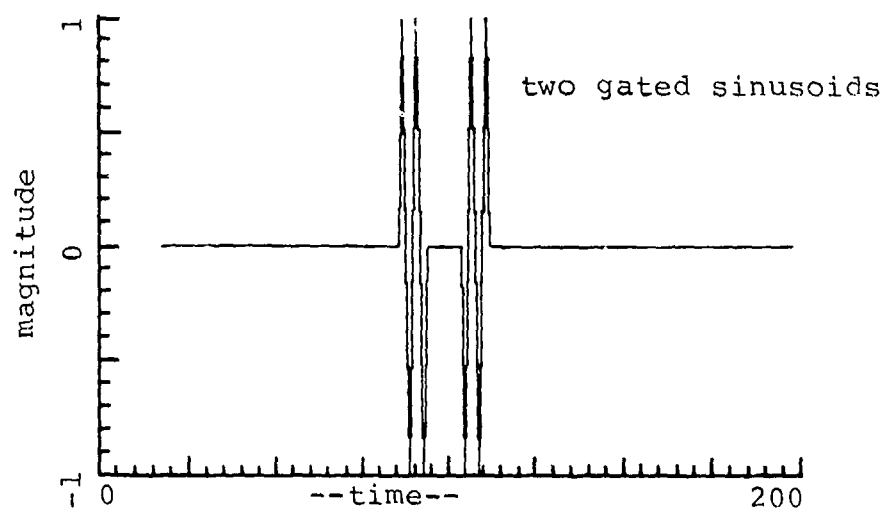
FOURIER TRANSFORM OF
PREDICTION ERROR

$$r(t) = \frac{E_N}{\sum_{n=0}^N \gamma_n^N e^{-itn(\Delta\omega)}}$$

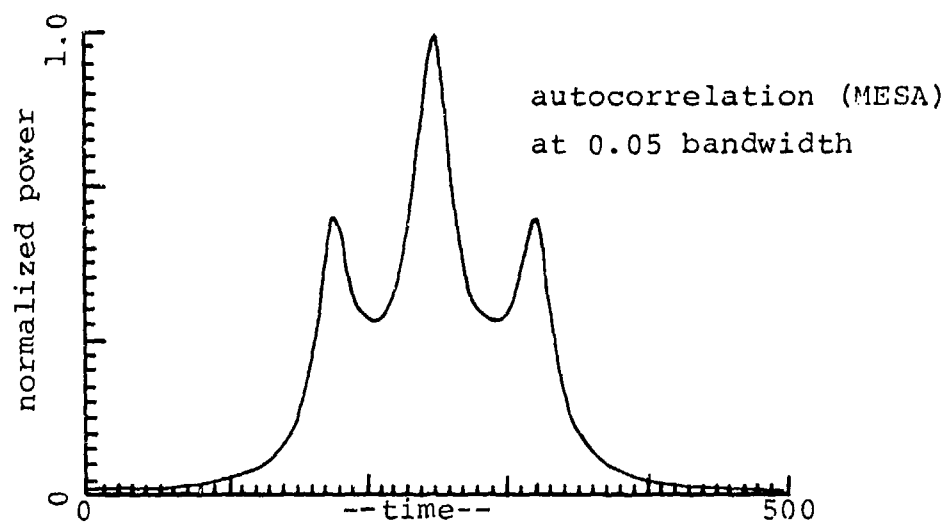
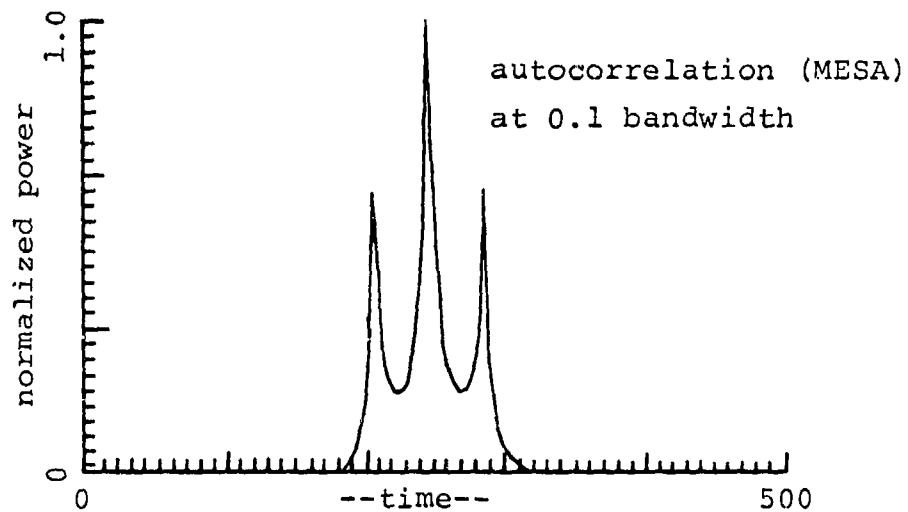
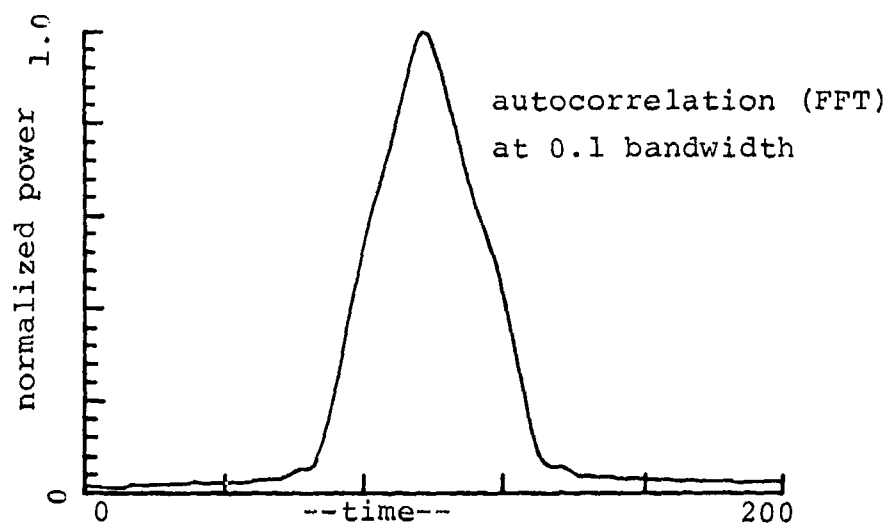
AUTOCORRELATION FUNCTION
AS PREDICTION ERROR
TRANSFORM

