BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 16, Number 1, January 1987

MAXIMUM ENTROPY AND THE MOMENT PROBLEM

H. J. LANDAU

Introduction. The trigonometric moment problem stands at the source of several major streams in analysis. From it flow developments in function theory, in spectral representation of operators, in probability, in approximation, and in the study of inverse problems. Here we connect it also with a group of questions centering on entropy and prediction. In turn, this will suggest a simple approach, by way of orthogonal decomposition, to the moment problem itself.

In statistical estimation, one often wants to guess an unknown probability distribution, given certain observations based on it. There are generally infinitely many distributions consistent with the data, and the question of which of these to select is an important one. The notion of entropy has been proposed here as the basis of a principle of salience which has received considerable attention. We will show that, in the context of spectral analysis, this idea is linked to a certain question of prediction by the trigonometric moment problem, and that all three strongly illuminate one another. The phenomena we describe are known, but our object is to unify them conceptually and to reduce the analytic intricacy of the arguments. To this end, we give a completely elementary discussion, virtually free of calculation, which shows that all the facts, including those concerning the moment problem, can be understood as direct consequences of orthogonal decomposition in a finite-dimensional space. We then describe how, in its continuous version, this leads to a view of second-order Sturm-Liouville differential equations, and conclude with some questions concerning the connection between combinatorial ideas and orthogonality in this problem.

Entropy and statistical inference. Suppose that we are interested in the distribution of some quantity, but know only the values of certain averages defined by that distribution, which are insufficient to specify it uniquely. For example, we might have tossed a six-sided die fifty times, wishing to find how often each face appeared, but were able to observe only the average value of these faces. What should we select as an appropriate distribution, on the strength of the available information? Various criteria have been proposed to

Received by the editors May 21, 1986.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 42A70; Secondary 42A05, 62M15, 94A17, 60G25.

0273-0979/87 $1.00 + $.25 per page

guide the choice; we focus on the following argument, which brings into play the notion of entropy, prominent in information theory.

Suppose that an experiment, which can have k different results, is performed in a (long) run of N repetitions. If the ith result occurred n_i times in the run, we denote by $f_i \equiv n_i/N$ its *realized frequency*; clearly, $f_i \geq 0$, $\Sigma f_i = 1$. A set $\{f_i\}$ of such numbers is called a *distribution* f, and the quantity $H_f \equiv -\sum_{i=1}^{k} f_i \log f_i$ is termed the *entropy* of that distribution.

There are k^N distinct possible outcomes for a run of N trials of the experiment. Of these, the number which give rise to a particular distribution of frequencies $f_1, \ldots, f_k$ is $W = N!/(Nf_1)! \cdots (Nf_k)!$, whence, for large N, Stirling's formula shows $(\log W)/N$ to be asymptotic to H_f. Thus distributions having entropy close to the maximum are realized in the greatest number of ways, hence occur most often in the list of all possible outcomes. Suppose now that we cannot directly observe the realized frequencies $f_1, \ldots, f_k$ in a run, but that some process of measurement has fixed the values of certain functions of the $\{f_i\}$. If, correspondingly, we restrict consideration to only those runs in which the frequencies satisfy these prescribed constraints, the same counting argument again shows that the vast majority have frequency distributions for which the entropy is close to the maximum attainable under the constraints.

In its limiting asymptotic form, this phenomenon was known already to Laplace. For constraints that are linear in the $\{f_i\}$, a more precise quantitative description, as a function of N and of the number of constraints, has been given by E. T. Jaynes in the *entropy concentration theorem* [**20**]. This result allows an accurate estimate of how sharply the entropy is concentrated, and shows that distributions whose entropy is close to the maximum predominate among all the possible outcomes even for relatively small values of N. Thus the frequency distribution having greatest entropy, subject to the constraints, can be viewed as the *most representative* of the class of candidates, and therefore a good guess for the unknown frequencies underlying the observed data.

