THE ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES, 45:1–71, 1981 January

STUDIES IN ASTRONOMICAL TIME SERIES ANALYSIS. I. MODELING RANDOM PROCESSES IN THE TIME DOMAIN

JEFFREY D. SCARGLE
Ames Research Center, NASA, Moffett Field
Received 1979 December 10; accepted 1980 May 14

CONTENTS

I. Introduction: Astronomical Time Series	3	1–A4
II. Modeling Random processes in the Time Domain	5	1–A7
a) Time Series and Random Processes	5	1–A7
b) White Noise; Independently Distributed Noise	9	1–A11
c) The Moving Average (MA) Model	9	1–A11
d) The Autoregressive (AR) Model	13	1–B13
e) The Relationship between the AR and MA Models	15	1–B4
f) Autoregressive-Moving Average (ARMA) Models	15	1–B4
g) AR Integrated MA (ARIMA) Models and Nonstationary Processes	17	1–B6
h) The Shot Noise Model	18	1–B7
III. The Structure of Pulses	18	1–B7
a) The Discrete Representation of Pulse Shapes	19	1–B8
i) One-sided Pulses	19	1–B8
ii) Two-sided Pulses	19	1–B8
b) Z-Transforms	19	1–B8
c) Convolution	20	1–B9
d) Factorization	20	1–B9
e) Delay (or Phase) Character	22	1–B11
f) Inverse Filters	23	1–B12
g) Correlation Functions and Power Spectra	28	1–C3
IV. Model Construction	28	1–C3
a) An Existence Theorem: the Wold Decomposition	28	1–C3
b) A Less Restrictive Existence Theorem	29	1–C4
c) Deconvolution via Independently Distributed Innovations	31	1–C6
d) Predictive Deconvolution of Time Series	35	1–C10
e) Predictive Deconvolution with the Absolute Value Norm	36	1–C11
V. Computational Methods	42	1–D3
a) Sampling	42	1–D3
b) Identification	42	1–D3
c) Computing the Innovation $R(A)$	42	1–D3
d) The Computation of $D_F(R)$	43	1–D4
e) Minimization of $D_F(A)$	43	1–D4
f) Computation of Subsidiary Quantities	45	1–D6
g) Gaps and Uneven Sampling	46	1–D7
i) No Coherence across the Gap	46	1–D7
ii) Coherence across the Gap	46	1–D7
iii) Arbitrary Sampling	46	1–D7
VI. Numerical Experiments	47	1–D8
a) Experiment 1: Comparison of Dependence Measures	47	1–D8
b) Experiment 2: Detailed Study of an AR(1, 1) Process	48	1–D9
c) Experiment 3: An AR(2, 1) Process	50	1–D11
d) Experiment 4: Gaussian Noise	52	1–D13
e) Experiment 5: A Sine Wave	52	1–D13
f) Experiment 6: 3C 273	53	1–D14
g) Discussion	58	1–E5
Appendix—The Algorithm	59	1–E6
Index	69	1–F2

LIST OF TABLES

Table 1	Properties of Dipoles and Their Inverses	24	1–B13
Table 2	Test Results with Various Dependence Measures (Innovations)	48	1–D9
Table 3	Test Results with Various Dependence Measures (Noise)	48	1–D9
Table 4	Deconvolution of the Process in Fig. 20	49	1–D10
Table 5	Deconvolution of the Process in Fig. 25	52	1–D13
Table 6	Deconvolution of the Process in Fig. 28	53	1–D13
Table 7	Deconvolution of the Light Curve of 3C 273	56	1–E3
Table A1	The FORTRAN Code	60	1–E7

LIST OF ILLUSTRATIONS

Fig. 1	A Real Time Series (3C 120)	4	1–A5
Fig. 2	An Artificial Time Series	6	1–A8
Fig. 3	Two Realizations of the Same Process	6	1–A8
Fig. 4	Four Degrees of Randomness	10	1–A12
Fig. 5	Noise Processes with Different Distributions	11	1–A13
Fig. 6	Schematic Representation of the MA	12	1–B1
Fig. 7	Schematic Representation of the AR	14	1–B3
Fig. 8	A Second-Order AR Process	15	1–B4
Fig. 9	Several More Second-Order AR Processes	16	1–B5

Fig. 10	Causal and Acausal Inverses	17	I–B6
Fig. 11	One- and Two-sided Pulses	19	I–B8
Fig. 12	Convolution of Filters	20	I–B9
Fig. 13	Convolution of Causal and Acausal Dipoles	21	I–B10
Fig. 14	The Concept of Minimum and Maximum Delay	23	I–B12
Fig. 15	The Wraparound Concept	25	I–B14
Fig. 16	Filter Inverses	26	I–C1
Fig. 17	The Unautocorrelated Pulse *P*	31	I–C6
Fig. 18	Prediction Error Filters	37	I–C12
Fig. 19	An Example of Absolute Value Minimization	38	I–C13
Fig. 20	Realization of an AR(1,1) Process	48	I–D9
Fig. 21	Estimated and Exact Innovations for This Process	50	I–D11
Fig. 22	Estimated and Exact Pulses for This Process	50	I–D11
Fig. 23	Estimated and Exact Innovations (Different Solution)	50	I–D11
Fig. 24	Estimated and Exact Pulses (Different Solution)	51	I–D12
Fig. 25	Realization of an AR(2,1) Process	51	I–D12
Fig. 26	Estimated and Exact Innovations for This Process	53	I–D14
Fig. 27	Estimated and Exact Pulses for This Process	53	I–D14
Fig. 28	Realization of the Same Process with Added Noise	54	I–E1
Fig. 29	Estimated and Exact Innovations for This Process	54	I–E1
Fig. 30	Estimated and Exact Pulses for This Process	54	I–E1
Fig. 31	Realization of Gaussian Noise; Estimated Innovation	54	I–E1
Fig. 32	Estimated Pulse for This Process	55	I–E2
Fig. 33	A Sine Wave with Added Noise; Estimated Innovation	55	I–E2
Fig. 34	Estimated Pulse for This Process	55	I–E2
Fig. 35	Light curve of 3C 273; Estimated Innovation	55	I–E2
Fig. 36	Minimum Delay and Mixed Delay Pulses for 3C 273	57	I–E4
Fig. 37	Innovations for the Minimum and Mixed Delay Solutions for 3C 273	57	I–E4
Fig. 38	Distribution of the Innovation for 3C 273	58	I–E5
Fig. 39	Grid for the Computation of the Cumulative Distribution	68	I–F1

ABSTRACT

This discussion of time series data produced by random physical processes emphasizes astrophysical data analysis. Several random process models phrased in the time domain are defined and discussed. The *moving average* (MA) model represents the data as a sequence of pulses occurring randomly in time, with random amplitudes. The *autoregressive* (AR) model represents the correlations in the process in terms of a linear function of its past values and is closely related to the differential equation describing the dynamics of the system. A given stationary process always has both a MA and an AR representation, and one can easily be transformed into the other using the discrete Fourier transform. The moving average form is usually more suitable for interpretation, as the pulses and pulse amplitudes often have direct physical significance. But the AR parameters are easier to determine from the time series data. Hence, the procedure is to determine the best AR model from the sampled data and then transform it to a MA for interpretation and comparison with theory. The technique for determining the AR parameters is based on interpreting the AR model as a filter which, when applied to the data, yields the sequence of pulse amplitudes. The parameters are adjusted to maximize the randomness of the pulse amplitudes—that is, to make them as statistically independent as possible. (It is not enough to make the amplitudes uncorrelated, or white.) This maximization is implemented by specifying that the joint cumulative probability function of the pulse amplitudes be as close as possible to the product of the individual cumulative distribution functions. A procedure for carrying this out is presented as a FORTRAN algorithm which has proven to be relatively stable numerically. Results of test cases are given to study the effects of adding noise and of different distributions for the pulse amplitudes. A preliminary analysis of the optical light curve of the quasar 3C 273 is given.

Subject headings: functions: numerical methods — quasars

I. INTRODUCTION: ASTRONOMICAL TIME SERIES

This mostly self-contained introduction to time domain models of intrinsically random physical processes is directed toward astronomers and scientists in related fields, particularly those involved in the analysis and interpretation of data. The goals are to develop an intuitive understanding for this view of random processes and to give specific numerical techniques for the analysis of time series data. Many of the concepts presented here have been developed in other literatures, especially those of geophysics, economics, and speech analysis. Appropriate references will be given; although the terminology and basic philosophy will be somewhat different, the reader is urged to consult these references. Of particular value are the following reviews, which parallel the present work in their viewpoint and emphasis on applications to data analysis: Wold (1964) (especially the two chapters by E. A. Robinson), Robinson (1962, 1967*b*), Box and Jenkins (1970), Kanasewich (1975), Claerbout (1976), and Granger and Newbold (1977). Reviews of stochastic processes in astronomy are given by Deeming (1970), Rothschild (1977), and Press (1978). A pioneering paper in the application of time domain models of random processes in astronomy is Fahlman and Ulrych's (1975) analysis of the optical light curve of 3C 273 (see also Ulrych and Clayton 1976; Ulrych and Bishop 1975). There are several books devoted to explicit computer codes for some of the operations discussed here (Simpson 1966; Robinson 1967*a*; Enochson and Otnes 1968). Texts are available on the following related topics: time series analysis (Hannan 1970; Anderson 1971), stochastic processes (Doob 1953; Parzen 1962; Bailey 1964; Papoulis 1965), prediction and optimization theory (Wiener 1949; Whittle 1963; Luenberger 1969), and probability theory (Feller 1957; Parzen 1960). There are also several interesting collections of related papers (Wax 1954; Rosenblatt 1963; Parzen 1967; Krishnaiah 1969). The 1974 December issue of the *IEEE Transactions on Automatic Control* was devoted to systems identification and time series analysis (see the papers by Hannan 1975, Akaike 1975, and Parzen 1974; see also Kailath 1974). For an extensive bibliography (roughly 10,000 entries) on time series and stochastic processes complete through 1959, as well as an interesting "graphic introduction to stochastic processes," see Wold (1965).

Much of the material in §§ II, III, and IV is borrowed from the literature cited above and is reviewed here because few astronomers have been exposed to this material. The emphasis in these sections on mixed delay acausal representations is unusual, although not new. What *is* original in this work is the procedure for determining pulse shapes and amplitude sequences based on the assumption that the latter are independently distributed. Many time series analysts feel that all information has been removed from data if one has found a filter which reduces the data to white (i.e., uncorrelated) noise. It does not seem to have been realized that there is still information remaining, that this is the *phase information* discarded when the complex absolute value of the Fourier transform is taken (to yield the power spectrum), and that this information can be extracted if one finds a filter which reduces the data to independently distributed noise. Also new is the use of cumulative distribution functions and the way in which they are estimated, although Parzen (1979) has recently emphasized a related distribution called the *quantile function*. Two other techniques for phase-sensitive deconvolution were presented at the Second Applied Time Series Analysis Symposium (Tulsa, Oklahoma, 1980 March 3–5) by Mendel (1980) of the University of Southern California and by Donoho (1981) of Harvard University. At the same meeting the author of this paper summarized the material contained here and outlined some of the ways in which astronomical data are different from those in geophysics, econometrics, speech analysis, and other areas (Scargle 1981). I have recently become aware of a paper by Benveniste, Goursat, and Ruget (1980), which also deals explicitly with this problem. It appears that their methods presuppose knowledge of the form of the distribution function of the input process R. The philosophy of the present work is that the determination of the unknown innovation is an interesting part of the overall problem.

Data from astronomy as well as from other physical and biological sciences often consist of a sequence of numbers, $\{X_1, X_2, X_3, \ldots, X_N\}$, obtained by measurement of quantity X at a set of times, $\{t_1, t_2, t_3, \ldots, t_N\}$. Such a sequence is a *time series*, and the data are *time series data*. The sample time series in Figure 1*a* illustrates a feature common in astronomical observations, brought about by practical considerations such as observing schedules, weather, equipment malfunction, etc.: the time points t_i are not evenly spaced. (It is then said that the sampling is uneven.) Several ways of graphically indicating to what degree the sampling is uneven are demonstrated in Figures 1*b*, 1*c*, and 1*d*. Sometimes it is assumed that X is actually constant, and the repeated measurements are made to reduce the uncertainty due to observational errors. Such data are not really time series data because the serial or sequential nature of the observations is irrelevant (i.e., the time ordering contains no useful information). This paper deals only with the situation where X may undergo real variations with time, and the sequential nature of the observations is crucial to the elucidation of the variations. The goal of the analysis—once the

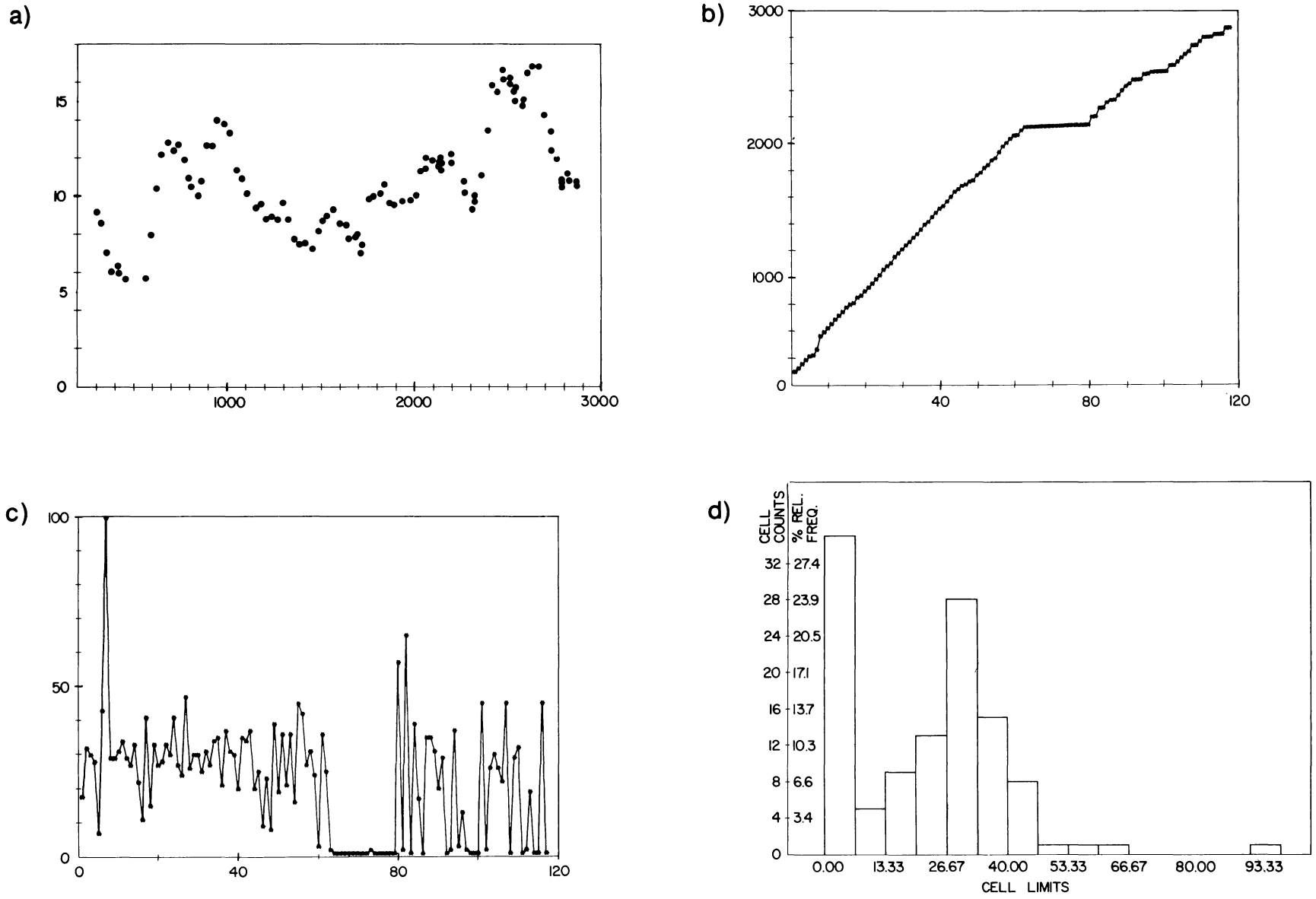

FIG. 1.—An example of time series data: the radiofrequency flux from the highly variable Seyfert galaxy 3C 120 (Algonquin Radio Observatory data from Medd, *et al.* 1972 and private communication). (*a*) The measured flux as a function of the time of observation, in days. (*b*) The times of observation t_i as a function of the index i. (This and the next two panels exhibit the degree of unevenness of the sampling: even spacing would yield a straight line.) (*c*) The differences $\Delta t_i = t_{i+1} - t_i$ as a function of i; this would be constant for even sampling. (*d*) The distribution function of Δt_i.

existence of such variations has been established—is the extraction of information about the physical process which gives rise to the variations.

This goal is usually approached by identifying a pattern in the observed variations and then trying to uncover the cause or explanation of the pattern, often in terms of a physical model. For example, the pattern may consist of a definite functional dependence of X on t, such as a linear variation or a harmonic oscillation partially hidden behind noise. One then attempts to fit a function (or model), the form of which is usually suggested by prior knowledge, physical understanding, or guesswork, to the model. This fitting is usually carried out by minimizing with respect to the model parameters a measure of the difference between the model and the observations. This measure is usually defined as the sum of some positive-definite function of the point-by-point difference between the model and the data. The most common such measure is the sum-of-squares of the X-differences, and the result is the ubiquitous least-squares procedure.

But what if there is no consistent pattern to the data? It may be, for example, that the data come from a physical system that is random. In some cases the process is intrinsically random because of quantum mechanical effects—for example, a radioactive decay process. In others, one should perhaps say the process is effectively random, because detailed knowledge of the initial conditions and of the governing physical laws might yield predictability (nonrandomness) for the system, but such knowledge may be virtually impossible or simply not practical. This situation is increasingly important in astrophysics, and examples could be cited from many areas, especially X-ray and radio astronomy. Is there any physical information to be extracted from such random data? The answer is yes, and the basic subject of this paper is the modeling of random processes to obtain concise and useful descriptions of the underlying physical processes. The discussion of the fundamental concept of *random process* in § II is oriented toward astrophysical data analysis and description in the time domain. Just as with deterministic processes, there is an infinite variety of possible forms or models which can be used to describe random processes. Familiar examples are shot noise models (Terrell and Olsen 1970; Terrell 1972), random walks (Wax 1954), diffusion models (Wax 1954), Markov chains (Doob 1953), discrete branching processes, birth and death processes, competition and predation, queueing processes (Bailey 1964), and other specialized techniques (e.g., Chandrasekhar and Münch 1951). In § II are descriptions of several types of models which are less familiar to astronomers, though ironically the models originated long ago in an astrophysical context (Yule 1927), namely the analysis of sunspot data. These models are emphasized here because of their direct physical interpretations (e.g., in terms of randomly occurring pulses [§ III]) and because of their very general applicability (§ IV). A common feature of these models is their simple and explicit *separation of the nonrandom from the random parts of the process*; this feature is responsible for their usefulness, because such a separation usually has a clear physical basis—i.e., the random and nonrandom parts correspond to fundamentally different aspects of the process. Such a separation is assured only for stationary processes (defined in § II*a*). We shall almost always assume that we are dealing with physical processes that satisfy the stationarity condition. For practical reasons we shall always assume that the time sampling is discrete (see § II*a*) rather than continuous. All processes will be assumed *ergodic*, such that time averages (determined from one realization) are the same as statistical averages (determined from an ensemble of realizations). In addition, non-Gaussian processes will play an important role, because Guassian processes cannot be unambiguously modeled in the way mentioned (see § IV). Model construction procedures are outlined in § IV; computational details appear in § V; and examples of the computations are presented in § VI. The Appendix contains a description of the algorithm, together with FORTRAN code, for the deconvolution of time series using cumulative distribution functions.

II. MODELING RANDOM PROCESSES IN THE TIME DOMAIN

This section begins with a brief account of the theory of random processes. Rather than a rigorous mathematical treatment, it is an informal heuristic discussion emphasizing a particular context—namely the interpretation of time series data produced by a physical process which is at least partly random. This situation is common in astrophysics as well as nearly all other quantitative sciences. Interpretation often means the construction of a model of the physical process. This section will discuss several ways of mathematically modeling a random process in the time domain. Frequency domain techniques, such as power spectrum analysis, are most useful when harmonic variations are present but are less suited to random variations. Two goals of this paper are to demonstrate the richness and usefulness of time domain analysis, and to indicate the type of problem for which it is superior to frequency domain analysis. The text by Box and Jenkins (1970) provides a good overview of this subject. The paper by Shinners (1974) is an interesting and practical discussion of the application of modeling techniques to human behavior.

a) *Time Series and Random Processes*

Consider a physical variable X that can be measured as a function of time t. In practice the values of t are not continuous but discrete because data-recording equipment is capable of sampling the observed quan-

tity only at a finite number of times, separated by some minimum time interval. There is thus a finite series of values of t, $\{t_i\}$, $i=1,2,3,\ldots,N$. The corresponding values of t, $\{t_i\}, i=1,2,3,\ldots,N$. Often the values of t can be chosen to be evenly spaced, so that $t_i = i\,\Delta t$, where Δt is the constant interval between the times of observation. In any case the set of numbers $\{X_i\}$ is called a *time series*. Figure 2 shows an example of a discrete, evenly spaced time series. Despite the name, time series are not limited to functions of time, which here stands for any independent variable of interest. Other examples are: position in space (three-dimensional), position on the sky (two-dimensional), and wavelength (one-dimensional). Because the term *time series* is used in all cases, it should be kept in mind that t may stand for a variable other than time, possibly of multiple dimensionality. Sometimes the term *sequential analysis* is used in place of *time series analysis* to emphasize the key property that the numbers X_i are sequentially related to each other. The dependent variable X may also be of multiple dimensionality.

A *process* is a rule or procedure that generates time series. That is, it is a prescription for determining the values of X for a given set of values of t and may or may not include a random element. Each such time series is called a *realization* of the process, and it is important to distinguish the process from a specific realization. The process can be identified with the set of all possible realizations of it. Figure 3 shows two more realizations of the same process which generated the time series in Figure 2.

The most interesting processes are those for which the rule generating the time series specifies probability distributions of the X_i, rather than specific values that are the same at every realization. In this case we have a *random process*, which can be thought of as a set of random variables, $\{X_i\}$. For precise definitions and discussions of random variables the reader is referred

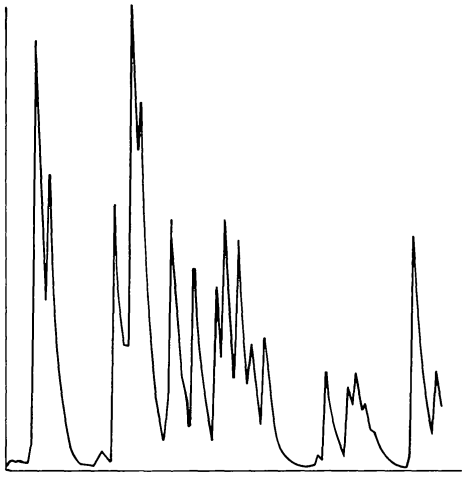

FIG. 2.—This artificial time series consists of a sequence of decaying exponential pulses occurring randomly in time in the sense that the amplitude of the pulse starting at any given time is a random variable. The sequence of pulse amplitudes was obtained by raising a sequence of random variables uniformly distributed on (0,1) to the ninth power. The horizontal axis represents time, which is discrete and evenly spaced, although straight lines have been drawn through the data points to give the curve more of the appearance of a continuous function. The apparent trend of diminishing amplitude with increasing time is spurious—the process generating these data is completely stationary.

to any text on probability or stochastic processes (e.g., Feller 1957; Parzen 1960, 1962). It is merely stated that a random variable, X_i, can be specified by giving its probability distributions, P_{X_i}, defined such that

$$P_{X_i}(x)\,dx = Pr\{x \le X_i \le x+dx\} \qquad (1)$$

in the usual limiting sense.[1] In many cases two random

[1] $Pr\{\bullet\}$ stands for the probability of event $\bullet$. In these definitions and elsewhere we shall use capital letters for the process (X) or random variable (X_i) and lower case for specific values of the random variable (e.g., x).

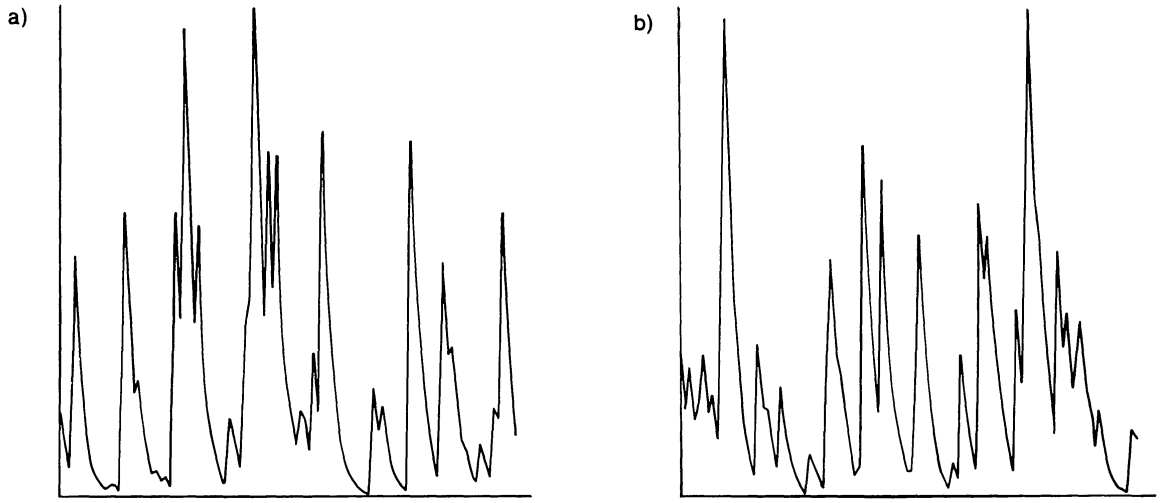

FIG. 3.—These two time series are different realizations of the same random process which generated the time series depicted in Fig. 2.

variables are related to each other; e.g., knowledge of the value of one may provide information about the other. There are two important definitions concerning the *degree* of such relatedness: two random variables, X and Y, are said to be

> *Independent* (of each other) if their joint probability distribution function equals the product of their individual probability distribution functions:
>
> $$P_{XY}(x, y) = P_X(x) P_Y(y),$$
>
> for all x and y,

and

> *Uncorrelated* if the expected value of their product equals the product of their expected values:
>
> $$\langle XY \rangle = \langle X \rangle \langle Y \rangle .$$

The joint probability distribution P_{XY} is defined by

$$P_{XY}(x, y)\, dxdy = Pr\{x \le X \le x + dx \text{ and } y \le Y \le y + dy\}. \quad (2)$$

Angle brackets are used for the *expected value* of the quantity enclosed:

$$\langle q \rangle = \int_{-\infty}^{\infty} P_X(x)\, q(x)\, dx. \quad (3)$$

The more familiar definition of uncorrelation is for the case where the processes are assumed (or made) to be zero-mean, so that $\langle XY \rangle$ also vanishes. Note that independence is the stronger of the two properties; it is easy to show that independence implies uncorrelation, but not vice versa. This is a key fact, and later we shall deal with variables that are uncorrelated with each other but are not independently distributed. There is a third property, intermediate between independence and uncorrelation:

> X has the *martingale difference property* (MDP) with respect to Y if the conditional expectation value of X (given the value of Y) is the same as the unconditional expectation value of X: $\langle X|Y \rangle = \langle X \rangle$.

The name martingale difference property (Segall 1976) is based on the fact that this kind of process is to a martingale as an independently distributed process is a process with independent increments. (Martingales and processes with independent or uncorrelated increments are usually defined in continuous time and will be of no concern here.) It can be shown that if X and Y are independent, they each have the MDP with respect to the other; in turn, if X has the MDP with respect to Y, then X and Y are uncorrelated.

Let us now be more precise with the definition of a process, which was already defined as a set of random variables. Take the set to be finite, with N members. The process is completely specified by giving any one of the following functions:

1) *The complete joint probability distribution function*,

$$P_{X_1, X_2, \ldots, X_N}(x_1, \ldots, x_N)\, dx_1 dx_2 \ldots dx_N = Pr\{x_1 \le X_1 \le x_1 + dx_1 \text{ and } x_2 \le X_2 \le x_2 + dx_2 \text{ and} \ldots \text{and } x_N \le X_N \le x_N + dx_N\}; \quad (4)$$

2) *The joint cumulative distribution function*,

$$F_{X_1, X_2, \ldots, X_N}(x_1, x_2, \ldots, x_N) = Pr\{X_1 \le x_1 \text{ and } X_2 \le x_2 \text{ and} \ldots \text{and } X_N \le x_N\}; \quad (5)$$

3) *The joint characteristic function*,

$$\phi_{X_1, X_2, \ldots, X_N}(u_1, u_2, \ldots, u_N) = \langle \exp i(u_1 X_1 + u_2 X_2 + \ldots + u_N X_N) \rangle. \quad (6)$$

Equations (5) and (6) are straightforward generalizations of the individual cumulative distribution function

$$F_X(x) = Pr\{X < x\} \quad (7)$$

and the characteristic function

$$\phi_X(u) = \langle \exp(iuX) \rangle \quad (8)$$

of a single random variable X. One can define what is called the *moment-generating function* by dropping the i in the definition of the characteristic function, but it does not always exist and is therefore of less theoretical importance. Nevertheless, it is of some practical use because of the concise way the nature of a variable can be expressed in terms of its moments.

We shall now distinguish several degrees of randomness. It is convenient to define these categories in terms of predictability. A process is said to be *deterministic* if, based on past observations, the future of the process can be predicted exactly (i.e., with zero error). An example of such a process is one with no probabilistic element at all, such as the sinusoid $X_i =$

$\sin(\omega t_i + \phi)$; in this case all realizations are the same. However, there are purely deterministic processes for which each realization is different. The above sinusoid would be an example if the phase ϕ were a random variable, fixed during each realization but chosen randomly each time—each realization would be exactly predictable once the phase had been determined by observation. An example of a deterministic process from astronomy would be a perfectly regular variable star.

A *random process*, on the other hand, is not perfectly predictable. Even if the rule generating the time series is known completely, it has a stochastic nature. Different realizations are therefore different and share only statistical properties (see Figs. 2 and 3). Discussions of the concept of prediction of time series can be found in texts by Whittle (1963), Robinson (1964*b*), Hannan (1970), and Granger and Newbold (1977). For the present purposes the important point is that while past observations may provide useful predictive information, for a random process there is nevertheless always some uncertainty or error in the predictions, even in the limit that the available data extend infinitely into the past. A case of particular importance is that in which past data provide no information about present or future values. (This must be made precise, because observations of the past provide some statistical information no matter how random the process: Because the process is stationary, the mean value derived from past data is the best prediction for X_n). In such cases there is no deterministic element, so the process can be called *purely random*. As with individual random variables there are three degrees of lack of determinism which it is crucial to distinguish.

The first is *independence*. A process is *independently distributed* (id) if all of the random variables are independent of each other. Then the past provides no information about the present. There are four equivalent conditions which are necessary and sufficient for $X_1, X_2, \ldots, X_N$ to be independent, i.e., that the process X be independently distributed (Parzen 1962):

1) *In terms of probability distributions*: for all real numbers $x_1, x_2, \ldots, x_M$,

$$P_{X_1, X_2, \ldots, X_M}(x_1, x_2, \ldots, x_M) = P_{X_1}(x_1)\, P_{X_2}(x_2) \ldots P_{X_M}(x_M). \quad (9)$$

2) *In terms of cumulative distribution functions*: for all real numbers $x_1, x_2, \ldots, x_M$,

$$F_{X_1, X_2, \ldots, X_M}(x_1, x_2, \ldots, x_M) = F_{X_1}(x_1)\, F_{X_2}(x_2) \ldots F_{X_M}(x_m). \quad (10)$$

3) *In terms of characteristic functions*: for all real numbers $u_1, u_2, \ldots, u_M$

$$\phi_{X_1, X_2, \ldots, X_M}(u_1, u_2, \ldots, u_M) = \phi_{X_1}(u_1)\phi_{X_2}(u_2) \ldots \phi_{X_M}(u_M). \quad (11)$$

4) *In terms of expectations*: for all functions

$$g_1, g_2, \ldots, g_M \; \langle g_1(X_1)\, g_2(X_2) \ldots g_M(X_M) \rangle = \langle g_1(X_1) \rangle \langle g_2(X_2) \rangle \ldots \langle g_M(X_M) \rangle, \quad (12)$$

provided all of the expectations indicated in this equation exist. These relationships must hold for $M = 2, 3, \ldots, N$. If, in addition, the X_i all have the same individual distributions, then X is said to be *identically and independently distributed* (iid). Independence is the strongest form of lack of relation and absence of predictability. The term *purely random* will be reserved for independently distributed processes.

A second and weaker description of a process is that it is *uncorrelated*. For a process with zero mean value, this means that the autocorrelation function vanishes for all except zero lag, that is,

$$\rho(X_n, X_m) \equiv \langle X_n X_m \rangle = \sigma^2 \delta_{n,m} \quad (13)$$

($\delta_{n,m}$ is the Kronecker delta, which vanishes if $n \neq m$, and is unity for $n = m$; $\sigma^2 = \langle X_n{}^2 \rangle$). Since $\langle X_n X_m \rangle$ is zero if X_n and X_m are independent of each other and can be nonzero otherwise, the autocorrelation function contains some information about dependence. Its vanishing implies a degree of lack of mutual dependence, but, as we shall see, not total absence.

The third description of a process involves the martingale difference property (see § II*a*). One says that a process X has the MDP if each X_n has the MDP with respect to the previous X_i, $i < n$. (Alternatively, the MDP could be with respect to all X_i, $i \neq n$.) This property is as fundamental to nonlinear estimation theory (Segall 1976) as are the concepts of uncorrelation and white noise to linear estimation. However, it does not seem to have been used very much in applied time series analysis. Although the results given below in § VI*a* are not very encouraging for the specific problem considered there, further development of this very easily applied technique is to be encouraged.

A process which is neither purely deterministic nor purely random could be called partially random. However, we will reserve this term for the case where the process has no deterministic component (in the sense made precise in § IV*a* below), but is also not uncor-

related; that is, loosely speaking, where it is a random process with some correlation present.

We will deal almost exclusively with stationary processes. Most discussions of stationary random processes assume that the mean value of all processes is zero because if it is not, the constant mean can be subtracted. If

$$X'_n = X_n - \langle X_n \rangle, \tag{14}$$

the new process X' has zero mean. However, this will not be done because there are cases where the positive definite nature of a signal is crucial (e.g., the examples in Figs. 2 and 3). This matter will be discussed further in § VI*f*.

Figure 4 shows examples of four types of processes: deterministic, random, uncorrelated, and independently distributed. Note particularly the process depicted in Figure 4*c*, which is uncorrelated but not independently distributed. (This process will be examined in detail in § IV*b*.) Another example of an uncorrelated but dependent process can be constructed as follows: Let X_1 be any zero-mean random variable. Define $X_2 = s_2 X_1$, where s_2 is randomly $+1$ or -1 with equal probability ($p = 1/2$). In general let $X_n = s_n X_1$, where the s_n are defined similarly to s_2, but are independent of each other and of s_2. It is easy to show that $\langle X_n X_m \rangle = 0$ for $m \neq n$, because $P_2(X_n, X_m)$ is an even function of at least one of its arguments. But the X_n are most definitely not independent, as $|X_n| = |X_1|$ for all $n > 1$. On the other hand, it is straightforward to show that if a process is independently distributed, then it is uncorrelated. Most data arise from a process which has a random aspect to it but is neither uncorrelated nor independently distributed; such is called a partially random process. In general a process can contain both deterministic and random components. Indeed, it can be shown that any stationary process[2] contains only these two components, and the separation between them can be written in a surprisingly simple and explicit form. This separation, called the *Wold decomposition*, will be discussed in detail in § IV.

It may seem strange, especially to the reader unfamiliar with the econometric approach to time series analysis (Wold 1964), that so much emphasis is put on prediction. But the relationship between prediction and statistical description is clear: a good prediction of the values of a process depends on good knowledge of its statistical properties. It will be seen that the concept of prediction must be extended to include the use of future data (i.e., estimation of X_n based on X_{n+1}, X_{n+2}, ...) as well as past data. That is, one pretends that X_n is unknown and tries to estimate or predict its value based on knowledge of the neighboring values $X_{n\pm1}, X_{n\pm2}, \ldots$. This approach leads to the concept of a two-sided (acausal) prediction-error filter, which forms the basis of the technique to be described in § IV for the extraction of information from time series data.

The ability to know when two random processes, say X and Y, are really the same is important. This does not mean that specific realizations of the processes are equal point-by-point (i.e., $X_n = Y_n$ for all n) because even different realizations of the same random process are not equal point by point. What is meant is that the probabilistic rules and sampling for X and Y are the same. Specifically, the joint probability functions listed in equations (4)–(6) must be identical.

b) *White Noise*; *Independently Distributed Noise*

Of special importance is the class of random processes R which satisfy all three of the following conditions: (1) $\langle R_n \rangle = 0$ (zero mean value), (2) $\langle R_n^2 \rangle = \sigma^2 < \infty$ (finite variance), and (3) $\langle R_n R_m \rangle = 0$ for $m \neq n$ (uncorrelated). Such a process is called *white noise*. Nothing is said in this definition about the probability distribution of R. There are many different kinds of white noise, according to the probability distribution. Gaussian, or normally distributed noise is very common, because of the fact expressed in the Central Limit Theorem.[3] It is also not necessarily true that the R_n be independently distributed, i.e., that R_n be statistically independent of R_m for $n \neq m$. White noise may be *independently distributed noise* or just *uncorrelated noise*. Both are "white" because the power spectrum of an uncorrelated process (and therefore of any independently distributioned process) is constant with frequency. Figure 5 and Figure 4*c* and 4*d* are examples of white noise with various distributions. Note further that only the second moment of R has been specified. The third and higher moments $\langle R_n R_m R_l \rangle$, etc., are not determined, although they are not completely arbitrary either, as they must conform to conditions (1)–(3) above.

c) *The Moving Average* (*MA*) *Model*

A *model* of a random process is an explicit mathematical description which is usually an attempt to describe a physical process in simple terms. It often involves a relatively small number of parameters, the values of which are to be determined by some procedure using the observed time series data (i.e., one or more realizations of the process). An extremely useful model is the *moving average*[4] (MA). A MA is a process

[2] A stationary process is one whose statistical properties do not depend on time. A strictly stationary process means that all of the joint probability distributions are invariant to a translation of time. There are other kinds of stationarity that are less restrictive, but we will not need to distinguish between them.

[3] The sum of independent random variables with any distributions tends to be normally distributed as the number of variables increases (Claerbout 1976, pp. 83–87).

[4] Unfortunately this term is also sometimes used for the procedure of smoothing data with a running mean, formally similar to the summation involved in the MA.