DERIVATION OF AUTOCORRELATION FUNCTION USING THE
PREDICTION ERROR TRANSFORM

SLIDE 6



SLIDE 7



SLIDE 8

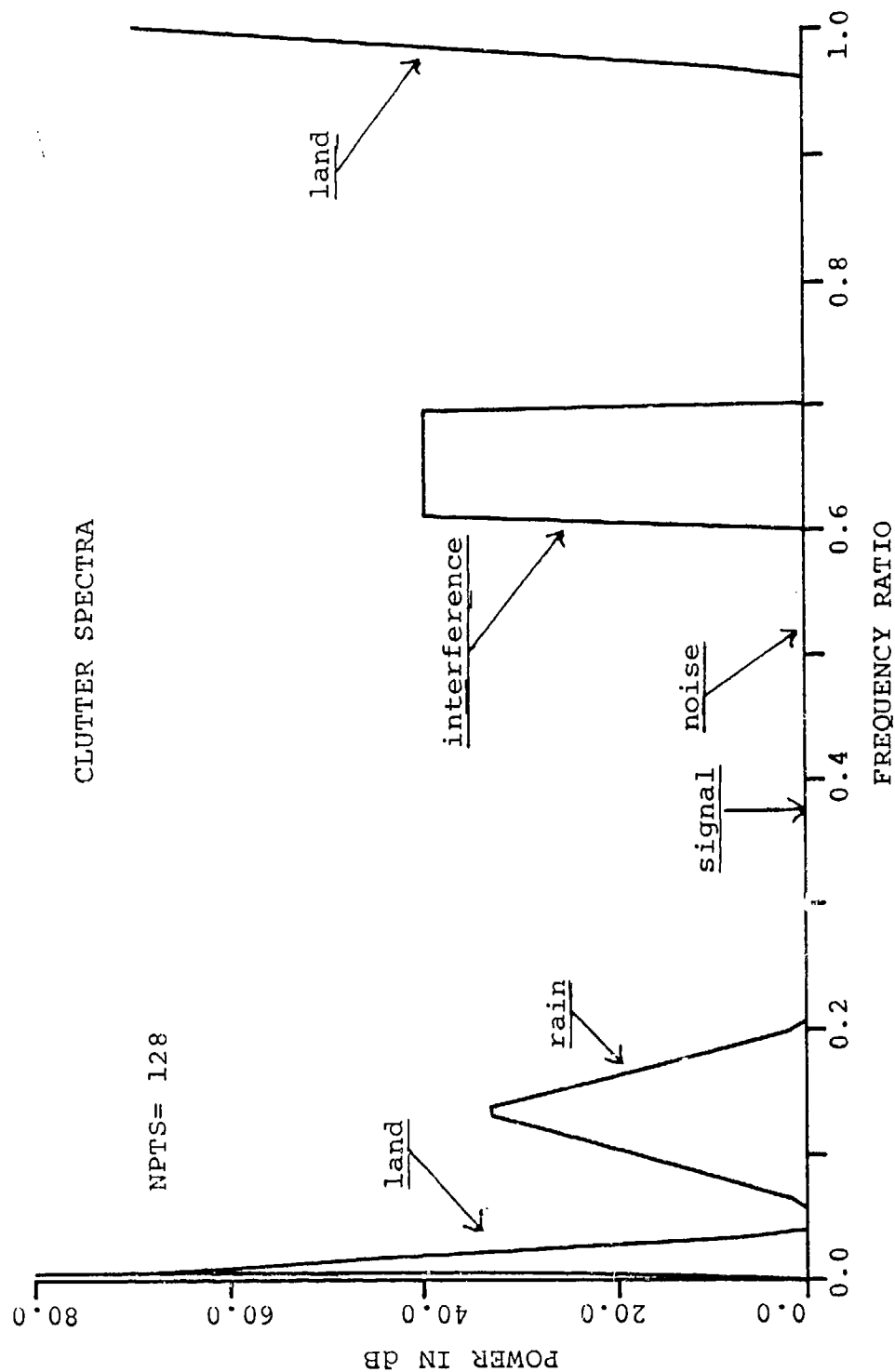
Signal Detection

The possibility of detecting signals by increasing the SNR with MESA was observed previously in the MESA antenna patterns shown in slides 2 and 3. In yet another demonstration of the whitening capability of MESA, radar clutter is simulated as shown in slide 9 where random phased clutter bands have power levels typical of ground, rain, and interference clutter. Such a clutter model has been used previously by Sawyers [3] in his demonstration of adaptive filtering. A signal having a 0 dB SNR is located between the clutter at the frequency ratio of .375 as denoted by the arrow in slide 9. The signal is detected as shown in slide 10, by applying MESA to several sets of 32 data samples and using 24 filter weights. The strong clutter bands are very effectively whitened by MESA such that the largest background peak in slide 10 is about 10 dB below the signal peak level. Similar results may be obtained for any signal location. For example in slide 11 a signal, located at the center of the interference clutter (.65), is equally well detected again with MESA applied to consecutive sets of 32 data samples using 26 filter weights. In both slides 10 and 11 the MESA filter weights are averaged over 30 consecutive sets of 32 data samples. While considerable averaging is used to achieve the results indicated in slides 10 and 11, less averaging of fewer filter weights may also achieve satisfactory signal detection, but with less resolution capability.

It is difficult to imagine that results comparable to those shown in slides 10 and 11 could be achieved with any conventional Fourier signal detection method.

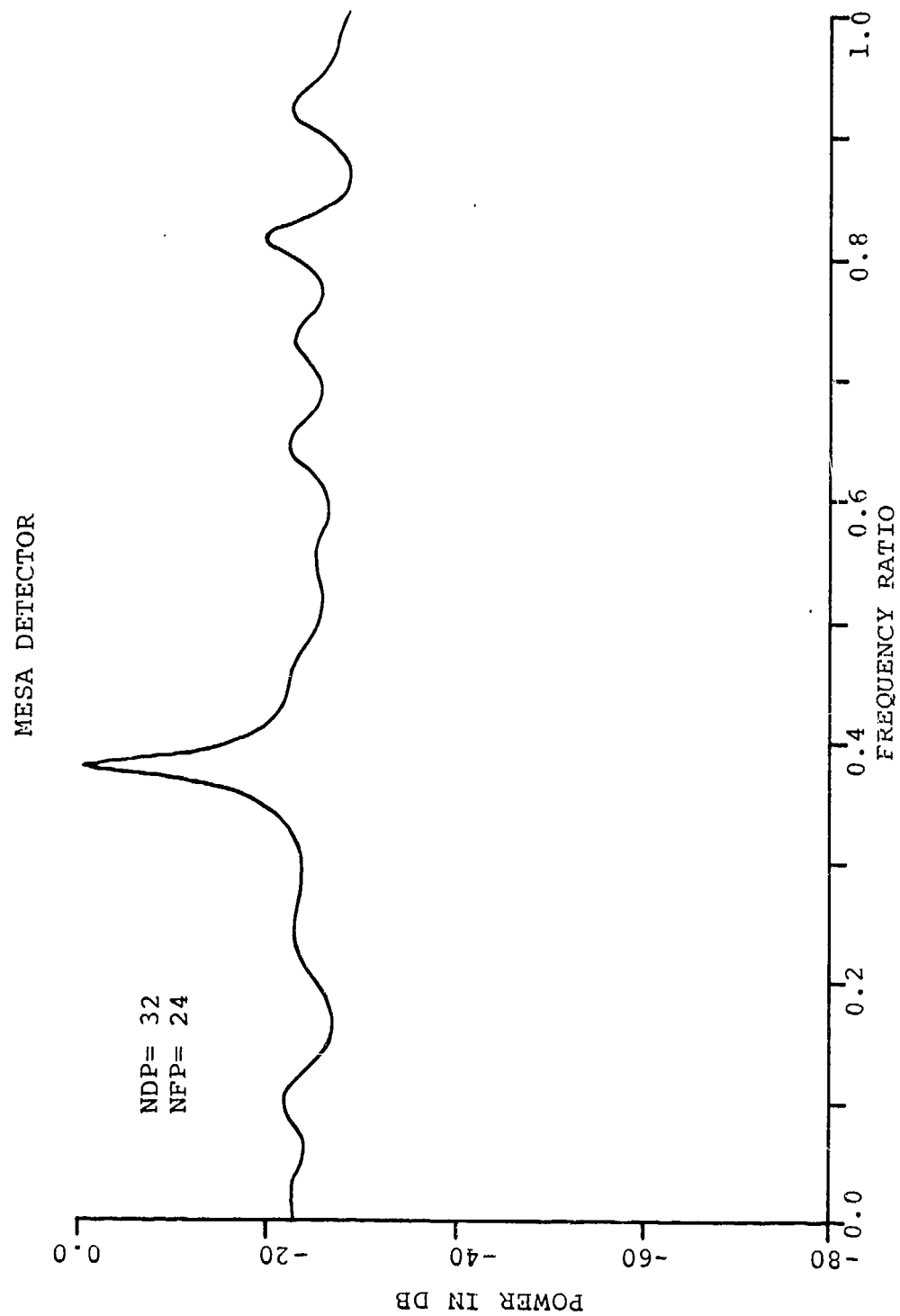
References

1. Burg, John P., "Maximum Entropy Spectral Analysis", presented at the 37th meeting of Society Explor. Geophys., Oklahoma City, Oklahoma (Oct. 1967)
2. King, W. R., "Stable MESA Antenna Patterns", NRL Memo Rept. 3998 (May 18, 1979)
3. Sawyers, J. H., "Applying the Maximum Entropy Method to Adaptive Filtering", Twelfth Asilomar Conference on Circuits, Systems, and Computers, page 198, (Nov. 6-8, 1978) IEEE cat. no. 78CHI369-8 C/CAS/CS