One can also express this in the language of information theory where, intuitively, the information associated with an outcome measures how surprising its occurrence is among the various possible outcomes. In these terms, as the entropy-maximizing distribution is encountered most often in the ensemble, it is the *least informative*, for a different one would exclude the bulk of possibilities and thereby convey more information—indeed, the argument goes, more than the data warrant.

In sum, the distribution with largest entropy, which fits the observations, is recommended as the most typical of the possibilities or, equivalently, as the most appropriate to the available information. The formulation here is deliberately combinatorial, so as to yield clear-cut conclusions, free of the complications attendant on interpreting probability as frequency. Nevertheless, it seems sensible—albeit with less precise justification—to use this distribution also to represent the probabilities in a probabilistic model of the experiment.

The reasonableness of entropy maximization as a selection mechanism is independently supported by the axiomatic analysis of J. E. Shore and R. W. Johnson [**36**], which establishes that criterion as the only method of inference satisfying certain consistency conditions. Moreover, the principle that nature

favors the states of largest entropy was introduced with striking success by Maxwell, Boltzmann, and Gibbs into statistical mechanics, where it continues to play an important role. The observation [**38**] that all of the probability distributions commonly encountered in statistics maximize the entropy, under suitable choice of linear constraints, further buttresses these ideas.

Finally, we note that, when generalized in the form of Kullback-Leibler information, entropy is connected with other statistical problems as well. For example, it can guide the updating of a prior guess of an unknown distribution in the light of new evidence [**37**]. It also enters into the analysis of the interesting EM algorithm, often used to find the maximum likelihood estimate, namely, that probability distribution in a given family which is most likely to have generated the observed data [**13, 12, 39**].

Stationary time series. A particular estimation problem of the preceding type, which occurs often and in various contexts, concerns a time series: that is, a sequence $\{X_k\}$ of random variables characterized by a family of distributions that specify probabilities for the joint occurrence of values of finite subsets of the variables. These probability distributions define expected value, denoted by $\mathscr{E}(\cdot)$, for functions of the random variables. Interpreting the index k as time, such a sequence is termed *stationary* if each of these distributions is independent of the choice of time-origin, that is, if

$$\text{Prob}\left\{|X_{i_1} - \nu_1| < \varepsilon_1, |X_{i_2} - \nu_2| < \varepsilon_2, \ldots, |X_{i_m} - \nu_m| < \varepsilon_m\right\}$$

is unchanged when $i_1, \ldots, i_m$ are shifted to $i_1 + j$, $i_2 + j, \ldots, i_m + j$, for any choice of j, m, i_k, ν_k, and ε_k, $1 \leqslant k \leqslant m$; it is called *stationary in the wide sense* if merely the expectation of quadratic functions, $\mathscr{E}(X_{i+j}\overline{X}_{k+j})$, is independent of j, for all i and k.

Stationary time series are used successfully to model a wide range of fluctuating phenomena, from samples of speech to geophysical measurements to economic variables. Typically, what is known in such applications, or can be estimated from observation by averaging, consists of *autocorrelation coefficients* $c_k \equiv \mathscr{E}(X_i\overline{X}_{i+k}) = \mathscr{E}(X_0\overline{X}_k)$, for a finite set of values $k = 0, \ldots, N$, and the question is how to select an appropriate probabilistic description of the process on the basis of this information.

In a remarkable, incisive analysis [**6, 9**], J. P. Burg introduced the criterion of maximum entropy into this problem and solved it explicitly: the entropy-maximizing distribution corresponds to the Gauss-Markov (autoregressive) process of minimal order (these terms to be defined later) having the given correlations. He also noted that his process produced the poorest prediction from the past; this gives a different sense in which it is least informative. Much has been written about this method, and several proofs are available [**10, 11, 18, 20, 34,** and references therein].

Here we will draw on basic considerations connected with the trigonometric moment problem to show, from first principles, that a natural finite-dimensional orthogonal decomposition underlies all of the results in this area, including those associated with the moment problem itself.