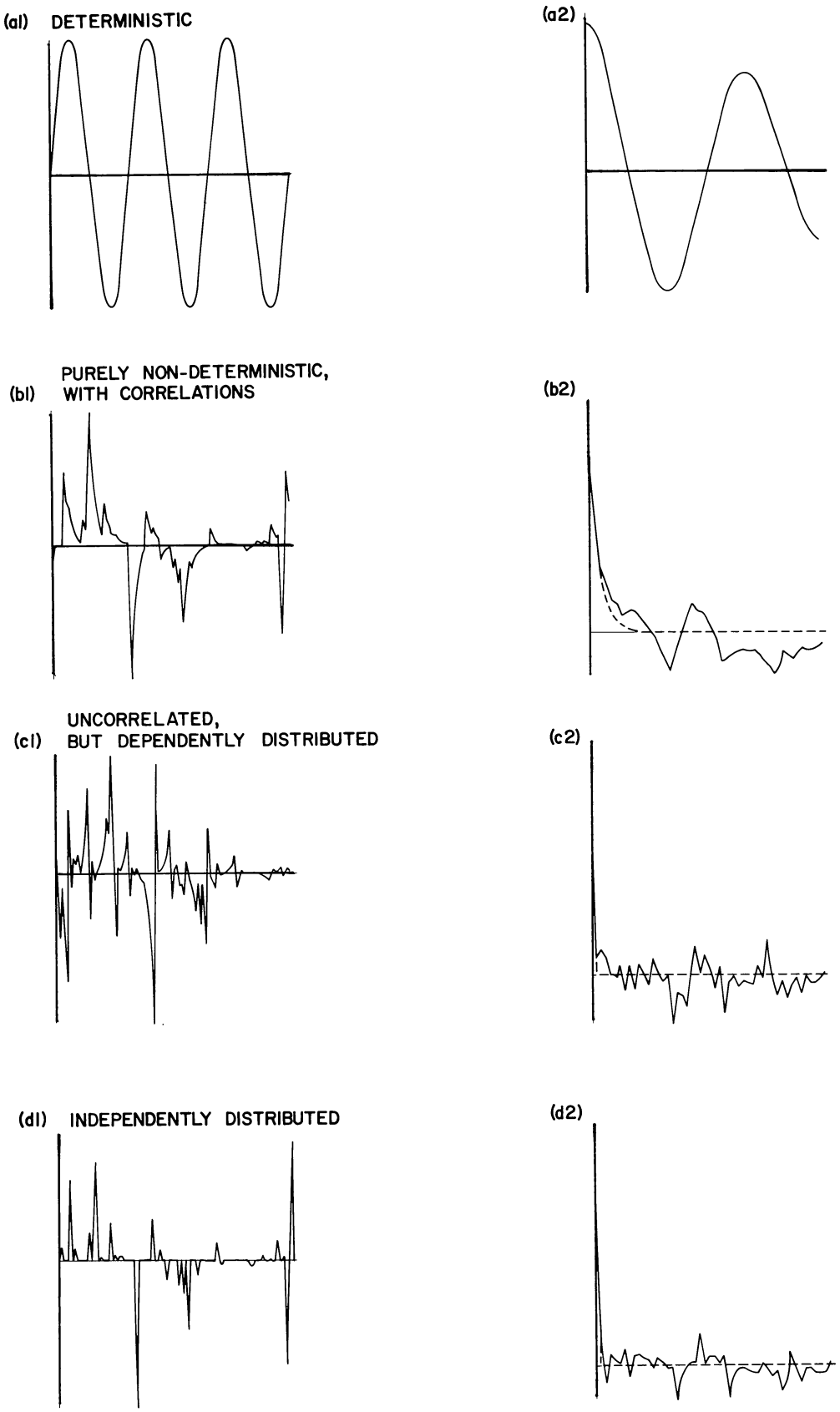

FIG. 4.—Time series produced by four different types of processes (*left*) and the corresponding autocorrelations (*right*). The dashed line is the theoretical autocorrelation, and the solid line is the estimate from the realization shown. The processes are: (*a*) a sine wave, (*b*) a moving average, (*c*) a moving average with the uncorrelated pulse shape shown in Fig. 17, and (*d*) independently distributed noise with a highly nonnormal distribution. (The autocorrelation of the sine wave in part (*a*) is damped because a finite realization was used to compute it.)

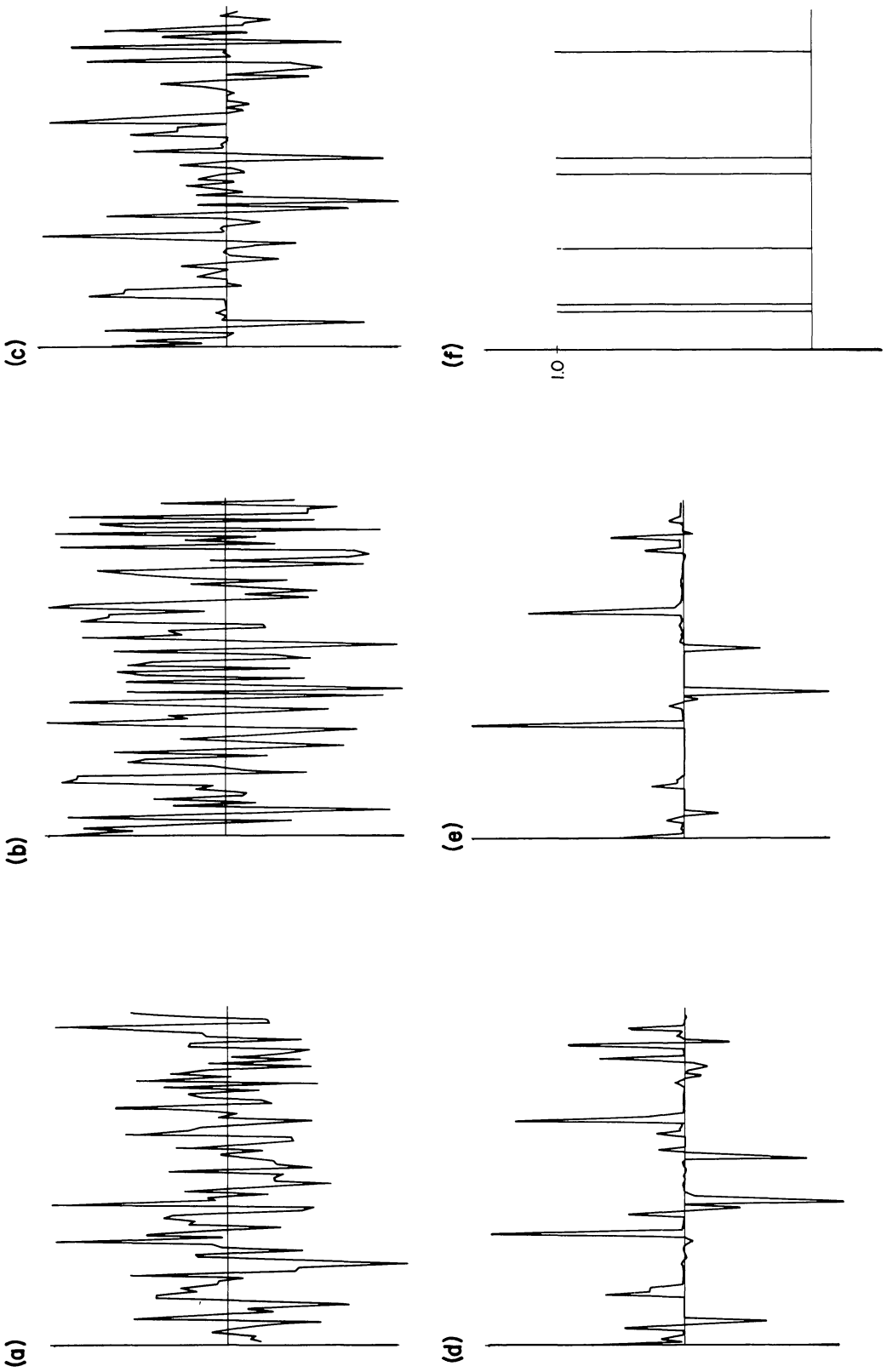

FIG. 5.—White noise with six different distributions (U^n = uniform raised to the nth power): (*a*) Gaussian, (*b*) U^1, (*c*) U^3, (*d*) U^9, (*e*) U^{19}, and (*f*) Poisson noise with constant amplitude.

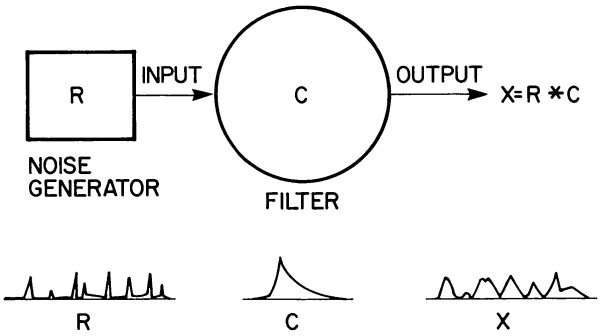

FIG. 6.—The moving average (MA) process depicted in terms of noise passed through a filter. The noise process shown is positive only, and the filter is roughly exponential in shape.

in the form $\Sigma_k C_k R_{n-k}$, where R is a white noise process and the C_k are constants. This summation is called a *convolution* (§ III*c*) and will be abbreviated C*R. The array of constants $C=\{C_k\}$ is called a *filter* or *linear system*. The reason for this terminology is that the above expression describes the output of an electrical filter into which is put a random sequence R of impulses (noise). That is, C_k regarded as a function of discrete time k describes the shape of a pulse that would result from an impulsive or delta-function input; C_k is the *impulse response* of the filter. This is easily seen by letting R_n be set equal to a delta function at $n=n'$ (i.e., $R_n=\delta_{n,n'}$), which then yields $X_n=C_{n-n'}$—that is the pulse $\{C_i\}$ with its origin, $i=0$, shifted to time n'. It is easily seen that if there are several or many nonzero values of R_n, each one produces a pulse at time n, of amplitude R_n. The net result is a sequence of overlapping pulses. The interpretation of the MA as filtered noise is illustrated in Figure 6. The time series in Figures 2 and 3 are also MAs. The closely related *shot noise process* will be discussed below, in § II*h*.

In most discussions of the MA the restriction is made that $C_n=0$ for $n<0$. This condition is called *causality*, and such a filter is said to be *causal* because a nonzero value at a negative time would correspond to a response of the filter at a time prior to the input. (The point $n=0$ will be called the *origin of time* for the pulse.) In some contexts this acausality would be unphysical, and it is convenient to restrict filters to respond only at and after the input; i.e., the filter can possess a memory but not premonition. However, for a number of reasons it is frequently useful or even necessary to relax this restriction. One reason is that it is often convenient to identify the origin of time for a pulse with a point near the peak rather than with the time of the cause of the pulse. For time series in which the independent variable is not time, the concept of causality is obviously of limited value. There is no Arrow of Space, or Arrow of Wavelength, as there is an Arrow of Time. Other reasons for dispensing with causality will be mentioned as they arise below. For the present, it should be simply noted that a filter is a set of numbers $\{C_n\}$ where n may take on negative as well as positive values. In practical computations, of course, n takes on a finite number of values, say $-q$, $-q+1$, $-q+2,\ldots,-2,-1,0,1,2,\ldots,p-1,p$. The case $q=0$ is the conventional one-sided or causal pulse and corresponds to a MA process of order p, abbreviated MA(p). The general case will be called a two-sided MA of order p, q, or MA(p, q).

An interpretation of the MA of interest in the economic applications (Wold 1964) is that the pulses represent the reaction or response of some system to news or information which arrives in discrete impulses. The effect of the news persists for some time (memory) but eventually dies out. This suggests a condition that the C_n get smaller as n gets large. In addition, it is convenient to allow the mean value of the input process R to be nonzero. For example, in some cases the pulse amplitudes must be positive because of their physical significance, as when the pulses are outbursts of radiation. If the mean value of the input is positive and the pulse shape has a positive "area" or total strength, the mean of the output is also positive, since $\langle X\rangle=\langle R*C\rangle=\langle R\rangle(\Sigma_k C_k)$.

The above statements are summarized in the following definition:

A *moving average* (MA) is a process X which can be written in the form:

$$X_n=\sum_{i=-\infty}^{\infty} C_i R_{n-i} \quad (X=C*R), \tag{15}$$

where R is an *uncorrelated* white noise process, possibly with nonzero mean:

$$\langle(R_n-\bar{R})(R_m-\bar{R})\rangle=\sigma^2\delta_{n,m} \quad (\bar{R}=\langle R_n\rangle) \tag{16}$$

and the C_i are constants satisfying $\Sigma_{-\infty}^{\infty}C_i^2<\infty$ (called *stability* of the filter C). If the C_i are zero for all negative (positive) values of i this is a *causal* (*purely acausal*) moving average. If neither is true, it is called a two-sided, or

acausal MA. A MA is said to be of order (p, q) if the range of i for which C_i is nonzero is from $-q$ to p.

The stability condition assures that the pulse dies out at infinity, and is written in the form given because ΣC_i^2 is the total energy output of an electrical filter if the input R represents the amplitude of the electric field at the input of the filter. The range of i may be finite or infinite. A finite MA is obviously stable.

It is important to note that R is random and C, if considered as a time series itself, is deterministic. That is, the process X has its random and its predictable aspects explicitly separated in the MA representation. Since R represents the new information arriving at the input of the system, it is called the *innovation*. We will be particularly interested in the class of MAs in which R is independently distributed but it should be remembered that the definition requires only that R be uncorrelated. Sometimes the terms "MA process" and "MA model" are used nearly interchangeably, but this is a loose usage. A MA process exactly satisfies the definition given above. A MA model is a representation or model which can be used to attempt a description of any process, whether or not it is actually a MA. For example, one can use a low-order MA model to approximate a process which is a higher-order (or infinite) MA or not a MA at all. The pulse shape $\{C_i\}$ is also assumed to be constant (independent of time, n). This will be seen below (in § IV*a*) to be less restrictive than it seems at first. A final point concerns normalization. If the switch $C \to \alpha C$, $R \to \alpha^{-1} R$ is made, then X obviously remains unchanged. Hence, in comparing different moving averages, it is convenient to remove this ambiguity by specifying in some sense the "size" of either R or C. Several possible choices are:

i) $C_0 = 1$, ii) $\sigma_R^2 = \langle R_n^2 \rangle = 1$,

iii) $\sum_i C_i^2 = 1$, iv) $\sum_i C_i = 1$,

v) $\sum_i |C_i| = 1$ vi) $\max_i C_i = 1$,

vii) $\max_i |C_i| = 1$.

For causal filters the conventional choice is (i). However, for acausal filters this choice would render the size of C dependent on the location of the time origin, which is to some extent arbitrary. (We will see another reason why this is a poor choice in § IV*e*.) The other six choices make the size of C invariant to a shift of the origin of time. The best choice of normalization seems to depend on the particular context.

To summarize: the moving average represents the deterministic part of a process with a constant filter, C, and the random part with an uncorrelated noise process, R. The process is the convolution of C with R, and can be viewed as a random sequence of pulses.

d) The Autoregressive (AR) Model

The MA model expresses the correlations in a process X in terms of memory, in the sense that the filter C remembers, for a while at least, the previous inputs R_i. There is another way of expressing such memory; that is, the process remembers its own behavior at previous times, or X_n remembers, or can be partially represented in terms of $X_{n-1}, X_{n-2}, \ldots$. If it is assumed that this representation involves a linear relationship, the memory can be represented by an expression of the form $B_1 X_{n-1} + B_2 X_{n-2} + B_3 X_{n-3} + \ldots$. This suggests writing

$$X_n = R_n + \sum_{k=1}^{\infty} B_k X_{n-k}, \tag{17}$$

where R_n is a random noise process just as before, and the B_k's are constant coefficients. The first term on the right-hand side of this equation represents the immediate response of the system to the random input, while the others are the memory. The conventional notation is to write $A_k = -B_k$, so that equation (17) becomes (with $A_0 = 1$)

$$R_n = \sum_{k=0}^{\infty} A_k X_{n-k} \tag{18}$$

or $R = A * X$. If this sum is finite, say from o to p, the process is called a (one-sided) AR process of order p, or AR(p). Note the symmetry of this relation with that for the MA (eq. [15]), namely $X = C * R$. The AR is the inverse of the MA in the sense that the filters C and A are convolutional inverses of each other. By analogy with the acausal or two-sided MA, the sum in the last equation may be extended to negative k; this gives the two-sided AR

$$R_n = \sum_{k=-\infty}^{\infty} A_k X_{n-k}. \tag{19}$$

The concept of a process's memory of its own future may seem unusual, but we are dealing with post–real–time data analysis or with cases in which the independent variable is not time, so that causality is not relevant. Also, this extension is necessary for consistency with the two-sided MA in equation (15). The name *autoregressive* arises because the expression just above equation (17) is in the form of a regression of X_n on itself evaluated at different times, so that equation (17) is a self- or autoregression.

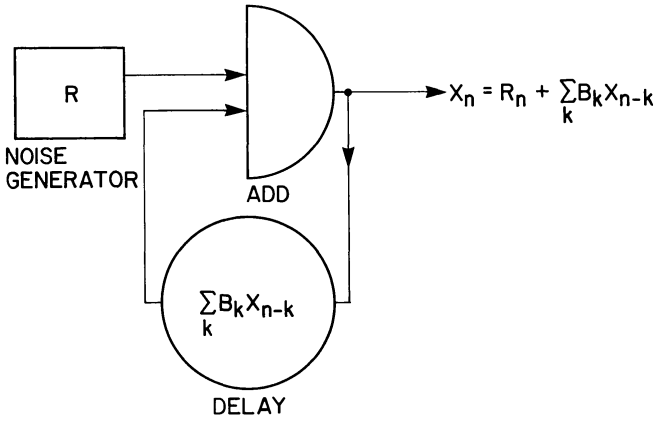

FIG. 7.—A circuit representing the autoregressive (AR) process. The signal is added to a delayed version of itself.

A schematic electric circuit representation of the AR process is shown in Figure 7. This circuit assumes a causal model, because there is no physical circuit that can generate future values. The discussion of normalization given above for MA models applies as well to AR models. Conventionally A_0 is set equal to 1; this will be done for some examples (such as the one to follow), but not generally.

An *autoregressive* (AR) *process* is one which can be written

$$A_0 X_n = R_n - \sum_{i \neq 0} A_i X_{n-i}, \tag{20}$$

or $R = A * X$, where R is an uncorrelated white noise process (as in the definition of the MA) and the A_i are constants satisfying $\Sigma i\ A_i^2 < \infty$(stability of A). The autoregressive filter A is purely causal, purely acausal, or two-sided depending on whether A_i is nonzero for only $i \geq 0$, for only $i \leq 0$, or for both $i \geq 0$ and $i \leq 0$. An AR is of order (p, q) if the range of i is from $-q$ to p.

An example of a second-order AR process is shown in Figure 8. Note that it has a sinusoidal appearance (and would probably be called quasi-periodic) even though it has no harmonic component nor any deterministic component. Figure 9 gives further examples of AR processes with quasi-harmonic appearance.

Actual physical random processes can often be well represented by an AR model with a small number of parameters A_i. Equation (20) is a difference equation which is the discrete version of the differential equation which describes the dynamics of the system (i.e., the equation of motion). Thus, the AR parameters can be interpreted as the coefficients of the linear differential equation of the system. The moving average pulse is the impulse response of this differential equation.

In fact, AR models can generally be rewritten in the form of moving averages. As an example, consider the simplest nontrivial AR process, namely the one-parameter process defined by:

$$X_n = R_n + \alpha X_{n-1}. \tag{21}$$

This corresponds to the AR filter $(1, -\alpha)$. Recursive substitution of the left-hand side of equation (21) into the right-hand side gives an explicit solution in the form of an infinite MA:

$$X_n = \sum_{k=0}^{\infty} \alpha^k R_{n-k}. \tag{22}$$

Thus, an input impulse at time n^*, of amplitude R_{n^*}, gives rise to the output pulse $\ldots 0, 0, 1, \alpha, \alpha^2, \alpha^3, \ldots$(multiplied by R_{n^*}). For $|\alpha| < 1$ this is an exponentially decaying pulse:

$$C_n = \begin{cases} 0, & n < n^* \\ \exp[(n - n^*) \ln \alpha], & n > n^*. \end{cases} \tag{23}$$

Note that we have converted this one-parameter AR process into an infinite but stable MA ($C_n \to 0$ quickly enough that the sum $\Sigma_{n=0}^{\infty} C_n^2$ converges). If $|\alpha| > 1$ the pulse given above is not stable, and further $C_n \to \infty$ exponentially as $n \to \infty$. To avoid this difficulty, let $n \to n+1$ and rewrite equation (21) as

$$X_n = \alpha^{-1} X_{n+1} - \alpha^{-1} R_{n+1}. \tag{24}$$

Recursive substitution with this equation leads to

$$X_n = -\sum_{k=1}^{\infty} \alpha^{-k} R_{n+k}. \tag{25}$$

The effect of a single impulse at time n^* is thus a growing exponential pulse of amplitude $-\alpha^{-1} R_{n^*}$ and growth constant α, terminating at time $n^* - 1$ (see Fig. 10). Thus, equation (21) has a stable solution for any α, *unless* $|\alpha| = 1$; in one case the pulse extends forward in time (i.e., is purely causal) and in the other it extends backward (is purely acausal).

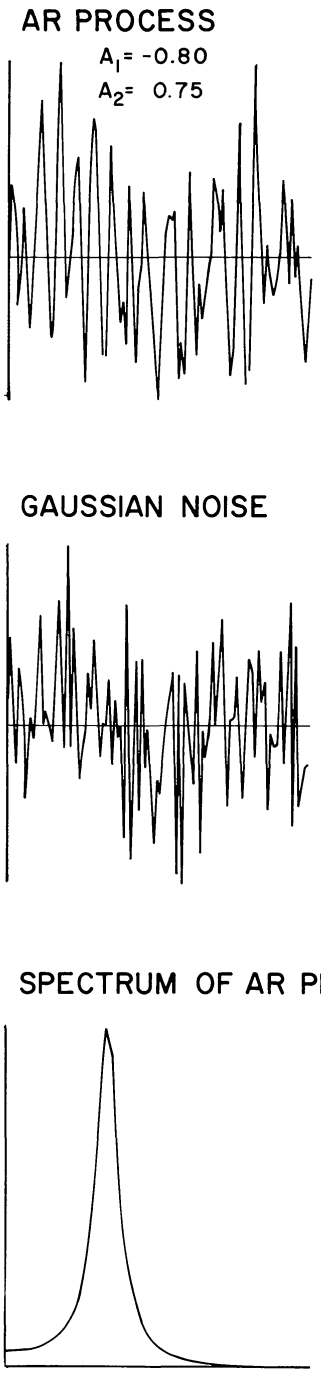

FIG. 8.—A realization of the second-order AR process $X_n = R_n + 0.8X_{n-1} - 0.75X_{n-2}$ (*top*). The middle curve is the realization of the Gaussian noise which drove the AR process. Since X is purely nondeterministic, the spectrum (*bottom*) is continuous, but it has a narrow peak corresponding to the quasi-sinusoidal appearance of the process.

e) *The Relationship Between the AR and MA Models*

In the example given in the previous section a simple AR model was converted into a MA. This is a general feature: *any AR model can be converted into a MA model and vice versa*. In the standard treatments of this subject special restrictions must be placed on the models for this to be true, and some otherwise well-behaved AR models, for example, are not convertible into (stable) MA's. But with the generalization to two-sided representations, convertibility holds without restriction. The fundamental reason for this is evident from the example in equations (21)–(25): $|\alpha| > 1$ led to a causal MA representation that diverged, and the restriction $|\alpha| < 1$ is usually imposed. But if two-sided representations are allowed, this restriction is unnecessary because there is a convergent acausal representation. The MA corresponding to an arbitrary AR process is usually two-sided. Unfortunately the direct approach of recursive substitution of the AR representation into itself is extremely awkward in the general case, because at each step there are choices to be made concerning the form of the substitution which have a complex dependence on the specific values of the AR parameters. However, the demonstration of how AR and MA models can be converted into each other, including the computation of the coefficients, is rendered simple by the introduction of Z-transforms, as will be shown in § III*f*.

To summarize, the MA representation emphasizes the memory of earlier values of the driving (input) process, while the AR emphasizes the memory of earlier values of the (output) process itself. Since the input determines the output, these representations are directly related and can be determined, one from the other.

f) *Autoregressive-Moving Average (ARMA) Models*

An obvious generalization is to allow the current value of the output, X_n, to depend explicitly on (i.e., to remember) values of *both* the output X and the input R at other times:

$$X_n = \sum_{k \neq 0} B_k X_{n-k} + \sum_k C_k R_{n-k}, \qquad (26)$$

or $A * X = C * R$, where A has the same relationship to the B_k as before. This is called a *mixed autoregressive-moving average model*, or an ARMA model. If the processes involved are finite and causal [e.g., AR(p) and MA(q)], the mixed process is denoted ARMA (p, q). (Generalization of this notation to the two-sided case is cumbersome and not necessary here.) Physically one can think of an ARMA process as representing a system, described by the AR parameters A, which is driven by an input which is itself a moving average process, rather than white noise. But as was indicated in the previous section, the distinction between system response as described by MA and AR models is merely a matter of interpretation. Hence, there is no rigid distinction between what portion of a process is AR and what part is MA. In fact, the AR part of an

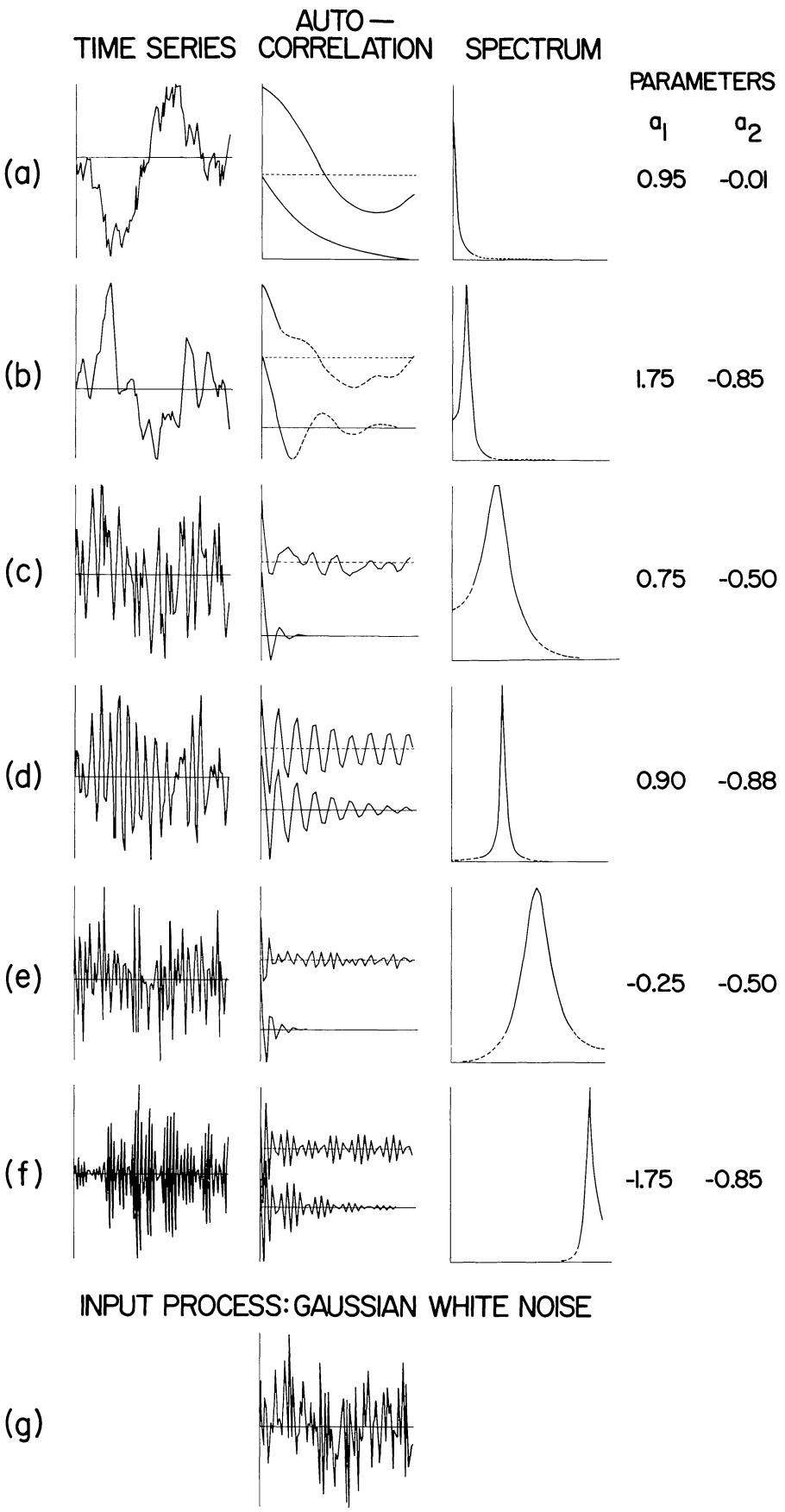

FIG. 9.—A series of AR processes of the form $X_n = R_n + a_1 X_{n-1} + a_2 X_{n-2}$, where R is independent Gaussian noise (g), and the values of a_1 and a_2 are shown at the right. The processes were chosen to exhibit various spectral peaks, but none has a deterministic harmonic component. The middle column shows the sample (*top*) and theoretical (*bottom*) autocorrelations for each process.

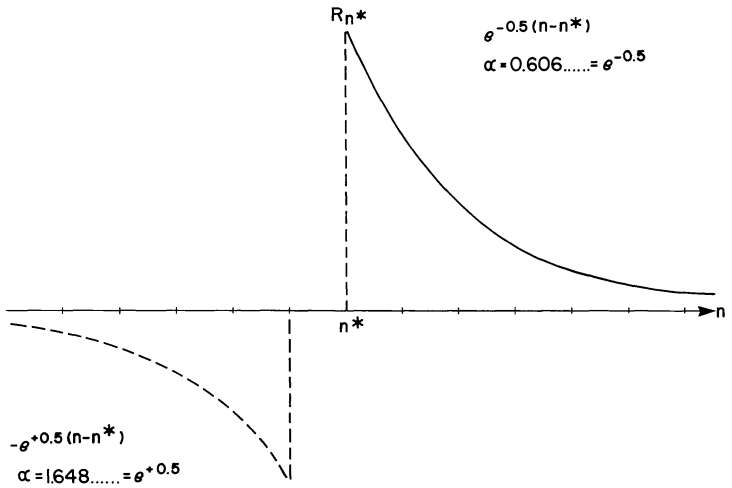

FIG. 10.—The exponential MA pulses inversely related to the first-order Ar process $(\mathbf{1}, -\alpha)$. *Solid line*: the causal pulse $(\mathbf{1}, \alpha, \alpha^2, \alpha^3, \ldots)$ with $1 > |\alpha| = 0.606$. *Dashed line*: the acausal pulse $(\ldots, -\alpha^{-3}, -\alpha^{-2}, \alpha^{-1}, \mathbf{0})$, with $1 < |\alpha| = 1.648$.

ARMA can be converted to a MA, yielding a pure MA. Similarly, an ARMA can also be converted to a pure AR. Furthermore, one could convert only part of the ARMA to MA (or AR), so that there is a great range of possible ARMA combinations to represent a given process.

It may be asked "What is the use of mixed representations at all, since they can all be converted to pure AR or MA?" The answer lies in a concept called *parsimony* of representation. The point is that some processes may be representable as an infinite-order AR or MA, but as a finite ARMA. The latter would then be a more compact or parsimonious representation. Parsimony can be of great importance in computing, where one is often searching for models involving the smallest number of parameters. But it should be stressed that parsimony is not necessarily of significance in the interpretation of the results of modeling. A good example is that given at the end of § II*d*, which has the most parsimonious representation as AR(1), but might well be most simply interpreted as MA(∞).

There are several discussions of the form of the autocorrelation functions and power spectra of low-order AR, MA, and ARMA processes which should be consulted by the reader interested in such functions (Box and Jenkins 1970; Stralkowski, Wu, and DeVor 1970, 1974).

g) AR Integrated MA (ARIMA) Models and Nonstationary Processes

The discussion so far has assumed that the process under discussion is stationary. This is an important restriction, for nonstationary processes do not have representations of the kind discussed up to this point. But a very special kind of nonstationarity can be incorporated in a simple modification of the AR, MA, or ARMA models. The general form is

$$A * (\nabla^d X) = C * R, \tag{27}$$

where ∇ represents the difference operator:

$$\nabla X_n = X_n - X_{n-1}, \tag{28}$$

and ∇^d stands for the dth difference operator, equivalent to operating with ∇ d times. If we let $W = \nabla^d X$ (so that W is an ARMA process) X can be obtained by integrating W d times. That is, $X = S^d W$, where S is the summation operator:

$$S(X_n) = \nabla^{-1} X_n = \sum_{i=-\infty}^{n} X_i. \tag{29}$$

Thus, X is said to be an autoregressive-integrated-moving average, or ARIMA, process.

Consider the simple case $d = 1$. While X is not stationary, its first difference is (Box and Jenkins 1970). The nonstationarity which this gives to X has the character of a floating mean value—the mean of the process is not constant with time but drifts. Similarly, a second-order ($d = 2$) ARIMA process is such that both the mean value and average slope wander as time goes on.

Finally, it is interesting to add a further generality in the form of a constant term in the equation:

$$A * (\nabla^d X) = C * R + D_0. \tag{30}$$

It can be seen that the meaning of the constant term D_o is to allow the process X to have a deterministic *trend* in the form of a polynomial of order d.

The ARMA and ARIMA representations can be quite useful in some specific applications. The current

discussion will center on the less complex AR and MA models for simplicity and because they seem to be sufficiently general for most astrophysical applications. The reader should consult Box and Jenkins (1970) for more details on ARMA and ARIMA models.

h) *The Shot Noise Model*

As already mentioned, the MA is closely related to the *shot noise model*, which is usually defined in continuous time as follows:

$$X(t)=\sum_i C(t-t_i), \tag{31}$$

where $C(t)$ is a given function of time (a continuous pulse shape) and the t_i are random points in time which are distributed according to the Poisson distribution. This process can be viewed as the output of a continuous linear system, with impulse response $C(t)$, resulting from an input consisting of a Poisson sequence of constant amplitude impulses

$$R(t)=\sum_i \delta(t-t_i). \tag{32}$$

The Poisson distribution results from randomly and independently placing the time points t_i. The probability of having k impulses in an interval Δt is

$$P_k(\Delta t)=\frac{e^{-\lambda\Delta t}(\lambda\Delta t)^k}{k!}, \tag{33}$$

where λ is a constant giving the mean rate of occurrence of the impulses, which here all have the same amplitude. If Δt is identified with the time interval in discrete time (see § II*a*) then equation (33) gives the probability distribution of pulse amplitudes, where k is to be identified with the amplitude. (The amplitudes are quantized in unit steps.) If time is sliced finely enough so that $\lambda\Delta t \ll 1$, then we have

$$P_k \approx \begin{cases} 1-\lambda\Delta t, & k=0 \text{ (no pulse)} \\ \lambda\Delta t, & k=1 \text{ (one unit amplitude pulse)} \\ 0, & k=2 \text{ (multiple pulses)} \end{cases}; \tag{34}$$

that is, most of the time a pulse does not occur, but occasionally a single pulse occurs, always with the same amplitude. It can be seen that the noise processes U^n, with large values of n, shown in Figure 5 and defined at the beginning of § VI, have approximately these properties (except that they are zero-mean processes and the amplitudes of the pulses are not always the same). Thus, an MA with pulse shape given by the discrete version of $C(t)$ and with the quantized probability distribution of the input R given by equation (33) (or in the limit $\lambda\Delta t \to 0$ by eq. [34]), with $k \to R$, is the discrete version of the shot noise model.

Some useful relations for the moving average, easily derived from the defining equations, are:

$$\langle X\rangle = \langle R\rangle\left(\sum_k C_k\right), \tag{35}$$

and

$$\sigma_X^2 = \langle (X-\langle X\rangle)^2\rangle = \sigma_R^2\left(\sum_k C_k^2\right). \tag{36}$$

These are somewhat different in form from the relations for the usual definition of the shot noise process. For example, if $\sigma_R^2=0$ in a moving average, pulses of uniform amplitude are occurring at every time, and X is constant ($\sigma_X^2=0$); this is not true for a Poisson distributed shot noise process where the variance of the amplitudes of the pulses is often taken to be zero. A related difference is that the concept *mean pulse rate* loses significance for an MA because it is automatically 1 per unit Δt. That is, pulses occur at every point of (discrete) time. The incidence of zero amplitude pulses is expressed in the distribution function of the innovation (as in eq. [34]) and is absorbed into the mean pulse amplitude.

For a good discussion of the shot noise model see Papoulis (1965). Terrell (Terrell and Olsen 1970, 1972; Terrell 1972) has applied this model, with exponential pulse shapes, to several astrophysical problems.

III. THE STRUCTURE OF PULSES

The separation of a process into a random part and a purely deterministic part, as exhibited in the moving average, is often of direct physical significance. The pulse may represent the unfolding of some process for which there is a physical theory. Knowledge of the pulse shape[5] may provide interesting numbers such as pulse width, rise and decay times, etc. The innovation, or random noise process R, represents the pulse amplitudes and contains information about pulse rates and the distribution of pulse amplitudes. To develop a feeling for the structure of pulses, this section discusses the representation of physical pulse shapes as filters, the algebra of filters, and a concept called the *phase character* (or sometimes *delay character*) of filters. These

[5]The terms *pulse shape*, *pulse*, (*moving average*) *filter*, *wavelet*, *impulse response*, *moving average representation*, and *moving average parameters* are all used in the literature to convey approximately the same meaning, and are interchangeable in many contexts. Here the term *impulse* will be reserved for a pulse, usually taken as the input to a filter, which is a delta function in time.

subjects are discussed extensively in various mathematical works (Robinson 1964*a*, 1967*a*, *b*: Treitel and Robinson 1966; Box and Jenkins 1970; Anderson 1971), which should be consulted for more details. The discussion here will be oriented toward the analysis and interpretation of astrophysical time series data and will emphasize two-sided filters, which have been neglected in much of the standard literature.

a) *The Discrete Representation of Pulse Shapes*

Suppose that a physical pulse is described by a continuous function of time, $C(t)$. An example would be the light curve produced by a nova or supernova. Let the values of C be specified (or sampled) at evenly spaced points in time, say $t_n = n\Delta t$, for some set of values of n; it is presumed that the points are close enough that the interesting structure in the pulse is resolved. Then the set of numbers or filter elements, $\{C_n\} \equiv \{C(t_n)\}$, is a discrete representation of the pulse shape $C(t)$.

i) *One-sided Pulses*

In many situations there is a moment before which C is identically zero. The classical example is the pulse which comes out of an electrical filter in response to an impulse at time t_o; in accordance with causality this output must be exactly zero at all previous times $t < t_o$. By identifying the origin of discrete time, $n=0$, with this moment, the filter elements need only be given explicitly for nonnegative indices, $n = 0, 1, 2, \ldots$. Such a filter is said to be *causal* or *one-sided*. The sum $\Sigma_{n=0}^{\infty} C_n^2$ can sometimes be associated with a physical quantity, such as the total energy in an electrical pulse; if so

$$\sum_{n=0}^{\infty} C_n^2 < \infty \tag{37}$$

must hold for any physical filter. This condition is called *stability* or *convergence*. In some cases, other stability conditions such as $\Sigma_{n=0}^{\infty} |C_n| < \infty$ are relevant (Robinson 1962). A filter which is both stable and causal is said to be *physically realizable*. We shall now see that some perfectly useful physical pulses are not causal.

ii) *Two-sided Pulses*

Consider the following scenario: a small signal grows with time, slowly at first, then more rapidly; reaching a peak, the signal begins to decay and eventually disappears. For example, take the specific form

$$C(t) = C_0 \begin{cases} e^{at}, & t < 0 \text{ (exponential growth)} \\ e^{-bt}, & t > 0 \text{ (exponential decay)} \end{cases} \tag{38}$$

or in discrete time:

$$C_n = C_0 \begin{cases} e^{an}, & n = \ldots, -3, -2, -1, 0 \\ e^{-bn}, & n = 0, 1, 2, 3, \ldots . \end{cases} \tag{39}$$

In this case it is not convenient to take the origin of time at the beginning of the pulse, which strictly speaking lies at $n = -\infty$. [Of course, it would always be possible to take the origin at some early time before which $C(t)$ is effectively zero, say to within the measurement accuracy. In the same sense almost all pulses can be taken to be of finite length.] A more important reason for considering noncausal filters is that, among causal filters, only the members of a very special class (called *minimum delay*, a term to be defined below) have stable, causal convolutional inverses. Since our methods for determining pulse shapes from time series data depend on first determining the inverse pulse shape, restriction to causal filters would imply the unnecessarily limiting restriction to minimum delay filters.