CLUTTER POWER SPECTRAL DENSITY

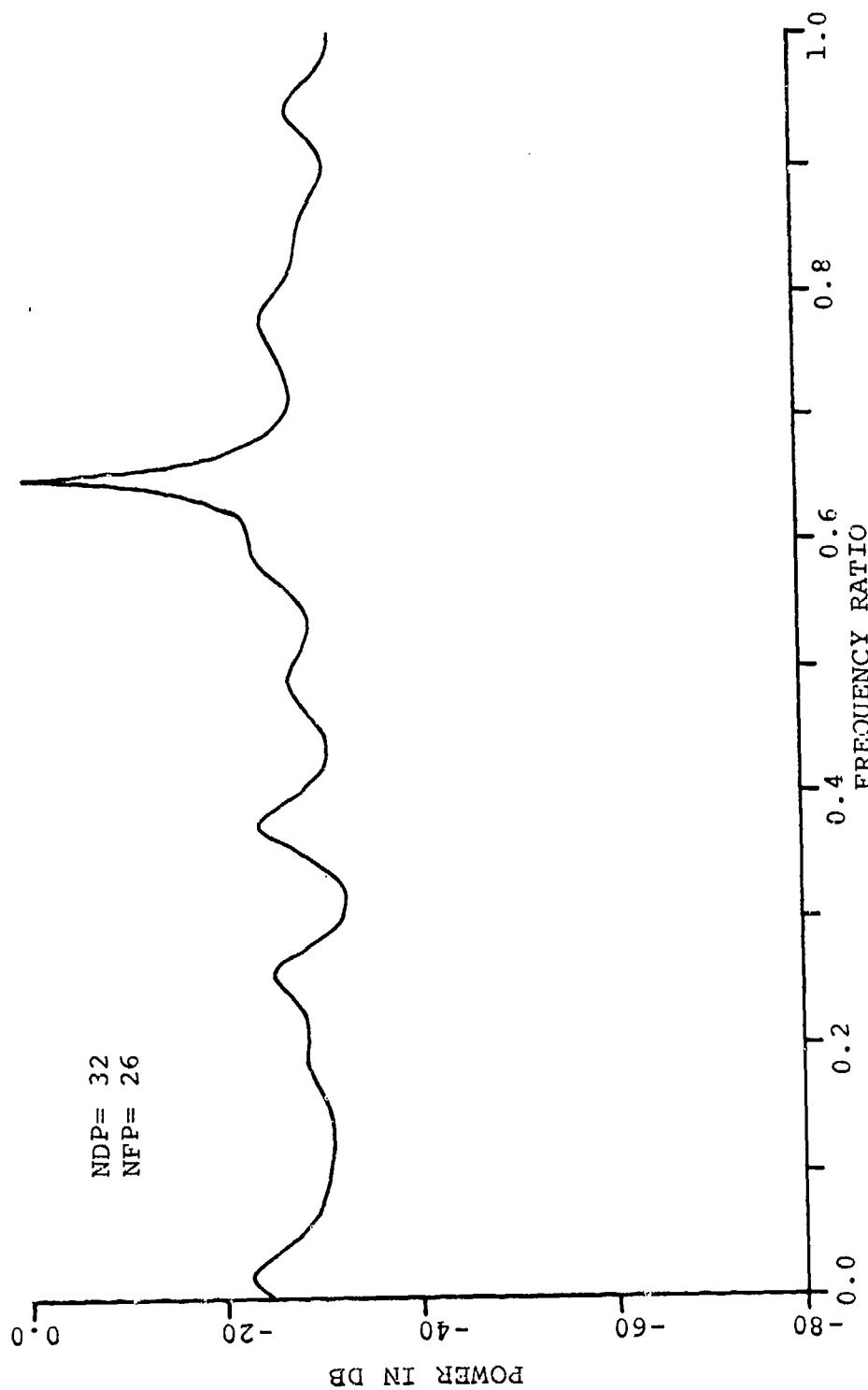
SLIDE 9



SIGNAL FREQUENCY = .375
SIGNAL LOCATED OUTSIDE CLUTTER REGIONS

SLIDE 10

MESA DETECTOR



SIGNAL FREQUENCY = .650
SIGNAL LOCATED INSIDE CLUTTER REGION

SLIDE 11

THE MAXIMUM ENTROPY METHOD APPLIED TO RADAR ADAPTIVE DOPPLER FILTERING

J. H. SAWYERS

Hughes Aircraft Company
Ground Systems Division
Fullerton, California

Abstract

The maximum entropy method of Burg is well known as a means of estimating high resolution power spectra from short time series data. In this paper, the Burg method is employed in the calculation of adaptive doppler filter coefficients for a pulse doppler radar operating in a nonstationary clutter environment. It is demonstrated by simulation that the adaptive doppler filters converge rapidly and accurately in severe models of clutter and thermal noise.

Introduction

High speed digital hardware technology has progressed to the stage of development where the implementation of many sophisticated radar signal processing techniques here-to-fore impractically difficult has become feasible. One area of signal processing that is receiving increased consideration is adaptive doppler filtering which is the subject of this paper.

Modern pulse doppler radar systems are required to operate in nonstationary clutter environments comprising land, weather, sea, chaff and other interference all of which can limit the ability of the radar to detect and track targets if the doppler filters are not matched to the clutter. The intent of this paper is to demonstrate the performance of an adaptive doppler filter bank (ADFB) that employs the maximum entropy method (MEM) of Burg [1] in deriving the filter coefficients in a nonstationary clutter environment.

The maximum entropy method of power spectrum estimation is well known in the field of geophysics, and in recent years has attracted attention in other scientific fields including radar. The acceptance occurs because MEM yields higher resolution power spectrum estimates from short time series data when compared to conventional methods of power spectrum estimation [2] - [4]. Implicit in conventional methods is a window, either weighted or unweighted, that treats missing data as zero thereby causing spectral sidelobes and a loss of resolution. On the other hand, the MEM spectrum is noncommittal with regard to missing data but corresponds to the most

random time series whose autocorrelation function agrees with the known or derived values. Since adaptive filtering is related to power spectrum estimation, incorporating MEM into adaptive filter designs is appropriate.

The paper begins with a brief description of adaptive doppler filtering by a pulse doppler radar. The next section outlines how the MEM is used to derive the adaptive coefficients for the finite-impulse-response (FIR) filters comprising the ADFB. The results of a Monte-Carlo simulation of selected adaptive doppler filters of the ADFB are given in the last section. The filters are subjected to severe clutter including land, rain, bandlimited interference, multiple narrowband point sources and thermal (white) noise. It is demonstrated that MEM adaptive doppler filtering is characterized by rapid convergence and overall excellent clutter rejection properties.

Adaptive Doppler Filtering

The pulse doppler radar in the search mode transmits a pulse train at the pulse-repetition-frequency (PRF) of $1/T$ pulses per second. The interpulse spacing, T , is based on the maximum unambiguous range and doppler frequency requirements of the particular radar. The unambiguous doppler frequency is bounded by $\pm 1/2T$. N pulses of the pulse train constitute a dwell, and after each dwell the radar antenna is directed to a different beam position. At each beam position, the ADFB processes the N doppler-shifted pulses received from each range resolution cell in the beam. The block diagram of the ADFB and target detection logic that processes these N pulses is shown in Figure 1.

In Figure 1, $\underline{x}_k$ is the complex digitized baseband signal vector of the N received pulses in a dwell on the k -th scan of a particular range resolution cell.[†] The $\underline{w}_k(n)$'s are the complex coefficient vectors of individual doppler FIR filters in the ADFB and are calculated by means of an adaptive algorithm using the MEM. The design criterion for the individual doppler filters is to provide maximum signal-to-clutter-plus-noise ratio. This dictates that the $\underline{w}_k(n)$'s be calculated by the following equation:

$$\underline{w}_k(n) = \hat{R}_k^{-1} \underline{u}(n). \quad (1)$$

In (1), $\hat{R}_k^{-1}$ is the inverse of the estimated correlation matrix of the clutter environment that existed on scan k , and $\underline{u}(n)$ is the steering vector that specifies the doppler frequency f_n at which the signal-to-clutter-plus-noise ratio (SCNR) is to be maximized. $\underline{u}(n)$ is given by

$$\underline{u}(n)^T = \begin{bmatrix} e^{-j\pi(N-1)f_n T} & e^{-j\pi(N-3)f_n T} & \dots & e^{j\pi(N-1)f_n T} \end{bmatrix}, \quad (2)$$

[†]The matrix notation used is as follows: Lower case letters are scalars; underlined lower case letters are vectors; and upper case letters are square matrices. The symbols *, T and $\dagger$ denote complex conjugate, transpose and complex conjugate transpose, respectively.

where the f_n 's are uniformly distributed in the interval $|f_n| \leq 1/2T$. The method of calculating $\hat{R}_k^{-1}$ is given in the next section.