The trigonometric moment problem. The trigonometric moment problem asks when a given sequence $1 = c_0, c_1, \ldots$ of complex numbers can be represented in the form

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} e^{ik\theta}\, d\mu(\theta), \qquad k \geq 0, \tag{1}$$

with some positive measure $d\mu(\theta)$, generally assumed to have an infinite number of points of increase. This representation illuminates an extraordinary range of subjects, including analytic and harmonic functions, the spectral theory of operators, prediction theory, and approximation; as M. G. Krein has pointed out, it also constitutes an archetypal inverse problem [**27, 14**]. Defining $c_{-k} = \bar{c}_k$, clearly a necessary condition for (1) is that

$$\sum\sum a_j \bar{a}_k c_{j-k} > 0, \tag{2}$$

for any choice of a finite number of nonzero $\{a_j\}$, since by (1) this quadratic form equals $(1/2\pi)\int_0^{2\pi}|\sum a_k e^{ik\theta}|^2\, d\mu(\theta)$. Requirement (2) turns out also to be sufficient. The classical proofs exploit the positivity in (2) mainly by means of convexity. To summarize broadly the lucid exposition of [**1**], one such interpretation, rich in connections with function theory, was introduced by Carathéodory, Toeplitz, Herglotz, and F. Riesz, who showed that, on associating $\{c_k\}$ with $f(z) \equiv c_0/2 + \sum_{k=1}^{\infty} c_k z^k$, sequences satisfying (2) correspond to analytic functions in $|z| < 1$ with positive real part; an integral representation of this convex family based on the Poisson formula then yields the desired form (1) for the coefficients. The related function $g(z) \equiv \{f(z) - f(0)\}/\{f(z) + \overline{f(0)}\}$ maps the unit disk into itself. By repeatedly applying Schwarz's lemma and a linear fractional transformation to $g(z)$, Schur derived an explicit characterization of such maps g, which in turn generates all moment sequences. It is interesting that this algorithm now figures in simulating certain physical systems, and in signal processing [**21, 23**]. Alternatively, an elegant argument due to M. Riesz begins with convexity of polynomials having the form $|\sum a_k z^k|^2$, uses (2) to define a positive linear functional on such polynomials, and extends this functional with its positivity preserved to the manifold of step functions, where it is given by a measure. Finally, the beautiful memoir of M. G. Krein [**26**], interpreting (2) to mean that the point $(1, \operatorname{Re} c_1, \operatorname{Im} c_1, \ldots, \operatorname{Re} c_k, \operatorname{Im} c_k)$ in $(2k+1)$-dimensional space lies in the convex hull of the curve generated by $(1, \cos\theta, \sin\theta, \ldots, \cos k\theta, \sin k\theta)$, $0 \leq \theta \leq 2\pi$, obtains (1) by a representation of points in convex bodies, and develops this idea into the far-reaching generalization of Tchebycheff spaces [**25**].

Here we suggest a different approach: we use (2) to define a scalar product for polynomials of degree n, and systematically apply orthogonal decomposition. The various features—orthogonal polynomials, recursions, reflection coefficients, quadrature formulas, prediction—of the problem, and their interrelationship, emerge naturally from this geometric setting.

A scalar product. Since the quadratic form (2) is positive definite, let us think of it as defining a scalar product on the linear space of finite sequences $(a_0, a_1, \ldots, a_k)$ or, equivalently, on polynomials. Specifically, for $n > 0$, let $1 = c_0, \ldots, c_n$ be given and satisfy (2). Let $\mathbf{C}_n$ denote the $(n+1) \times (n+1)$

matrix $[c_{j-k}]$, $j, k = 0, \ldots, n$, with $c_{-k} = \bar{c}_k$; this is a *Toeplitz* matrix—that is, one whose entries are constant along each diagonal—and is Hermitian. Let $a = (a_0, \ldots, a_n)$ and $b = (b_0, \ldots, b_n)$ represent $(n+1)$-dimensional vectors, with $(a, b) \equiv \sum a_i \bar{b}_i$; and associate to such a vector a the (trigonometric) polynomial of degree n