In many cases when a filter is written explicitly as an array of filter elements, such as $(\ldots, C_{-2}, C_{-1}, C_0, C_1, C_2, \ldots)$, the location of the origin of time is obvious (C_0 in this example). But in some cases it is not obvious from the indexing or from the context, and a boldface symbol will be used to locate the origin [e.g., $(\mathbf{1}, -a)$ denotes $C_0 = 1, C_1 = -a$]. Figure 11 illustrates the basic difference between one- and two-sided pulses.

b) *Z-Transforms*

We now introduce a powerful tool for the analysis of pulses, the Z-transform. It is a tremendous time saver in the manipulation of filters as well as in the proofs of certain relationships between filters. Consider a pulse or filter $C = \{C_n\}$, $n = -q, -q+1, \ldots, -2, -1, 0, 1, 2, \ldots, p-1, p$, containing $p+q+1$ elements. The Z-transform of C is defined as the following function of the dummy

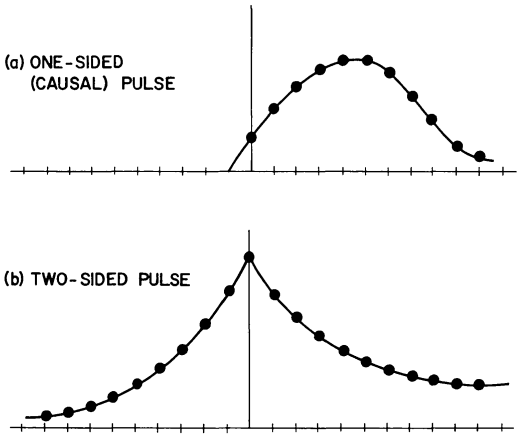

FIG. 11.—Schematic representation of the difference between (*a*) causal pulses and (*b*) acausal pulses.

complex variable z:

$$C(z)=\sum_{n=-q}^{p} C_n z^n. \tag{40}$$

This is simply a polynomial or power series in positive and negative powers of z. In the case p or $q=\infty$, we assume that the series converges on the complex plane within some annulus including the unit circle.

The coefficients determine the filter (and vice versa); that is, $C(z)$ determines the C_n and vice versa. The transform will sometimes be denoted with the operator Z thus: $C(z)=Z(C)$. The inverse transform will be denoted Z^{-1}, and can be thought of as the operation of identifying the coefficients in a series expansion of $C(z)$. The Z-transform has the following alternative interpretations:

1. A representation of the time behavior of pulses in which z represents the unit delay operator (and z^{-1} represents the unit advance operator).
2. A discrete analog of the Laplace transform: if $f(t)$ is replaced by $\Sigma_n f(t_n)\,\delta(t-t_n)$, where $t_n=n\,\Delta t$, then the Laplace transform of f becomes the Z-transform ($z=e^{-s}$, where s is the Laplace transform variable).
3. Similarly, a version of the discrete Fourier transform (DFT) with $z=e^{i\omega}$.
4. A generating function for the filter C.

The Z-transform maps from the time domain to a transform domain. The operations of shifting in time are denoted with the *unit delay operator*, D, and the *unit advance operator* A:

$$\left.\begin{aligned} D(X_n)&=X_{n-1}; \quad D^j(X_n)=X_{n-j};\\ A(X_n)&=X_{n+1}; \quad A^j(X_n)=X_{n+j}. \end{aligned}\right\} \tag{41}$$

In the transform domain D^j corresponds to multiplication by z^j and A^j corresponds to division by z^j. The definitions, theorems, and proofs involved in the use of the Z-transform closely parallel those for integral transformations (such as the Laplace and Fourier transforms) of continuous functions. The Z-transform will be demonstrated in applications in the rest of this paper. Further details can be found in various sources (e.g., Jury 1964; Gold and Rader 1969; Oppenheim and Schafer 1975; Rabiner and Gold 1975).

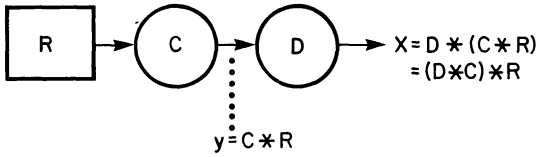

FIG. 12.—The convolution $C*D$ interpreted as filters C and D connected in series.

c) *Convolution*

Consider the effect of putting a signal R into a filter C and connecting the output, Y, into a second filter D. That is, C and D are placed in series (see Fig. 12). By definition:

$$Y_n=\sum_k C_k R_{n-k}, \tag{42}$$

so

$$\begin{aligned} X_n&=\sum_k D_k Y_{n-k}=\sum_k D_k \sum_l C_l R_{n-k-l}\\ &=\sum_k D_k \sum_m C_{m-k} R_{n-m}=\sum_m B_m R_{n-m}, \end{aligned} \tag{43}$$

where

$$B_m \equiv \sum_k D_k C_{m-k}, \tag{44}$$

which is easily shown to be the same as

$$B_m=\sum_k C_k D_{m-k}. \tag{45}$$

Thus, the action of two filters in succession (series) can be completely represented by a single filter, called the *convolution* of the two, written as

$$B=C*D. \tag{46}$$

It is readily verified that *the Z-transform of the convolution of two filters is the product of their Z-transforms*:

$$B(z)=C(z)\,D(z). \tag{47}$$

This is the most important reason for the utility of the Z-transform. Furthermore, convolution is commutative and associative:

$$A*B=B*A, \tag{48}$$

$$A*(B*C)=(A*B)*C. \tag{49}$$

It should be noted that the output of the MA is formally the convolution between the input noise process and the pulse shape, although the physical interpretation is somewhat different in this case (convolution of a process with a filter instead of two filters with each other).

d) *Factorization*

As will be demonstrated shortly, any finite filter with more than two nonzero elements can be broken down into the convolution of a number of shorter filters. In

particular, a filter of length $n+1$ can be written as the convolution of n filters of length 2. Such filters have two and only two successive elements nonzero and are called *couplets* or *dipoles*: (C_n, C_{n+1}). Since many of the important properties of pulses are invariant to a shift in time, it is convenient to take $n=0$, and denote the dipole as (C_0, C_1). This is acceptable if all pulses are shifted so that their first nonzero element is at $n=0$ (i.e., causality), but to allow factorization of two-sided filters acausal dipoles of the form(C_{-1}, C_0) must also be introduced. Figure 13 depicts causal and acausal dipoles and shows how convolutions generate longer filters.

Now consider the filter $\{C_n\}, n=-q,\ldots,p$, where q and p are nonnegative integers. (This is not the most general case, as the index set might contain only positive terms (e.g.,..., 0,0, C_2, C_3,0,0,...), but such cases can be handled with the same methods.) The function

$$P(z)=z^q C(z) \tag{50}$$

(where $C[z]$ is the Z-transform of $\{C_n\}$) is a polynomial of degree $p+q$, with nonnegative powers of z only. Hence by the fundamental theorem of algebra it can be written

$$P(z)=C_{-q}\prod_{i=1}^{p+q}\left(1-\frac{z}{z_o^i}\right), \tag{51}$$

where the z_o^i are the complex zeros of $P(z)$. With a little algebra it can be shown from this expression that

$$C(z)=\left[C_{-q}\prod_{i=p+1}^{p+q}(-z_o^i)^{-1}\right] \times\left[\prod_{i=1}^{p}\left(1-\frac{z}{z_o^i}\right)\right]\left[\prod_{i=p+1}^{p+q}\left(1-\frac{z_o^i}{z}\right)\right]. \tag{52}$$

With the definition

$$a_i=\begin{cases}(-z_o^i)^{-1}, & i=1,2,\ldots,p\\ -z_o^i, & i=p+1,p+2,\ldots,p+q,\end{cases} \tag{53}$$

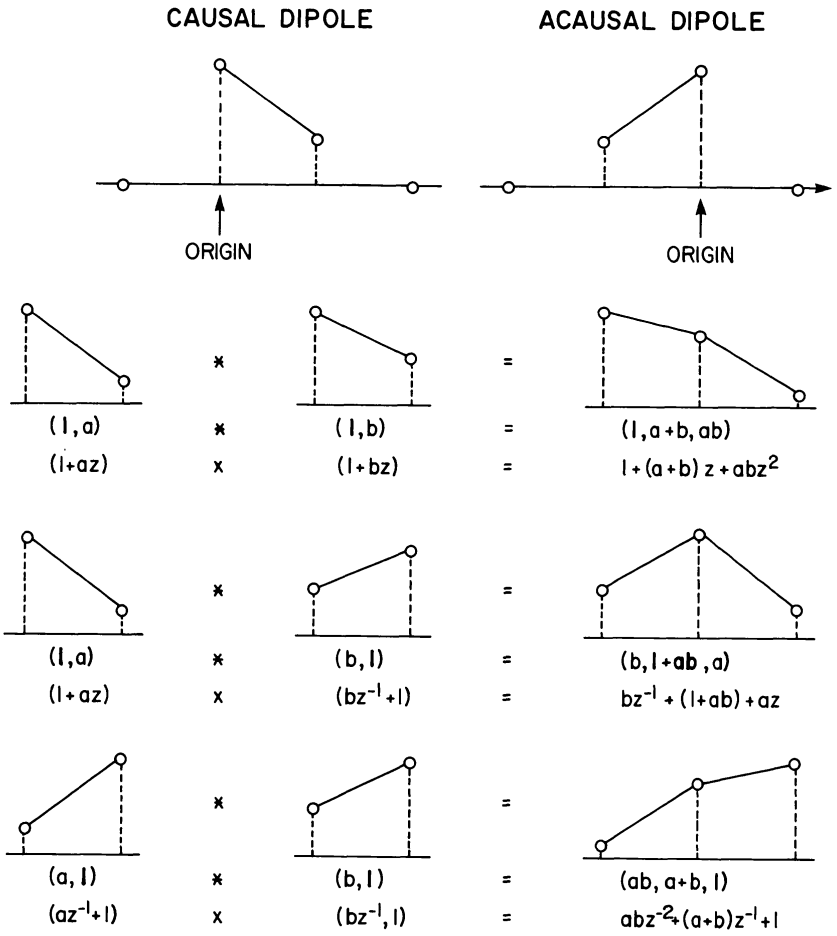

FIG. 13.—Graphical representation of causal and acausal dipoles (*top*) and their convolutions in various combinations. Shown with the filter convolution equations are the corresponding Z-transform relations.

the inverse Z-transform of this equation gives

$$C = K[(\mathbf{1}, a_1) * (\mathbf{1}, a_2) * \ldots * (\mathbf{1}, a_p)] * [(a_{p+1}, \mathbf{1}) * \ldots * (a_{p+q}, \mathbf{1})], \tag{54}$$

where K is the first factor in square brackets in equation (52). The first p dipole factors are causal and the last q are acausal. Since the ordering of the z_o^i has not yet been specified, there are many possible distinct factorizations of this form, depending on which z_o^i are assigned to the causal factors and which to the acausal factors. As will be shown in the next two sections, among the many choices possible for the origin of time in the original filter and for the assignment of the z_o^i, there is a single choice which has the property that each causal (acausal) dipole has a convergent causal (acausal) inverse. (The inverse is defined below in § III*f*.) It is obtained simply by making $|a_k| < 1$ for all k, which can be achieved unless $|z_o^k| = 1$ for some k. This can be considered as the unique factorization of the original filter C, although it really represents merely the simplest of many possible factorizations. If the original filter is causal, then $q = 0$ and the above analysis shows that there is only one factorization into causal dipoles; this is the "unique" factorization which is usually discussed.

e) *Delay (or Phase) Character*

In electrical engineering the frequency response of a filter describes the degree to which an AC signal at a given frequency will be attenuated on passing through the filter. Another effect of a filter is to cause frequency-dependent phase shifts of signals. For the present applications, rather than view these effects in the frequency domain, it is more convenient to use the time domain.

Consider first a causal dipole $(\mathbf{C_0}, C_1)$ as in § III*d*. This filter is defined to be *minimum delay* (or *minimum phase*) if $|C_1| < |C_0|$; it is *maximum delay* (or *maximum phase*) if $|C_1| > |C_0|$. These names are derived from the way in which energy is delayed at the output of the filter, as will be detailed below. Since delay properties are not affected by an overall shift in time, an acausal dipole $(C_{-1}, \mathbf{C_0})$ is minimum delay if $|C_0| < |C_{-1}|$ and maximum delay with the opposite inequality. The case $|C_0| = |C_{\pm 1}|$ is somewhat singular in that the inverse does not converge (see below); hence this case must be handled separately.

Now consider a filter $C = \{C_i\}$ of arbitrary length, say $n+1$. Again because of time-shift invariance only the causal case need be considered. That is, if the filter is not causal, its causal equivalent should be used. The causal equivalent of a filter is simply the filter shifted so as to bring its first nonzero element (which may not exist if the filter is infinite) to $i = 0$. From the previous section we know that there is a unique factorization into n causal dipoles. Each dipole is either minimum delay or maximum delay. If all the dipole factors are the former, the entire pulse is said to be a *minimum delay pulse*; if the factors are all maximum delay, so is the entire pulse. If there are some of each, we have a *mixed delay pulse*. Thus, the delay character of the pulse is specified by the delay character of the dipole factors of its causal equivalent. The physical meaning of these concepts is as follows. Introduce the quantity

$$P_i = \sum_{k=-\infty}^{i} C_k^2; \tag{55}$$

this is the integrated energy—the energy which has come out of the filter up to and including time i—due to a delta function input at time 0 (for electromagnetic signals energy = [amplitude]^2). This function rises from zero for $i < 0$ (since by assumption $C_i = 0$ for $i < 0$), monotonically, to its final maximum at $i = n+1$, and thereafter remains constant at a value $P_\infty = P_{n+1} = \sum_{i=-\infty}^{\infty} C_i^2$, which corresponds to the total energy output of the filter. Corresponding to filter C there is a family of filters (all of length $n+1$) which is generated by reversing all possible subsets of the dipole factors of C. (The reverse of $[\mathbf{C_0}, C_1]$ is $[C_1^*, \mathbf{C_0^*}]$, where the superscript $*$ represents complex conjugation of the possibly complex filter elements. Correspondingly, the reverse of any filter is obtained by reflection about the origin of time and by complex conjugation of all of the filter elements. The time reverse of any array $X = \{X_n\}$ will be denoted $\tilde{X} = \{X_{-n}^*\}$.) Since there are n such factors, this family has 2^n members, including the original filter itself, although they are not all necessarily distinct. It will be evident from the discussion in § III*g* that the power spectra and autocorrelations of the members of the family are identical. Hence a causal filter of arbitrary delay properties can be converted into a minimum delay causal filter (by making all dipole factors minimum delay and shifting the total filter with the delay operator to make it causal) without altering its autocorrelation function. The family of filters mentioned above may be defined as the set of pulses of length $n+1$ with the same autocorrelation and spectrum as C. Further, the total energy P_∞ of all these filters is the same, so the partial energy curves of these filters all begin and end at the same points (see Fig. 14). Between these points the curves are quite different and even cross each other. But it can be shown that there is one curve which everywhere lies above all the others—and it corresponds to the single minimum-delay member of the family of pulses. That is, the energy output of the minimum delay filter is delayed as little as possible, among all possible filters with the same spectrum, in that at each moment of time the integrated energy is maximum. Similarly the unique maximum delay pulse has a partial energy out-

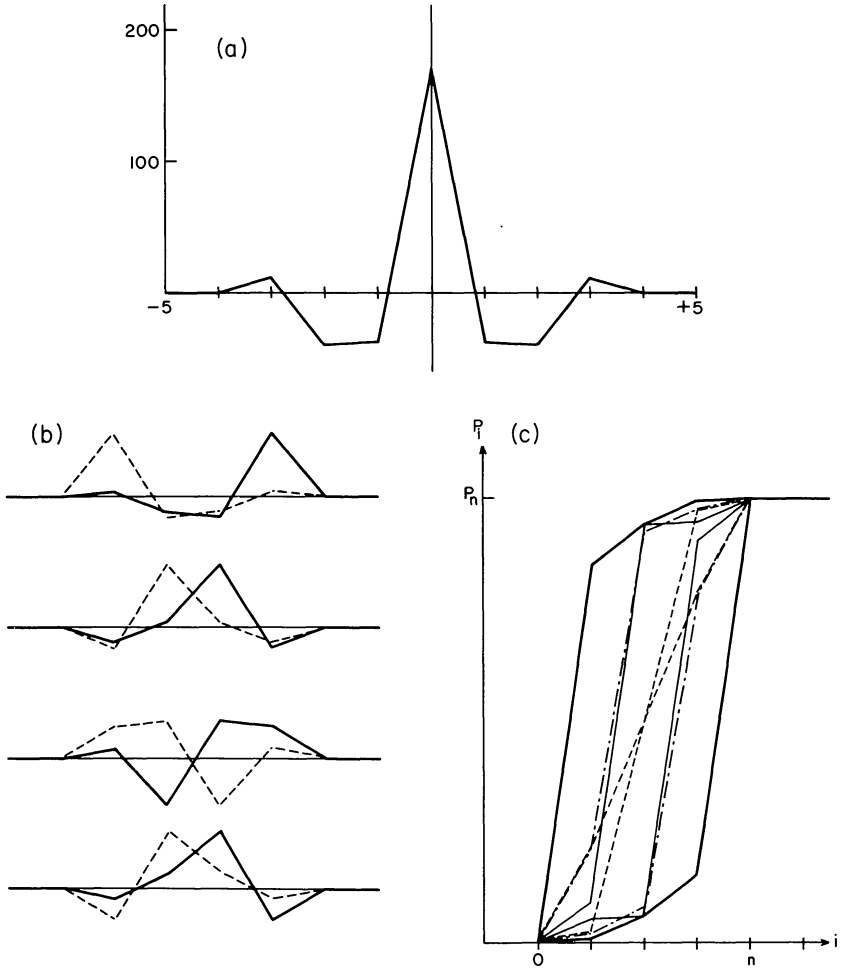

FIG. 14.—The concepts of minimum and maximum delay. (*a*) A short autocorrelation function. (*b*) The set of eight pulses which share this autocorrelation. (*c*) a plot of the eight corresponding partial energy curves: the uppermost curve corresponds to the minimum delay pulse (*dashed line*, topmost part of [*b*]) and the lowest curve corresponds to the maximum delay pulse (*solid line*, topmost part of [*b*]).

put which lies below all the other curves and corresponds to delaying the energy as much as possible.

Minimum delay pulses begin suddenly and decline slowly. In fact the minimum delay pulse rises as sharply and declines as gradually as possible, consistent with the given autocorrelation. The maximum delay pulse is the time reverse of the minimum delay and has the reverse of these properties. Further discussions of the physical and mathematical meaning of minimum delay are in the geophysical literature (Robinson 1962, 1963, 1964*a*, 1966, 1967*b*; Smylie, Clarke, and Ulrych 1973; Berkhout 1973; Schoenberger 1974).

f) *Inverse Filters*

The filter which assumes the role of unity for convolution is the delta function,

$$\delta = \{\delta_{n,0}\} = (\ldots, 0, 0, \mathbf{1}, 0, 0, \ldots), \tag{56}$$

since convolution with it leaves any filter unchanged. Then given any filter C we can ask whether there is an inverse, C^{-1}, such that $C * C^{-1} = \delta$. The answer is obtained by applying the Z-transform to this equation:

$$C(z)C^{-1}(z) = 1, \tag{57}$$

so

$$C^{-1} = Z^{-1}\left[\frac{1}{C(z)}\right], \tag{58}$$

where Z^{-1} denotes the inverse Z-transform. Hence finding the inverse of C is reduced to finding the coefficients in the series expansion of the reciprocal of the Z-transform of C. Such expansions always involve choices as to whether to use positive or negative powers of z. The choice is made on the basis that the resulting inverse filter should converge, as will now be explained. Consider the dipole factorization given in § III*d*. It is easily seen that the inverse of the filter is the convolution of the inverses of its dipole factors, so the problem is reduced to finding the inverse of a dipole. Consider

first causal dipoles which, except for a constant factor, can be written $(\mathbf{1}, -a)$. The Z-transform is $(1-az)$. Which expansion of $(1-az)^{-1}$ converges[6] depends on the magnitude of a:

$$(1-az)^{-1}$$

$$= \begin{cases} 1+az+(az)^2+(az)^3+\ldots, & \text{if } |a|<1; \\ -\left[(az)^{-1}+(az)^{-2}+(az)^{-3}+\ldots\right], & \text{if } |a|>1. \end{cases} \tag{59}$$

Thus, the Z-transform of the inverse of a minimum (maximum) delay causal dipole must be expanded in positive (negative) powers of z if the result is to converge. If $C=(\mathbf{1}, -a)$

$$C^{-1} = \begin{cases} (\mathbf{1}, a, a^2, a^3, \ldots), & |a|<1; \\ (\ldots, -a^{-3}, -a^{-2}, -a^{-1}, \mathbf{0}, 0, 0, \ldots), & |a|>1. \end{cases} \tag{60}$$

(See Fig. 10 and the associated discussion in § II*e*.) Similarly, a maximum (minimum) delay acausal dipole gives a convergent expansion in negative (positive) powers of z. It is easy to prove (e.g., with Z-transforms) that a minimum delay causal dipole has the special simplifying property that its inverse is also minimum delay and causal. The same holds for the convolution of arbitrarily many such dipoles. Similarly, the inverse of a maximum delay acausal pulse is maximum delay and acausal (see Table 1). Because of this, it is convenient to arrange the factorization so that all factors are in one of these forms. This can always be accomplished as follows: suppose P of the zeros of $C(z)$ satisfy $|z_o^i|>1$ and the remaining Q zeros satisfy $|z_o^i|<1$ (assume all $|z_o^i|\neq 1$). Then shift the time origin of C so that in the notation of § III*d* $p=P$ and $q=Q$. Then assign the P zeros which lie outside the unit circle in the complex plane to the p causal dipoles in the factorization (eq. [52])—these will be minimum delay. The Q zeros inside the unit circle are assigned to the q acausal dipoles, which are then maximum delay. This factorization represents the filter as the convolution of two factors:

$$F=(\mathbf{1}, a_1)*(\mathbf{1}, a_2)*\ldots*(\mathbf{1}, a_p) \tag{61}$$

(p factors, minimum delay, causal), and

$$G=(a_{p+1}, \mathbf{1})*(a_{p+2}, \mathbf{1})*\ldots*(a_{p+q}, \mathbf{1}) \tag{62}$$

(q factors, maximum delay, acausal), so that $C=K(F*G)$ and $C^{-1}=(K^{-1})(F^{-1}*G^{-1})$, where K is as defined above. Note that F^{-1} and G^{-1} have the same

[6]Convergence at $z=1$ is implied, because we are really interested in the convergence of the coefficients of z^n in the expansion of the Z-transform. This allows use of the DFT, because $|z|=|\exp(-i\omega)|=1$ on the unit circle.

TABLE 1

PROPERTIES OF DIPOLES AND THEIR INVERSES

Case	C	$C(z)$	Zero of $C(z)$	Comments	$A\equiv C^{-1}$	$A(z)$	Zero of $A(z)$	Comments
I	$(\mathbf{1}, -a), \lvert a\rvert<1$	$1-az$	a^{-1} Outside unit circle	Causal; minimum delay	$+(\mathbf{1}, a, a^2, \ldots)$	$(1-az)^{-1}$	∞ Outside unit circle	Causal; minimum delay[a]
II	$(\mathbf{1}, -a), \lvert a\rvert>1$	$1-az$	a^{-1} Inside unit circle	Causal; maximum delay	$-(\ldots, a^{-2}, a^{-1}, \mathbf{0})$	$(1-az)^{-1}$	∞ Outside unit circle	Acausal; maximum delay[a]
III	$(-a, \mathbf{1}), \lvert a\rvert>1$	$1-az^{-1}$	a Outside unit circle	Acausal; minimum delay	$-(\mathbf{0}, a^{-1}, a^{-2}, \ldots)$	$z/(z-a)$	0 Inside unit circle	Causal; minimum delay[a]
IV	$(-a, \mathbf{1}), \lvert a\rvert<1$	$1-az^{-1}$	a Inside unit circle	Acausal; maximum delay	$+(\ldots, a^2, a, \mathbf{1})$	$z/(z-a)$	0 Inside unit circle	Acausal; maximum delay[a]

[a]The delay properties of these pulses are, strictly speaking, not defined under the definition given in the text because they are not factorable into a finite number of dipoles. The characterization given here applies to finite (truncated) versions of the pulses. Cases II and III might be classified oppositely according to the zeros of the infinite Z-transforms, but this would be misleading; e.g., the zero of $A(z)$ in case III can be removed by left-shifting A by one unit.

delay and causality properties as do F and G, respectively. It can be shown that the Laurent series thus generated for $C^{-1}(z)$ converges within an annulus in the complex plane which includes the unit circle, and it is the coefficients of the various powers of z in this series which give the elements of C^{-1}.

In many of the standard treatments of this subject only causal filters are allowed. It then results that a filter has a convergent inverse if and only if the zeros of its Z-transform all lie outside the unit circle; otherwise the forward expansion diverges and the acausal backward expansion is not permitted. In other words, only minimum delay (causal) pulses have (causal) inverses, and then the inverse is also minimum delay. This problem was apparently first discussed by Wold (1938*b*). Two-sided filters always have a convergent inverse (unless a zero lies exactly on the unit circle).

In practice, a very convenient way to evaluate inverses is to replace the Z-transforms in equation (58) with the discrete Fourier transform (DFT). A code for this procedure is contained in the Appendix. Specifically, given a set of filter elements $\{C_i\}$, one evaluates the DFT of C, takes the reciprocal term by term, and then obtains the inverse DFT. *This procedure automatically provides the correct convergent expansion of a two-sided filter*—without explicit evaluation of the zeros of the Z-transform of the pulse! For example, consider the pulse $C=(\mathbf{1}, -a)$. The DFT procedure yields the inverse $(\mathbf{1}, a, a^2, a^3, \ldots)$. If $|a|<1$ this is obviously the correct inverse, interpreted as a causal pulse. Many terms may be necessary to get a good representation of the pulse shape, especially if $|a|$ is close to 1. If $|a|>1$, the above inverse, interpreted as a causal pulse, is divergent (or unstable). The trick is to note that for any finite number of terms, $(\mathbf{1}, a, a^2, \ldots, a^n)$, there will be one largest term, a^n. The inverse should then be renormalized to make this element unity: $(\boldsymbol{a}^{-n}, \ldots, a^{-2}, a^{-1}, 1)$, and then interpreted as an expansion backward in time, $(a^{-n}, \ldots, a^{-2}, a^{-1}, \mathbf{1})$. This is the correct (acausal) inverse if $|a|>1$. The same procedure works in the general case, in which the inverse pulse extends both forward and backward in time.[7] In general some zeros must be appended to the original pulse before applying the DFT inverse because the inverse is almost always longer than the filter itself. For two-sided pulses this is also needed to ensure that the backward and forward tails of the inverse pulse do not overlap, due to the wraparound feature of the DFT. (Envision the arrays pasted on the surface of a cylinder, with the right-hand and left-hand ends abutting. Any set of entries on the right end can be transferred to the left end without affecting the DFT. This is illustrated in Fig. 15.) Examples of inverses calculated in this way are shown in Figure 16.

While the inverse as defined here is unique, there are other inverses which can be defined. Noting, for example, that the exact inverse of most filters will be infinitely long, one can ask: What *finite* filter, of fixed length, is closest to being an inverse to C in the sense that the sum of the residuals from the delta function

$$\sum_{i=-q}^{p} |(C*C^{-1})_i - \delta_i|^2,$$

is minimum? The solution to this problem is the *truncated approximate (least squares) inverse* of C, and is discussed extensively by Robinson (1964*a*, 1967*a*; see also Treitel and Robinson 1966). One could just as well ask for the truncated inverse which minimizes the absolute value residuals (see Claerbout and Muir 1973 for an interesting discussion of some of the properties of this inverse). Inverses may also be evaluated by various techniques which involve determination of the zeros of the Z-transform of the filter (see, e.g., Steiglitz 1974),

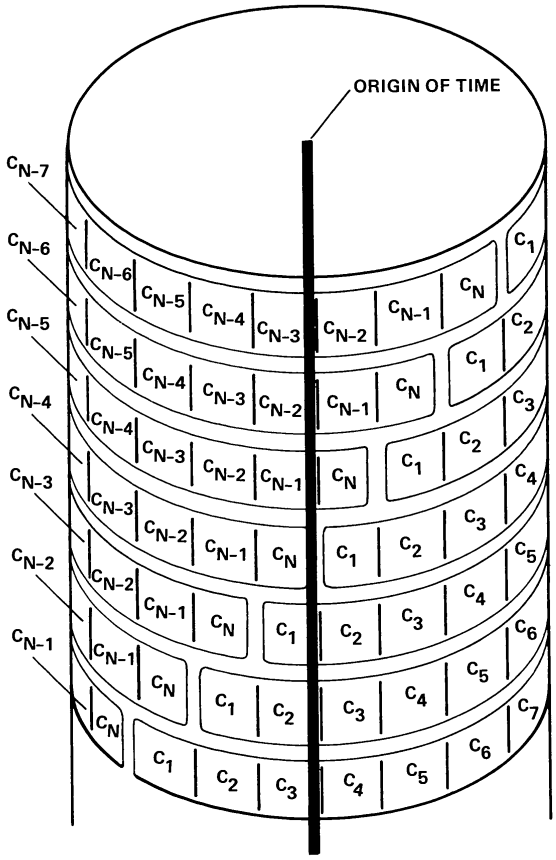

FIG. 15.—The wraparound feature of pulse shapes. All the pulses shown are equivalent in the sense that their inverses (and DFTs) are identical, except that they are similarly rotated with respect to each other.

[7]In this case the time origin does not appear at a fixed place in the inverse and must be identified by some other means. This inability to pinpoint the origin of time in the calculated inverse is the price paid for not having to determine the zeros of $C(z)$. Specifically, if we knew how many zeros lie inside and outside of the unit circle, we could then locate the origin. Frequently, but not always, the origin is located at the peak of the inverse pulse.

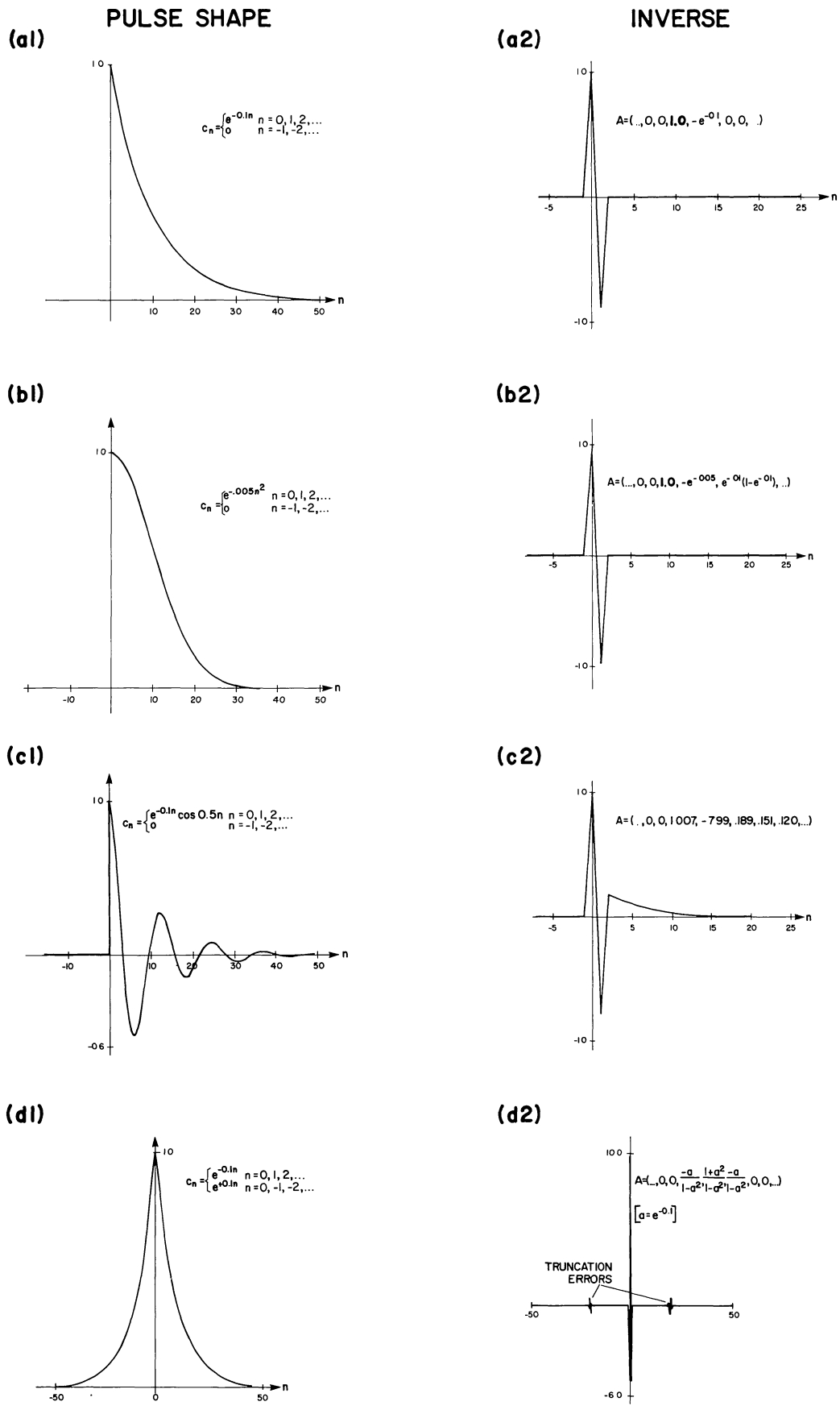

FIG. 16.—A sample zoo of pulse shapes (*left*) and the corresponding inverses (*right*) as determined with the discrete Fourier transform.

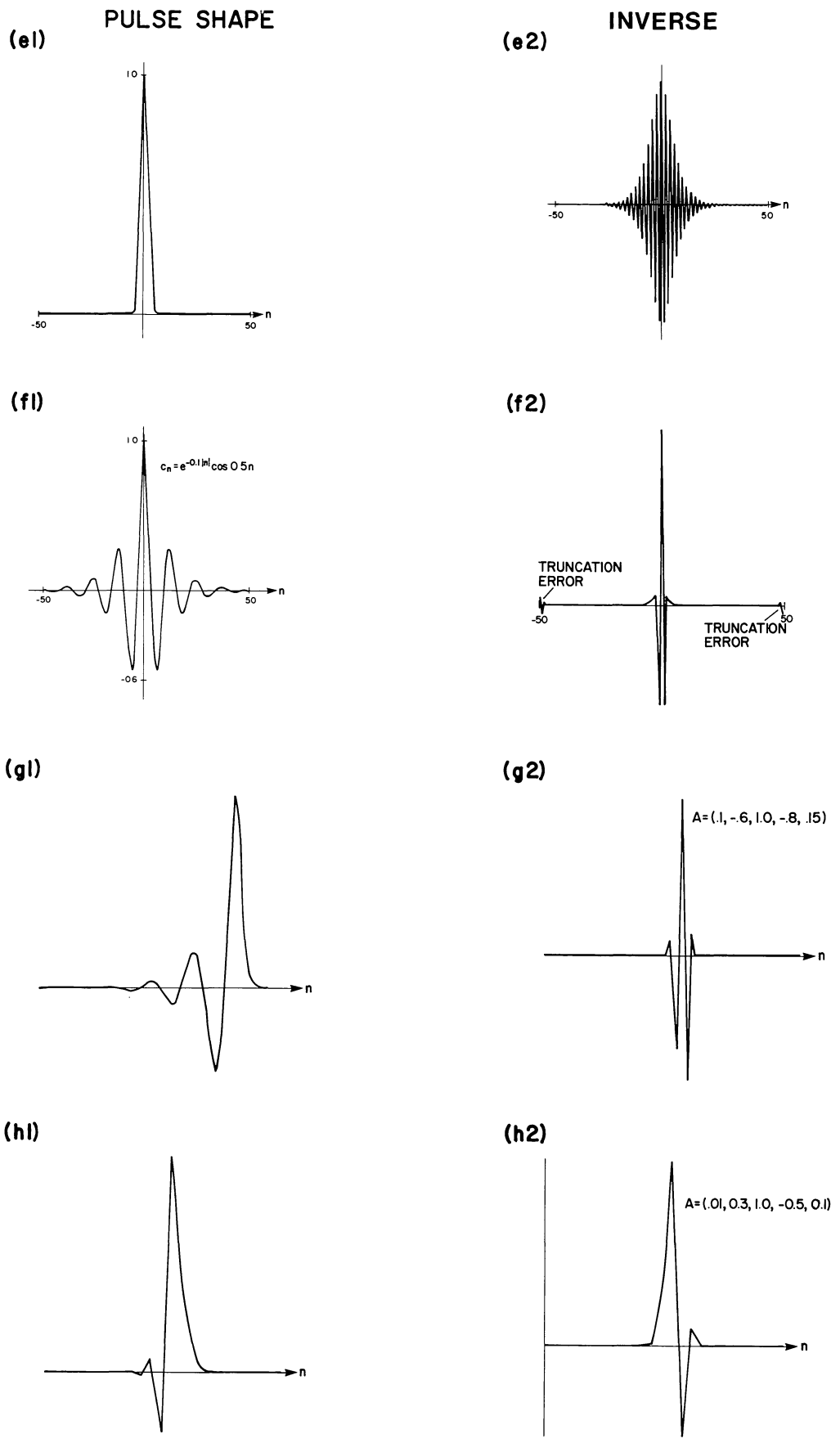

FIG. 16.—*Continued*

but this approach is computationally quite laborious compared to the DFT method.

g) *Correlation Functions and Power Spectra*

The autocorrelation function of a process X is defined as

$$\rho_X(n,m)=\langle(X_n-\bar{X})(X_m-\bar{X})\rangle, \quad (63)$$

where $\bar{X}\equiv\langle X_n\rangle$. Section II*a* outlined its significance. The power spectrum is the Fourier transform of the autocorrelation and also is equal to the squared magnitude of the Fourier transform of the time series itself. We shall give, without proof, expressions which are readily derived from the definitions.