In the detection process, the magnitudes of the outputs from the N doppler filter-detector combinations are calculated by the following:

$$v_k(n) = \left| \mathbf{w}_{k-1}^\dagger(n) \mathbf{x}_k \right| \quad (3)$$

The maximum $v_k(n)$ is automatically selected and compared to a predetermined threshold setting that is based on false alarm considerations. If $v_k(n)_{\max}$ exceeds the threshold, a target detection is declared at the corresponding doppler frequency f_n , and the data $\mathbf{x}_k$ is inhibited in the adaptive algorithm from contributing to the update of $\hat{R}_k^{-1}$.

The MEM Adaptive FIR Filter [5]

In radar adaptive doppler filtering, it is necessary to calculate and update $\hat{R}_k^{-1}$ periodically in order to account for changing clutter conditions. The computational procedure used by the MEM adaptive filter algorithm in Figure 1 to accomplish this is as follows:

The last coefficients of each l -th order prediction filter, $a_{l,l}(k)$, $l = 1, \dots, L$, where $L \leq N/2$, is calculated from $\mathbf{x}_k$ and stored. The procedure for calculating the $a_{l,l}(k)$'s from the $\mathbf{x}_k$'s is given elsewhere [6,7]. After a selected number of scans, $k = m, \dots, n$, the average of the $a_{l,l}(k)$'s are calculated by

$$\tilde{a}_{l,l} = \frac{1}{n-m+1} \sum_{k=m}^n a_{l,l}(k) \quad (4)$$

for $l = 1, \dots, L$. Next, the L -th order prediction filter coefficients, $\tilde{a}_{L,i}$'s, are calculated from the $\tilde{a}_{l,l}$'s by means of the recursion

$$\tilde{a}_{l,i} = \tilde{a}_{l-1,i} + \tilde{a}_{l,l} \tilde{a}_{l-1,l-i}^* \quad (5)$$

$\hat{R}_k^{-1}$ is then calculated from the $\tilde{a}_{L,i}$'s by a simple algorithm that results from choosing $L \leq N/2$. The elements, $z_{i,j}$, of the upper half of the matrix $\tilde{p}_L \hat{R}_k^{-1}$, where $\tilde{p}_L$ is the average of the L -th order prediction error powers [1], are calculated by

For $i = 1, \dots, N/2$ and $j = i, \dots, L+1$

$$z_{i,j} = \tilde{a}_{L,i-1} \tilde{a}_{L,j-1}^* + z_{i-1,j-1} \quad (6a)$$

where $\tilde{a}_{L,0} = 1$ and $z_{i,0} = z_{0,j} = 0$.

For $i = 2, \dots, N/2$ and $j = L+2, \dots, i+L$

$$z_{i,j} = z_{i-1,j-1} \quad (6b)$$

For $i = 1, \dots, N/2-1$ and $j = i+1, \dots, N/2$

$$z_{j,i} = z_{i,j}^* \quad (6c)$$

For $i = 1, \dots, N/2$ and $j = i+L+1, \dots, N$

$$z_{i,j} = 0 \quad (6d)$$

Combining (6), (2) and (1) yields the desired $\underline{w}_k(n)$'s.

Performance Analysis of the MEM ADFB

For examples showing the performance of the MEM ADFB, we choose $N = 32$ and $L = 15$. The performance is evaluated by Monte-Carlo simulation of radar returns from three clutter environments. The power density spectrum of the first, illustrated in Figure 2a, consists of land, rain, band-limited interference and thermal noise. The second, illustrated in Figure 6a, consists of five narrowband point sources, doppler offset land clutter and thermal noise. The third is thermal noise only.

The power density spectrums obtained from the Yule-Walker (YW) equations [3] using the first $L = 15$ lags of the exact autocorrelation functions obtained from the spectrums of Figures 2a and 6a are given in Figures 2b and 6b, respectively.

The simulated time series representing the radar return from the clutter environment is generated by taking the discrete Fourier transform of a random phase line voltage spectrum corresponding to those of Figures 2a and 6a. The phase is assumed to be independently distributed from line-to-line and dwell-to-dwell. Added thermal noise is obtained from a Gaussian random number generator. The clutter is assumed stationary over the period of adaptation in regard to the use of equation (4).

The optimum response functions and corresponding SCNR's for selected filters of the ADFB, obtained by means of a closed-formed solution from the spectrums of Figures 2a and 6a, are given for reference in Figures 3a, 4a, 5a and 7a. Similarly, the responses and corresponding SCNR's for the YW spectrums are given in Figures 3b, 4b, 5b and 7b. The definition of SCNR is based on the assumption of unity signal voltage and unity thermal noise power per pulse at the input to the ADFB. It should be noted that the power density spectrums and the filter response functions are periodic; one period is shown.

The simulated MEM power density spectrums obtained from one and two dwells corresponding to the exact spectrums of Figures 2a and 6a are given in Figures 2c, 6c, 2d and 6d, respectively. The corresponding adaptive response functions and SCNR's for selected filters of the ADFB are given in c and d of Figures 3, 4, 5 and 7. The figures are arranged for easy comparison.

The thermal noise only filter response functions are given in Figure 8. Note here that the convergence time is also fast as in the previous examples employing strong clutter.

In comparing the results of the analysis as given in Figures 2 through 8, excellent results are obtained after only one dwell of a particular range resolution cell, regardless of the clutter environment. However, in those cases where the main lobe of the response function is close to strong clutter, for example Figure 4, possibly two or more dwells are required to improve the SCNR by one dB or so. If the main lobe of the adaptive doppler filter is entirely within heavy clutter, Figure 5 for example, the target must be quite large in order to be detected.

While adaptation can occur on targets as well as clutter, thus potentially inhibiting their detection, targets can be seen as they move from one resolution cell to the next. In addition, target adaptation effects can be ameliorated by performing the averaging, using (4), over a number of resolution cells in deriving the adaptive filter coefficients.

Conclusions

This analysis has shown in somewhat of a limited manner the potential benefits offered by MEM adaptive doppler filtering: rapid convergence in arbitrary clutter environments and adaptive response functions that closely approach the optimum.

References

- [1] J. P. Burg, "Maximum entropy spectral analysis," Ph. D. dissertation, Stanford University, Stanford, CA, May 1975.
- [2] R. T. Lacoss, "Data adaptive spectral analysis methods," Geophysics, vol. 36, pp. 661-675, Aug. 1971.
- [3] T. J. Ulrych and T. N. Bishop, "Maximum entropy spectral analysis and autoregressive decompositions," Reviews of Geophysics and Space Physics, vol. 13, pp. 183-200, 1975.
- [4] M. Kaveh and G. R. Cooper, "An empirical investigation of the properties of the autoregressive spectral estimator," IEEE Trans. Info. Theory, vol. IT-22, No. 3, pp. 313-323, May 1976.
- [5] J. H. Sawyers, "Applying the maximum entropy method to adaptive digital filtering," Conf. Rec. Twelfth Asilomar Conf. on Circuits, Systems and Computers, pp. 198-202, Nov. 6-8, 1978.
- [6] M. Andersen, "On the calculation of filter coefficients for maximum entropy spectral analysis," Geophysics, vol. 39, No. 1, pp. 69-72, Feb. 1974.
- [7] S. Haykin and S. Kesler, "The complex form of the maximum entropy method for spectral estimation," Proc. IEEE, vol. 64, No. 5, pp. 822-823, May 1976.