For a moving average $X=R*C$, where R is assumed stationary and with spectrum $S_R(\omega)\equiv 1$, we have

$$\rho_X(n,m)=\rho_X(n-m)=\sigma_R^2\rho_C(n-m)-\bar{X}^2, \quad (64)$$

where $\sigma_R^2=\langle R_n^2\rangle$ is the variance of the innovation and ρ_C is the autocorrelation of the pulse, defined by

$$\rho_C(n-m)=\sum_k C_k C_{k+n-m}. \quad (65)$$

It can be seen that the autocorrelation is the convolution of the pulse with its reverse. For zero-mean processes (e.g., with $\langle R_n\rangle=0$) the autocorrelation of the MA is proportional to the autocorrelation of the pulse shape. Similarly, for this case the spectrum of the process is equal to the spectrum of the pulse shape:

$$S_X(\omega)=|C(\omega)|^2, \quad (66)$$

where $C(\omega)$ is the Fourier transform of the pulse:

$$C(\omega)=\sum_k C_k e^{ik\omega}, \quad (67)$$

and the normalization of R is such that $S_R(\omega)=1$. In terms of Z-transforms we have

$$S_X(\omega)=C^*(z^{-1})C(z), \quad \text{where } z=e^{i\omega}. \quad (68)$$

For an AR process, $R=A*X$, it is easy to show that

$$S_X(\omega)=\frac{1}{|A(\omega)|^2}, \quad (69)$$

where $A(\omega)$ is the Fourier transform of the AR filter:

$$A(\omega)=\sum_k A_k e^{ik\omega}. \quad (70)$$

Finally, for an ARMA process, $A*X=R*C$,

$$S_X(\omega)=\frac{|C(\omega)|^2}{|A(\omega)|^2}. \quad (71)$$

It is readily verified from these formulas (or directly from the definitions) that both the spectrum and autocorrelation of X are unchanged by time reversal of C, a result alluded to in § III*e*.

IV. MODEL CONSTRUCTION

The tools are now at hand to construct stochastic models from time series data. In outline the procedure is: (1) obtain data from one or more realizations of the process of interest; (2) decide on the form of the model to be fit to these data; (3) use the data to generate estimates of the model parameters; and (4) if necessary, transform the resulting model to a form more easily interpreted physically. (The last step recognizes that the form most suited to computations may not be the most suitable for comparison with physical models. Typically a low-order AR model is easiest to compute, and the corresponding MA has the simplest physical interpretation. See § V*f*.) The stage will be set by presenting an existence theorem which justifies the concern in §§ II and III for the MA and AR models, by asserting that any stationary process can be represented with these models. Then explicit methods for the estimation of the parameters in these models will be developed. We assume that all processes of interest are stationary.

a) *An Existence Theorem: The Wold Decomposition*

Moving average models were introduced in § II as a rather arbitrary way of representing memory or correlations. The question arises as to what processes can be represented in this seemingly very special form. The surprising answer, first demonstrated in 1938 by the econometrician Herman Wold (1938*a*), is that any stationary process can be so represented. The simple explicit form, known as the *Wold Decomposition*, is given in the following theorem.

> The *Wold Decomposition Theorem*: Given any stationary process, X, there exist:
>
> 1. a purely deterministic process D,
> 2. an uncorrelated zero-mean noise process R, and
> 3. a moving average filter C,
>
> such that $X=R*C+D$.

This is a decomposition of X into a deterministic part (D) and a random part ($R*C$). The random part may contain correlations and can in turn be deconvolved into a moving average, in which the correlations are represented by the deterministic filter C and the purely random part is contained in the white noise process R. If the MA is restricted to be causal, this decomposition/deconvolution is unique (except for a constant

factor which can be exchanged between R and C). It is not unique without the causality condition, because there are other noncausal MA representations. This nonuniqueness is the subject of the following subsection. If, in addition, X has an absolutely continuous (Titchmarsh 1939, p. 364) spectral distribution function (i.e., X is itself not deterministic), then C is minimum delay, and therefore has a convergent, causal, minimum delay inverse A. This fact assures the existence of a unique autoregressive representation of the *detrended* process $X-D$, in the form $A*(X-D)=R$, where $A=C^{-1}$. Thus the Wold theorem establishes that any stationary process, with its deterministic part (including the mean value) removed, can be represented as an MA, AR, or a mixed ARMA process (see § IIf). The estimation of D (or the related problem of *detrending* the data) is a nontrivial problem which will not be discussed here, except to remark that the spectrum of D is generally discrete (lines), whereas that of $R*C$ is continuous.

For a thorough discussion and proof of this theorem see Hannan (1970, p. 137) or Robinson (1964b, p. 126). The following informal proof conveys the spirit of these rigorous works. Consider a given stationary process X, which for simplicity will be taken to have zero mean. The forward predictor of order p is defined as

$$\hat{X}_n^{(p)} = \sum_{k=1}^{p} B_k X_{n-k}, \tag{72}$$

for any set of numbers B_k, $k=1,2,\ldots,p$. This linear expression is designed to forecast the value of X_n, based on the previous values $X_{n-1}, X_{n-2},\ldots, X_{n-p}$. The quality of the prediction of course depends on the values of the B_k. Those values which give the best predictions form the *optimum predictor* of order p. More specifically, the optimum least-squares predictor of order p is defined as that which minimizes the mean square prediction error,

$$E(B) = \sum_n \left[X_n - \hat{X}_n^{(p)} \right]^2, \tag{73}$$

with respect to the parameters in B. The optimum predictor is the limit as $p\to\infty$. A very important process is that defined by

$$R_n = X_n - \hat{X}_n^{(\mathrm{opt})}, \tag{74}$$

the error made by the optimum predictor at time n. This random process is to be identified with the white noise process R in the definitions of AR, MA, and ARMA processes (§ II) and is called the *innovation* of the process X (Kailath 1968; Parzen 1969). The error at time n is due to the new pulse starting at that time, because the effects of pulses starting at previous times are completely incorporated into the optimum prediction. That $\langle R_n \rangle = 0$ follows immediately from the vanishing of $\langle X_n \rangle$ and the definition of R. It can be shown (Wold 1938a) that

$$\langle X_{n-k} R_n \rangle = 0, \quad \text{for all } k>0. \tag{75}$$

Intuitively this is so because R_n is the error made at time n by a predictor optimized on all prior data (i.e., $X_{n-1}, X_{n-2},\ldots$), so there can be no correlation of R with these data. Otherwise the correlations could be used to improve the already optimum predictor. It follows that $\langle R_n R_m \rangle = 0$ for all $m\neq n$; for, taking $m>n$ without loss of generality,

$$\begin{aligned} \langle R_m R_n \rangle &= \langle R_m (X_n - \hat{X}_n) \rangle \\ &= R_m X_n - \sum_k B_k R_m X_{n-k} = 0 \end{aligned} \tag{76}$$

because all of the terms are of the form in equation (75). This makes the R_n a kind of orthogonal set, and the process X can be expanded in the series

$$X_n = \sum_{k=1}^{\infty} C_k R_{n-k} + D_n, \tag{77}$$

where D is a residual process, orthogonal to R. By the usual technique of multiplying this equation by R_m and taking expectation values, the expansion coefficients can be found:

$$C_k = \langle X_n R_{n-k} \rangle. \tag{78}$$

(This formula is an alternate way of computing the MA parameters and has some advantages over the direct inversion $C=A^{-1}$.) The final step, that D is deterministic, is a consequence of the vanishing of the prediction error for D. The details of this proof can be found in the above references. Caines and Sethi (1979) give an interesting discussion of causality and the Wold theorem.

b) A Less Restrictive Existence Theorem

The moving average filter of the Wold representation is (1) convergent (or stable): $\Sigma_k C_k^2 < \infty$; (2) causal: $C_k = 0$ for $k<0$; (3) minimum delay (see § IIIe); and (4) constant (C_k independent of time). Extending Robinson's (1962) terminology, we call any filter with these properties a *minimum delay wavelet*. It is indeed a curious feature of the Wold theorem that an arbitrary stationary process can be represented in such a special form. *What about an MA process with a pulse that does not have these properties? The Wold decomposition exactly represents such a process with an MA model which **does** have these properties*. For example, it represents a mixed-delay MA in terms of minimum delay wavelets. It would seem that such representations are misrepresentative. Some processes seem to have better representations than the one provided by the Wold theorem.

But how can this be? The answer lies in the fact that, while too restrictive with the pulse C, the Wold decomposition is too liberal with regard to the innovation. It would be preferable, at least for physical processes consisting of independent pulses, to restrict the innovation to be independently distributed (not just uncorrelated) and to allow the pulse to be mixed-delay and acausal, rather than assuming causality. (Incidentally, there is presumably a similar extension in which constancy of the pulse is dispensed with, for one can construct a stationary MA with nonconstant pulses. Stability cannot be dispensed with because it corresponds to finiteness of observable quantities.)

A key point is that a given stationary process can be represented by any member of a large family of MA models. The members of the family share a common autocorrelation and power spectrum but have different delay/causality properties; the corresponding innovations have different degrees of randomness ranging from uncorrelated to independently distributed. The Wold theorem singles out the unique minimum delay wavelet representation because only causal filters are permitted. The existence theorem for the more general representations is as follows (Scargle 1977):

> *The Extended Decomposition Theorem*: Given any stationary process X, there exist:
>
> 1. a purely deterministic process D,
> 2. a family of uncorrelated, zero-mean noise processes, $\{R^{(i)}\}$, and
> 3. a family of two-sided moving average filters, $\{C^{(i)}\}$,
>
> such that $X = D + C^{(i)} * R^{(i)}$. The filter family is the set of all filters which have the same autocorrelation function as X; one of them is minimum delay, and one maximum delay, and the rest are mixed delay.

The proof is simple. Since X is stationary, the Wold theorem applies and assures the existence of a unique, causal, moving-average representation,

$$X - D = C^w * R^w, \tag{79}$$

where C^w is a minimum delay wavelet. It was shown above (§ III*d*; see also Robinson 1964*b*; Smylie, Clarke, and Ulrych 1973) that there is a family of filters which share a given autocorrelation and which can be obtained from each other by all possible combinations of time-reversal of the dipole factors. We define the family $\{C^{(i)}\}$ as the set of all filters which have the same autocorrelation as C^w. If C^w is finite, of length $N+1$, then there are 2^N (not necessarily distinct) members of this set. One is minimum delay (C^w itself), one is maximum delay (the reverse of C^w), and the rest are mixed delay. For each $C^{(i)}$ define $A^{(i)} = [C^{(i)}]^{-1}$ and $R^{(i)} = A^{(i)} * (X - D)$. Then

$$C^{(i)} * R^{(i)} = C^{(i)} * A^{(i)} * (X - D) = X - D, \tag{80}$$

establishing the desired representation. A direct calculation of the autocorrelation of $R^{(i)}$ shows that it is the same as that of R^w, namely $\sigma^2 \delta_{nm}$, and this completes the proof. The uniqueness of this family is also readily demonstrated. Note that the representations in this theorem are not just similar, they are *exactly equivalent*. They differ only in the way in which the random and deterministic parts are assigned to the innovation and to the pulse.

It is possible that a theorem stating that one and only one of the $R^{(i)}$ is always independently distributed can be proved. I do not know whether this theorem is true. There are theorems dealing with the existence of nonlinear representations with independently distributed innovations (Rosenblatt 1971) or innovations with the martingale difference property (Seagall 1976). Therefore, it seems likely that further restrictions beyond stationarity must be imposed on a process to ensure the existence of a linear MA with an independently distributed innovation. Our point of view will be to *assume* the independence of the noise driving the observed process. Then one of the family of representations will certainly be independently distributed; this one will be regarded as the correct one, as it most completely and faithfully separates the random and nonrandom parts of the process. It is easily seen that the innovations of the other representations can be written as linear combinations of the id one at different lags (cf. eq. [86] below) and are therefore dependently distributed. Although exactly equivalent to the correct one, they will be considered incorrect representations because their innovations are not purely random.

A concrete example will help clarify these matters. Consider the exponential pulse

$$C_k = \begin{cases} 0, & k < 0 \\ e^{-bk}, & k \geq 0 \end{cases} \quad (b > 0), \tag{81}$$

which has been invoked in astronomical shot noise models (e.g., Terrell and Olsen 1970). This minimum delay wavelet is the inverse of the simplest possible nontrivial AR filter, that is, the one-parameter model used as an example in § III*f*, with $a = e^{-b}$. Let R be an id noise process, and consider the moving average $X = R * C$. The inverse of C is the dipole $(1, -a)$. Hence the family of MA filters for this process has only two members, namely

$$C^w = (1, a, a^2, a^3, \ldots), \tag{82}$$

and

$$\tilde{C}^w=(\ldots,a^3,a^2,a,\mathbf{1}). \tag{83}$$

The corresponding inverses are $(\mathbf{1},-a)$ and $(-a,\mathbf{1})$. The MA representations are $X=C^w*R$ (precisely the form used to define X) and $X=\tilde{C}^w*R'$, where

$$R'=(-a,\mathbf{1})*X=(-a,\mathbf{1})*(\mathbf{1},-a)^{-1}*R=P*R, \tag{84}$$

with

$$P=(-a,\mathbf{1})*(\mathbf{1},-a)^{-1}. \tag{85}$$

The pulse P is fundamental in the algebra of dipoles: convolution with P of a filter that has the dipole factor $(\mathbf{1},-a)$ reverses that factor. With the aid of Z-transforms the following explicit forms can be derived:

$$R'_n=(1-a^2)\sum_{k=0}^{\infty}a^kR_{n-k}-aR_{n+1}, \tag{86}$$

and

$$P_k=\begin{cases}(1-a^2)a^k, & k\geq 0\\ -a, & k=-1\\ 0, & k<-1.\end{cases} \tag{87}$$

It might be surmised from inspection of equation (86) that R'_n and R'_{n+1} are correlated because they have many terms in R in common. However, a straightforward calculation yields

$$\langle R'_nR'_m\rangle=\sigma^2\delta_{nm}=\langle R_nR_m\rangle, \tag{88}$$

and

$$(\tilde{P}*P)_n=\delta_{n0}, \text{ i.e., } (\ldots,0,0,\mathbf{1},0,0,\ldots). \tag{89}$$

Figure 4 shows an example of processes related in this way: that in Figure 4*d*1 is independently distributed, and Figure 4*c*1 is the same process (same realization) filtered with P. The pulse P is graphed in Figure 17. It is perhaps surprising to find a pulse other than the delta function itself which has a delta function autocorrelation. There are many such pulses. They are sometimes called *all-pass filters*. The filter $D^N\tilde{A}*A^{-1}$, for arbitrary A of order N or less, has this property (D is the unit delay operator defined in eq. [41]). (Radar design is one application where unautocorrelated pulses are sought Boehmer 1967.) Note that our process, constructed as randomly occurring, decaying exponential pulses, can also be represented as randomly occurring, *growing* exponentials! These representations are mathematically equivalent, as $C^w*R=\tilde{C}^w*R'$. But

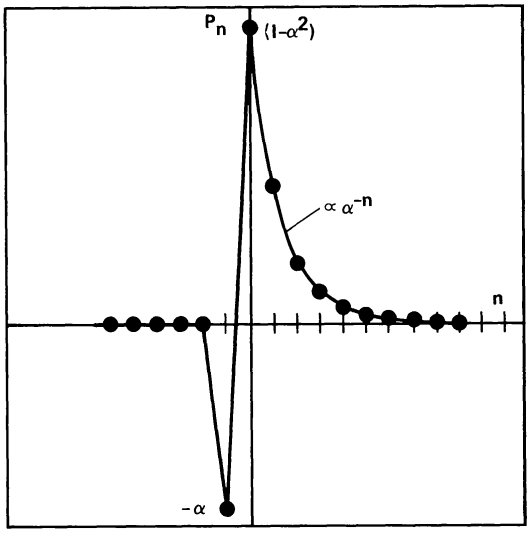

FIG. 17.—The pulse $P=\{P_n\}$, described in the text, which has a delta-function autocorrelation. This was the pulse in the moving average shown in Fig. 4(*c*).

C^w*R is a better representation because it is the same one used to construct the process in the first place. It is better in the sense that its innovation is independently distributed and not merely uncorrelated, as is the innovation R'.

c) Deconvolution via Independently Distributed Innovations

The previous subsection assures that a MA representation exists. While it is not automatic that this linear superposition of constant pulse shapes is physically significant, it frequently is. That is, random processes which occur in nature often consist of the summation of independent pulses. Since the moving average model represents a process as the convolution of a pulse shape C with an innovation R, the process of deducing the model (C,R) from time series data is called *deconvolution*. (Sometimes this term is used if C is known, but here it will always be assumed that C is to be determined.) The goal is to disentangle the overlapping pulses from each other, revealing the underlying pulse shape and information about the amplitudes of the pulses.

Most of the standard deconvolution techniques (§ IV*d*) are based on least-squares modeling or the autocorrelation function and are therefore insensitive to the information needed to determine the phase character of the pulses. Such techniques cannot distinguish among the representations in the extended decomposition theorem. Further, if the driving process R is normally distributed (Gaussian) noise, it can be shown (Parzen 1962, p. 90) that the process $X=R*C$ is also normal and therefore completely characterized by its mean value and its autocorrelation function. In this

case no technique can recover the phase information. The pulses in an MA driven by Gaussian noise overlap so much that the phase information is irretrievably lost. However, many physical processes are not normally distributed, and for these the problem arises as to how to determine the pulse shape with the correct phase property. This is the

> *Fundamental problem*: Given data sampled from the moving average process $X = R * C$, where R is independently distributed noise and C is a (not necessarily minimum delay) pulse, find estimates of the pulse shape C and amplitude sequence R.

The standard techniques determine the minimum delay pulse which has the same autocorrelation as C. But if R is not Gaussian, the correct pulse shape can be recovered. The key fact is that *the innovation corresponding to the correct pulse is independently distributed*, while the other members of the family of innovations in the extended decomposition are not independent. This fact follows if the actual pulses are independent of each other—an assumption which has to be justified on physical grounds for each case under study. In astronomy this justification often derives from the notion that the pulses arise in different physical regions that do not communicate effectively with each other.

The procedure to be described here is a direct search for an independently distributed innovation. We seek the model (AR, MA, or ARMA) which, of all models consistent with the sampled data, has the least dependence in the distribution of the innovation. Begin by writing, in terms of the data X, the innovation as a function of the model parameters (eqs. [15], [19], [26], and [27]):

$$R = \begin{cases} A * X & \text{(AR model)}, \\ C^{-1} * X & \text{(MA model)}, \\ A * C^{-1} * X & \text{(ARMA model)}, \\ A * C^{-1} * (\nabla^d X) & \text{(ARIMA model)}. \end{cases} \tag{90}$$

Because of its simplicity and practicality the AR model is the prototype in this discussion, but the others can be treated in much the same way. The explicit form of R in this case is

$$R_n = \sum_{k=-q}^{p} A_k X_{n-k}, \quad n = p+1, p+2, \ldots, N-q, \tag{91}$$

where A is of order (p, q): $A = (A_{-q}, \ldots, A_{-1}, A_0, A_1, \ldots, A_p)$. Then construct a measure of the dependence of the process R, and minimize it with respect to the model parameters. There is no one correct way of defining a suitable dependence measure. Corresponding to each of the definitions of independence given in § II*a* there is the following quantity which could be used as a measure of the dependence of the process R:

> 1. $P_M(r_1, r_2, \ldots, r_M) - P_1(r_1)P_1(r_2)\ldots P_1(r_M)$ using probability distributions;
> 2. $F_M(r_1, r_2, \ldots, r_M) - F_1(r_1)F_1(r_2)\ldots F_1(r_M)$ using cumulative probability functions;
> 3. $\phi_M(u_1, u_2, \ldots, u_M) - \phi_1(u_1)\phi_1(u_2)\ldots\phi_1(u_M)$ using characteristic functions; and
> 4. $\langle g_1(R_1)g_2(R_2)\ldots g_M(R_M)\rangle - \langle g_1(R_1)\rangle\langle g_2(R_2)\rangle\ldots\langle g_M(R_M)\rangle$ using expectations.

In these expressions a simplified notation is used to give the order of the statistical functions. If R were independently distributed these four expressions would all vanish for all values of the appropriate independent variables (the r or the u) or for all functions g_i, and for all values of the integer M. There is a variety of ways one might choose to estimate the statistical functions in these expressions or to assess the departure of the chosen expression from zero. Of these many ways of proceeding, different ones will undoubtedly be suitable for different kinds of problems. Extensive experimentation has led to one procedure which has worked well in a variety of test cases. This procedure is offered as a fairly general purpose one, but the reader may wish to consider other approaches to dependence minimization for his data analysis problems.

How are the individual and joint probability functions in the above expressions to be evaluated? First, equation (91) (or, more generally, eq. [90]) generates R from the sampled data X, as a function of the model parameters. The resulting values of R are then used to estimate the function of interest, in the form of an average. Assume that R is *ergodic*, so that the desired ensemble averages can be computed as time averages. For example, to estimate $Q_2(r_1, r_2)$, where Q stands for P, F, or ϕ, evaluate the average

$$\langle \hat{Q}_2(r_n, r_{n+1})\rangle_n = \frac{1}{N-(p+q+1)} \sum_{n=p+1}^{N-q-1} \hat{Q}_2(r_n, r_{n+1}). \tag{92}$$

The way in which the estimators (such as $\hat{P}_2$) are calculated is different for each of the four forms (1)–(4) above and will be described below. Because time averages are used, no distinction can be made between the various second-order functions, such as $Q_2(r_1, r_2)$, $Q_2(r_2, r_3), \ldots, Q_2(r_k, r_{k+1})$. Such a distinction is unnecessary, however, because the assumption that R is

stationary means that all of these are equal anyway. The next step would be to consider third-order functions, such as $Q_3(r_1, r_2, r_3)$, which are awkward to deal with numerically. Fortunately (Papoulis 1965), the added information by going from second to third order is contained in simpler expressions, such as $Q_2(r_n, r_{n+2})$, or in general $Q_2(r_n, r_{n+m})$. The corresponding time-average is

$$\langle \hat{Q}_2(r_n, r_{n+m}) \rangle_n = \frac{1}{N-(p+q+m)} \sum_{n=p+1}^{N-q-m} \hat{Q}_2(r_n, r_{n+m}). \quad (93)$$

Hence, expressions higher than the second order never need be considered.

The final dependence measure is the sum of expressions such as equation (93), from $m=1$ to some maximum value, m^* (see, e.g., eq. [98] below). What should this range of values be? Unless $m^* \ll N$, the small number of terms in sums such as equation (93) will make the estimates ill determined. Numerical experiments of the kind described in § V, mostly with cumulative probability functions, have yielded the following: For simple models of order one or two, the single lag $m=1$ may be sufficient in the sense that no further information is added by including larger lags. But for higher order models m^* must be larger than 1 if all of the information about the process is to be extracted from the data. The best choice for m^* appears to be roughly equal to the number of free parameters, i.e., $m^* \approx p+q$. A rationale for this empirical result is lacking, although it is not unreasonable.

The approaches using probability distribution functions (PDF), cumulative probability functions, and characteristic functions were tested on problems with known pulse shapes and innovations. The dependence measure based on cumulative probability functions proved by far the best (see § VI*a*), and the details of this approach will now be given. A straightforward estimate of the cumulative probability function of R_n is (see the definition in eq. [7]):

$$F_1(x) = \frac{1}{N^*} \sum_{N=1}^{N^*} H(x-R_n), \quad (94)$$

where $N^* = N-(p+q)$ and the R_n have been reindexed as described in § V*c*. The quantity $H(x)$ is the unit step function:

$$H(x) = \begin{cases} 0, & x<0 \\ 1, & x \geq 0. \end{cases} \quad (95)$$

The sum in equation (94) is just the number of R_n which are $\leq x$, and therefore $1/N^*$ times the sum is an estimate of $Pr\{R_n \leq x\}$. F_1 is a step function, consisting of equal steps (of amplitude $1/N^*$) at each of the R_n. Similarly, the second-order joint cumulative probability function for lag m is estimated with

$$F_2^m(x,y) = \frac{1}{(N^*-m)} \sum_{n=1}^{N^*-m} H(x-R_n)H(y-R_{n+m}). \quad (96)$$

In this expression the sum is just the number of pairs (R_n, R_{n+m}) such that

$$R_n \leq x \quad \text{and} \quad R_{n+m} \leq y \quad (97)$$

(see the definition of F_N given in eq. [5]). The dependence measure is taken to be

$$D_F(A) = \sum_{m=1}^{m^*} D_F^m(A) = \sum_{m=1}^{m^*} \int\int |F_2^m(x,y) - F_1(x)F_1(y)|^2 \, dx \, dy. \quad (98)$$

The evaluation and minimization of this expression are described in § V.

So far we have ignored the possible contamination of the observations by noise. As will be discussed at the beginning of § VI, there is a bias in the estimated model parameters in such cases. Although the effect appears to be small in the numerical experiments considered there, one should be aware of this potentially significant bias in cases of low signal-to-noise ratio.

An analog of equation (98) with probability distributions replacing probability functions is

$$D_P(A) = \sum_{m=1}^{m^*} \int\int |P_2^m(x,y) - P_1(x)P_1(y)|^2 \, dx \, dy. \quad (99)$$

However, numerical tests have shown this dependence measure to be inferior to D_F. The reasons are readily understood. The estimates of P_1 and P_2 involve the construction of intervals or bins in both R_n and (R_n, R_{n+m}) space, and then counting the number of points in the bins. This procedure has several difficulties. First, the results are rather sensitive to the sizes and positions of the bins, and there is no obvious way to choose these optimally. Indeed, it appears that the optimum bins depend on the distribution of R, which of course is not known *a priori*. A second difficulty lies in the quantized nature of bin occupation: A sufficiently small change in A only moves the R points around within the bins, and leaves the number of points in the bins (and therefore the estimates of P_1 and

P_2) unchanged. Hence the derivative of the penalty function, D_P, is highly discontinuous. This effect foils minimization methods which use gradients, and it also appears to produce a forest of local minima which makes the global minimum very elusive. The author achieved some success in alleviating these problems by weighting the points according to their distance from the bin center (a Gaussian dependence proved superior to exponential or linear), to remove the quantum effect. Even so, there were still numerous local minima in typical problems. The expression in equation (98), because it uses cumulative functions, requires no binning and is a smoothly varying function of A. For low-order models it possesses a single minimum to which the minimizer converges rapidly, independently of the starting value. This result holds even with $m^* = 1$. When the order of the model is larger, local minima invariably appear unless m^* is increased (see § V).

Another problem with D_P concerns the treatment of the points that spill outside the chosen R interval. Again some success was achieved with empirical remedies, namely the application of a penalty for such spills (to be added to D_P) or defining the edges of the bins, in an R-dependent way, to include the maximum and minimum R-values. But these stop-gap remedies were only partially successful at producing a well-behaved dependence function. It is also awkward to have so many adjustable parameters (number, size and location of bins, weighting functions, spill penalties, etc.) to be chosen arbitrarily or optimized using trial cases. In comparison, the function D_F, given in equation (98) and evaluated as described in § V*d*, is very well behaved and free of undetermined parameters or functions.

The characteristic function is intrinsically a continuous function of the A_k, so the method based on these functions also avoids some of the problems discussed above. The first and second orders are

$$\phi_1(u) = \langle \exp(iuR_n) \rangle , \tag{100}$$

and

$$\phi_2^m(u_1, u_2) = \langle \exp[i(u_1 R_n + u_2 R_{n+m})] \rangle . \tag{101}$$

The corresponding condition for independence is

$$G_\phi^m(u_1, u_2) \equiv \phi_2^m(u_1, u_2) - \phi_1(u_1)\phi_1(u_2) = 0, \tag{102}$$

for $m = 1, 2, \ldots$. Since this function must vanish for all u_1 and u_2, there are various expressions which could be adopted as the dependence measure, the most obvious being the integral

$$D_\phi^m = \int\int_{-\infty}^{\infty} |G_\phi^m(u_1, u_2)|^2 \, du_1 \, du_2 . \tag{103}$$

This measure, even with weighting functions thrown into the integrand, did not give very promising results and was numerically awkward. A much simpler procedure comes out of the Taylor series expansion of the function G_ϕ^m in equation (102). Write

$$\phi_1(u) = \sum_{k=0}^{\infty} \left[\frac{d^k \phi_1(u)}{du^k} \right]_{u=0} \frac{u^k}{k!} , \tag{104}$$

and

$$\phi_2^m(u_1, u_2) = \sum_{k,j=0}^{\infty} \left[\frac{\partial^{k+j} \phi_2^m(u_1, u_2)}{\partial u_1^k \partial u_2^j} \right]_{u_1 = u_2 = 0} \frac{u_1^k u_2^j}{k! j!} . \tag{105}$$

The quantities in square brackets are $i^k \mu_k$ and $i^{k+j} \mu_{k,j}^m$, respectively, where the μ are the moments

$$\mu_k = \langle R_n^k \rangle , \tag{106}$$

and

$$\mu_{k,j}^m = \langle R_n^k R_{n+m}^j \rangle . \tag{107}$$

Since all powers of u_1 and u_2 must vanish in equation (102),

$$\mu_{k,j}^m = \mu_k \mu_j , \tag{108}$$

for all k, j, and m. Accordingly, the expression

$$D_{mom} = \sum_m \sum_{k,j} |(\mu_{k,j} - \mu_k \mu_j) w(k,j)|^2 \tag{109}$$

can be taken to represent the degree of dependence of the process R ($w[k, j]$ is a weighting function). For simple models the single value $m = 1$, and just a few terms k, $j = 1, 2$, seem to suffice. The term $k = j = 1$ corresponds to the autocorrelation function, and the terms $k = 1$, $j = 2$, and $k = 2$, $j = 1$ are related to the "time skewness function" of Frenkiel and Klebanoff (1967) applied to a related problem by Weisskopf *et al.* (1978). The numerical tests showed that moments and characteristic functions have some merit for this problem, but again local minima were bothersome, and no choice of weights for the u or the μ could be found that yielded consistently satisfactory results.

The author has not experimented with expectations of arbitrary functions (method [4] in the list above), mostly because the infinite arbitrariness in choosing the function sets is so imposing.

Finally, while it would not necessarily yield independently distributed innovations, a procedure based on maximizing the martingale difference property (§ II*a*) was considered. In fact, the implementation is straightforward and easy. Select a set of R bins, denoted $\mathcal{R}_i$,

and then evaluate the conditional expectation value

$$\langle R_n | R_{n+m} \in \mathcal{R}_i \rangle = \frac{1}{\mathcal{N}} \sum_{\substack{n \text{ such that} \\ R_{n+m} \in \mathcal{R}_i}} R_n, \quad (110)$$

where $\mathcal{N}$ is the number of n such that $R_{n+m} \in \mathcal{R}_i$ and $\in$ stands for "is an element of." The measure of the martingale characteristic would then be

$$D_{\text{MDP}} = \sum_i |\langle R_n | R_{n+m} \in \mathcal{R}_i \rangle - \langle R_n \rangle|^2, \quad (111)$$

where

$$\langle R_n \rangle = \frac{1}{N} \sum_{n=1}^{N} R_n \quad (112)$$

is just the (unconditional) expectation value of R. As will be seen in § VI*a*, this procedure does not appear to be very effective.

d) *Predictive Deconvolution of Time Series*

Predictive deconvolution (Peacock and Treitel 1969) or *predictive decomposition* (Robinson 1967*b*) refer to the use of linear prediction (Makhoul 1975), usually based on past data only, to yield information which allows representation of a process in terms of elementary building blocks (such as white noise processes, MA or AR filters, and deterministic processes). Since least-squares methods are almost always used, and these cannot recover phase information, only a brief sketch will be given. This discussion is intended to clarify the relation of predictive techniques to the material presented above, and also to motivate a technique (§ IV*e*) which is a simple extension of linear least-squares prediction and which *can* recover pulse phase information. More details than are given here can be found in an extensive literature (Kolmogorov 1941; Mann and Wald 1943; Wiener 1949; Bode and Shannon 1950; Durbin 1959, 1960; Walker 1962; Robinson 1964*b*; Gersch 1970; Akaike 1971, 1974; Chow 1972*a*; Kashyap 1974; Shinners 1974; Åström and Söderström 1974; Gertler and Bányász 1974; Gersch and Foutch 1974; Graupe, Krause, and Moore 1975; Tong 1975, 1976, to name a few), and especially the reviews by Robinson (1967*a*) and Box and Jenkins (1970). The reader interested in the new techniques only should skip to § V at this point.

The basic principle of predictive decomposition is that a model which gives good predictions of the behavior of a process undoubtedly is a good representation of the process. Thus one takes a model with a simple structure and adjusts it (by adjusting the values of the model parameters) until some measure of the error the model makes when tested against the available time series data is minimized. This procedure is called *optimizing* the model. The goal is not prediction per se, but representation of the statistical properties of the process. The hope is that the optimization will extract all of the information about the process that is contained in the data at hand.

The basics of the predictive approach are as follows. The term *linear prediction* used above simply means that the predictor is taken in the form[8]

$$\hat{X}_n = B_1 X_{n-1} + B_2 X_{n-2} + \ldots + B_k X_{n-k} \quad (113)$$

(see the autoregressive memory discussed in § II*d* and in the proof of the Wold decomposition theorem in § IV*a*). That is, this expression is to be used to predict[9] the value of X_n, based on knowledge of the previous values $X_{n-1}, X_{n-2}, \ldots$ only. The numbers B_i are related to the AR parameters and are to be determined by minimizing the prediction errors, in a sense to be defined. The error in prediction at time n is

$$E_n = X_n - \hat{X}_n = X_n + \sum_{i=1}^{k} A_i X_{n-i} = \sum_{i=0}^{k} A_i X_{n-i}, \quad (114)$$

where we have taken $A_i = -B_i$ and $A_0 = 1$. In other words, the expression

$$E = A * X \quad (115)$$

is the sequence of prediction errors as a function of time, and for this reason A is sometimes called a *prediction-error filter*. Suppose we take the sum of the squares of the prediction errors, that is

$$E(A) = \sum_n E_n^2, \quad (116)$$

as the measure of the errors which is to be minimized. In practice the length of the prediction filter is taken to be much less than the length of data ($k \ll N$), so that a large number ($N-k$) of trial predictions can be evaluated. The minimization equations are

$$\frac{\partial E}{\partial A_i} = 0 \quad (i = 1,2,3,\ldots), \quad (117)$$

[8]The caret (ˆ) is placed over quantities which are estimated or predicted, based on data and (usually) a set of parameters such as the B_i. It is to be distinguished from the angle brackets, used to denote the expected value, which is a statistical average, depending on the whole process (theoretical expectation) or on a realization of it (sample expectation).

[9]It should be emphasized that the word "predict" is not meant in the literal sense, as it would be, for example, if we were interested in real-time analysis of a manufacturing process we wished to control. Rather, we consider $\hat{X}_n$ to be the guess or estimate we would make for the value of X_n if we didn't know it, based on knowledge of values of X at other times. Conventionally, the restriction to the use of past data is imposed, but in general use can be made of past and future. (A two-sided prediction-error filter is sometimes called an interpolation operator.)

or

$$\left(\sum_{k=0}^{\infty} A_k X_{n-k}\right) X_{n-i} = 0, \qquad (118)$$

the expectation value of which is

$$\sum_{k=0}^{\infty} A_k \rho(k-i) = 0, \quad i = 1,2,3,\ldots, \qquad (119)$$

where ρ is the autocorrelation function. These are the standard Yule-Walker equations (Ulrych and Bishop 1975). The procedure is to use the data to compute an estimate of ρ, then solve equation (119) for the coefficients $A_1, A_2, \ldots$. The resulting A is minimum delay. Ulrych and Bishop give specifics and FORTRAN programs for carrying out this solution.

From the solution for A and the data, it is straightforward to calculate an estimate of the innovation from the relation in equation (115). Indeed, the sequence of prediction errors E_n corresponding to the optimum A is an estimate of the innovation R. That is, R is both the sequence of optimum prediction errors and the sequence of pulse amplitudes. This equivalence can be understood by noting that, with the correct A, there is no prediction error at time n due to pulses starting before n; the error is totally due to the *new* pulse, of amplitude R_n; hence $E_n = R_n$. This estimate, of course, is of the innovation corresponding to the specific realization of the processes which has been sampled, but therein is also contained information about averaged quantities, such as the pulse rate (which for a continuous distribution of amplitudes is expressible in the distribution function of pulse amplitudes). The Yule-Walker equations can be generalized to the case of two-sided filters, but this exercise is useless because it provides no added information.

A recursive procedure for determining A is due to Burg (1968, 1975) and is discussed by Ulrych and Bishop (1975), Fahlman and Ulrych (1975), Kanasewich (1975, pp. 260–283), Ulrych and Clayton (1976), and others. The sum of the squares of the forward and backward prediction errors of a one-sided prediction-error filter, namely

$$E_p = \frac{1}{2(N-p)} \sum_{n=p+1}^{N} \left[\left(\sum_{k=0}^{p} A_k X_{n-k}\right)^2 + \left(\sum_{k=0}^{p} A_{p-k} X_{n-k}\right)^2\right], \qquad (120)$$

is minimized with respect to A. The first term inside the brackets corresponds to the error made by the filter in predicting X_n based on the p preceding values $X_{n-1}, X_{n-2}, \ldots, X_{n-p}$. Since least-squares modeling cannot distinguish one sense of the direction of time from the other, Burg (1968) introduced the idea that one should include the backwards predictions, which are represented by the second term in equation (120). This term is the sum of the squares of the postdiction errors, made by the same filter (reversed) based on the subsequent values $X_{n+1}, X_{n+2}, \ldots, X_{n+p}$. The terms in the backward and forward contributions to E_p, when expanded out, are identical except for end effects. Thus Burg's idea is most important for short segments of data for which end effects are most important. This procedure explicitly assumes that the process X is intrinsically symmetric, in that forward and backward predictions need not be distinguished, and of course this is not generally true.

The limits of the n-sum are chosen such that no datum outside the sample range, $n = 1,2,\ldots,N$, is ever called for; that is to say, the estimate is noncommittal about the unsampled data. (In some formulations of such problems the unsampled data is set to zero.) Therefore the resulting parameter values are "maximum entropy" estimates (Burg 1968; Lacoss 1971; Ulrych 1972; Ables 1974). Hence A can be used to compute an estimate of the power spectrum (eq. [69]) of X (Burg 1967; Akaike 1969*a*, *b*, 1970*b*; Parzen 1968, 1969) which is called a maximum entropy method (MEM) spectrum. The nature of the predictions and the ranges of the summations are depicted in Figure 18. Details of the method are given by Ulrych and Bishop (1975) and more completely by Andersen (1974). The first of these references describes a convenient recursive solution to this least-squares problem. This is the Levinson (1947) recursion, also discussed by Durbin (1960) and Burg (1975). The resulting A is minimum delay, as with the Yule-Walker solution. Ulrych and Bishop discuss various practical matters, give a FORTRAN program for the determination of the AR coefficients as well as the spectrum, and outline the use of the final-prediction-error (FPE) criterion for the determination of the length of the (one-sided) AR filter.

This procedure is very efficient at determining the AR coefficients from time series data generated by simple processes where there is little noise present. It should probably be used if it is known *a priori* that the pulse is minimum delay. In astronomy this is seldom the case.

e) Predictive Deconvolution with the Absolute Value Norm

The choice of the sum-of-squares of the errors, in equations (116) and (120), is not the only possibility. Least-squares modeling is used because it gives maximum likelihood parameter estimates (Box and Jenkins 1970). It is also convenient because of the simplicity with which the minimization can be expressed in terms of the autocorrelation function (eq. [119]). But some other measure of the errors could be substituted for the

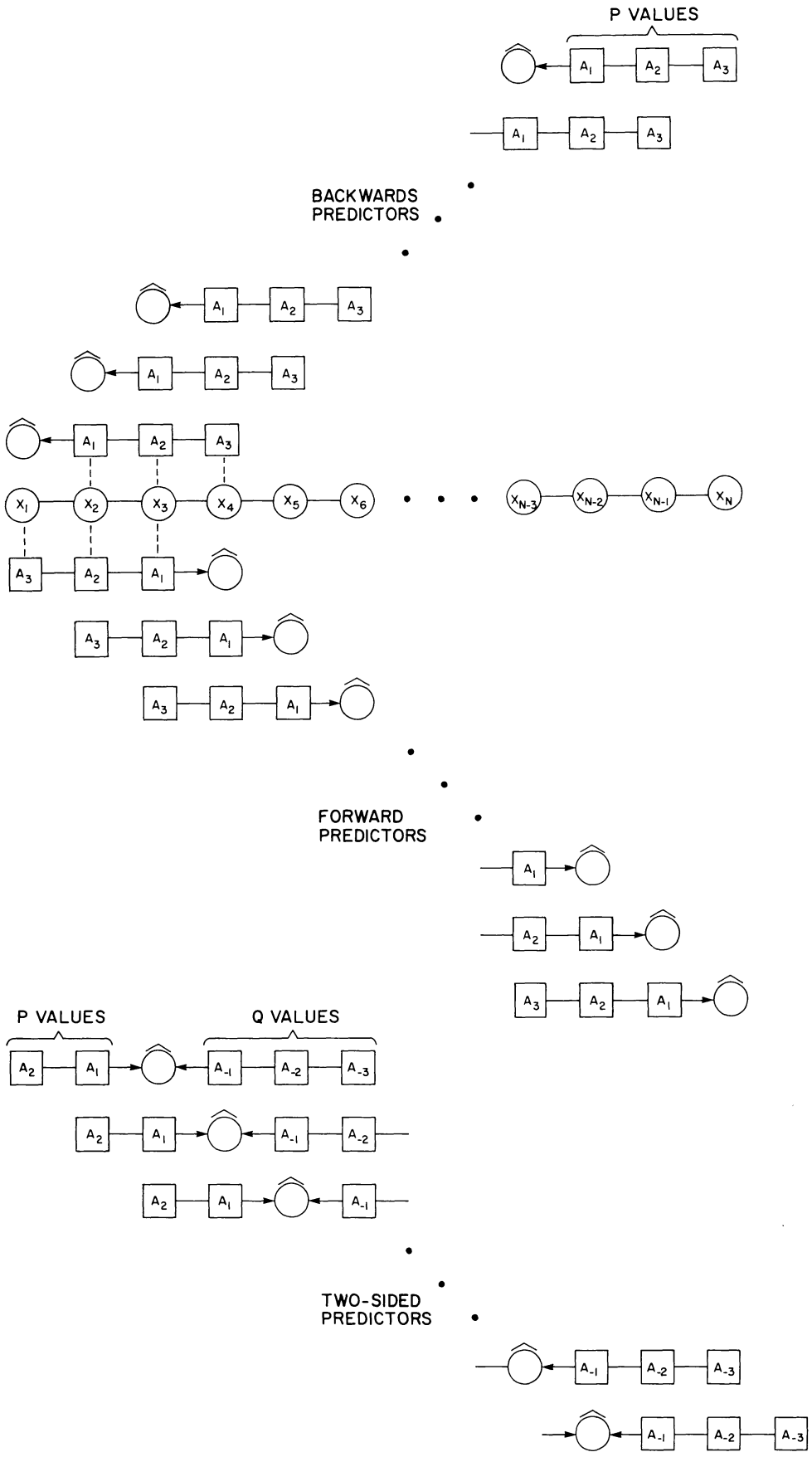

FIG. 18.—A prediction filter can be depicted as a string of numbers A_k which is to be overlaid on the time series to obtain the cross product $\hat{X}_n = \Sigma_{k \neq 0} A_k X_{n-k}$, which is the predicted value of X_n.

mean square error. The AR parameters could be determined by minimizing the more general form $E(A)=\|E\|$, where $\|E\|$ denotes an arbitrary error norm.[10] For example, consider the L_α norm:

$$L_\alpha(E)=\left(\sum_n |E_n|^\alpha\right)^{1/\alpha}. \tag{121}$$

The mean square error corresponds to $\alpha=2$. The usefulness of the choice $\alpha=1$ (Claerbout and Muir 1973; Scargle 1977) will now be demonstrated. Consider the MA process $X=R*C$, where R is an independently distributed process and C is the two-point pulse $(\mathbf{1}, c)$; if $|c|<1$ C is minimum delay, and if $|c|>1$ C is maximum delay. Introduce a two-point forward prediction-error filter $A=(\mathbf{1}, a)$:

$$E_n=X_n+aX_{n-1}, \tag{122}$$

i.e., the form of the prediction is simply

$$\hat{X}_n=-aX_{n-1}. \tag{123}$$

The best value of a minimizes $L_\alpha(E)$, which is equivalent to minimizing

$$[L_\alpha(E)]^\alpha=\sum_n |X_n+aX_{n-1}|^\alpha, \tag{124}$$

$$=\sum_n |(R_n+cR_{n-1})+a(R_{n-1}+cR_{n-2})|^\alpha, \tag{125}$$

$$=\sum_n |R_n+(a+c)R_{n-1}+acR_{n-2}|^\alpha. \tag{126}$$

This last expression is difficult to deal with because of the pulse overlap manifested in its three terms. But progress can be made if the pulse overlap is neglected, because its effects should average out. The prediction-error due to a single, isolated pulse at time n is

$$E_f=|R_n|^\alpha(1+|a+c|^\alpha+|ac|^\alpha). \tag{127}$$

But also consider the reversed or backward prediction-error filter $\tilde{A}=(a,\mathbf{1})$, which leads to the error

$$E_b=|R_n|^\alpha(|a|^\alpha+|1+ac|^\alpha+|c|^\alpha). \tag{128}$$

[10] Random processes can be considered as elements of a normed linear space [for L_2 this is a Hilbert space with the inner product $(X,Y)=\langle XY\rangle$]. A norm satisfies the three following conditions: (*a*) $\|X\|=0$ if and only if $X=0$; (*b*) $\|aX\|=|a|\|X\|$; and (*c*) $\|X+Y\|\le\|X\|+\|Y\|$. These are pleasant but not necessary properties for a measure of the errors or residuals in model fitting. For example, the skew "norm" of Claerbout and Muir (1973) does not satisfy (*b*) or (*c*), but it is still a useful penalty function for residuals.

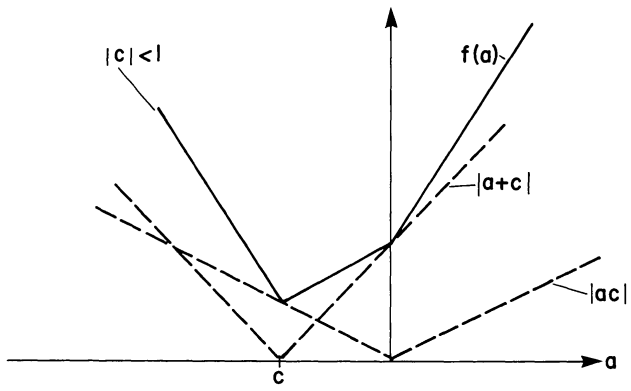

FIG. 19.—Graphs of terms which arise in absolute value minimization (*dashed lines*) showing how their sum (*solid line*) is a polygonal figure with vertices directly over the zeros of the individual terms. The minimum must occur at one of these vertices. More generally, the solution to a set of overdetermined linear equations in the "least absolute value" sense always solves one or more of the equations exactly.

It happens that $E_f=E_b$ if and only if $\alpha=2$. That is, least squares prediction is identical in the forward and backward directions and would yield the same result if the time series were reversed. The quantities E_f and E_b can be easily minimized if $\alpha\neq1$. For $\alpha=1$ consider the graph of $|a+c|+|ac|$ in Figure 19. Each of the two terms is a simple absolute value curve with a slope discontinuity at the point where its argument is zero. Hence the sum is piecewise linear, with vertices at these zeros. The minimum must fall at one of the vertices,[11] and simple comparison of the two values shows that for $\alpha=1$,

$$a_{f,\min}=\begin{cases}-c & \text{if } |c|<1\\ 0 & \text{if } |c|>1\end{cases}. \tag{129}$$

For $\alpha\neq1$

$$a_{f,\min}=-c(1+|c|^{\beta+1})^{-1}, \tag{130}$$

with $\beta\equiv(\alpha-1)^{-1}$, and a similar analysis of the backward case gives

$$a_{b,\min}=\begin{cases}-c|c|^{-1}(1+|c|^{\beta+1})^{-1}, & (\alpha\neq1);\\ 0, & \text{if } |c|<1\\ -c^{-1}, & \text{if } |c|>1\end{cases}\quad(\alpha=1). \tag{131}$$

[11] If $|c|=1$, the line between the vertices is horizontal and the minimum occurs everywhere along this line. This degenerate case is not important, because such a pulse has no stable AR representation anyway. It should be remembered, however, that absolute value minima are not always unique.

The values at minimum are related as follows:

$$E_f(a_{f,\min}) = 1 + |c|^2 < E_b(a_{b,\min}) = 1 + |c|, \quad \text{if } |c| < 1; \tag{132}$$

and

$$E_b(a_{b,\min}) = |c|^{-1} + |c| < E_f(a_{f,\min}) = 1 + |c|, \quad \text{if } |c| > 1. \tag{133}$$

Hence the minimum for either forward or backward prediction is at

$$a_{\min} = \begin{cases} -c, & \text{if } |c| < 1 \\ -c^{-1}, & \text{if } |c| > 1 \end{cases} \quad (\alpha = 1), \tag{134}$$

which gives for the optimum A, for $\alpha = 1$,

$$A = \begin{cases} (\mathbf{1}, -c), & \text{if } |c| < 1 \\ (-c^{-1}, \mathbf{1}), & \text{if } |c| > 1. \end{cases} \tag{135}$$

These solutions, for all values of α, are to be compared to the exact inverse of the pulse from which the process was formed,

$$A = C^{-1} = \begin{cases} (\mathbf{1}, -c, c^2, -c^3, \ldots), & \text{if } |c| < 1; \\ (\ldots, -c^{-3}, c^{-2}, -c^{-1}, \mathbf{1}), & \text{if } |c| > 1. \end{cases} \tag{136}$$

The two-term L_1 solution in equation (135) agrees with the first two terms of this exact result. For $|a| \ll 1$ the filter $(\mathbf{1}, a_{f,\min}) = [\mathbf{1}, -a + o(a^2)]$ is approximately correct for any $\alpha > 1$, and similarly for $|a| \gg 1$ the filter $(a_{b,\min}, \mathbf{1}) = [-a^{-1} + o(a^{-2}), \mathbf{1}]$ is a good approximation. The inequalities in equations (132) and (133) hold for $1 < \alpha < 2$, but the opposite sense inequalities hold for $\alpha > 2$ (with equality for $\alpha = 2$, as already noted). Thus any L_α norm with $1 \leq \alpha < 2$ makes the correct decision between minimum delay and maximum delay, but $\alpha \geq 2$ is unsatisfactory. The best choice is $\alpha = 1$, for at least in this example the resulting parameter values are then most accurate.

This demonstration of the phase-determining ability of the absolute value norm is for the simple case of a first order AR process. More general cases are difficult to treat analytically, but there are many good numerical techniques for absolute value minimization (Barrodale and Roberts 1973, 1974; Osborne and Watson 1971; Barrodale, Roberts, and Hunt 1970; Barrodale and Young 1966; Robers and Ben-Israel 1969; Barrodale 1970; Ekblom and Henriksson 1969; Rice and White 1964; Maria and Fahmy 1974; and Claerbout and Muir 1973). Numerical tests (Scargle 1977, and § VI*a*) show that the L_1 norm does work for more complicated cases, as long as the driving process R is at least moderately nonnormal. But a difficulty arises when two-sided filters are introduced, as they must be for this problem.

In the above example, permitting either forward or backward prediction was crucial to the phase determination. In more complicated cases, for example when the pulse is mixed delay, the obvious generalization is to allow A to be two sided. For example, if $A = (a, \mathbf{1}, b)$, the predictor is

$$\hat{X}_n = -aX_{n+1} - bX_{n-1}, \tag{137}$$

and the prediction-error sequence is

$$E_n = X_n - \hat{X}_n = aX_{n+1} + X_n + bX_{n-1}, \quad \text{i.e., } E = A * X. \tag{138}$$

It is here that the liberal interpretation of the word "prediction" noted above first comes into play. The general forms are

$$A = (a_{-q}, \ldots, A_{-1}, \mathbf{A_0}, A_1, \ldots, A_p) \tag{139}$$

and

$$E_n = \sum_{k=-q}^{p} A_k X_{n-k}. \tag{140}$$

The corresponding mean error in the L_α norm is

$$E(A) = \frac{1}{N-p-q} \sum_{n=p+1}^{N-q} |E_n|^\alpha, \tag{141}$$

where as usual the sum is such that the filter A never extends outside the sampled data (see Fig. 18). The optimization problem is to find the minimum of this expression with respect to the parameters $A_{-q}, \ldots, A_p$. It seems natural to not regard A_0 as a free parameter, but to fix it at the value 1 because of the special nature of the prediction point. The condition $A_0 = 1$ can be thought of as a normalization condition imposed on A to avoid the trivial minimum at $A_i = 0$, all i. However, this normalization choice is inappropriate for two-sided deconvolution problems. Consider the MA process $X = R * C$, where C is some particular two-sided pulse. The mean L_2 prediction error is

$$E(A) = \frac{1}{N-p-q} \sum_{n=p+1}^{N-q} \left[\sum_{k=-q}^{p} A_k (R * C)_{n-k} \right]^2, \tag{142}$$

and therefore

$$\frac{\partial E(A)}{\partial A_i} = \frac{2}{N-p-q} \sum_n \left[\sum_k A_k (R*C)_{n-k} \right] (R*C)_{n-i}, \tag{143}$$

for $i \neq 0$. Now insert the desired solution $A = C^{-1}$:

$$\left. \frac{\partial E(A)}{\partial A_i} \right|_{A=C^{-1}} = \frac{2}{N-p-q} \sum_n R_n (R*C)_{n-i}, \tag{144}$$

$$= \frac{2}{N-p-q} \sum_n R_n \sum_k R_k C_{n-i-k}, \tag{145}$$

the expectation value of which is

$$\left\langle \left. \frac{\partial E(A)}{\partial A_i} \right|_{A=C^{-1}} \right\rangle = \frac{2}{N-p-q} \sum_n \sum_k \langle R_n R_k \rangle C_{n-i-k}, \tag{146}$$

$$= 2\sigma_R^2 C_{-i}, \tag{147}$$

since the R_n's are mutually uncorrelated. This expression is not zero unless C_{-i} vanishes, so that in general the A which is the correct inverse pulse, namely C^{-1}, does not solve the optimization problem with the constraint $A_0 = 1$. On the other hand, if only one-sided pulses are allowed $C_{-i} = 0$ for all $i > 0$, and the desired A does make the above derivatives zero; this A does solve the optimization problem. The choice $A_0 = 1$ is correct for causal pulses but not for two-sided ones.

What can be done to ensure that the solution of the minimization problem is the correct inverse pulse? If A_0 is an arbitrary function of the other A_n's, rather than held constant, the above analysis yields, instead of equation (147), the set of equations

$$C_{-i} + C_0 \frac{\partial A_0}{\partial A_i} = 0, \tag{148}$$

for $i \neq 0$, which integrates to

$$\sum_i A_i C_{-i} = 1, \tag{149}$$

as a necessary condition that optimization of A yield the inverse of C. (Note that for causal filters this reduces to $A_0 C_0 = 1$, and the conventional constraints $A_0 = C_0 = 1$ are correct.) Unfortunately, equation (149) cannot be considered as a simple constraint on A because it involves the unknowns C_i. An obvious possible remedy is to compute iteratively, starting with a guess for C, imposing the constraint in equation (149) on the minimization to produce a new A and a new $C = A^{-1}$. The convergence and uniqueness properties of this iteration have been studied in numerous simple cases. For low-order processes with little noise it converges very rapidly to a unique minimum which is very much better than the solution with $A_0 = 1$. But for more difficult problems there tend to be oscillations. In some cases these can be damped out very effectively by adopting a suitable averaging scheme for the update of A. But a way has not been found to predict ahead of time which of several such averaging procedures will succeed on a given set of data, nor a single procedure that is successful on all data. (Curiously, although the above derivation is for L_2, the same results can be demonstrated for L_1 using the methods of Rice 1964.)

An interesting feature of the above iteration is that, since none of the A_n's are constrained to equal one, the identification of the prediction point becomes vague. Indeed the very concept of a specific point singled out as the prediction point loses much of its significance. But let us call the element of A largest in absolute value the prediction point, just because it is often true with the constraint $A_0 = 1$ that $|A_i| < 1$ for all $i \neq 0$. It is found that as the iteration proceeds, this point is not fixed but moves around within the filter and eventually converges to a fixed point. This is favorable as it eliminates what would otherwise be an arbitrary parameter (the location of the prediction point, or MPT in the terminology in the appendix). The length of A, however, is still arbitrary. For least-squares problems Akaike (1970*a*) has shown how the length can be determined in an objective, automatic way, based on the FPE criterion. This technique introduces the quantity

$$\mathrm{FPE}_M = \frac{(N+M)}{(N-M)} S_M^2, \tag{150}$$

where S_M^2 is the sum-of-squares of the residuals (i.e., of the innovation), N is the number of data points, and M is the number of free parameters in the model (including one for the mean value if this has been subtracted from the data before analysis). Starting from small values, M is increased until the FPE (for final prediction error) stops decreasing and begins to increase. One can interpret the factor $(N+M)/(N-M)$ as the statistical penalty that should be paid for using more free parameters. Without such a penalty, the residuals would always decrease as the number of parameters increases, so that the FPE going through a minimum is the signal that diminishing returns has set in. Other techniques have been proposed (see especially Gray, Kelley, and McIntire 1977), but none seems easily generalized to the L_1 case. However, empirically it has proven satisfactory to simply replace S_m^2 in equation (150) with S_M^1, the sum of the absolute value residuals. No theoretical justification for this procedure has been found, and it

should be regarded as an empirical result with quite meagre support in numerical experiments. Short of using this, the magnitude of the residuals and the values of the model parameters must be inspected as the complexity of the model increases. It has been said that so much judgment is necessary in such matters that the procedure should not be attempted for the first time (Granger and Newbold 1977). This seems extreme, but some limit must be placed on the order of the model to avoid the pitfalls of fitting too many parameters.

This section concludes with the one analytical result uncovered for the L_1 problem which is as close as possible to showing that absolute-value optimization of a two-sided AR filter yields the correct deconvolution of a MA process driven by independently distributed noise. First the following lemma is established:

> Lemma: If X and Y are zero-mean, independently distributed processes, then
>
> $$\langle|X+Y|\rangle \geq \max(\langle|X|\rangle, \langle|Y|\rangle), \tag{151}$$
>
> with equality if X or Y is the null process.

First note that if X and Y are independent

$$\langle|X+Y|\rangle = \int\int dx\,dy P_2(x,y)|x+y|$$

$$= \int\int dx\,dy P_X(x)P_Y(y)|x+y|, \tag{152}$$

$$= \int dx\left[-\int_{-\infty}^{-x}(x+y)P_X(x)P_Y(y)\,dy + \int_{-x}^{\infty}(x+y)P_X(x)P_Y(y)\,dy\right], \tag{153}$$

$$= \int_0^{\infty} dx\left[-2\int_{-\infty}^{-x}(x+y)P_X(x)P_Y(y)\,dy + \int_{-\infty}^{\infty}(x+y)P_X(x)P_Y(y)\,dy\right] + \int_{-\infty}^{0} dx\left[-\int_{-\infty}^{\infty}(x+y)P_X(x)P_Y(y)\,dy + 2\int_{-x}^{\infty}(x+y)P_X(x)P_Y(y)\right]. \tag{154}$$

The first and last of the four terms in this equation can be written as the integral of a nonnegative quantity, as follows

$$Q = 2\left[\int_0^{\infty} dx\int_{-\infty}^{-x}|x+y|P_X(x)P_Y(y)\,dy + \int_{-\infty}^{0} dx\int_{-x}^{\infty}|x+y|P_X(x)P_Y(y)\,dy\right] \geq 0. \tag{155}$$

In fact this quantity vanishes only in degenerate cases, the most important ones being P_X or $P_Y \equiv 0$. The second and third terms in equation (154) simplify:

$$\langle|X+Y|\rangle = \int_0^{\infty} dx[xP_X(x)+\langle Y\rangle P_X(x)] + \int_{-\infty}^{0} dx[-xP_X(x)-\langle Y\rangle P_X(x)]+Q, \tag{156}$$

$$= \langle|X|\rangle + \langle Y\rangle\int_{-\infty}^{\infty}\mathrm{sign}(x)P_X(x)\,dx+Q, \tag{157}$$

$$= \langle|X|\rangle + Q \quad (\text{since } \langle Y\rangle = 0), \tag{158}$$

and so except in the degenerate cases in which $Q=0$

$$\langle|X+Y|\rangle > \langle|X|\rangle. \tag{159}$$

Since X and Y are interchangeable in the above analysis, the result stated in the lemma follows.

Turning to the main issue, consider the process $X = R*C$; we wish to show that $\langle|A*X|\rangle$ is minimum if $A = C^{-1}$ (subject to the condition in eq. [149]). Write

$$A = C^{-1} + \delta A, \tag{160}$$

so that

$$A*X = (C^{-1}+\delta A)*C*R, \tag{161}$$

$$= R + (\delta A * C)*R, \tag{162}$$

or

$$(A*X)_n = R_n + \sum_k a_k R_{n-k}, \tag{163}$$

where

$$a_k = \sum_m \delta A_m C_{k-m}, \tag{164}$$

and equation (149) gives

$$a_0 = 0. \tag{165}$$

Hence

$$\langle|(A*X)_n|\rangle = \left\langle \left| R_n + \sum_{k\neq 0} a_k R_{n-k} \right| \right\rangle . \quad (166)$$

But since R is independently distributed, the sum in this equation is distributed independently of R_n, and the lemma applies, to give

$$\langle|A*X|\rangle \geq \langle|R|\rangle , \quad (167)$$

with equality if $a_k = 0$ for all k, a condition equivalent to $\delta A_k = 0$ for all k. Hence $A = C^{-1}$ gives a minimum (not necessarily unique), and we have established the

> *Theorem*: If $X = R*C$, with R independently distributed noise, then $A = C^{-1}$ is a solution of the optimization problem $\min_A \langle|A*X|\rangle$ subject to $\Sigma_k A_k C_{-k} = 1$.

It must be cautioned that this minimization problem is not specified in the usual way, because the constraint explicitly involves the solution, and the theorem is to be understood in the sense indicated in its proof. The practical value of this result is in the iterative method which it leads to, as described earlier.

V. COMPUTATIONAL METHODS

The goal of this section is to provide enough computational details so that the reader can apply the techniques described above to his own data. In outline all of the methods to be discussed proceed in the same way, as follows:

a) Obtain the data.

b) Decide on the form of the model (AR, MA, ARMA, ARIMA,...).

c) Provide a way of computing the innovation R as a function of the model parameters.

d) Choose the property of R to be minimized, and provide a scheme for evaluating the corresponding norm $D(R)$.

e) Minimize $D(R)$ with respect to the model parameters.

f) Compute the physically interesting quantities from the optimum model found in the previous step.

The following subsections explain these steps in turn.

a) *Sampling*

Assume that the sampling is in even intervals of the independent variable (time, position, wavelength,...) so we have a set of measured numbers X_n, $n = 1, 2, \ldots, N$. This is not a fundamental limitation, however, as the techniques described here can be readily generalized to data with gaps or uneven sampling or both (§ V*g*).

b) *Identification*

Choosing the best form (which is traditionally called *identification*) of the model is not always straightforward, and there is a large and complex literature on this problem (see, e.g., Box and Jenkins 1970; Parzen 1974; or Granger and Newbold 1977). We will not attempt to summarize the ideas in this literature, but the following general comments are appropriate. Many astronomical time series can be well represented as low-order AR processes, and this discussion therefore emphasizes AR models. Remember that a given process can be represented in a variety of ways (§ II*e*), so identification should not be viewed as finding the True Model, but as finding a simple, physically suggestive model which adequately represents the observations. Also keep in mind that this step is not irrevocable, once taken. Rather, the results of subsequent steps often suggest some revision in the form of the model.

c) *Computing the Innovation* $R(A)$

The relation used depends on the form of the model (eq. [90]). For an ARIMA model, the data are differenced d times and then an ARMA model is fit. The most direct way of computing R is to carry out the operation in equation (90) with the discrete Fourier transform:

$$R = \mathcal{F}^{-1}\left[\frac{\mathcal{F}(X)\mathcal{F}(A)}{\mathcal{F}(C)} \right] . \quad (168)$$

Note that C enters this calculation effectively as its inverse, so that even here the MA part of the model is converted into an autoregressive representation. The only points that are not straightforward in implementing this expression with the DFT are: complex arithmetic must be used in the multiplications and divisions, and the arrays A and C must be zero-extended to the same length as the data before applying the transformation. It would seem that the result would be of the same length (i.e., N points), but to avoid spurious end effects the array R must be truncated somewhat, depending on the length of A and C. These end effects arise whether the innovation is calculated with the DFT or directly evaluated with a summation (see, eq. [91] for the pure AR case). In either case the innovation is defined at slightly fewer than N points. This is the reason for the limits N1D and N2D in the FORTRAN code provided in the appendix. For the pure AR case, R is defined at $p+q$ (=the length of the AR filter -1) fewer than N points. But the values are not N1D$=p+1$ and N2D$=N-q$, as would be expected from equation (91), simply because negative values of indices are not permitted in FORTRAN. In the code in the appendix R is computed as outlined above, using the DFT. Alternatively, the sum in equation (91) may be directly evaluated; for small values of p and q this procedure is

faster than the use of the DFT. However, the evaluation of R is a minor part of the computation of D. For convenience the R_n are reindexed as $\{R_n, n=1,2,\ldots, N^*\}$, where $n=1$ corresponds to $n=p+1$ in equation (91), and $N^*=N-(p+q)$. It is important that A be prohibited from running over the ends of the data (see Fig. 18), to avoid the numerically harmful end effects (i.e., to preserve the "maximum entropy" condition, § IV*d*).

d) *The Computation of* $D_F(R)$

The choices for the property of R to be minimized include dependence (§ IV*c*), the martingale difference property § IV*c*), the mean-square prediction error (§ IV*d*), and the mean absolute prediction error (§ IV*e*). Another example is a measure of simplicity called the *varimax norm* (Kaiser 1958; Wiggins 1977; Ooe and Ulrych 1979). In turn there are several ways to implement each of these. For example, we saw above that dependence could be measured in terms of differential cumulative probability distributions, moments, characteristic functions, or expectations of arbitrary functions. Since the scheme involving cumulative distribution functions proved much the most satisfactory, details of the other approaches have been omitted. The remarks about them in § IV should enable the interested reader to construct algorithms implementing the other approaches. Test results with all of the methods save those using moments (messy) and expectations of arbitrary functions (not tested), will be given in § VI for comparison.

The function to be minimized is defined in equation (98). Because of the step-function nature of the estimates of F_1 (eq. [94]) and F_2^m (eq. [96]), this integral can be evaluated exactly with a finite double sum, as we shall now see. It is convenient to introduce an ordered version of the R_n; i.e., define an index transformation $i=f(n)$ such that if $R_i'=R_{f(n)}'=R_n$, then the R_i' form an ordered set:

$$R_1' \le R_2' \le R_3' \le \cdots \le R_{N^*-1}' \le R_{N^*}'. \quad (169)$$

As long as R_{n+m} is associated with its correct neighbor in the unordered set, namely R_n, then the integrand in equation (98) is unchanged by this ordering. The integral may be written as

$$D_F^m(A)=\sum_{i=1}^{N^*-1}\sum_{j=1}^{N^*-1}|F_2^m(R_i', R_j') - F_1(R_i')F_1(R_j')|^2\,\Delta R_i'\,\Delta R_j', \quad (170)$$

where $\Delta R_i'=R_{i+1}'-R_i'$. This sum is over a two-dimensional (unevenly spaced) grid of rectangles with area $\Delta R_i'\,\Delta R_j'$, and with edges at the values R_i', $i=1,2,\ldots, N^*$ (see Fig. 39 in the Appendix). From definitions (94) and (96) it can be seen that $F_2^m(R_i', R_j')$, $F_1(R_i')$, and $F_1(R_j')$ are all constant over each of these rectangles and therefore so is the summand, $F_2^m - F_1F_1$. Hence, the sum in equation (170) is an exact evaluation of equation (98). Of course, the expressions for F_1 and F_2^m are inexact estimates of the corresponding quantities. However, they exactly represent all the information contained in the given realization of R; this is not true of the estimates of P_1 and P_2, since there is always some loss of information in a binned histogram. This is probably the main reason for the superiority of the cdf approach. The advantage of R', the ordered version of R, is that the summand can be computed recursively, for example, with

$$F_2^m(R_i', R_j')=F_2^m(R_{i-1}', R_j') + \frac{1}{(N^*-m)}H\left[R_j'-R_{f^{-1}(i)+m}\right], \quad (171)$$

(H is the step function defined in eq. [95].) This relation follows from the fact that no more than one new step in F_2^m begins at a given value of R_i', corresponding to a given row in the matrix (R_i', R_j'). Further discussion of this recursion is in the appendix.

e) *Minimization of* $D_F(A)$

The minimum of D_F, with respect to the filter elements A, can be found with any of several standard numerical techniques; the simplex method is described here because it is the one the author happened to use, not because it has been proven to be more suited to this problem. The following warning should be issued with the simplex method (Nelder and Mead 1965; Powell 1964): After the convergence criteria have been satisfied, a restart should be made to check the possibility that the simplex has become degenerate or is otherwise unable to progress toward the true minimum. A *restart* is a reinitiation of the iteration with a new simplex at the point to which the procedure appears to have converged (see the Appendix). Another caution is that D_F may have more than one local minimum. With numerical techniques it is never possible to be certain that the global minimum has been achieved. But the expression for D_F in equation (98) is far superior in this regard to all of the other methods tried and to other ways of estimating the cumulative probability functions. For data generated by a simple process and fit by simple models (e.g., AR[1, 1]), D_F has never been found to have more than one minimum, and the simplex rapidly converges to the (global) minimum from essentially any starting value. The convergence is also very sure in that a restart is never needed. (Nevertheless it is wise to try a restart in all cases, even if it is expected that it will not be necessary.) As the order of the fitted model is increased three symptoms eventually appear:

(1) local minima abound; (2) restarts are frequently necessary (i.e., false convergence becomes common); and, not surprisingly, (3) convergence is generally slower.

Experiments have shown that the first two of these problems are eliminated if m^* is increased sufficiently, typically to a value slightly less than $p+q$. Because the time to evaluate D_F is roughly proportional to m^*, the computation time increases as m^* is increased, but the reward in sureness of convergence, elimination of spurious local minima, and accuracy of the solution is certainly worth the price. For a given data set, the procedure found to be best is as follows:

(A) *Fit a very low-order model*, such as AR(1, 1), with $m^*=1$. Unique and sure convergence has always obtained at this step, but caution suggests that one: (*a*) experiment with a variety of starting values, such as $A=(0,\mathbf{1},0)$, $(1,\mathbf{1},1)$, or the $(a_{-1},\mathbf{1},a_1)$ which gives the L_1 minimum (i.e., minimum of $\Sigma_n|R_n|$); (*b*) try a variety of sizes for the initial simplex; and (*c*) always try restarts, with moderately large simplexes. Hopefully these steps will not be necessary, and the results will be the same for all starting solutions and simplexes. However, since ill-conditioning tends to grow with the model complexity, confidence in the good behavior of the procedure at this stage is essential. If there are convergence or uniqueness problems at this early stage, there are several possibilities: (i) The process is not stationary, the ∇ should be applied one or more times before modeling is attempted; (ii) an even simpler model should be used to start with, such as AR(0, 1) or AR(1, 0); (iii) a totally different form should be tried, such as MA or ARMA; or (iv) the value of m^* should be increased (see step D below).

(B) *Increase the order of the model.* A good way is to compare the results for $p\rightarrow p+1$ and for $q\rightarrow q+1$, using as starting values the solution from step A with zero for the new parameter. Of these two models, adopt the one which gives the lower value of D_F. (Remember that restarts and multiple initial solutions are never out of place. The appearance of false minima turned up by restarts or multiple minima turned up by various initial solutions are symptoms that m^* is too small and should be increased.)

(C) *Step B should be repeated until there is indication that the correct order has been reached*, for example, until (*a*) the parameters from the lower order solution do not change, and the new parameter is relatively small, or (*b*) the residuals stop decreasing with increasing order—more properly, the residuals should decrease only as much as would be expected from the mere fact that another parameter is varied.

(D) *Increase the value of m^* and repeat steps A–C.* If the results do not change significantly with m^* it can be presumed that the value used is large enough.

The format of Tables 4–7 follows this scheme.

Determining the correct order of the model is important. If the order is taken too small, there will be residual serial correlation in the estimated innovation, indicating that not all of the information about the process has been extracted. In spectral analysis the symptom of too small an order is that the spectrum is heavily smoothed—the frequency resolution has been degraded by using too few parameters. In deconvolution the pulse shape is similarly oversmoothed. In principle, taking the order too large is not as harmful because the extra parameters will be very small (provided there are enough data). In practice, however, even a few too many parameters cause numerical difficulties and add greatly to the cost of the computations. If the number of parameters becomes of order N (heaven forbid!), the estimates all become unstable because there are too few terms in the corresponding sums. In general, too many parameters show up as large spurious spikes in the power spectrum or as wild oscillations or other erratic behavior in the pulses. There are many approaches to the order problem in the classical least-squares arena (e.g., Chow 1972*a*, *b*, *c*; Anderson 1963; Jenkins and Watts 1969; Akaike 1970*a*; Gailbraith 1971; Lindberger 1972; Parzen 1974; Jones 1974; Graupe, Krause, and Moore 1975; and Tong 1975). Also an innovative approach has been developed by Gray, Kelly, and McIntire (1977). It is not surprising that the same difficulties confront modeling with independently distributed innovations, as the models are identical. The steps A–D above, based on experience with both test cases and real data, are offered as guidelines only. It is hoped that an objective technique such as the FPE (see § IV*e*) can be developed. Toward this goal, the quantity FPE_M in equation (150), with S_M^2 replaced by D_F, is routinely tabulated (see § VI). In some cases this quantity can be helpful in

deciding when the order is correct, but it is far from infallible. When using the suggestion given above (step B) for increasing the order of the model, the FPE will be systematically underestimated, because the smaller of two values of D, corresponding to the two choices for the location of the new parameter, is selected. This could cause the quasi-FPE criterion to overestimate the order of the model, as occurs in the examples in § VI.

A note about multiple minima: For a given total order (e.g., $p+q$ for the model AR[p, q]) there will be distinct minima for each of the possible choices of p and q. (For example, if $p+q=3$, the four possibilities are AR(0,3), AR(1,2), AR(2,1), and AR(3,0).) With the current algorithm the prediction point cannot move during the minimization, so that all of these choices are separate problems. It would be helpful if a scheme to allow automatic migration of the prediction point could be developed, as with the L_1 minimization with a pseudo-constraint (§ IV*e*). Then all of these problems (with a given total order, $p+q$) could be solved together with a single minimization. In lieu of such a procedure one must simply compare the minima for the various choices. Some judgment can be used here; for example, if a model of the form AR(1,2) yields the grand minimum for $p+q=3$, it is unlikely that AR(4,0), or even AR(3,1), will give the grand minimum for $p+q=4$.

All of these matters will be illustrated in the examples in § VI.

f) *Computation of Subsidiary Quantities*

The point of this section is that the model parameters estimated in steps (1) to (5) are not necessarily the most interesting numbers in the physical interpretation of the data. For example, as already mentioned, the AR parameters are often the most easily and directly calculated, but the MA pulse shape is the quantity for which there is a physical theory. (For example, if quasar light fluctuations are due to supernovas, the pulse shape should resemble the supernova light curve.) Hence one of the transformations that is useful is $A \rightarrow C$. The direct way to carry this out is to compute

$$C = A^{-1} = \mathcal{F}^{-1}\left(\frac{1}{\mathcal{F}(A)}\right), \tag{172}$$

using the discrete Fourier transform, as discussed at length in § III*f* and explicitly shown in the Appendix. But there is another way of evaluating the MA parameters, namely with the relation

$$C = X * \tilde{R} = \tilde{A} * X * \tilde{X} \text{ (with } \langle X \rangle = \langle R \rangle = 0), \tag{173}$$

where the tilde over a variable indicates the time reverse of that variable. Indeed, this is the form used in the constructive proof of the (Wold) existence theorem for the MA pulse (see eq. [78]). It can be thought of as the "superposed epoch" method (e.g., Gosling *et al.* 1972) because the convolution in equation (173), rewritten as

$$\hat{C}_n = \sum_k X_{n+k} R_k, \tag{174}$$

represents the operation of shifting each pulse to bring its origin to a common point in time and then averaging with a weight proportional to the pulse amplitude. All of the other, overlapping pulses are added in, too; their contribution averages to zero because they are uncorrelated with each other, but the pulse which has been shifted to the common origin always adds in phase. The cancellation of the random overlapping pulses requires that the mean of X be zero, which explains the need for $\langle X \rangle$ and $\langle R \rangle$ to be zero in equation (173). This relation can be proved by noting that if $X = R * C$, then

$$X * \tilde{R} = C * (R * \tilde{R}). \tag{175}$$

But the expectation value of $R * \tilde{R}$ is a delta function, so that the expected value of the right-hand side of equation (175) is just C. Of course the estimate of $R * \tilde{R}$ for any realization is not exactly a delta function, but will contain zero-mean noise for nonzero lags. (One can use the symmetry of $R * \tilde{R}$ to aid in distinguishing this noise from the tails of the pulse.) The estimate in equation (173) has several advantages over the simpler form in equation (172): A^{-1} is a smoothed estimate, especially if it has a small number of parameters, and to some extent it conceals the uncertainties in the pulse shape. Because equation (173) invokes the data directly, the resulting pulse is less smoothed than A^{-1} and thus provides a better feeling for the variance of the values of the elements C_n. Another shortcoming of the direct inversion is that it is nonlinear in A and thus is a biased estimate. For example, if X were white noise, the expected value of A is a delta function (at least for some ways of determining it; see § VI*d*). But A^{-1} contains quadratic and other even powers of the A_k which do not have zero expectation value, hence $\langle A^{-1} \rangle$ is not a delta function as it should be. In practice this bias is not important for most problems.