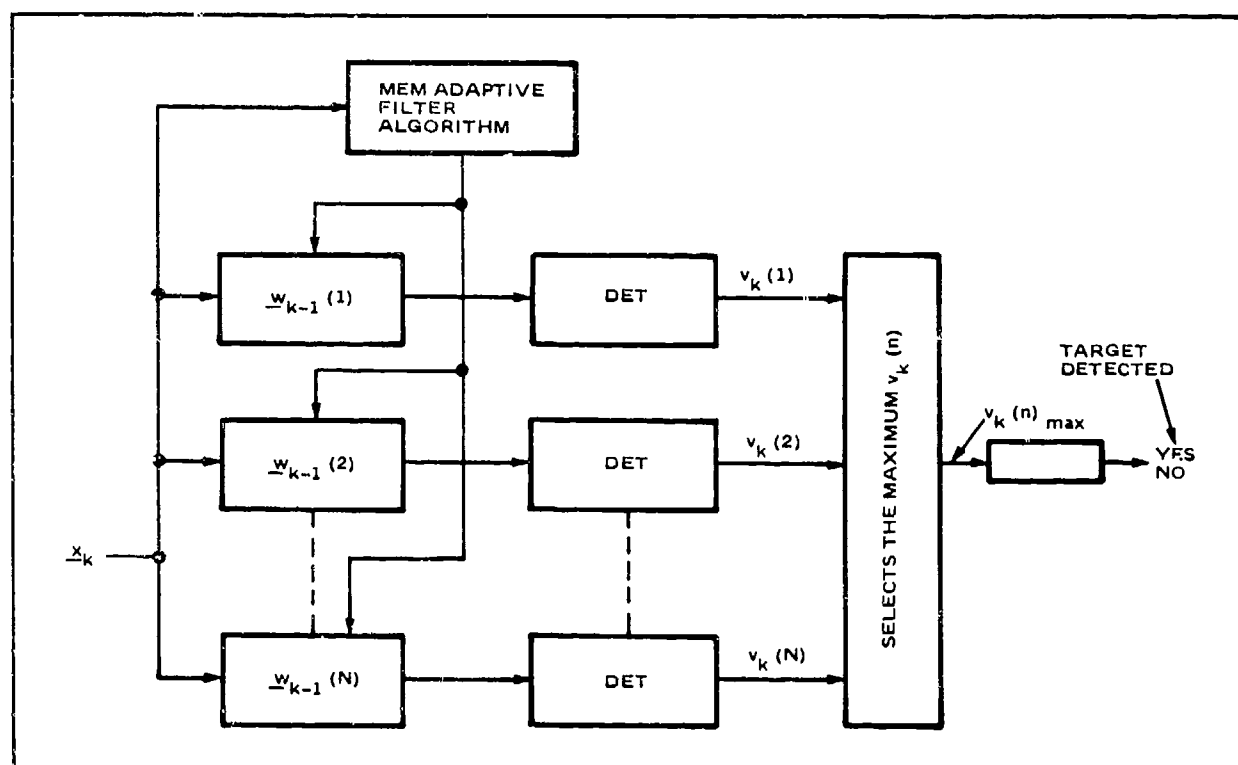


Figure 1. The Adaptive Doppler Filter Bank (ADFB) and Target Detection Logic

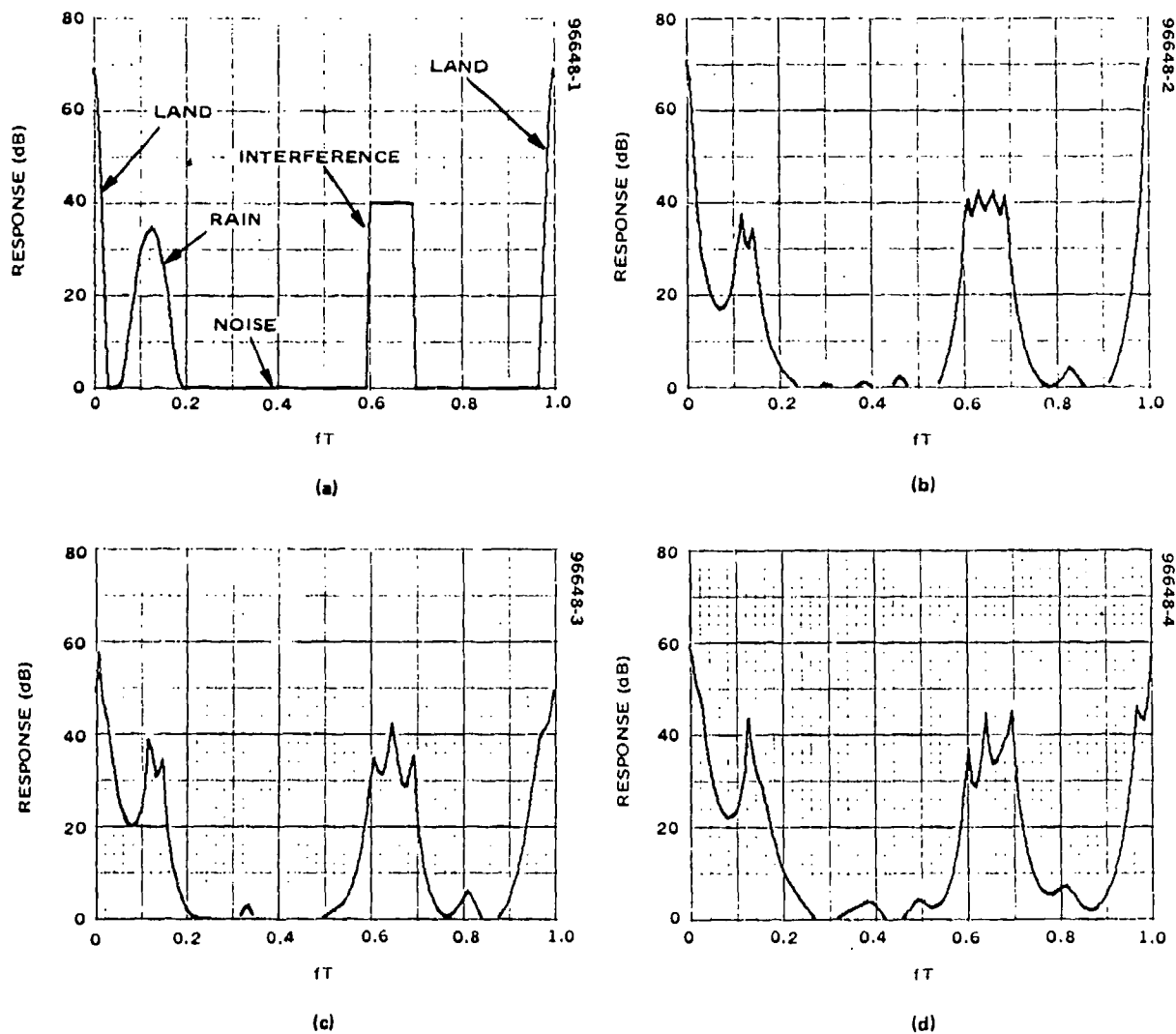


Figure 2. (a) The power density spectrum of land clutter, rain, band-limited interference and thermal noise. The clutter-to-noise ratio, CNR = 50.1 dB. (b) The YW spectrum using the first 15 lags of the autocorrelation function derived from (a). (c) The simulated MEM spectrum after one dwell. (d) The simulated MEM spectrum after two dwells