Another interesting quantity is the estimate of the innovation,

$$\hat{R} = \hat{A} * X, \tag{176}$$

which is computed every time $D(R[A])$ is. Of course $\hat{R}$ is a sample estimate and refers to the pulse amplitudes in the particular realization of the data at hand. It is the best (optimum) estimate of the amplitudes with

which the pulses, C, occurred to produce the observed realization. Note that since $\hat{R}=\hat{A}*X$, if $\hat{C}=\hat{A}^{-1}$ it follows that $X=\hat{R}*\hat{C}$ exactly. That is, the model has sufficiently many degrees of freedom to reproduce the sampled data exactly. There is thus never any question of how well the data is fit. The questions are: How random (independent) is the estimated pulse amplitude sequence? How physically reasonable is the estimated pulse shape? The amplitudes may be less interesting than their distribution, so it is often useful to construct a histogram which is an estimate of the amplitude distribution.

One can also readily compute the autocorrelation function and power spectrum of X, directly from A (see eqs. [65] and [69]).

g) *Gaps and Uneven Sampling*

Any technique based on prediction-error filters can be readily adapted to data which do not have the simple sampling assumed in § V*a*, for there are ways of generalizing the concept of the output of such filters with the input data unevenly sampled.

Consider first even sampling with one or more gaps. The case of one gap is easily generalized to an arbitrary number. We describe one gap in terms of two index sets for the independent variable:

$$\{X_n;\, n\in S_1, n\in S_2\}. \tag{177}$$

For example, a gap of length m could be represented with $S_1=(1,2,\ldots,N_1)$ and $S_2=(N_1+m+1, N_1+m+2,\ldots,N_2)$. There are two subcases as given in the following two paragraphs.

i) *No Coherence across the Gap*

There are situations where the length of the gap is unknown (so that the second segment cannot be phased relative to the first), the gap is not an integer number of the sampling intervals, or where it is believed (or assumed) that the process is not coherent across the gap. For example, in a pure MA process there is no coherence across a gap wider than the total extent of the pulse. Even if the pulse is infinite, the coherence will diminish rapidly as the gap exceeds, say, twice the FWHM of the pulse. The case of no coherence is the easiest to handle. One simply redefines the function D as a sum over the index sets taken separately. That is, if $D^i(A)$ is the norm evaluated on the data for index set i, treated as if these were the only data available, then define

$$D(A)=\sum_i D^i(A), \tag{178}$$

where the sum is over all the relevant index sets. The minimization of this total D is exactly as before.

ii) *Coherence across the Gap*

It is rare that information is coherent across anything but a small gap, the most notable exception being signals consisting of phase coherent sinusoids or other deterministic functions. If it is desired to retain such information, the technique just outlined cannot be used, as the filters are never applied to data on both sides of the gap simultaneously. The basis of a method for such cases has been suggested independently by several workers: Use (one-sided) prediction error filters to fill the gap(s), and then optimize a new filter on this interpolated data. There are various choices as to how to merge the predictions (one from the right and one from the left) at the center of the gap. An example of this technique is given by Ulrych and Clayton (1976).

iii) *Arbitrary Sampling*

Consider the case where there are not just a few gaps in otherwise even sampling, but where the time points, $\{t_n\}$, are arbitrary (see §§ I and II). Discrete AR representations are applicable only to the special case where the sampling times are evenly spaced, because the optimization requires sliding the filter along the data (see Fig. 18). But the simple generalization to continuous filters allows arbitrary sampling. The prediction error, given in the discrete case by equation (114), is

$$R_n=X_n+\int X(s)A(t_n-s)\,ds, \tag{179}$$

and the integral is replaced by a sum, yielding

$$R_n=X_n+\sum_{k\neq n} X(t_k)A(t_n-t_k)\,\Delta t_k. \tag{180}$$

Since $A(t)$ is continuous, it does not matter that the intervals t_n-t_k are not all the same. To parameterize the function A, so that the optimization can be carried out with respect to a set of discrete parameters rather than a continuous function, introduce the expansion

$$A(t)=\sum_k A_k\phi_k(t), \tag{181}$$

where the $\phi_k(t)$ are a set of continuous functions which must be specified. The problem has been reduced to the same form as before—the innovation defined (by eqs. [180] and [181]) in terms of a discrete set of parameters, $\{A_k\}$. The optimization can be carried out as before, and the pulse shape and amplitude sequence, autocorrelation function, or power spectrum can be evaluated much as before. The author has carried out limited experiments with the choice $\Delta t_n=1/2(t_{n+1}-t_{n-1})$ and the expansion (eq. [181]) given as either a Fourier series or a power series. While encouraging, the

results will not be described here as there was moderate dependence on the choice of the functions $\phi_k(t)$, the number of terms kept in equation (181), the length of the operative time interval (i.e., the number of values over which the k-sum in eq. [180] is evaluated), etc. Good methods of selecting these must be developed.

VI. NUMERICAL EXPERIMENTS

The best way to evaluate a deconvolution procedure is to try it out on artificially generated data of known characteristics. All of the test problems described here are low-order autoregressive processes, with specific choices for A. The time series were actually generated by filtering an innovation R with the inverse $C=A^{-1}$ (thus representing the process as a high-order moving average). The innovations are of the form $R=U^n$; U is a sequence of independent uniformly distributed random numbers. In the earlier examples (Fig. 5) the interval for U was taken to be $(-1/2,1/2)$, but to simulate positive-only amplitudes we now take it to be $(0,1)$. U^n means simply that U is raised to the nth power, term by term. As seen in Figure 5, a large value of n gives a few large amplitudes and approximates the shot noise process. The other limit, small n, corresponds to much pulse overlap (i.e., many large amplitudes instead of a few) and takes on the appearance of a normal process. The higher the value of n, the less pulse overlap there is and the easier deconvolution should be. In the extreme case of normally distributed R the overlap is so great (to the point that $X=R*C$ is also normally distributed) that no method can recover phase information, and the deconvolution problem as meant here (i.e., with correct phase) is intrinsically unsolvable. Any technique should give progressively worse results as n is decreased and should be completely unable to recognize phase properties as R approaches normalcy. These expectations are borne out by the experiments about to be described. White noise with several variances is added to some of the test data sets, so that the time series is of the form

$$X=U^n*A^{-1}+\sigma_N^2 N, \qquad (182)$$

where N is Gaussian noise of unit variance.

As mentioned earlier, this noise may produce a bias. To see this, simply convolve equation (182) with the correct filter A; the extra term proportional to $A*N$ contains correlations which contaminate the innovation. Therefore, the A which optimally removes correlations and dependencies from the innovation $A*X$ will not necessarily be the correct one. As will now be seen, the bias is small in the simple examples which follow. Nevertheless, this bias can in principle be removed if one has an estimate of the amount of observational noise present. Work is in progress on a procedure to eliminate it.

a) Experiment 1: Comparison of Dependence Measures

The dependence measures introduced in § IV*c* were tested on the process defined in equation (182), with $A=(-0.2,\mathbf{1},-0.3)$. The corresponding inverse pulse $C=A^{-1}$ is a two-sided exponential which rises somewhat more rapidly than it decays. Table 2 presents results for a sequence of innovations ranging from $n=40$ (highly nonnormal, pulses essentially isolated, easy for almost any technique) to $n=1$ (nearly normal, much pulse overlap, difficult for any technique). No noise was added. Note that L_1 optimization (with $A_0=1$) is exact[12] for large n but degenerates quickly as the pulse overlap increases. The iterative L_1 procedure (§ IV*e*) degenerates much more slowly as n decreases and would have made an impressive entry in Table 2. However, difficulties with convergence on more difficult problems make this technique, as implemented, unacceptable as a general-purpose method. Surprisingly the martingale difference property method fails badly, even for the easy U^{40} problem. This failure is unfortunate in view of the simplicity of the technique. Further development of the MDP approach may be fruitful.

The results shown for probability distribution functions (PDF) were calculated with five equally spaced and equal bins in R space (25 in $[R_n, R_{n+m}]$ space), chosen to float with the changing values of the minimum of $R(A)$ and maximum of $R(A)$, as this was empirically found to be better than having fixed bins. For some problems it is preferable to choose the R bins so that roughly equal numbers of points fall in them. Gaussian weight functions for the bins were used to combat the quantization problem outlined in § IV*c*. The results are substantially dependent on the number and placement of the bins, and at best the test answers are less accurate than those obtained with cumulative probability functions (CPF). In addition, the convergence properties of D_P, although better than those of the other dependence measures (based on characteristic functions, moments, and the MDP), are much worse than those of D_F.

Table 3 displays the results of similar tests dealing with the effects of additive noise on the computations. With R fixed at U^9, various levels of noise were added according to formula (182). In both comparisons the cumulative probability function method is superior to each of the others. The problem with $R=U^4$ and only 100 data points is very difficult, and compared to any other method tested the current one does amazingly well. Tables 2 and 3 do not represent enough trials to

[12] This is a general property of L_1, and is related to the fact that the L_1 optimum solution of an overdetermined set of linear algebraic equations always solves a subset of the equations exactly, as was realized by Laplace (see Claerbout and Muir 1973).

TABLE 2

TEST RESULTS
Innovations with Various Distributions: $R = U^n$
Pulse Shape: Two-sided exponential $(-0.2, \mathbf{1}, -0.3)^{-1}$
Length of Data: $N = 100$ (Averages of four such realizations)

n	CPF Method[a]	PDF Method[a]	MDP Method[a]	L_1-optimization
40...	−0.200, **1**, −0.300	−0.195, **1**, −0.296	−0.194, **1**, −0.419	−0.200, **1**, −0.300
9...	−0.202, **1**, −0.309	−0.207, **1**, −0.294	−0.219, **1**, −0.251	−0.230, **1**, −0.306
4...	−0.191, **1**, −0.305	−0.169, **1**, −0.250	−0.041, **1**, −0.453	−0.318, **1**, −0.328
1...	−0.201, **1**, −0.348	−0.582, **1**, −0.017	−0.257, **1**, −0.148	−0.509, **1**, −0.503

[a]Maximum lag, $m^* = 1$.

TABLE 3

TEST RESULTS
Various amounts of additive Gaussian noise: σ_N = noise variance (pulse peak = 1)
Pulse shape: Two-sided exponential $(-0.2, \mathbf{1}, -0.3)^{-1}$
Length of data: $N = 100$ (Averages of four such realizations)
Innovation: $R = U^9$

σ_N	CPF Method[a]	PDF Method[a]	L_1-Optimization
0.00...	−0.202, **1**, −0.309	−0.207, **1**, −0.294	−0.230, **1**, −0.306
0.01...	−0.202, **1**, −0.300	−0.230, **1**, −0.282	−0.239, **1**, −0.317
0.05...	−0.184, **1**, −0.261	−0.130, **1**, −0.258	−0.232, **1**, −0.339
0.10...	−0.169, **1**, −0.200	+0.003, **1**, −0.133	−0.183, **1**, −0.351

[a]Maximum lag, $m^* = 1$.

be definitive, but they indicate trends confirmed by other computations which are not presented here.

b) *Experiment 2: Detailed Study of an AR(1, 1) Process*

This experiment is an intensive study of a process similar to that in experiment 1. The aim is to study in detail a relatively difficult problem, namely deconvolution of the AR(1, 1) process

$$X = U^3 * A^{-1} + 0.05N, \qquad (183)$$

where $A = (-0.2, \mathbf{1}, -0.3)$ is the same as in experiment 1. This choice combines a moderately high noise level (see Table 3) and a low value of n (see Table 2), and presents a rather difficult problem. The solid line in Figure 20 is a realization of this process with $N = 100$.

Table 4 is a summary of the results of minimizing D_F with five different values of m^*. In all cases the starting solution was (0, 1, 0), and convergence to the AR(1, 1) solution shown in the table was rapid. In no case did restarts lead to significant changes in either of the parameters. The procedure was then to optimize both AR(1, 2) and AR(2, 1) filters, using as starting values the AR(1, 1) solution with a zero appended. What is shown in the next line of the table is the third-order ($M = 3$) solution which had the smaller value of the

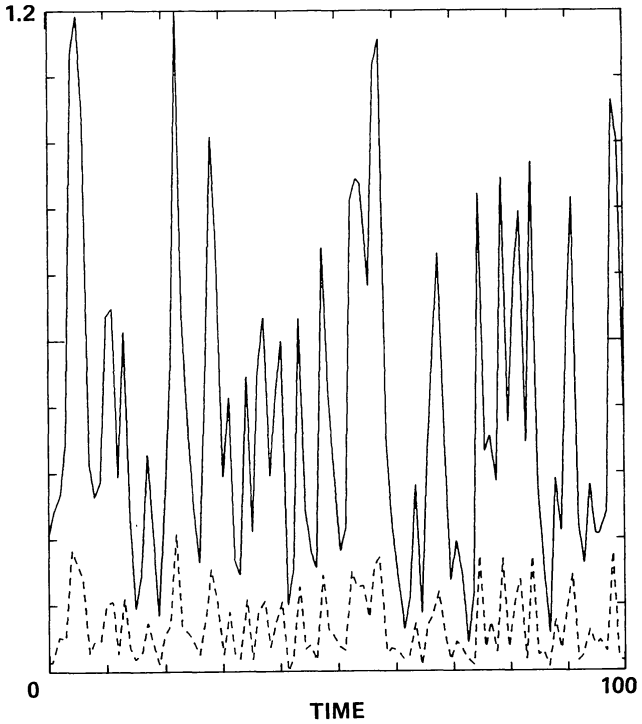

FIG. 20.—Realization of the AR(1, 1) process given by equation (183), with U^3 innovation and added Gaussian noise. The dashed line is the estimate of the innovation or pulse amplitude sequence.

TABLE 4

DECONVOLUTION OF $(-0.2, \mathbf{1}, -0.3)^{-1} * U^3 + 0.05N$

A (1)	$D(A)$ (2)	M (3)
$m^* = 1$		
$0, \mathbf{1}, 0$	2.1078	...
$-0.164, \mathbf{1}, -0.304$	0.1555	2
$-0.207, -0.112, \mathbf{1}, -0.312$	0.1117	3
$-0.076, -0.137, -0.105, \mathbf{1}, -0.302$	0.0810	4
$m^* = 2$		
$0, \mathbf{1}, 0$	1.2218	...
$-0.166, \mathbf{1}, -0.308$	0.1444	2
$+0.033, -0.183, \mathbf{1}, -0.276$	0.1437	3
$+0.003, -0.177, \mathbf{1}, -0.283, -0.010$ ...	0.1328	4
$m^* = 3$		
$0, \mathbf{1}, 0$	0.9270	...
$-0.148, \mathbf{1}, -0.331$	0.1557	2
$+0.048, -0.188, \mathbf{1}, -0.281$	0.1502	3
$+0.013, -0.177, \mathbf{1}, -0.282, -0.004$ ...	0.1495	4
$-0.010, +0.020, -0.145, \mathbf{1}, -0.292, +0.008$ ...	0.1339	5
$m^* = 4$		
$0, \mathbf{1}, 0$	0.8238	...
$-0.205, \mathbf{1}, -0.316$	0.1945	2
$+0.070, -0.245, \mathbf{1}, -0.283$	0.1868	3
$+0.014, -0.197, \mathbf{1}, -0.302, -0.004$ ...	0.1967	4
$-0.010, +0.013, -0.153, \mathbf{1}, -0.295, +0.024$ ...	0.1842	5
$m^* = 5$		
$0, \mathbf{1}, 0$	0.7934	...
$-0.235, \mathbf{1}, -0.318$	0.2126	2
$+0.071, -0.264, \mathbf{1}, -0.275$	0.2069	3
$+0.066, +0.076, -0.266, \mathbf{1}, -0.256$	0.2184	4
$-0.002, +0.050, -0.222, \mathbf{1}, -0.261, -0.005$ ...	0.2228	5

minimum D_F of these two cases. This process is then repeated. At each step, the filter may grow to the left or to the right, according to which produces the smaller D_F. Let us examine the convergence in this process, starting with $m^* = 1$. The quantity tabulated in the second column is

$$D = D_F(\min)\left(\frac{N+M}{N-M}\right)\frac{1}{m^*}, \qquad (184)$$

by analogy with equation (150), thus including the penalty for the number of parameters in the model. It is hoped that this quantity might have the property that makes the FPE useful: As a function of M (the number of free parameters), a minimum of D indicates that the correct order M has been reached. But for $m^* = 1$ this quantity keeps on decreasing with M, giving no indication of reaching a minimum. Also the values of the new parameters are *not* small, so there is no indication of convergence at all. This situation is greatly improved for $m^* = 2$, as the new parameters ($+0.033$ and -0.010) are relatively small. In addition, while D does not reach a minimum, it decreases quite slowly with M. One might guess that the correct order is AR(1, 1) (i.e., $M = 2$) from the entries in table 4 for $m^* = 2$. The improvement continues for $m^* = 3$. Starting at $m^* = 4$ there is a minimum in D, at $M = 3$ (the correct order is $M = 2$), and the value of the extra parameter A_{-2} (which should be 0) is small, 0.07 in both cases. Starting with $m^* = 4$, and especially at $m^* = 5$, the values of the parameters change significantly from the values they had for lower m^*. It appears from this experiment that if m^* is too low (1 or perhaps 2 in this example) or too large (5 or perhaps 4) the results are not as good as they are for an intermediate value. This result is as expected: If m^* is small, some of the information to be gained by reducing the dependence in R at larger lags is lost. If m^* is too large, the information will be diluted as the minimization will try to reduce dependencies at large lags where there are none to reduce.

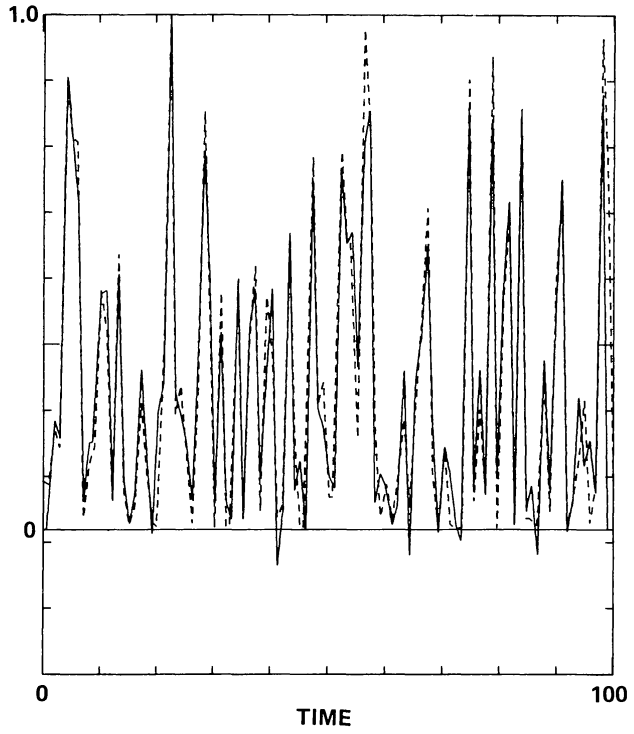

FIG. 21.—Comparison of the estimated (*solid line*) and exact (*dashed line*) innovations for the process shown in Fig. 20.

This suggests taking $m^* = 3 \pm 1$ in the present experiment. Figures 20–22 show results for the $M = 3$, $m^* = 3$ solution (which is very similar to $M = 4$, $m^* = 3$ and to $M = 2$ or 3, $m^* = 2$). The dashed line in Figure 20 is the estimated innovation. Figure 21 compares this with the exact innovation from which the realization of X was constructed. This estimate and the corresponding pulse (compared with the exact one in Fig. 22) are very accurate. Figures 23 and 24 present similar comparisons for the somewhat different solution corresponding to $M = 3$, $m^* = 5$ (which is similar to $M = 3$, $m^* = 4$), which might have been selected from Table 4 if the quasi-FPE criterion were taken seriously. This solution yields slightly poorer reproductions of the innovation and the pulse shape (although the latter is difficult to see in comparing Figs. 22 and 24).

c) Experiment 3: An AR(2, 1) Process

The goal of this test is to see what happens if the process is more complicated. In particular, we will see to what extent the quasi-FPE criterion (a minimum in the function $D[M]$ given in eq. [184]) is useful in determining the order of a higher-order process. The process chosen is again given by the basic form in equation (182), with $A = (-0.3, \mathbf{1}, -0.2, -0.3)$, $n = 9$,

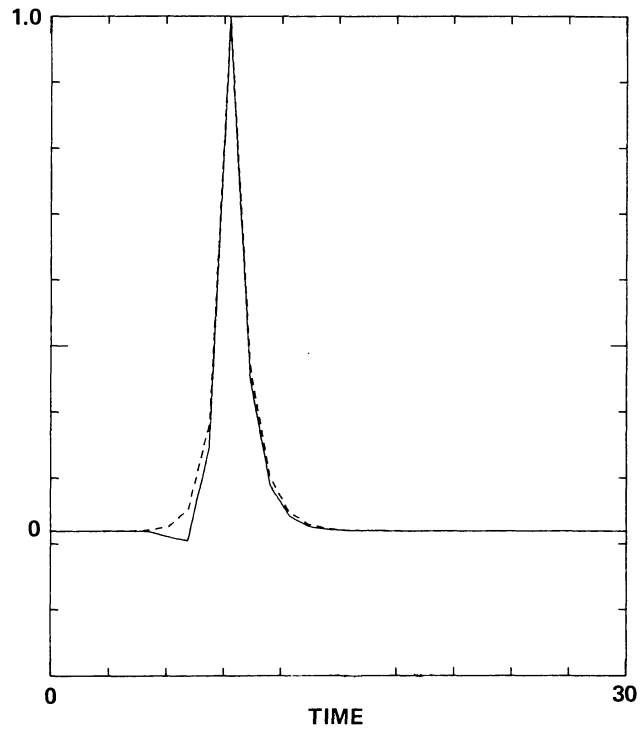

FIG. 22

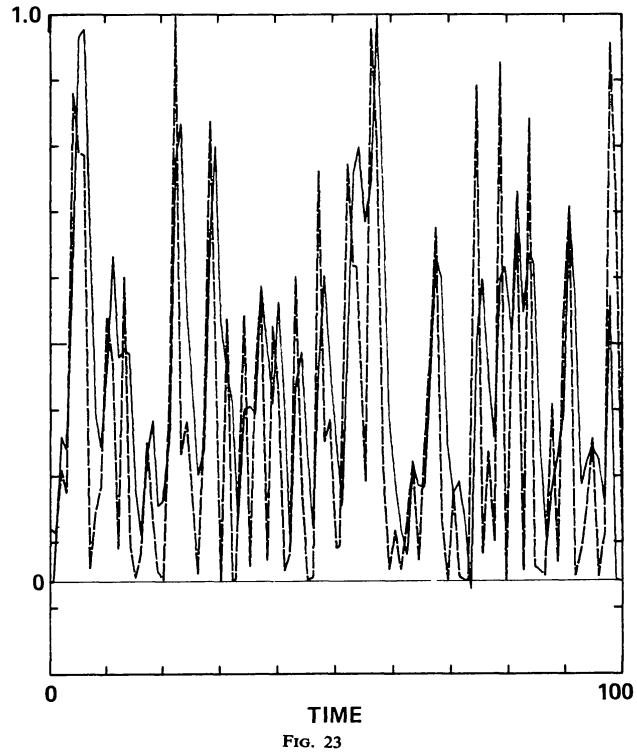

FIG. 23

FIG. 22.—Comparison of the estimated (*solid line*) and exact (*dashed line*) pulse shapes for the process shown in Fig. 20. The solution shown is $A = (0.048, -0.188, \mathbf{1}, -0.281)$ (obtained with $m^* = 3$).

FIG. 23.—Comparison of the exact (*dashed line*) and estimated (*solid line*) innovation for the process shown in Fig. 20, but corresponding to a different solution, namely $A = (0.071, -0.264, \mathbf{1}, -0.275)$ obtained with $m^* = 5$. This result illustrates that the value of m^* can be too large.

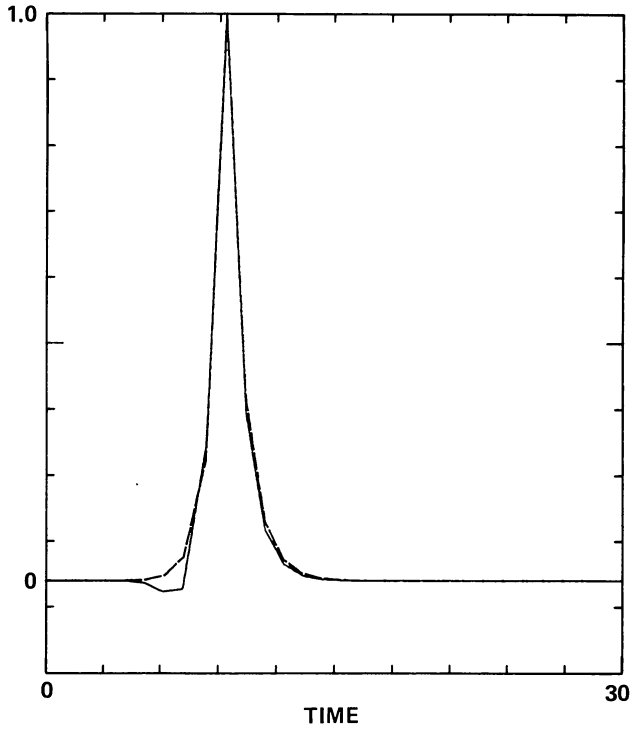

FIG. 24.—Comparison of the exact pulse (*dashed line*) with the pulse derived from the solution mentioned in the caption to Fig. 23.

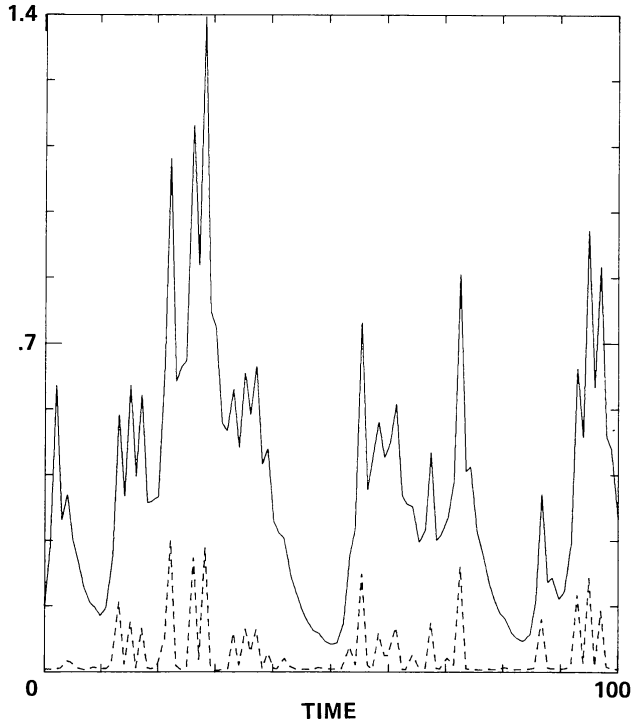

FIG. 25.—Realization of the AR(2, 1) process described in the text (U^9, no noise). The dashed line is the estimated innovation corresponding to the solution $A=(-0.27, 1, -0.202, -0.323)$ obtained for both $m^*=2$ and $m^*=3$.

and $\sigma_N=0$. This should be an easy problem because the innovation (U^9) is so highly nonnormal and because there is no noise. This was done purposely, to minimize the confusion due to noise and excessive pulse overlap, thus isolating the order-determining problem. The realization studied here is plotted in Figure 25. Table 5 summarizes the minimization, in the same format as in Table 4. Because the AR filter generating the process is longer, a larger range of values of m^* has been included. As before, the starting solution was the simple AR(1, 1) with $A=(0, 1, 0)$. The results for $m^*=1$ are very poor, as might be expected, as A ties together values separated by up to three lags, so a lag of one appears to be inadequate. As expected, the results are much improved for $m^*=2$ and 3. For $m^*=3$ and 4, the result is essentially perfect, in that the quantity D goes through a minimum at the correct order, the parameter values are almost the same for the two values of m^*, and the values of the higher-order parameters ($M=4, 5, \ldots$) are very small. For $m^*=10$ the minimum in D occurs at $M=4$, too large by 1, but again the extra parameter is very small, so that this solution is essentially identical (e.g., in terms of the corresponding pulse shape) to the solutions for lower values of m^* which are of the correct order. Figure 26 shows the innovation and Figure 27 the pulse shape estimates from the $M=3$ solution for $m^*=2$ or $m^*=3$ (the values for A are essentially identical, and, for example, the pulse shapes would be indistinguishable in Fig. 27). In each case the estimate is compared with the exact quantity. Both the pulse shape and the innovation are reproduced very accurately. Note that there is a large amplitude pulse which occurred very near the beginning of the realization. The pulse actually occurred prior to the first point of the estimated innovation, but it is shown in Figure 26 to stress the point that pulses very near the end and beginning of the realization are not represented accurately because of end effects. Nevertheless the part of any such pulse that extends into the realization is included in the determination of the model parameters.

Table 6 and Figures 28, 29, and 30 (for $M=4$) present the deconvolution of the same realization just discussed, but with added Gaussian noise of variance 0.05. These results show that the accuracy of the parameters determined above for this third order process is not due to the absence of noise. The innovation shows increased variance, including the appearance of small negative amplitudes which are not present in the actual innovation. It appears that the effect of additive noise is to add noise to the estimated innovation, but it is uncertain whether the distribution of the noise in the innovation is also Gaussian. Figure 30 shows that the basic shape, including the secondary peak, of the pulse is retained, but the tail of the pulse is altered somewhat.

TABLE 5

DECONVOLUTION OF $(-0.3, 1, -0.2, -0.3)^{-1} * U^9$

A	$D(A)$	M
$m^* = 1$		
0, 1, 0	4.232	...
−0.256, 1, −0.336	0.1331	2
−0.274, 1, −0.107, −0.577	0.0906	3
−0.288, 1, +0.017, −0.874, +0.106	0.0747	4
−0.368, −0.207, 1, +0.119, −0.761, +0.256	0.0327	5
−0.347, −0.220, 1, −0.017, −0.563, +0.362, −0.164	0.0319	6
$m^* = 2$		
0, 1, 0	4.026	...
−0.132, 1, −0.462	0.5114	2
−0.269, 1, −0.202, −0.323	0.1002	3
−0.262, 1, −0.194, −0.316, −0.060	0.0967	4
−0.260, 1, −0.120, −0.409, −0.128, +0.174	0.0924	5
−0.256, 1, −0.138, −0.274, −0.197, −0.101, +0.318	0.0698	6
$m^* = 3$		
0, 1, 0	3.547	...
−0.170, 1, −0.441	0.3683	2
−0.268, 1, −0.202, −0.323	0.0884	3
−0.273, 1, −0.201, −0.321, −0.0001	0.0921	4
−0.256, 1, −0.233, −0.324, +0.043, −0.024	0.0938	5
−0.307, −0.226, 1, −0.180, −0.261, +0.204, −0.359	0.0686	6
$m^* = 4$		
0, 1, 0	3.168	...
−0.129, 1, −0.482	0.3565	2
−0.272, 1, −0.201, −0.322	0.0749	3
−0.276, 1, −0.199, −0.318, −0.002	0.0780	4
−0.279, 1, −0.210, −0.318, +0.019, −0.003	0.0799	5
−0.237, 1, −0.242, −0.376, +0.127, +0.067, −0.192	0.0698	6
$m^* = 10$		
0, 1, 0	1.8225	...
−0.126, 1, −0.497	0.2519	2
−0.274, 1, −0.201, −0.323	0.0919	3
−0.271, 1, −0.202, −0.322, −0.003	0.0869	4
−0.274, 1, −0.219, −0.316, +0.016, −0.008	0.0924	5
−0.282, 1, −0.224, −0.321, −0.040, +0.002, −0.029	0.0866	6

d) *Experiment 4*: *Gaussian Noise*

One can consider independently distributed noise as the convolution of an independently distributed innovation with a delta function. When applied to noise, the deconvolution procedure should produce a delta function pulse. This experiment was designed to test the procedure on independent Gaussian noise. The solid line in Figure 31 is the analyzed noise. the minimization was done for the single value $m^* = 2$. The quasi-FPE did not clearly indicate convergence, but this hardly matters because all of the solutions were close to delta functions. The dashed line in Figure 31 is the estimated innovation (plotted with a different scale) and, as desired, is very nearly the same as the data itself. The pulse shape shown in Figure 32 is the inverse of the best third-order solution $A = (-0.061, +0.072, 1, +0.134)$ and is not far from the desired delta function ($|C_n|$ is < 0.1 for all $n \neq 0$).

e) *Experiment 5*: *A Sine Wave*

The technique we have been discussing was designed for random processes, and it could easily break down

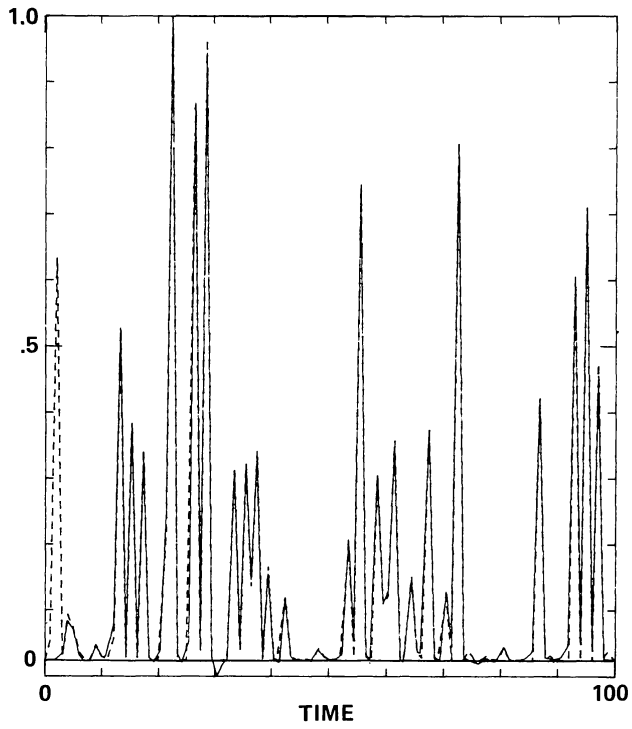

FIG. 26.—Comparison of the estimated (*solid line*) and exact (*dashed line*) innovations for the process shown in Fig. 25. The solution is the one given in the caption for that figure.

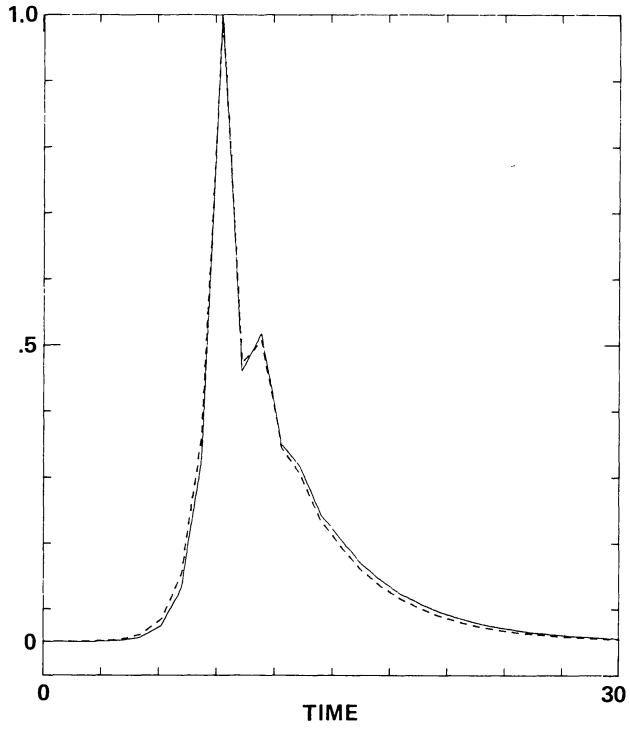

FIG. 27.—Comparison of the estimated (*solid line*) and exact (*dashed line*) pulse shapes for the process shown in Fig. 25 (solution as in previous figure).

in the presence of a deterministic part to the data. This experiment tests this possibility, using a sine wave as an example of a deterministic process. If a sine wave is considered as a MA pulse (which would be *unstable*, as the coefficients do not converge), the corresponding AR filter has a zero on the unit circle. (Compare to the case $A=[1, 1]$, with $C=A^{-1}=[1, -1, 1, -1, 1, -1, 1, \ldots]$.) When applied to a pure sine wave the simplex minimizer had convergence difficulties, and D_F dropped by a factor of 10^{30} during the minimization. The pulse shapes that it was leaning toward, however, were more or less sinusoidal. Since the pure sine wave is a singular case, a small amount of noise was added, so that the data were given by

$$X_n=\sin(0.5n)+0.0025N, \qquad (185)$$

where as before N is unit variance Gaussian noise. This addition removed the convergence problems, and the solution $A=(-0.419, -0.070, 1, -0.813)$ was obtained with $m^*=3$. Figure 33 shows the data as a solid line. In this case the interpretation of the innovation (dashed line in Fig. 33) is not straightforward. A sine wave is a single pulse, not a random sequence of pulses. But this model appears to represent a sine wave as a random sequence of the pulses shown in Figure 34 (i.e., the inverse of the above solution for A), basically a damped sine wave. Remember that because of the way the innovation is calculated, the data are exactly reproduced (except near the ends) by the expression $R*A^{-1}$, so the innovation in Figure 33, convolved with the pulse (not all of which is shown in Fig. 34), reproduces the data.

f) *Experiment 6: 3C 273*

Data on the optical variation of the Quasar 3C 273 (Kunkel 1967) have been analyzed by a number of

TABLE 6

DECONVOLUTION OF $(-0.3, \mathbf{1}, -0.2, -0.3)^{-1}*U^9+0.05N$

$m^*=3$	$D(A)$	M
0, **1**, 0	0.8238	...
+0.085, **1**, −0.695	0.3124	2
−0.362, **1**, +0.031, −0.424	0.0872	3
−0.310, **1**, −0.075, −0.333, −0.088	0.0855	4
−0.441, +0.059, **1**, −0.231, −0.201, −0.015	0.0761	5

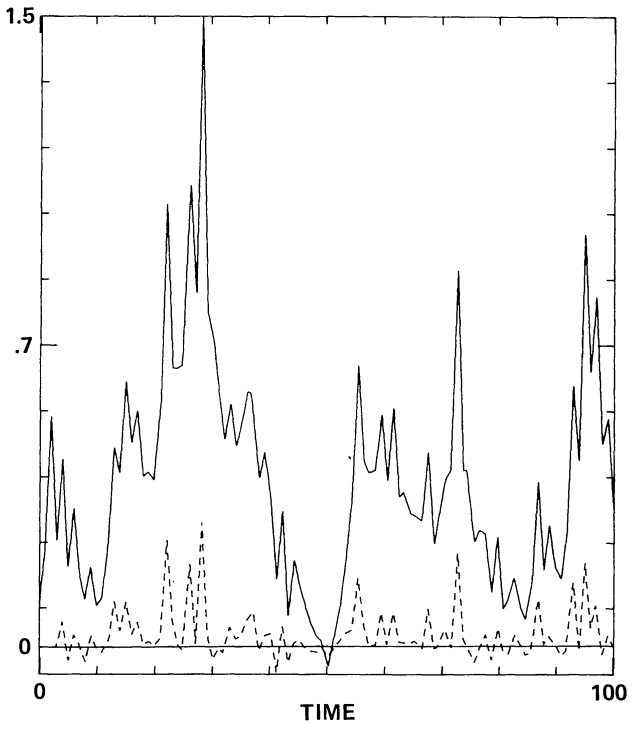

FIG. 28

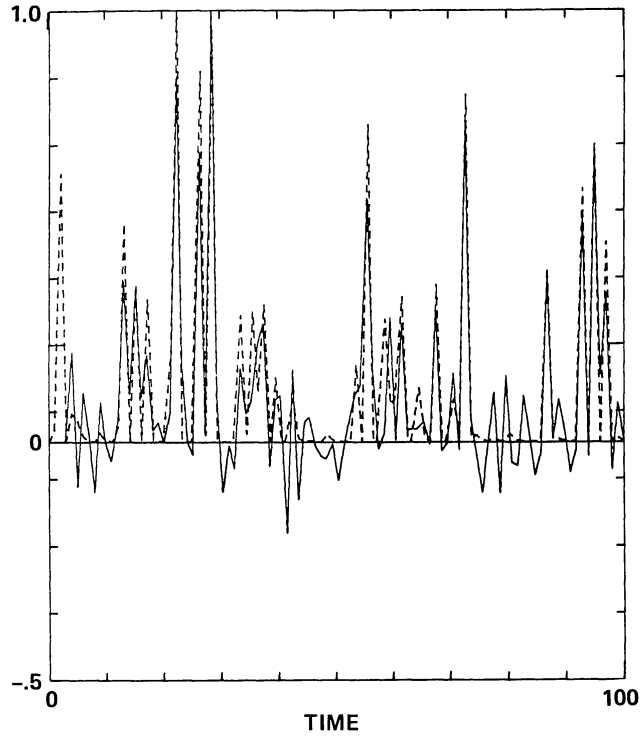

FIG. 29

FIG. 28.—The same realization shown in Fig. 25, but with added noise of variance 0.05. The dashed line is the innovation derived from the solution $A = (-0.310, \mathbf{1}, -0.075, -0.333, -0.088)$ (obtained for $m^* = 3$).

FIG. 29.—Comparison of the estimated (*solid line*) and exact (*dashed line*) innovations for the process shown in the previous figure.

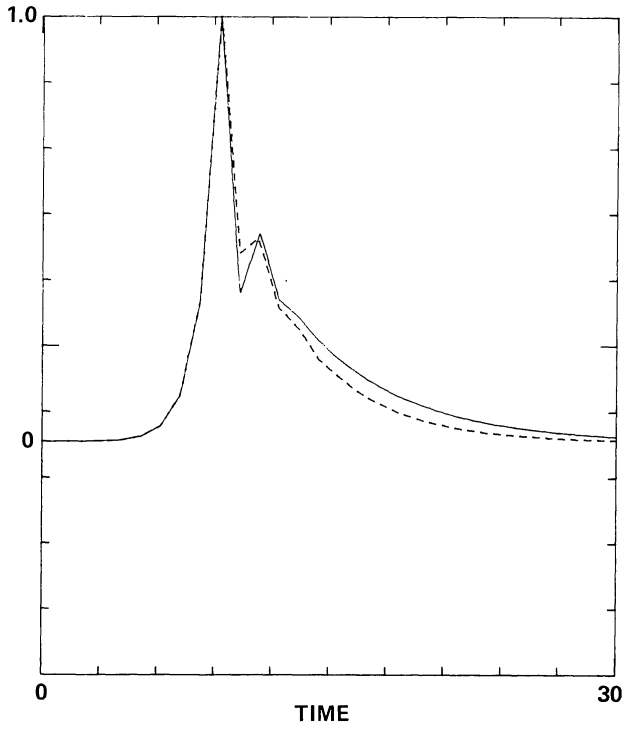

FIG. 30

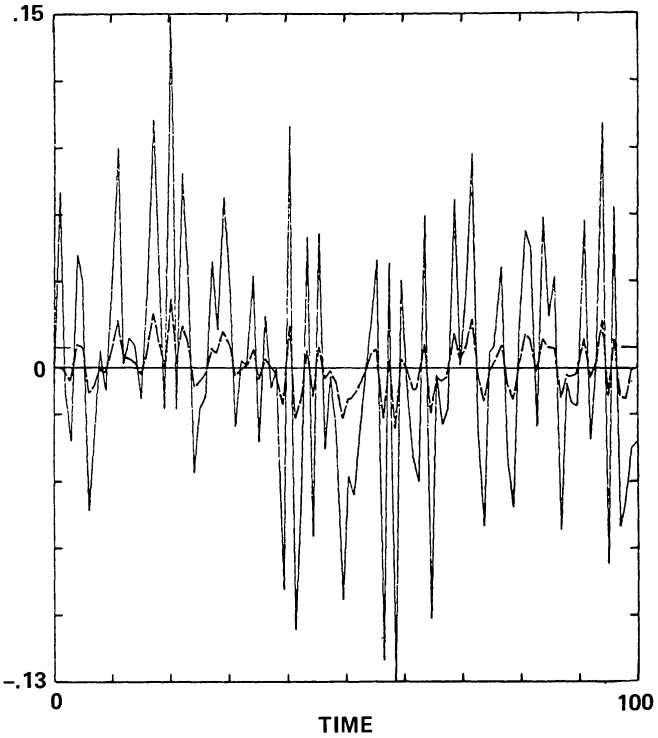

FIG. 31

FIG. 30.—Comparison of the estimated (*solid line*) and exact (*dashed line*) pulse shapes for the process shown in Fig. 28, with the solution quoted in the caption for that figure.

FIG. 31.—Independently distributed Gaussian noise ($N = 100$), analyzed in the same way as the data shown in the previous figures. The estimated innovation (*dashed line*) is essentially identical to the data (plotted on a different scale).

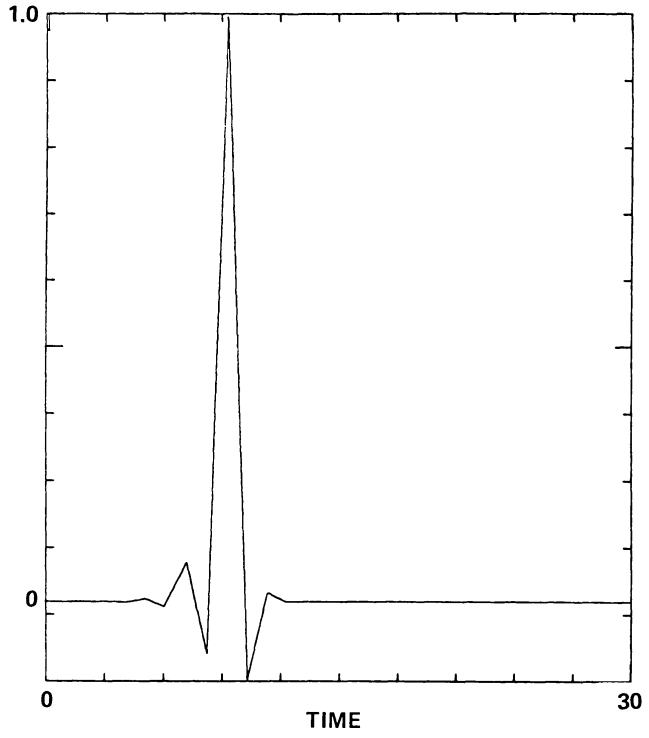

FIG. 32

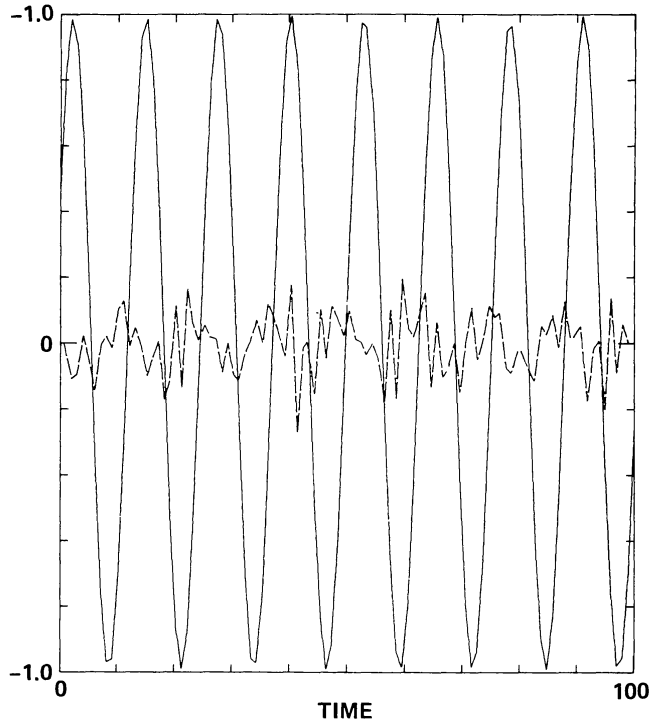

FIG. 33

FIG. 32.—The pulse shape derived for the data shown in the previous figure, corresponding to the A given in the text. The ideal solution would be a delta function. The horizontal scale of this figure is ~3 times that in Fig. 31.

FIG. 33.—A sine wave with small added noise, analyzed in the same way as the moving averages in the previous figures. The estimated innovation (*dashed line*) appears to be random.

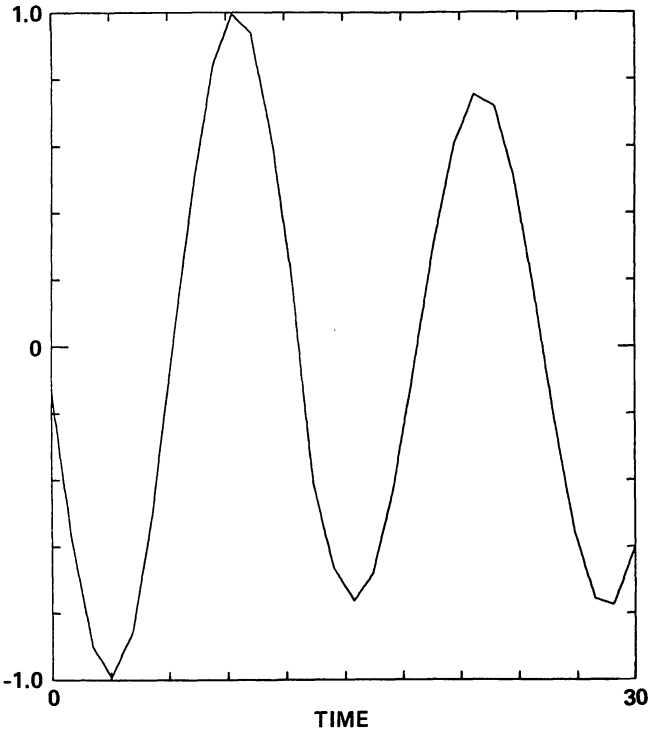

FIG. 34

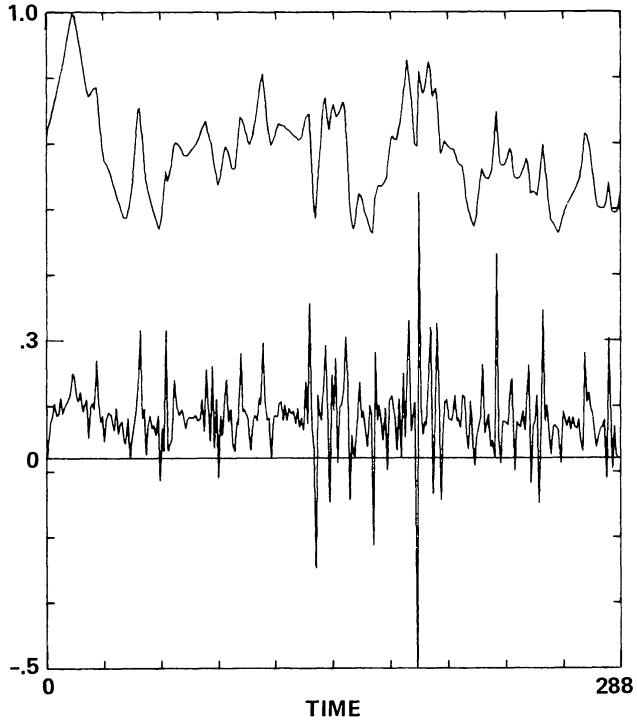

FIG. 35

FIG. 34.—The pulse shape obtained in the analysis of the data shown in Fig. 33. Only the first part of this gradually damped sine wave is shown (scale is as in Fig. 32).

FIG. 35.—The historical light curve of 3C 273 (*top*), derived directly from the magnitudes given by Kunkel (1967). The intensity is on a linear scale in arbitrary units, and the time covered is 28,800 days. The estimated innovation shown is for $A = (-0.081, 0.265, -0.740, 1, -0.419)$ obtained for $m^* = 3$.

workers looking for periodicities and for pulses (the closest in philosophy to the present work are Fahlman and Ulrych 1975, 1976). A future paper will give the details of the analysis of these data using the CPF-method, but preliminary results will be given here to demonstrate the application of the technique to real data. In particular the issue of determining the amount of a possible constant component to the light curve is raised. The point is that there are two contributions to the mean value of the data: (1) a background constant, due for example to light from a source other than the one which is pulsed, and (2) the mean value of the (positive only) pulsed component. If the pulses are sparse enough, there will be a part of the time series where the contribution from pulses can be neglected, and then the minimum value of the curve, $\min_n(X_n)$, would be a good estimate of the background constant. But in general there can be enough pulse overlap at all times that this procedure will overestimate the constant. Indeed the deconvolution is nontrivial only when there is much pulse overlap. In such cases it is known only that the constant lies between 0 and $\min_n(X_n)$. We will see below that this problem in some circumstances can be solved with the current technique.

Figure 35 depicts the light curve in linear intensity units, while Table 7 tabulates the results of the minimizations. This is a relatively long time series ($N=292$ in the original data; the first four points were discarded so that $N=288$ because the FFT algorithm requires that the largest prime factor of N be ≤ 23). Since the number of operations scales as N^2, the reductions are moderately time consuming. For example, the run with $N=288$, $m^*=3$, and $M=2,3,4,5,6$ took 1,140 CPU seconds on the NASA-Ames CDC 7600. It will be noted that D does not go through a minimum, although for $m^*=3$ it is virtually stationary for $M=4$ and 5. Also, the parameters A_{-3} and especially A_{+2} are small. This suggests that the $M=4$ solution is to be adopted, but further computations with larger m^* will be necessary before this can be made definite. The pulse shape is shown in Figure 36 and is compared with the minimum delay pulse determined by Fahlman and Ulrych (1975) with $M=3$ (as determined by a legitimate FPE criterion). The innovation for this solution is shown in the lower part of Figure 35 and is compared to the innovation from the minimum delay solution in Figure 37. Both innovations have substantial numbers of negative amplitudes.

The author has carried out numerical experiments similar to those discussed by Fahlman and Ulrych (1976), confirming their contention that such behavior can have two causes: (1) nonconstancy of the pulse shape, or (2) use of a minimum delay solution, if the actual pulse is not minimum delay. The point in (1) is that the pulse shape may actually be changing, say in a random but stationary way, rather than being constant. The MA representation is still exactly correct, as long as X is stationary, but it uses a single pulse shape. This shape is a kind of time average of the actual pulse shape. (It is not simply representable as a time average, however; the deconvolution procedure yields some kind of nonlinear average of A, then C is the corresponding inverse.) When a pulse with a shape close to this average is convolved with the optimum A, a delta function results, as desired. But if the shape is somewhat different from the average, this convolution pro-

TABLE 7

DECONVOLUTION OF THE LIGHT CURVE OF 3C 273 ($N=288$)

A	$D(A)$	M
$m^*=1$		
0,1,0	1.832	...
−0.535,1,−0.519	0.001 677	2
+0.181,−0.696,1,−0.450	0.000 6955	3
$m^*=2$		
0,1,0	1.321	...
−0.516,1,−0.500	0.000 2493	2
−0.503,1,−0.540,+0.028	0.000 2122	3
−0.495,1,−0.569,+0.078,−0.028	0.000 2108	4
$m^*=3$		
0,1,0	0.809 5	...
−0.523,1,−0.528	0.000 6587	2
+0.146,−0.655,1,−0.462	0.000 3273	3
−0.081,+0.265,−0.740,1,−0.419	0.000 2883	4
−0.082,+0.264,−0.741,1,−0.421,+0.002	0.000 2879	5
−0.029,+0.212,−0.713,1,−0.469,+0.081,−0.063	0.000 2683	6

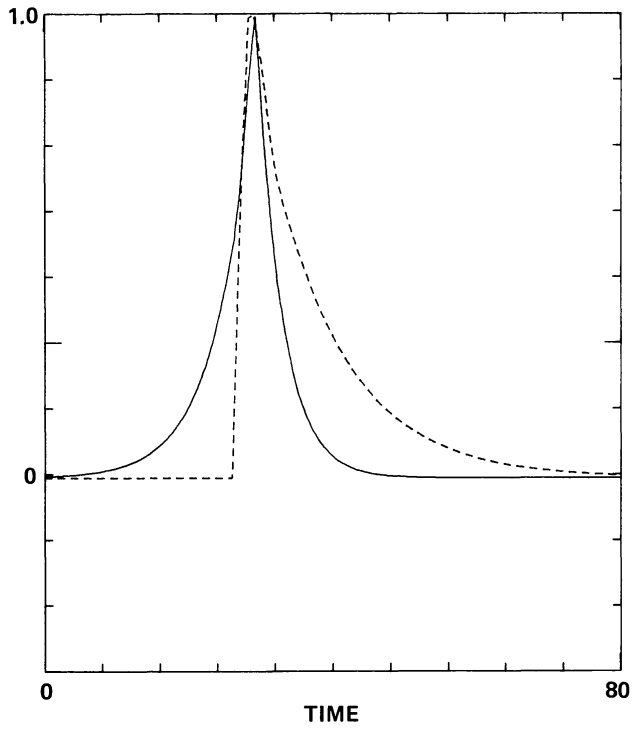

FIG. 36.—Comparison of the pulse shape for 3C 273 derived from the solution given in the caption to Fig. 35 (*solid line*), which is mixed delay, with the minimum delay pulse (*dashed line*) as derived by Fahlman and Ulrych (1975). The mixed delay pulse is nearly symmetric.

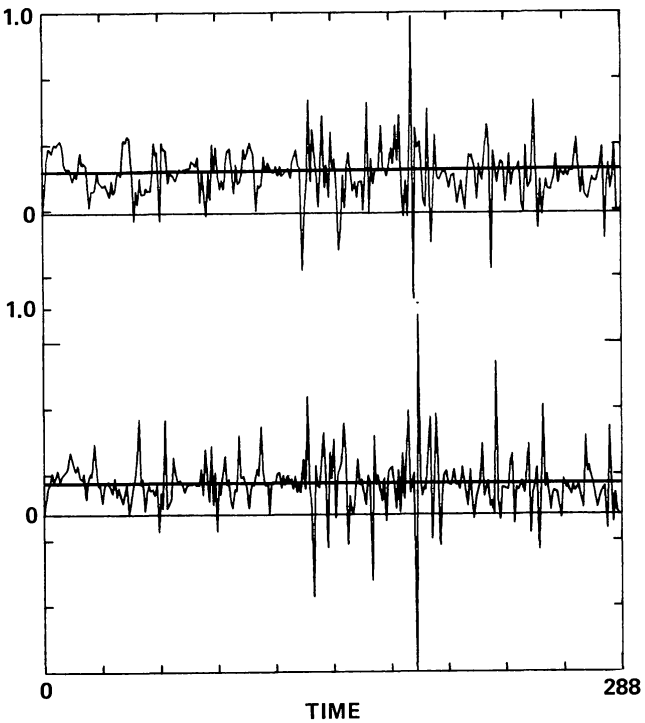

FIG. 37.—Comparison of the innovations derived from the minimum delay (*top*) and mixed delay (*bottom*) solutions as in Fig. 36. Note that the negative spikes are typically associated with nearby positive spikes; however, the pattern of this association seems to be different in the two innovations.

duces something other than a delta function. Simulations consisting of two or three distinct pulse shapes occurring randomly and independently show that the resulting amplitude usually consists of a first-negative-then-positive (or *vice versa*) spike, like the discrete version of the derivative of a delta function. Such spikes can be seen in the innovations in Figure 37. The form of the spike appears to be sensitive to the delay character of A, as the simultaneous spikes in the two innovations are sometimes quite different.

Effect (2) is quite similar, because the optimum minimum delay A is not the correct inverse of a mixed-delay pulse, and its convolution with the actual mixed-delay pulse will also produce other than a delta function. From the fact that the mixed-delay result shown here contains roughly the same amount of negative amplitude as the minimum delay result (Fig. 37), it appears that in 3C 273 the pulse shapes are indeed varying, and the negative amplitudes found by Fahlman and Ulrych (1976) are not due to the minimum delay assumption. (It is possible, but unlikely, that there is an additional source of negative amplitudes.)

There is one facet of the distribution function approach (either cumulative or differential) which is very useful, namely that it is completely insensitive to an additive constant in the data. The only factor that enters into the expressions for D_F or D_P is the shape of the joint and individual distribution functions. Adding a constant merely shifts the position of the functions on the R axis and does not change their shapes. Hence D is invariant to a shift in X, a property not shared by other deconvolution techniques. This invariance is important because it means that it is possible to estimate the size of the background component...if something is known about the distribution of the amplitudes. First, note that a constant in the data shows up as a constant in the estimated innovation, if one has the correct inverse pulse. For, letting K be the constant unit process:

$$K_n = 1, \quad (n = 1,2,3,\ldots,N), \tag{186}$$

we can write

$$X = R * C + aK, \tag{187}$$

where a is some unknown constant. The estimated innovation is

$$\hat{R} = A * X = R * (C * A) + aA * K$$

$$= R * (C * A) + \left(a \sum_k A_k\right) K, \tag{188}$$

and if $A = C^{-1}$,

$$\hat{R} = R + \left(a \sum_k A_k\right) K, \quad \text{QED.} \tag{189}$$

The second term on the right is a constant, but it is not yet obvious how to determine its value (and hence the value of a), because we know only $\langle \hat{R} \rangle$, and not $\langle R \rangle$. If it was known, or one wished to assume, that $\langle R \rangle = 0$, then

$$a = \frac{\langle \hat{R} \rangle}{\sum_k A_k} . \tag{190}$$

But the case $\langle R \rangle \neq 0$ is of particular importance in astronomy. For example, suppose that the actual amplitudes are positive only (as with light pulses), with a distribution which is either finite at $R=0$ or goes smoothly to zero (so that some pulses have amplitudes close to zero, but none are negative). Then

$$a\left(\sum_k A_k\right) = \min_n (\hat{R}_n) \tag{191}$$

could be used to estimate a. However, observational errors produce a variance in $\hat{R}$ which would make this estimate biased toward too small a value of a. This bias could be eliminated if the center of symmetry of the (presumably Gaussian) distribution of these observationally induced errors in R could be recognized. But an even larger problem with the estimate in equation (191) for the innovation of 3C 273 is the incidence of the large negative amplitude spikes. One must turn to more qualitative aspects of the distribution of $\hat{R}_n$. Specifically, the innovations in Figure 37 appear to have a definite background level (possibly better seen in the mixed-delay solution case), indicated in the figure with horizontal lines. This level corresponds to the peak in the distribution of $\hat{R}$ (which is Fig. 38), and is probably best estimated with the *median* of $\hat{R}$ (to avoid the bias in the mean value which the real pulses might produce). In the case of 3C 273 the mean and median are not very different, as the entire distribution is nearly symmetric (there is possibly a slightly significant bias on the positive side of the distribution shown in Fig. 38). In summary, the mean level for 3C 273 cannot yet be determined unambiguously because of the effect of the negative amplitudes in the innovation, but the levels shown in Figure 37 are reasonable guesses for this background of nonpulsed light.

In some other deconvolution methods the mean value of X is removed, and this is an example of a shift which may alter the deduced pulse shape. In particular, the optimum prediction-error filter method is usually applied to data that has had the mean subtracted out, because the form of equation (114) implies that, since the mean prediction error should vanish, either $\langle X \rangle$ or $\Sigma_i A_i$ must vanish. If the sum of the A_i vanishes, $A(z)$ has a zero on the unit circle, and A itself is not invertible because A^{-1} does not converge. Indeed, it is found in numerical trials that if the mean of X is left in, the resulting values of A_i sum to zero and A cannot be inverted. But if $\langle X \rangle$ is 0, A is well behaved. This is probably the basis on which Fahlman and Ulrych (1976) state that their analysis "...only makes use of the variance in the light curve. Hence the pulse shape... is unaffected by the presence of a background." However, one is not justified in subtracting out the mean just because the analysis breaks down otherwise.

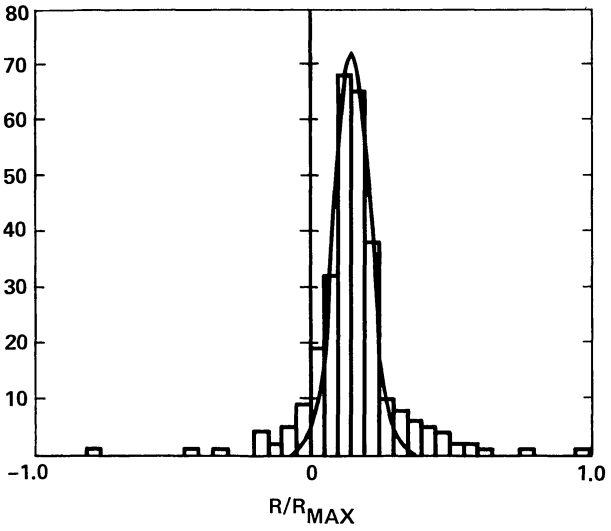

FIG. 38.—Histogram showing the distribution of the pulse amplitudes shown in Fig. 37 (*bottom*). A Gaussian curve fitting the central few bins is drawn for comparison. The overall distribution is definitely not purely Gaussian. It may have a Gaussian component, possibly connected with the observational scatter in the data. There may be a small asymmetry favoring the positive amplitudes, but the negative amplitudes (which are probably due to pulse shape variation) are nearly as numerous—this prevents the zero level of the amplitudes from being determined unambiguously. The results for the innovation derived from the minimum delay solution (top curve in Fig. 37) are very similar.

g) *Discussion*

The minimization of D_F appears to be a powerful deconvolution technique for moving average, autoregressive, or shot noise processes where the pulses are statistically independent of each other. An estimate of the pulse shape which is not constrained to have the minimum delay shape can be obtained, as well as an estimate of the amplitudes which the pulses had in the realization at hand. With the latter, the distribution of the pulse amplitudes can be studied. If a feature in the distribution corresponding to the zero level of the amplitudes can be recognized, the background level of nonpulsed signal can be determined.

It is well known that the fitting of sums of exponentials with unknown decay constants, as well as amplitudes, to data (e.g., radioactive decay data) is a very ill-conditioned problem. Since the exponential is in a sense the elementary pulse shape (see eqs. [21] through [23]) the deconvolution of MA processes is not unre-

lated to this problem. One of the difficulties is that the data can be nearly equally well represented by somewhat different models (different in form and in the values of the model parameters). The search for the best dependence measure (see § IV*c* and § VI*a*) was basically a quest for a procedure which minimizes the indeterminacy in the model fitting. In this respect, the one adopted (D_F) is generally superior to the others considered. It makes full use of the data at hand and has a well-defined and unique minimum in situations where the other measures have many shallow minima. The following points should be considered by anyone using this technique:

1) As with conventional time-domain modeling, the identification of the form of the model (even within the context of ARIMA models) is an important problem which does not have a precise general solution.

2) Since any stationary process has MA, AR, and ARMA representations, the successful modeling of time series data with a specific model does not guarantee that the structure of the physical process has been correctly interpreted.

3) Since the data are always exactly reproduced by the model, the meaning of successful modeling is not based on the smallness of the residuals between the sampled and modeled values of X, but rather on the degree to which the resulting amplitudes are independently distributed (e.g., as measured by the smallness of D_F).

4) As with conventional modeling, including spectral analysis, trends in the data can affect the results in very significant ways. There is no totally objective and automatic procedure for removing trends. There is no dependable way that an apparent trend can be distinguished from a statistical fluctuation in the underlying random process. Detrending should be done cautiously, and one should be suspicious of apparent trends.

5) The algorithm provided in the appendix is quite time consuming, especially for long arrays of data. Only minor efforts to speed up the computations have been made. Improvements in the algorithm can undoubtedly be made. Hopefully there is some approximation that can be used for large N. Some time can be saved for *low-order* AR models by computing the convolution $A * R$ by direct summation instead of FFT. The FFT version is given in the Appendix because it can be readily generalized to ARMA models, and because this convolution is not a major part of the computation.

Since this work was completed, John Deeter of the University of Washington found a way of evaluating D_F in a recursive way that requires a number of operations proportional to $N \log N$, as opposed to the N^2 required by the version of FUNK given in the Appendix. (This is the same advantage that the FFT has over the direct DFT.) His algorithm suffers some loss of precision relative to the brute-force double sum, and on short-word machines double precision may be necessary. This algorithm is currently being tested, and will be described in a future publication.

I am grateful to numerous colleagues for he pful discussions, including Dick Miller, Paul Swan, John Szabo, Tad Ulrych, and John Deeter. John Szabo developed a time series analysis system for the NASA Ames time sharing system, which was used for most of the computations presented here. John Deeter provided me with his code for an $N \log N$ version of the basic algorithm to evaluate D_F. I am grateful to Ed Groth and Dick Miller for numerous corrections and clarifications of an early version of the manuscript.

APPENDIX

THE ALGORITHM

The FORTRAN code given below (Table A1) is a nearly self-contained program which will enable the reader to use the deconvolution technique (based on cumulative probability functions). The only missing element is the FFT routine, which is a standard one, available in most program libraries.

The MAIN program reads the value of m^*, the data, the length of the AR filter (LAC), the position within the filter of the prediction point (MPT), the initial guessed solution (AOLD), and the number of times the order of the model is to be increased (NUMIT). The Fourier transform of the data is put in the arrays XR and XI, for that is the form in which the data will be referenced henceforth. The subroutine F2DC carries out the minimization, starting with a given solution, and returns the resulting minimum value of D_F (RES). This is done first with the input guessed solution, and then the order is increased in steps of one as indicated. The two minimum values RES1 (corresponding to A[new] = {A[old], 0}) and RES2 (corresponding to A [new] = {0, A[old]}) are compared, and the smaller is selected. This procedure is terminated arbitrarily by the value of NUMIT. The correct order must be determined by inspection of the behavior of D_F (minimum) with increasing order, and inspection of the way in which the values of the parameters change, as discussed in the text.

Subroutine F2DC carries out the minimization, trying restarts until the solution settles down. A criterion has been shown in terms of the minimum D_F, but one could also use criteria in terms of the changes in the parameters. The