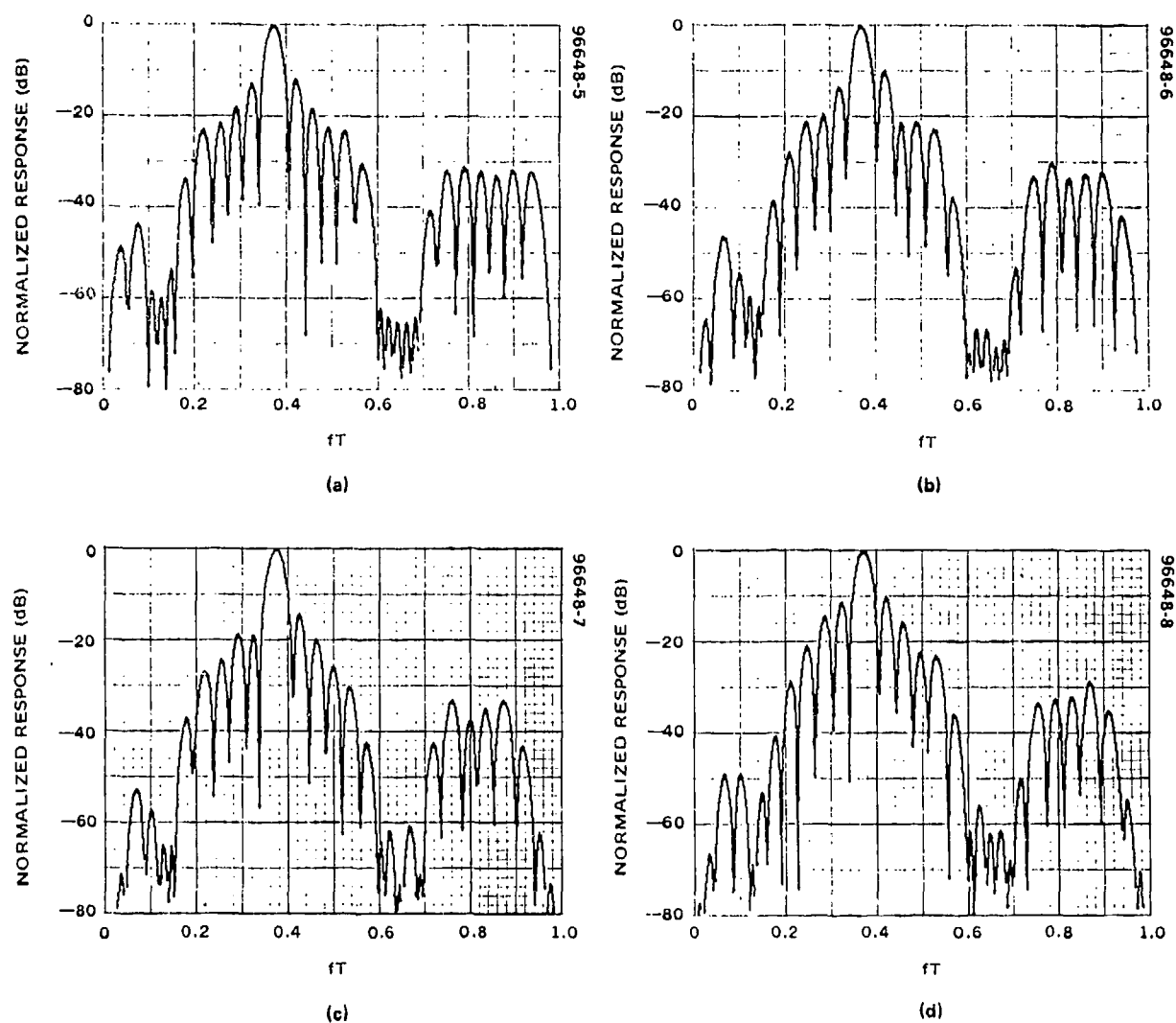


Figure 3. The Adaptive Doppler Filter Responses to the Spectrums of Figure 2. $f_n T = 0.375$. (a) The optimum closed-form response. SCNR = 14.8 dB. (b) The YW response. SCNR = 14.7 dB. (c) The MEM response after one dwell. SCNR = 14.0 dB. (d) The MEM response after two dwells. SCNR = 14.5 dB.

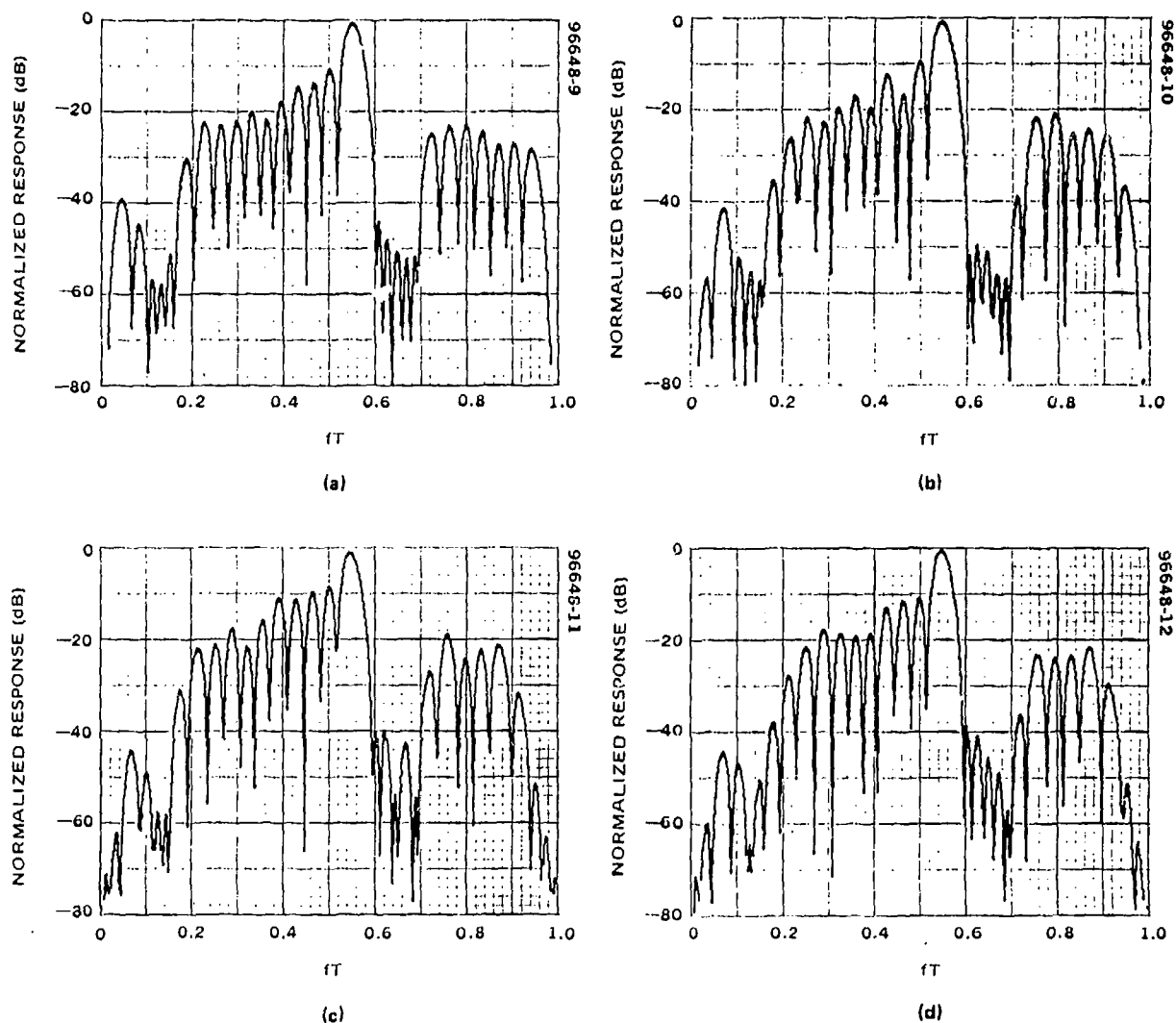


Figure 4. The Adaptive Doppler Filter Responses to the Spectrums of Figure 2. $f_n T = 0.5625$. (a) The optimum closed-form response. SCNR = 11.9 dB. (b) The YW response. SCNR = 11.2 dB. (c) The MEM response after one dwell. SCNR = 9.4 dB. (d) The MEM response after two dwells. SCNR = 10.6 dB.

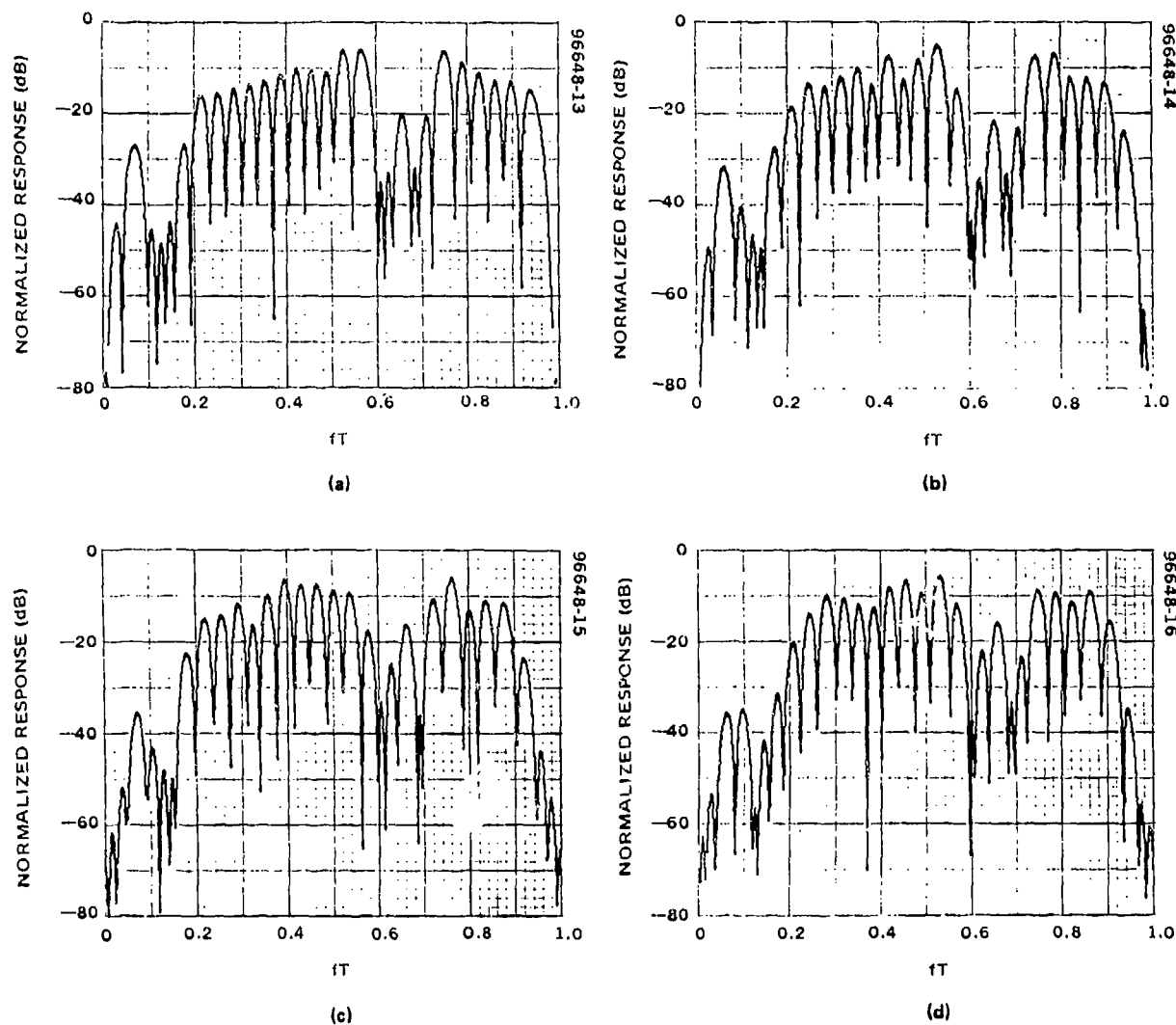
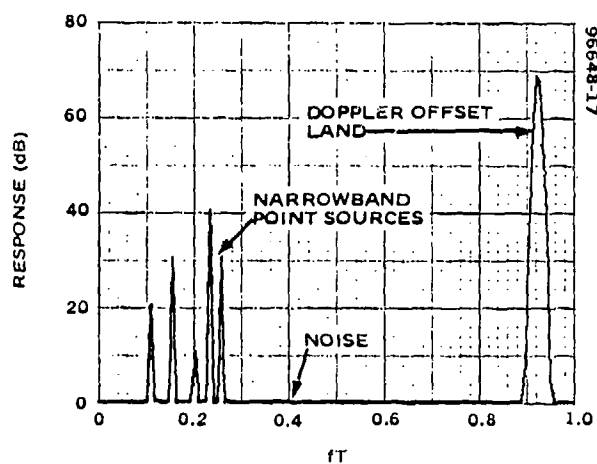
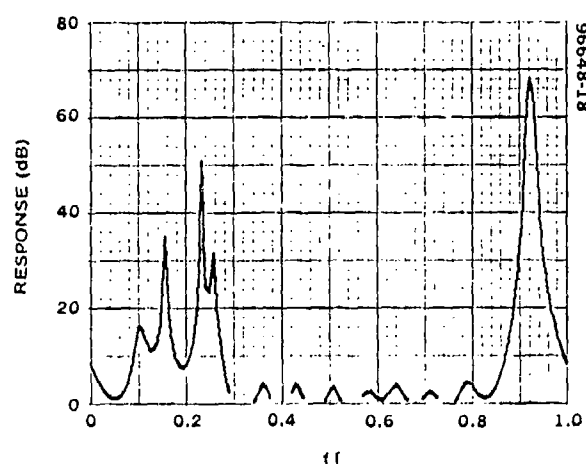


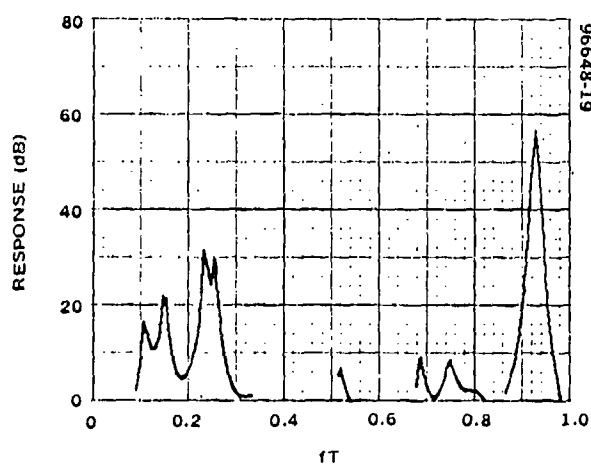
Figure 5. The Adaptive Doppler Filter Responses to the Spectrums of Figure 2. $f_n T = 0.65625$. (a) The optimum closed-form response. SCNR = -23.2 dB. (b) The YW response. SCNR = -23.5 dB. (c) The MEM response after one dwell. SCNR = -29.6 dB. (d) The MEM response after two dwells. SCNR = -24.4 dB.



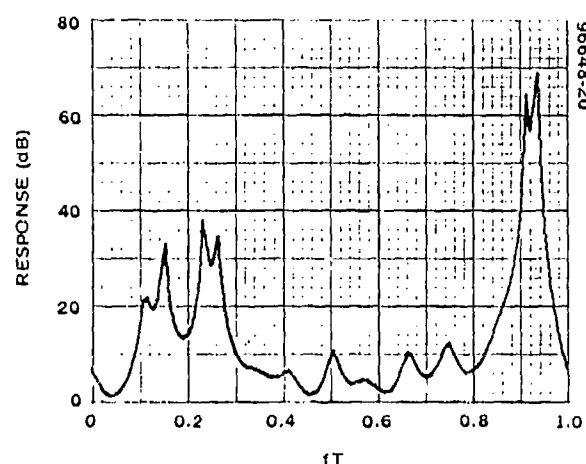
(a)



(b)



(c)



(d)

Figure 6. (a) The power density spectrum of five narrowband point sources, doppler offset land clutter and thermal noise. $CNR = 50$ dB. (b) The YW spectrum using the first 15 lags of the autocorrelation function derived from (a). (c) The simulated MEM spectrum after one dwell. (d) The simulated MEM spectrum after two dwells.