TABLE A1

THE FORTRAN CODE

```
C**                    MAIN PROGRAM
C**
      COMMON/F2DVEC/XR(1000),XI(1000),RR(1000),RI(1000)
      COMMON/F2DSCA/FACN,FACR,FAC1
      COMMON/F2DINT/LDAT,NUMR,NR,MPT,LAC,N1D,N2D,MAXLAG
      COMMON/INOV/R(1000)
      DIMENSION AOLD(20),A1(20),A2(20)
      DIMENSION DATA(1000)
C**
C**  INITIALIZE ARRAYS
C**
      DO 1 I=1,1000
      XR(I)=0.0
      XI(I)=0.0
      RR(I)=0.0
1     RI(I)=0.0
C**
C**  READ DATA
C**
      READ(8,50)MAXLAG
      READ(8,50)LDAT
      READ(8,51)(DATA(I),I=1,LDAT)
50    FORMAT(3I3)
51    FORMAT(6E12.5)
C**
C**  CALCULATE THE FFT OF THE DATA
C**
      DO 2 I=1,LDAT
2     XR(I)=DATA(I)
      CALL FFT(XR,XI,LDAT,LDAT,LDAT,-1)
C**
C**  READ THE PARAMETERS OF THE INITIAL MODEL
C**
      READ(8,50)LAC,MPT,NUMIT
      READ(8,51)(AOLD(I),I=1,LAC)
C**
C**  CARRY OUT THE FIRST MINIMIZATION
C**
      CALL F2DC(AOLD,RES)
      IF(NUMIT.EQ.0)STOP
C**
C**  DO MINIMIZATIONS WITH INCREASING MODEL ORDER
C**
      DO 20 IT=1,NUMIT
C**
C**  A1 IS THE OLD MODEL EXTENDED TO THE RIGHT (A,0)
C**  A2 IS THE OLD MODEL EXTENDED TO THE LEFT (0,A)
C**
      DO 10 I=1,LAC
      A2(I+1)=AOLD(I)
10    A1(I)=AOLD(I)
      A1(LAC+1)=0.0
      A2(1)=0.0
      LAC=LAC+1
      CALL F2DC(A1,RES1)
      MPT=MPT+1
      CALL F2DC(A2,RES2)
C**
C**  SELECT THE BETTER OF A1 AND A2
```

```
C**
      IF(RES1.LT.RES2)GO TO 12
      DO 11 I=1,LAC
11    AOLD(I)=A2(I)
      GO TO 20
12    DO 13 I=1,LAC
13    AOLD(I)=A1(I)
      MPT=MPT-1
20    CONTINUE
      STOP
      END
```

```
      SUBROUTINE F2DC(A,RES)
      COMMON/F2DVEC/XR(1000),XI(1000),RR(1000),RI(1000)
      COMMON/F2DSCA/FACN,FACR,FAC1
      COMMON/F2DINT/LDAT,NUMR,NR,MPT,LAC,N1D,N2D,MAXLAG
      DIMENSION A(20)
C**
C**  CARRY OUT THE BASIC MINIMIZATION
C**
      CALL F2D(A,RES)
      RESOLD=RES
C**
C**  NOW DO RESTART MINIMIZATIONS (UP TO 3) UNTIL
C**  THE SOLUTION DOES NOT CHANGE SIGNIFICANTLY
C**
      DO 1 I=1,3
      CALL F2D(A,RES)
      DIFRES=(RESOLD-RES)/RESOLD
      RESOLD=RES
      IF(DIFRES.LT.1.0E-4)GO TO 3
1     CONTINUE
      PRINT 2
2     FORMAT(15H DID NOT SETTLE)
C**
C**  CALCULATE THE PULSE SHAPE
C**
3     DO 4 I=1,LDAT
      RR(I)=0.0
4     RI(I)=0.0
      DO 5 I=1,LAC
5     RR(I)=A(I)
      CALL FFT(RR,RI,LDAT,LDAT,LDAT,-1)
      DO 6 I=1,LDAT
      TEM=RR(I)**2+RI(I)**2
      RR(I)=RR(I)/TEM
6     RI(I)=-RI(I)/TEM
      CALL FFT(RR,RI,LDAT,LDAT,LDAT,+1)
C**
C**  NORMALIZE AND SHIFT THE PULSE SO THAT THE
C**  PEAK IS NEAR THE CENTER OF THE ARRAY
C**
      IMAX=0
      CMAX=0.0
      DO 7 I=1,LDAT
      TEST=ABS(RR(I))
      IF(TEST.GT.CMAX)IMAX=I
7     IF(TEST.GT.CMAX)CMAX=TEST
      CMAX=RR(IMAX)
      DO 8 I=1,LDAT
      INDEX=I-1+IMAX-LDAT/2
      IF(INDEX.LT.1)INDEX=LDAT+INDEX
      IF(INDEX.GT.LDAT)INDEX=INDEX-LDAT
```

```
8       RI(I)=RR(INDEX)/CMAX
C**
C**   PLOT THE PULSE SHAPE
C**
        CALL PLOT(RI,LDAT)
        FPE=RES*FLOAT(LDAT+LAC)/FLOAT(MAXLAG*(LDAT-LAC))
        RETURN
        END
```

```
        SUBROUTINE F2D(A,RES)
        COMMON/F2DVEC/XR(1000),XI(1000),RR(1000),RI(1000)
        COMMON/F2DSCA/FACN,FACR,FAC1
        COMMON/F2DINT/LDAT,NUMR,NR,MPT,LAC,N1D,N2D,MAXLAG
        DIMENSION P(21,20),Y(21),X(21),A(20)
        DATA SCALF,IPR/1.0,5/
C**
C**   PRINT INPUT MODEL AND SET UP PARAMETER VALUES
        PRINT 50,(A(I),I=1,LAC)
        NACT=LAC-1
        NPOINT=NACT+1
        N1D=LAC+1
        N2D=LDAT
        NUMR=N2D-N1D+1
        NR=NUMR-1
        FACR=1.0/FLOAT(NUMR*NUMR)
        FAC1=1.0/FLOAT(NR)
        FACN=1.0/FLOAT(LDAT)
        PSUM=0.0
        J=0
C**
C**   SET UP THE INITIAL SIMPLEX
C**
        DO 1 I=1,LAC
        IF(I.EQ.MPT)GO TO 1
        J=J+1
        TEMP=A(I)
        PSUM=PSUM+ABS(TEMP)
        P(1,J)=TEMP
1       CONTINUE
        FNUM=FLOAT(NACT)
        TES=ABS(PSUM)
        IF(TES.LE.1.0E-3)PSUM=0.15
        QSC=-SCALF*PSUM/FNUM
        TEMP=SQRT(FNUM+1.0)-1.0
        DEN=FNUM*SQRT(2.)
        PN=(TEMP+FNUM)*QSC/DEN
        QN=TEMP*QSC/DEN
        DO 3 I=2,NPOINT
        DO 2 J=1,NACT
2       P(I,J)=P(1,J)+QN
3       P(I,I-1)=P(I,I-1)+PN-QN
C**
C**   CALCULATE THE FUNCTIONAL VALUES FOR THE INITIAL SIMPLEX
C**
        DO 5 I=1,NPOINT
        DO 4 J=1,NACT
4       X(J)=P(I,J)
5       Y(I)=FUNK(X)
C**
C**   NOW DO THE MINIMIZATION
C**
        ITER=0
        IPRINT=IPR
        CALL AMOEBA(P,Y,NPOINT,ITER,IPRINT)
```

```
C**
C**   STORE AND PRINT THE RESULTS
C**
      J=0
      DO 10 I=1,NACT
      IF(I.EQ.MPT)J=J+1
      J=J+1
10    A(J)=P(IPRINT,I)
      A(MPT)=1.0
      RES=Y(IPRINT)
      PRINT 50,(A(I),I=1,LAC)
50    FORMAT(5G14.6)
      RETURN
      END
```

```
      SUBROUTINE AMOEBA(P,Y,NPOIN,ITER,IPRIN)
C**
C**   MINIMIZATION USING A SIMPLEX METHOD
C**     ITER IS A COUNTER FOR ITERATIONS
C**     EVERY IPRIN ITERATIONS THE SMALLEST FUNCTIONAL VALUE
C**     AND THE PERCENTAGE SPREAD OF THE VALUES (ERR) ARE
C**     PRINTED
C**     ON RETURN, THE VALUE OF IPRIN IS THE INDEX CORRESPONDING
C**     TO THE MINIMUM FUNCTIONAL VALUE
C**
      DIMENSION P(21,20),Y(21),PR(20),PRR(20),PBAR(20),PINV(20)
      EQUIVALENCE(PINV,PRR),(YPRR,YPINV)
      DATA ALPHA,BETA,GAMMA/1.0,0.5,2.0/
      DATA TOL,NSTOP/1.0E-03,150/
      NVAR=NPOIN-1
519   CONTINUE
1     ILO=1
      IHI=1
      INHI=1
      DO 10 I=1,NPOIN
      YI=Y(I)
      IF(YI.GE.Y(ILO)) GO TO 10
      ILO=I
10    CONTINUE
      DO 11 I=I,NPOIN
      YI=Y(I)
      IF(YI.LE.Y(IHI))GO TO 11
      IHI=I
11    CONTINUE
      IF(IHI.EQ.1)INHI=2
      DO 12 I=1,NPOIN
      IF(I.EQ.IHI)GO TO 12
      YI=Y(I)
      IF(YI.LE.Y(INHI))GO TO 12
      INHI=I
12    CONTINUE
      IF(MOD(ITER,IPRIN).NE.0) GO TO 209
      ERR=100.*(Y(IHI)-Y(ILO))/Y(ILO)
121   PRINT 205,Y(ILO),ERR
205   FORMAT(1P,G13.4,F6.3)
206   DIF=Y(IHI)-Y(ILO)
      RAT=DIF/Y(INHI)
      IF(RAT.LE.TOL)GO TO 80
      IF(ITER.GE.NSTOP)GO TO 84
      IF(IGO.NE.0) GO TO 80
209   ITER=ITER+1
      DO 21 I=1,NVAR
21    PBAR(I)=0.
      DO 23 I=1,NPOIN
      IF(I.EQ.IHI) GO TO 23
```

```
      DO 22 J=1,NVAR
22    PBAR(J)=PBAR(J)+P(I,J)
23    CONTINUE
      DO 24 I=1,NVAR
24    PBAR(I)=PBAR(I)/NVAR
      DO 25 J=1,NVAR
25    PR(J)=(1.+ALPHA)*PBAR(J)-ALPHA*P(IHI,J)
      YPR=FUNK(PR)
258   IF(YPR.LE.Y(ILO)) GO TO 30
      IF(YPR.GE.Y(IHI)) GO TO 40
      IF(YPR.GE.Y(INHI)) GO TO 38
26    DO 27 J=1,NVAR
27    P(IHI,J)=PR(J)
      Y(IHI)=YPR
      GO TO 1
30    DO 31 J=1,NVAR
31    PRR(J)=GAMMA*PR(J)+(1.-GAMMA)*PBAR(J)
      YPRR=FUNK(PRR)
      YTEST=Y(ILO)
      IF(YPRR.GE.YTEST) GO TO 26
319   DO 32 J=1,NVAR
32    P(IHI,J)=PRR(J)
      Y(IHI)=YPRR
      GO TO 1
38    DO 39 J=1,NVAR
39    P(IHI,J)=PR(J)
      Y(IHI)=YPR
40    DO 41 J=1,NVAR
41    PINV(J)=BETA*P(IHI,J)+(1.-BETA)*PBAR(J)
      YPINV=FUNK(PINV)
      IF(YPINV.GE.Y(IHI)) GO TO 50
      DO 42 J=1,NVAR
42    P(IHI,J)=PINV(J)
      Y(IHI)=YPINV
      GO TO 1
50    DO 55 I=1,NPOIN
      IF(I.EQ.ILO) GO TO 55
      DO 53 J=1,NVAR
      PR(J)=0.5*(P(I,J)+P(ILO,J))
53    P(I,J)=PR(J)
      Y(I)=FUNK(PR)
55    CONTINUE
60    GO TO 1
80    IPRIN=ILO
      RETURN
84    PRINT 841
841   FORMAT(' DID NOT CONVERGE')
      IPRIN=ILO
      RETURN
      END
```

```
      FUNCTION FUNK(PAR)
C**
C**   VERSION USING CUMULATIVE DISTRIBUTION FUNCTIONS (JUNE 1979)
C**
      DIMENSION PAR(20),IIND(1000)
      DIMENSION ROW(1000),IRANK(1000),NP1(1000)
      COMMON/F2DVEC/XR(1000),XI(1000),RR(1000),RI(1000)
      COMMON/F2DSCA/FACN,FACR,FAC1
      COMMON/F2DINT/LDAT,NUMR,NR,MPT,LAC,N1D,N2D,MAXLAG
      COMMON/INOV/R(1000)
C**
C**   INITIALIZE ARRAYS
C**
      DO 2 I=1,LDAT
      R(I)=0.0
```

```
      RR(I)=0.0
2     RI(I)=0.0
C**
C**   PUT FOURIER TRANSFORM OF A INTO (RR,RI)
C**
      JJ=0
      DO 20 I=1,LAC
      IF(I.EQ.MPT)GO TO 20
      JJ=JJ+1
      RR(I)=PAR(JJ)
20    CONTINUE
      RR(MPT)=1,0
      CALL FFT(RR,RI,LDAT,LDAT,LDAT,-1)
C**
C**   DERIVE INNOVATION (=A*X) WITH FOURIER TRANSFORMS
C**
      DO 3 I=1,LDAT
      QR=XR(I)*RR(I)-XI(I)*RI(I)
      QI=XR(I)*RI(I)+XI(I)*RR(I)
      RR(I)=QR
3     RI(I)=QI
      CALL FFT(RR,RI,LDAT,LDAT,LDAT,1)
      DO 4 I=1,LDAT
4     RR(I)=RR(I)*FACN
C**
C**   SHIFT, ORDER, AND DIFFERENCE INNOVATION
C**
      DO 5 I=N1D,N2D
      INDX=1-MPT+1
      IF(INDX.LE.0)GO TO 49
      R(INDX)=RR(I)
49    CONTINUE
      INDX=I-N1D+1
      RR(INDX)=RR(I)
5     CONTINUE
      DO 51 I=1,NUMR
51    IIND(I)=I
      CALL ORDER(RR,IIND,IRANK,NUMR)
      DO 52 I=1,NUMR
      INDY=IIND(I)
      RI(I)=RR(INDY)
      IRANK(INDY)=I
52    CONTINUE
C**
C**   THE RELATIONS BETWEEN THE INDICES FOR THE ORIGINAL (OLD)
C**   AND ORDERED (NEW) INNOVATION ARRAYS ARE AS FOLLOWS
C**         OLD=IIND(NEW)
C**         NEW=IRANK(OLD)
C**   NOW CALCULATE DR
C**
      DO 54 J=1,NUMR
      RR(J)=RI(J+1)-RI(J)
54    CONTINUE
C**
C**   NOW INTEGRATE THE FOLLOWING EXPRESSION
C**   DR(I)DR(I+LAG)(F2(R(I),R(I+LAG))-F1(R(I))F1(R(I+LAG)))**2
C**   "ROW" IS ROW OF THE MATRIX REPRESENTING
C**   THE CUMULATIVE DISTRIBUTION FUNCTION OF (R(I),R(I+LAG))
C**
C**   THE RESULTS ARE SUMMED FOR SEVERAL LAGS
C**
      FUNK=0.0
      DO 80 LAG=1,MAXLAG
      FAC1=1.0/FLOAT(NUMR-LAG)
C**
C**   INITIALIZE ARRAYS
C**
      DO 58 I=1,LDAT
      ROW(I)=0.0
```

```
58      NP1(I)=NUMR
C**
C**   FIND THE INDEX IN ORDERED INNOVATION
C**   CORRESPONDING TO THE GIVEN LAG
C**
        DO 59 J=1,NUMR
        INDY=IIND(J)+LAG
        IF(INDY.GT.NUMR)GO TO 59
        NP1(J)=IRANK(INDY)
59      CONTINUE
C**
C**   CARRY OUT SUMMATION IN EQUATION (170)
C**
        FSUM=0.0
        DO 64 J=1,NR
        DR=RR(J)
        IJUMP=NP1(J)
        FAC2=FLOAT(J)*FACR
        DO 60 I=1,NR
        IF(I.GE.IJUMP)GO TO 61
        FSUM=FSUM+DR*RR(I)*(ROW(I)-
        FAC2*FLOAT(I))**2
60      CONTINUE
        GO TO 64
61      CONTINUE
        DO 62 K=I,NR
        ROW(K)=ROW(K)+FAC1
        FSUM=FSUM+DR*RR(K)*(ROW(K)-
        FAC2*FLOAT(K))**2
62      CONTINUE
64      CONTINUE
        FUNK=FUNK+FSUM
80      CONTINUE
        RETURN
        END
```

```
        SUBROUTINE ORDER(D,II,JJ,N)
        DIMENSION II(N),JJ(N),D(N)
        K=1
10      KK=K+K
        IF(K.GE.N) RETURN
        CALL SORT(D,II,JJ,K,KK,N)
        K=KK
        IF(K.GE.N) GO TO 15
        KK=K+K
        CALL SORT(D,JJ,II,K,KK,N)
        K=KK
        GO TO 10
15      DO 16 I=1,N
16      II(I)=JJ(I)
        RETURN
        END
```

```
        SUBROUTINE SORT(D,II,JJ,K,KK,N)
        DIMENSION II(K,1),JJ(KK,1)
        M=N/KK
        IF(M.LE.0) GO TO 25
        DO 20 J=1,M
        I=J+J
20      CALL MERGE(D,II(1,I-1),K,II(1,I),K,JJ(1,J))
25      LEFT=N-KK*M
```

```
      IF(LEFT.LE.0) RETURN
      M1=M+1
      MM1=M+M1
      IF(LEFT.LE.K) GO TO 30
      LEFT=LEFT-K
      MM2=M1+M1
      CALL MERGE(D,II(1,MM1),K,II(1,MM2),LEFT,JJ(1,M1))
      RETURN
30    CALL MOVE(II(1,MM1),JJ(1,M1),LEFT)
      RETURN
      END
```

```
      SUBROUTINE MOVE(X,Y,N)
      INTEGER X,Y
      DIMENSION X(1),Y(1)
      NA=IABS(N)
      IF(NA.LE.0.OR.NA.GT.10000) RETURN
      IF(N) 10,30,20
10    DO 15 I=1,NA
15    Y(I)=-X(I)
      RETURN
20    DO 25 I=1,NA
25    Y(I)=X(I)
30    RETURN
      END
```

```
      SUBROUTINE MERGE(D,X,N,Y,M,Z)
      INTEGER X,Y,Z
      DIMENSION X(N),Y(M),Z(1),D(1)
      NM=N+M
      J=1
      I=1
      JGO=1
      IF(M.EQ.0) JGO=3
      IF(N.EQ.0) JGO=2
      DO 30 K=1,NM
      JX=X(J)
      IY=Y(I)
      GO TO (10,25,20),JGO
10    IF(D(JX).GT.D(IY)) GO TO 15
      Z(K)=JX
      IF(J.EQ.N) GO TO 17
      J=J+1
      GO TO 30
15    Z(K)=IY
      IF(I.EQ.M) GO TO 19
      I=I+1
      GO TO 30
17    JGO=2
      GO TO 30
19    JGO=3
      GO TO 30
20    Z(K)=JX
      J=J+1
      GO TO 30
25    Z(K)=IY
      I=I+1
30    CONTINUE
      RETURN
      END
```

---

program is written so that if three restarts are not sufficient, "DID NOT SETTLE" is written and the program continues. The rest of the program, from statement 3 on, is merely to evaluate the pulse shape C inverse to the converged A (in practice this should be printed or plotted, so that it can be seen how the pulse shape is changing as the procedure continues to higher orders). Also calculated is the quasi-FPE quantity given in equation (184). This number should also be printed.

Subroutine F2D sets up some constants that are needed in FUNK, computes the initial simplex using formulas given by Jacoby, Kowalik, and Pizzo (1972), calls the minimization routine, AMOEBA, and prints out the resulting AR filter. The program AMOEBA directly implements the simplex procedure as given in the references cited in the text. The criterion for convergence is in terms of the relative magnitudes of the maximum and minimum functional values on the simplex; this could be experimented with, as there are other equally valid convergence criteria.

Function FUNK is the guts of the program, as it provides the values, as a function of the AR parameters, of the measure of independence D_F which is to be minimized by AMOEBA. The evaluation of the innovation has been discussed in the text (§ Vc). The ordering of the innovation is important for an efficient evaluation of D_F and is carried out with sorting (SORT), moving (MOVE), and merging (MERGE) routines, all controlled by the main ordering program ORDER. These routines are based on material in the volume by Knuth (1973) and are such that the number of operations increases as $N \log N$. The only part of the procedure which produces an N^2 dependence is the summation over the two-dimensional grid.

The structure of the recursion for the summand in equation (170) (see eq. [171]) can be understood by reference to Figure 39. This figure shows the two-dimensional grid of the reordered values R'_n, with $R'_1 = \min_n(R_n)$ and $R'_{N^*} = \max_n(R_n)$. A given R'_i is paired with the R'_j which was its mth removed neighbor in the original (unordered) set $\{R_n\}$:

$$\left\{ \begin{matrix} R'_i \leftrightarrow R_n \\ R'_j \leftrightarrow R_{n+m} \end{matrix} \right\}. \tag{A1}$$

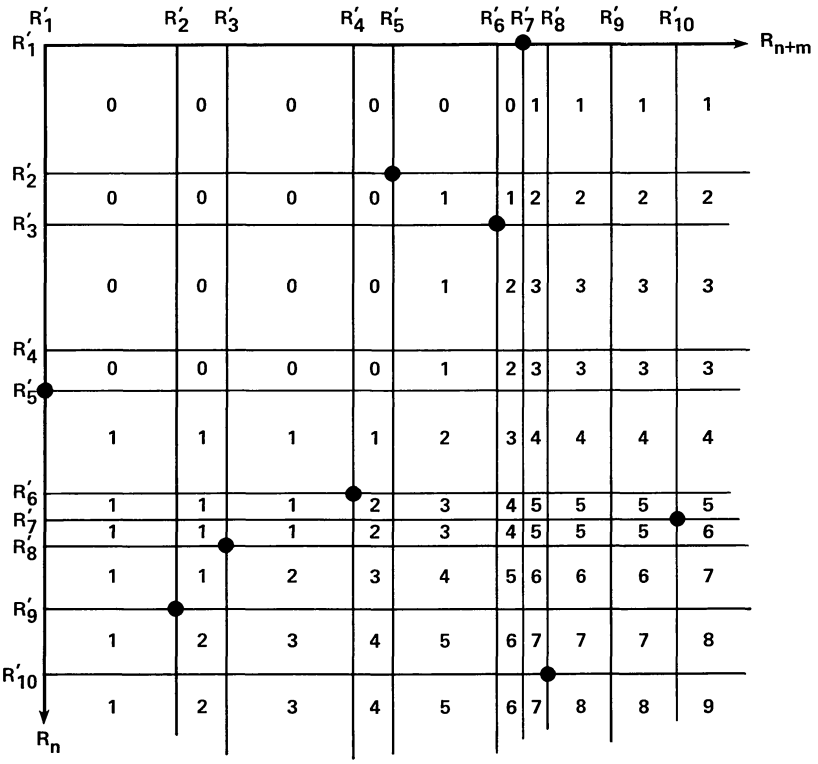

FIG. 39.—The two-dimensional grid used in the computation of the estimate of the joint cumulative distribution function. This example is for $m=1$ and $N^*=10$. Each of the N^*-m $(=9)$ pairs (R_n, R_{n+m}) is indicated with a dot at the intersection of the grid lines for these values (but labeled in terms of the ordered version of the innovation, R'). In this example, the original sequence was $(R'_9, R'_2, R'_5, R'_1, R'_7, R'_{10}, R'_8, R'_3, R'_6, R'_4)$. The numbers are the counts of the dots above and to the left of the box in which the number appears. The counts in each row are always 0 or 1 more than the counts in the row above: 0 for boxes to the left of the dot in the row, and 1 for boxes to the right (see eq. [171]). To get the function F_2, the counts must be normalized by the final count, $N^*-m=9$.

This pairing is indicated by the dots at the grid points in the figure. In the example shown, R'_1 is paired with R'_7, R'_2 with R'_5, and so forth. Each R'_i is of course paired with no more than one R'_j. For $m=1$, the R'_i equivalent to R_{N^*} has no mate, because R_{N^*+1} is not defined. Similarly for the R'_j equivalent to R_1. Hence there is one row and one column without a dot. (Similar results hold for larger values of m.)

Now $F_2^m(R'_p, R'_q)$ is $1/(N^*-m)$ times the number of pairs (dots) above and to the left of the point (R'_p, R'_q) (see eqs. [96] and [97]). A running count of this number is kept for successive rows in the grid. Since there is only one (or no) new point per row, this row count increases by unity for all squares to the right of the new point in the row. This relation is expressed in the recursion formula (171). The figure shows an example with $N^*=10$. The number in each box is the number of dots above and to the left of the box. The entries in the last row and column of the grid are never utilized but are shown to indicate how the normalization works: $F_2^m(x, y)$ for $x \geq R'_{N^*}$ and $y \geq R'_{N^*}$ is equal to the total number of dots $(=N^*-1)$ divided by N^*-1. The individual cumulative distribution is trivial in the system of the ordered R'_is:

$$F_1(R'_i)=\frac{i}{N^*}. \tag{A2}$$

INDEX

Absolute value (L_1), 13, 19, 25, **36–39**, 41–43, 45, 47–48
Acausal, 9, **12–14**, 19, 21–22, 24–25
Advance operator, **20**
All pass filter, **31**
Autocorrelation (function), **8**, 16, 17, 22–23, 28, 30, 31, 46
Autoregressive (AR) (model, process, representation), 2, **13–18**, 28–30, 32, 42–45, 47–59
  integrated moving average (ARIMA), **17–18**, 32, 42
  moving average (ARMA), **15**, 17, 29, 32
Bins, 33–34, 43, 47
Causal, **12**, 13–15, 19, 21–22, 24–25, 28, 29–30, 35, 40
Central limit theorem, 9
Characteristic function (see also joint characteristic function), **7**, 8, 32–33, 34, 43, 47
Computation (numerical experiments), 47–69
Constant component, 56, **57–58**
Convolution, 13, **20**, 23, 24, 31, 56
Cumulative distribution function (see also joint cumulative distribution function), **7**, 8, 32, 33, 43, 47–48, 64–66, 68–69
Decomposition, see Wold decomposition
Deconvolution, 28, **31**, 35, 47
  tables, 48, 49, 52, 56
Deterministic, 7–**8**, 9, 13, 28–29, 30, 35, 53
Delay character (also phase character), 3, 18–19, **22–23**, 29–30, 31–32, 39, 47, 56–57
  operator, 20
Dependence (dependently distributed, dependence measure), 8–9, **32–35**, 43, 47–48, 64–66
Difference operator (∇), **17**
Dipole (couplet), **21**, 22, 31
Discrete Fourier transform (DFT), see Fourier transform
Ergodic, 5, 32
Estimates, statistical, 29, 33, 35
Expected value, **7**, 8, 29, 32, 43
Factorization (into dipoles), **20–22**, 23–24
Filter (see also pulse shape), 9–10, **12–13**, **18–28**
  continuous, 46
Final prediction error (FPE), 36, 40, 44–45, 49, 52, 56
Fourier transform, 2, 20, 24, 25–28, 45, 56, 59, 60–61, 65
Frequency domain, 16, 20, 22
Gaps, see Sampling
Gaussian noise, see Noise, Gaussian
  process (normal process), 5, 31–32, 47–48, 52, 53
Identically and independently distributed (iid), 8
Identification (see also order), 3, 28, 40–41, **42**, 59
Impulse, 12, 14, **18**
Independent (independently distributed), **7** (random variables), **8** (processes), 9–10, 30, 31–32, 41–42, 59
Independently distributed innovations, 13, 30, 31–35, 52
  noise, see Noise, independently distributed
Innovation, **13**, 18, 29–32, **42–43**, 45–46, 50–55, 57–58, 68–69
Inverse (convolutional), 19, **23–28**, 29, 30, 40, 41–42, 45, 58
Joint characteristic function, **7**, 8, 9, 32
  cumulative distribution function, **7**, 9, 32, 33, 43, 64–66, 68–69
  probability distribution function, **7**, 9, 32, 33
Lag (m^* = maximum lag), 33, 34, 43, 44, 48–56, 60–69
Least-squares, 5, 25, 31, 35, 36
Linear system, see Filter
Local minimum, 34, 43–44, 59
Martingale difference property (MDP), **7**, 30, 34–35, 47–48
Maximum delay (or phase), **22–23**, 24–25, 30
  entropy method, 36, 43
Mean value, 8, 12, 17, 58
Memory, 12, 13, 28, 35
Minimization (optimization, deconvolution), 32, 33, 35–36, 38–45, 46, 48–49, 51–53, 56, 59, 62–64
Minimum delay (or phase), 19, **22–23**, 24–25, 29–30, 36, 38, 56–57
Mixed delay, **22**, 29–30, 39, 57
Models, 2, 5, **9**, 13, 28–42
Moment, 9, 34, 43
  generating function, **7**
Moving average (MA) (model, process, representation), 2, **9–13**, 14–15, 17–18, 28–32, 38, 42
Negative amplitude, 51, 56–58
Noise, 9, 47–48, 52, 53–54
  Gaussian, 9, 11, 16, 31–32, 47, 48, 52, 54–55
  independently distributed, 9, 16, 30, 32, 52
  uncorrelated, see Noise, white
  uniformly distributed (U), 6, 11, 18, 47–53
  white, 9, 11, 14, 15, 35, 47
Nonstationary, 6, 17
Norm, 36, **38–41**
Normalization, pulse, **13**, 14, 39–40, 61–62
One-sided (pulses, filters, representations; see also causal), **19**, 35, 37, 40, 46
Optimization, see Minimization
Order (of a process), 12–14, 40–41, 44–45, 50–51, 59
Ordering (according to magnitude), 43, 65, 66–67, 68
Origin of time, 12, **19**, 22, 25
Parsimony, **17**
Partial energy curve, **22–23**
Periodic signals (quasi-periodic signals), 14–16
Phase character, see delay character
Physically realizable, **19**
Poisson process, 18
Prediction (predictive deconvolution, predictive decomposition), 3, 8, 9, 29, **35–39**, 46
  error (see also Innovation), prediction error filter, 29, 35–39, 43, 46, 58
Probability distribution (see also Joint probability distribution), **6**, 8, 32, 33, 43, 47–48
Process, 3, **6**, **7**
Pulse shapes (see also Filter, Impulse), 12–13, 18–28, 31, 45, 50–51, 53–58, 61–62
  exponential, 6, 14, 17, 19, 26, 30–31, 47, 58
  rate, 18, 36
  amplitude (see also Innovation), 12, 18, 31, 36, 58
  amplitude distribution, 58
Purely random, **8**

Quasar, 3C 273, 53, 56–58
Random process (stochastic process), 3, **5–9**
Realizable, see Physically realizable
Realization (realization of a specific process), **6**, 8, 10, 11, 15–16, 48, 51, 54, 55
Restart, **43**, 44, 59, 61
Reverse, time, **22**, 23, 30–31, 36
Sampling, 3–6, 18, 19, **42**, 46
Sequential analysis, 6
Shot noise (model, process), 5, 12, **18**, 47
Simplex, 43–44, 62–64, 68
Sinusoidal signal, 52–53, 55
Skewness, time skewness function, see Time skewness
Skew-norm, 38
Spectrum, 16, 17, 22, 28, 29, 30, 36, 46, 59
Stability, filter (convergence), **12–13**, 14, **19**, 24, 29, 53
Stationary, 2, 6, **9**, 17, 28, 30, 33
Stochastic process, see Random process
Summation operator (S), **17**
Time domain, 3, 5, 20, 22
  series (see also Realization), 3, 5–6, 55
  skewness, 34, 36, 38–39
Trend (detrending), 6, 17, 29, 59
Two-sided filters (see also Acausal), 9, 13, 14, 15, **19**, 21, 25, 30, 39–40, 47
Uncorrelated (see also Noise, white), 7, **8**, 9–10, 28, 30, 31
Uneven sampling, see Sampling
Uniformly distributed noise, see Noise, uniformly distributed
Unstable, see Stability
Varimax norm, 43
Wavelet, 18, **29**
White noise, see Noise, white
Wold decomposition (Wold theorem and extension), 9, **28–29**, 30
Wraparound, **25**
Yule-Walker equations, **35–36**
Zero (of *Z*-transform), 21, 24–25
Z-transform, 15, **19–20**, 21–25, 28, 31

REFERENCES

Ables, J. 1974, *Astr. Ap. Suppl.*, **15**, 383.
Akaike, H. 1969*a*, *Ann. Inst. Stat. Math.*, **21**, 243–247.
_____. 1969*b*, *Ann. Inst. Stat. Math.*, **21**, 243–247.
_____. 1970*a*, *Ann. Inst. Stat. Math.*, **22**, 407–419.
_____. 1970*b*, *Ann. Inst. Stat. Math.*, **22**, 219–223.
_____. 1971, *Ann. Inst. Stat. Math.*, **23**, 163–180.
_____. 1974, *Ann Inst. Stat. Math.*, **26**, 363–387.
_____. 1975, *IEEE Trans.*, **AC-19**, 716–723.
Andersen, N. 1974, *Geophysics*, **39**, 69–72.
Anderson, T. W. 1963, in *Time Series Analysis*, ed. W. Rosenblatt (New York: John Wiley), pp. 425–446.
_____. 1971, *The Statistical Analysis of Time Series* (New York: John Wiley).
Åström, K. J., and Söderström, T. 1974, *IEEE Trans.*, **AC-19**, 769–773.
Bailey, N. T. J. 1964, *The Elements of Stochastic Processes with Applications to the Natural Sciences* (New York: John Wiley).
Barrodale, I. 1970, in *Approximation Theory*, ed. A. Talbot (New York: Academic).
Barrodale, I., and Roberts, F. D. K. 1973, *SIAM J. Numerical Anal.*, **10**, 839–848.
_____. 1974, *Assoc. for Comp. Mach.*, **17**, 319–320.
Barrodale, I., Roberts, F. D. K., and Hunt, C. R. 1970, *The Computer Journal*, **13**, 382–386.
Barrodale, I., and Young, A. 1966, *Numerische Mathematik*, **8**, 295–306.
Benveniste, A., Goursat, M., and Ruget, G. 1980, *IEEE Trans.*, **AC-25**, 385.
Berkhout, A. J. 1973, *Geophysics*, **38**, 657.
Bode, H. W., and Shannon, C. E. 1950, *Proc. I.R.E.*, **38**, 417–425.
Boehmer, A. M. 1967, *IEEE Trans.*, **IT-13**, 156–167.
Box, G. E. P., and Jenkins, G. M. 1970, *Time Series Analysis, Forecasting and Control* (San Francisco: Holden-Day).
Burg, J. P. 1967, paper presented at the 37th Annual International S.E.G. Meeting, Oklahoma City, 1967 October 31.
_____. 1968, *A New Analysis Technique for Time Series Data*, paper presented at the NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics, NATO, Enschede. This unpublished paper was followed up with several other unpublished papers: *Recommendations Concerning Maximum Entropy Spectral Estimation* (1973), *General Principles for Estimation of Covariance Matrices*, and *Time and Space Adaptive Deconvolution Filters* (with Don C. Riley). See Burg (1975).
_____. 1975, *Maximum Entropy Spectral Analysis*, Ph.D. thesis, Stanford University.
Caines, P. E., and Sethi, S. P., 1979, *IEEE Trans.*, **AC-24**, 113.
Chandrasekhar, S., and Münch, G. 1951, *Ap. J.*, **112**, 380.
Chow, J. C. 1972*a*, *IEEE Trans.*, **AC-17**, 268–269.
_____. 1972*b*, *IEEE Trans.*, **AC-17**, 386–387.
_____. 1972*c*, *IEEE Trans.*, **AC-17**, 707–709.
Claerbout, J. F. 1976, *Fundamentals of Geophysical Data Processing with Applications to Petroleum Prospecting* (McGraw-Hill, Inc.).
Claerbout, J. F., and Muir, F. 1973, *Geophysics*, **38**, 826.
Deeming, T. J. 1970, *A.J.*, **75**, 1027.
Donoho, D. 1981, in *Second Applied Time Series Analysis Symposium*, ed. D. Findley (New York: Academic Press).
Doob, J. L. 1953, *Stochastic Processes* (New York: Wiley).
Durbin, J. 1959, *Biometrika*, **46**, 306.
_____. 1960, *Rev. Int. Inst. Statist.*, **28**, 233.
Ekblom, H., and Henriksson, S. 1969, *SIAM J. Appl. Math.*, **17**, 1130–1140.
Enochson, L. D., and Otnes, R. K. 1968, *Programming and Analysis for Digital Time Series Data* (Washington: Navy Publication and Printing Service Office).
Fahlman, G. G., and Ulrych, T. J. 1975, *Ap. J.*, **201**, 277.
_____. 1976, *Ap. J.*, **209**, 663.
Feller, W. 1957, *An Introduction to Probability Theory and Its Applications*, Vol. **1** (2d ed., New York: John Wiley).
Frenkiel, F. N., and Klebanoff, P. S. 1967, *Physics of Fluids*, **10**, 507.
Gailbraith, J. N. 1971, *Geophysics*, **35**, 25–265.
Gersch, W. 1970, *IEEE Trans.*, **AC-15**, 583–588.
Gersch, W. and Foutch, D. A. 1974, *IEEE Trans.*, **AC-19**, 898–903.
Gertler, J., and Bányász, C. 1974, *IEEE Trans.*, **AC-19**, 816–820.
Gold, B., and Rader, C. M. 1969, *Digital Processing of Signals* (New York: McGraw-Hill).
Gosling, J. T., Hundhausen, A. J., Pizzo, V., and Asbridge, J. R. 1972, *J. Geophys. Res.*, **77**, 5442.
Granger, C. W. J., and Newbold, P. 1977, *Forecasting Economic Time Series* (New York: Academic Press).
Graupe, D., Krause, D. J., and Moore, J. B. 1975, *IEEE Trans.*, **AC-20**, 104–107.
Gray, H. L., Kelley, G., and McIntire, D. 1977, Southern Methodist University, Dept. of Statistics, Tech. Report No. 126.
Hannan, E. J. 1970, *Multiple Time Series* (New York: John Wiley).
_____. 1975, *IEEE Trans.*, **AC-19**, 706–715.
Jacoby, S. L. S., Kowalik, J. S., and Pizzo, J. T. 1972, *Iterative Methods for Nonlinear Optimization Problems* (New Jersey: Prentice-Hall).
Jenkins, G. M., and Watts, D. G. 1969, *Spectral Analysis and Its Applications* (San Francisco: Holden-Day, Inc.).
Jones, R. H. 1974, *IEEE Trans.*, **AC-19**, 894–897.
Jury, E. I. 1964, *Theory and Application of the Z-Transform Method* (New York: John Wiley).
Kailath, T. 1968, *IEEE Trans.*, **AC-13**, 646–655.
_____. 1974, *IEEE Trans.*, **IT-20**, 146–181.
Kaiser, H. F. 1958, *Psychometrica*, **23**, 187.
Kanasewich, E. R. 1975, *Time Sequence Analysis in Geophysics* (Edmonton, Alberta: University of Alberta Press).
Kashyap, R. L. 1974, *IEEE Trans.*, **AC-19**, 13–21.
Knuth, D. E. 1973, *The Art of Computer Programming*. Vol. **3**, *Searching and Sorting*. (Reading, Mass.: Addison-Wesley).
Kolmogorov, A. N. 1941, *Bull. Moscow State University, Math.*, **2**:6, 1 (*Math. Rev.*, **5**, 101; **13**, 1138).
Krishnaiah, P. R. 1969, *Multivariate Analysis*, Vol. **2** (New York: Academic Press).
Kunkel, W. E. 1967, *A. J.*, **72**, 1341.

Lacoss, R. T. 1971, *Geophysics*, **36**, 661–675.
Levinson, H. 1947, *J. Math. Phys.*, **25**, 261.
Lindberger, N. A. 1972, *IEEE Trans.*, **AC-17**, 689–691.
Luenberger, D. G. 1969, *Optimization by Vector Space Methods* (New York: John Wiley).
Makhoul, J. 1975, *Proc. IEEE*, **63**, 561.
Mann, H. B., and Wald, A. 1943, *Econometrica*, **11**, 173–220.
Maria, G. A. and Fahmy, M. M. 1974, *IEEE Trans.*, **ASSP-22**, 15–21.
Medd, W. J., Andrew, B. H., Harvey, G. A., and Locke, J. L. 1972, *Mem. R.A.S.*, **77**, 109.
Mendel, J. 1980, unpublished.
Nelder, J. A., and Mead, R. 1965, *The Computer Journal*, **8**, 308.
Ooe, M., and Ulrych, T. J. 1979, *Geophysical Prospecting*, **27**, 52.
Oppenheim, A. V., and Schafer, R. W. 1975, *Digital Signal Processing* (Englewood Cliffs: Prentice-Hall).
Osborne, M. R., and Watson, G. A. 1971, *The Computer Journal*, **14**, 184–188.
Papoulis, A. 1965, *Probability, Random Variables, and Stochastic Processes* (New York: McGraw-Hill).
Parzen, E. 1960, *Modern Probability Theory and Its Application* (New York: John Wiley).
_____. 1962, *Stochastic Processes* (San Francisco: Holden-Day).
_____. 1967, *Time Series Analysis Papers* (San Francisco: Holden-Day).
_____. 1968, Stanford Tech. Rept. No. 11.
_____. 1969, in *Multivariate Analysis*, Vol. **2**, ed. P. R. Krishnaiah (New York: Academic Press), p. 309.
_____. 1974, *IEEE Trans.*, **AC-19**, 723.
_____. 1979, *Journal Stat. Assoc.*, **74**, 105.
Peacock, K. L., and Treitel, S. 1969, *Geophysics*, **34**, 155.
Powell, M. J. D. 1964, *The Computer Journal*, **7**, 155.
Press, W. H. 1978, *Comments Ap.*, **7**, 103.
Rabiner, L. R. and Gold, B. 1975, *Theory and Application of Digital Signal Processing* (Englewood Cliffs: Prentice-Hall).
Rice, J. R. 1964, *The Approximation of Functions*, Vol. **1** (Reading, Mass.: Addison-Wesley).
Rice, J. R., and White, J. S. 1964, *SIAM Review*, **6**, 243–256.
Robers, P. D., and Ben-Israel, A. 1969, *J. Approx. Theory*, **2**, 323–336.
Robinson, E. A. 1962, *Random Wavelets and Cybernetic Systems* (New York: Haffner).
_____. 1963, *Proceedings of the Symposium on Time Series Analysis*, ed. M. Rosenblatt (New York: John Wiley).
_____. 1964*a*, in *Econometric Model Building*, ed. H. Wold (Amsterdam: North-Holland Publishing Co.), p. 37.
_____. 1964*b*, in *Econometric Model Building*, ed. H. Wold (Amsterdam: North-Holland Publishing Co.), p. 111.
_____. 1966, *Geophysics*, **31**, 482.
_____. 1967*a*, *Multichannel Time Series Analysis with Digital Computer Programs* (San Francisco: Holden-Day).
_____. 1967*b*, *Geophysics*, **32**, 418–484.
Rosenblatt, M. 1963, *Proceedings of the Symposium on Time Series Analysis* (New York: John Wiley).
_____. 1971, *Markov Processes. Structure and Asymptotic Behavior* (New York: Springer Verlag).
Rothschild, R. 1977, in *Recognition of Compact Astrophysical Objects*, ed. H. Ogelman and R. Rothschild (NASA SP-421).
Scargle, J. D. 1977, *IEEE Trans.*, **IT-23**, 140.
_____. 1981, in *Second Applied Time Series Symposium*, ed. D. Findley (New York: Academic Press).
Schoenberger, M. 1974, *Geophysics*, **39**, 828–833.
Segall, A. 1976, *IEEE Trans.*, **IT-22**, 275.
Shinners, S. M. 1974, *IEEE Trans.*, **SMC-4**, 446.
Simpson, S. M., Jr. 1966, *Time Series Computations in Fortran and FAP*, Vol. **1**, *A Program Library* (Reading, Mass.: Addison Wesley).
Smylie, D. E., Clarke, K. C., and Ulrych, T. J. 1973, in *Methods of Computational Physics*, Vol. **13** (New York: Academic Press, Inc.), p. 391.
Steiglitz, K. 1974, *An Introduction to Discrete Systems* (New York: John Wiley).
Stralkowski, C. M., Wu, S. M., and DeVor, R. E. 1970, *Technometrics*, **12**, 669.
_____. 1974, *Technometrics*, **16**, 275.
Terrell, N. J. 1972, *Ap. J. (Letters)*, **174**, L35.
Terrell, N. J., and Olsen, K. H. 1970, *Ap. J.*, **161**, 399.
Terrell, N. J., and Olsen, K. H. 1972, in *External Galaxies and Quasi Stellar Objects*, ed. D. S. Evans, I.A.U.
Titchmarsh, E. C. 1939, *The Theory of Functions* (London: Oxford University Press).
Tong, H. 1975, *IEEE Trans.*, **IT-25**, 476.
_____. 1976, *IEEE Trans.*, **IT-22**, 493.
Treitel, S., and Robinson, E. A. 1966, *IEEE Trans.*, **GE-4**, 25.
Ulrych, T. J. 1972, *J. Geophys. Res.*, **77**, 1396.
Ulrych, T. J., and Bishop, T. N. 1975, *Rev. Geophys. Space Phys.*, **13**, 183.
Ulrych, T. J., and Clayton, R. W. 1976, *Phys. Earth Planet. Int.*, **12**, 188.
Walker, A. M. 1962, *Biometrika*, **49**, 117.
Wax, N. 1954, *Selected Papers on Noise and Stochastic Processes* (New York: Dover).
Weisskopf, M. C., Sutherland, P. G., Katz, J. I., and Canizares, C. R. 1978, *Ap. J. (Letters)*, **223**, L17.
Whittle, P. 1963, *Prediction and Regulation by Linear Least-Square Methods* (Princeton: D. Van Nostrand).
Wiener, N. 1949, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications* (Cambridge: MIT Press).
Wiggins, R. A. 1977, *Minimum Entropy Deconvolution*, Proc. IEEE Computer Society, **7**.
Wold, H. 1938*a*, *A Study in the Analysis of Stationary Time Series*, 2d ed. (Uppsala: Almqvist and Wiksell).
_____. 1938*b*, *Skandinavisk Aktuarietidskrift*, **21**, 208 (in English).
_____. 1964, *Econometric Model Building* (Amsterdam: North-Holland).
_____. 1965, *Bibliography on Time Series and Stochastic Processes* (Cambridge: MIT Press).
Yule, G. U. 1927, *Phil. Trans. Roy. Soc. London, A*, **226**, 267.

JEFFREY D. SCARGLE: Mail Stop 245-3, Theoretical and Planetary Studies Branch, NASA Ames Research Center, Moffett Field, CA 94035