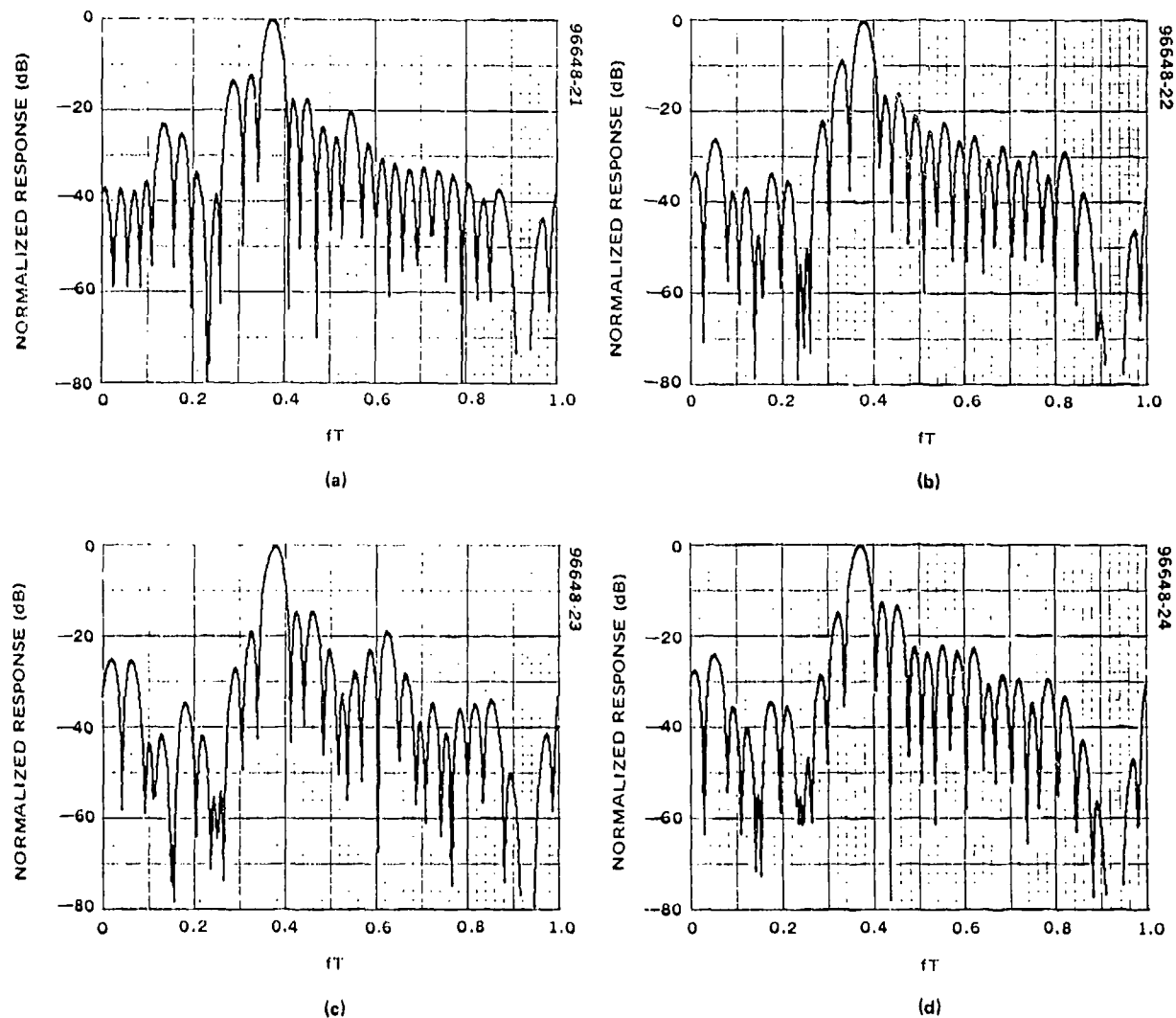


Figure 7. The Adaptive Doppler Filter Responses to the Spectrums of Figure 6. $f_n T = 0.375$. (a) The optimum closed-form response. SCNR = 15.0 dB. (b) The YW response. SCNR = 14.6 dB. (c) The MEM response after one dwell. SCNR = 14.3 dB. (d) The MEM response after two dwells. SCNR = 14.6 dB.

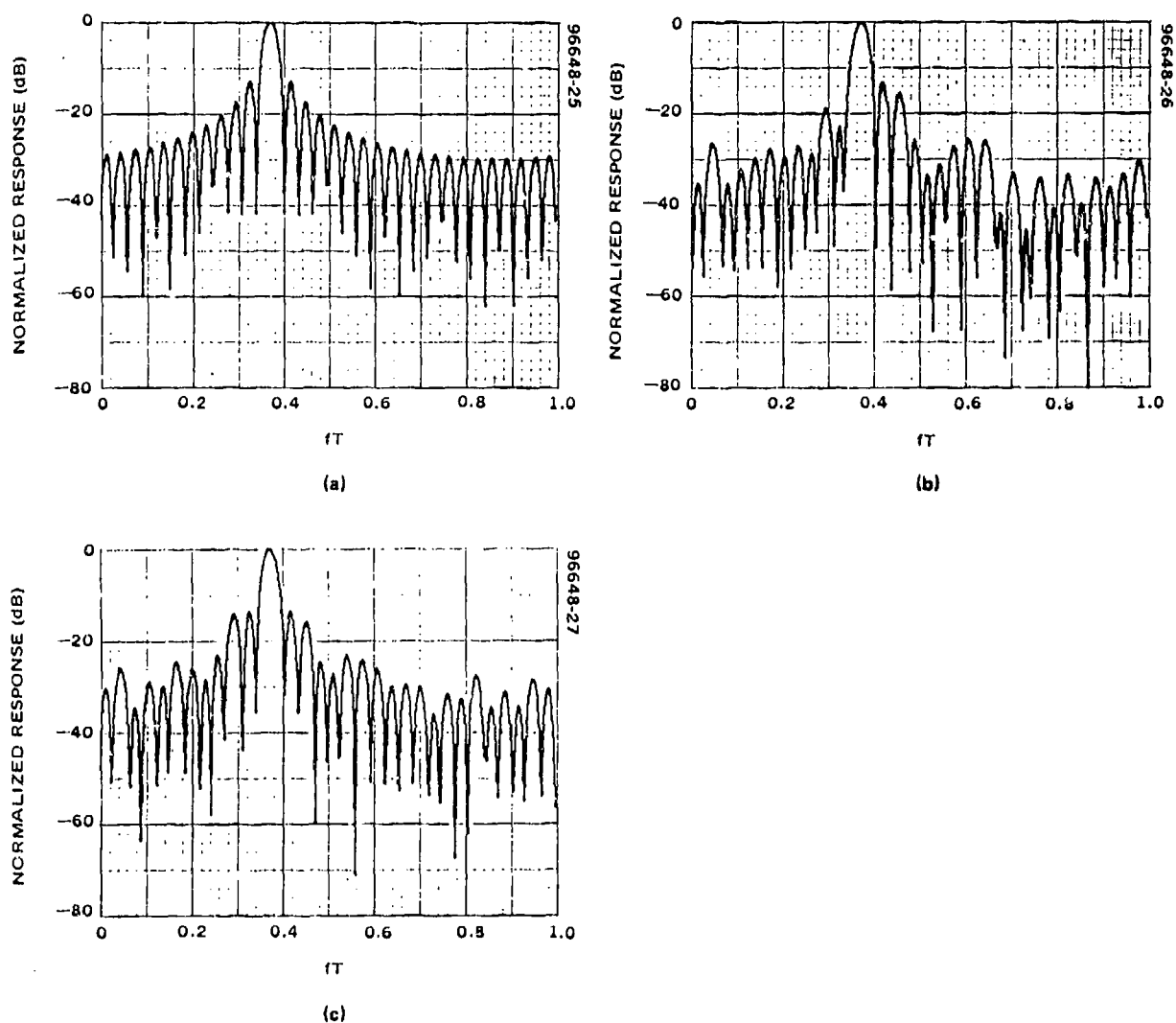


Figure 8. The Adaptive Doppler Filter Response in Thermal Noise. $f_n T = 0.375$. (a) The optimum closed-form response. SNR = 15.1 dB. (b) The MEM response after one dwell. SNR = 14.7 dB. (c) The MEM response after two dwells. SNR = 14.9 dB.

*MISSION
of
Rome Air Development Center*

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control Communications and Intelligence (C³I) activities. Technical and engineering support within areas of technical competence is provided to ESD Program Offices (POs) and other ESD elements. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.