$$A = \sum_{k=0}^{n} a_k e^{ik\theta} = \sum_{k=0}^{n} a_k z^k, \qquad |z| = 1,$$

denoting by Π_n the $(n+1)$-dimensional space of these polynomials. Then in view of (2) we can introduce a scalar product in Π_n by setting, for $B = \sum_0^n b_k z^k$,

$$[A, B] \equiv (a, \mathbf{C}_n b) = (\mathbf{C}_n a, b) = \overline{[B, A]} \tag{3a}$$

$$= \sum \sum a_j \bar{b}_k c_{j-k}, \tag{3b}$$

$$\|A\|^2 \equiv [A, A].$$

The reason for applying this scalar product to polynomials, rather than to the vectors themselves, is that the Toeplitz nature of $\mathbf{C}_n$ is now succinctly expressed by

$$[zA, zB] = [A, B], \tag{4}$$

for $A, B \in \Pi_{n-1}$. As a matter of notation, we will write both column and row vectors in row form, refer to a as the *coefficient vector* of A, to a_n as the *leading coefficient* of A, and will say that A and B are *orthogonal* whenever $[A, B] = 0$.

A basis: orthogonal polynomials. As progressively more coefficients of the sequence $(c_0, c_1, \ldots)$ are given, the corresponding matrices $\mathbf{C}_n$ define the scalar product on the increasing family of subspaces $\{\Pi_n\}$ without altering it where previously specified. Thus it is natural, in choosing an orthogonal basis for Π_n, to select the coordinate elements so as to span subspaces of polynomials of successively increasing degree. This can be done by the Gram-Schmidt process applied to $1, z, \ldots, z^n$. Specifically, suppose $P_k(z) \in \Pi_k$ has coefficient vector τ_k, with leading coefficient t_k, such that

$$\mathbf{C}_k \tau_k = (0, \ldots, 0, 1). \tag{5}$$

Then by (3a), $[S_{k-1}, P_k] = 0$ for each $S_{k-1} \in \Pi_{k-1}$, hence P_k is a scalar multiple of the desired kth basis element. To normalize it, we note that, again by (3) and (5),

$$0 < \|P_k\|^2 = (\tau_k, \mathbf{C}_k \tau_k) = t_k, \tag{6}$$

so that $P_k(z)/\sqrt{t_k}$, which has degree k and leading coefficient $\sqrt{t_k}$, is the kth unit element of the basis—that is, the kth (normalized) orthogonal polynomial. We remark that this orthogonality can also explain the frequent occurrence of the vectors τ_k of (5) in methods for inverting Toeplitz matrices, a topic we consider later.

PROPOSITION 1. *All the zeros of $P_k(z)$ lie in $|z| < 1$.*

PROOF. This seemingly analytic fact depends only on orthogonality (and on the fundamental theorem of algebra). For suppose γ to be a zero of P_k, so that $P_k(z) = (z - \gamma)S_{k-1}(z)$, or

$$P_k(z) + \gamma S_{k-1}(z) = zS_{k-1}(z), \tag{7}$$

with some $S_{k-1} \in \Pi_{k-1}$. Then, since P_k is orthogonal to Π_{k-1}, on taking norms in (7) and using (4) we find

$$\|P_k\|^2 + |\gamma|^2 \|S_{k-1}\|^2 = \|zS_{k-1}\|^2 = \|S_{k-1}\|^2,$$

whence $1 - |\gamma|^2 = \|P_k\|^2/\|S_{k-1}\|^2 > 0$, as required.

Evaluation polynomials. If the moment problem has a solution, then on substituting (1) into (3b) we obtain a representation of $[S, T]$ in the form

$$[S, T] = \frac{1}{2\pi} \int_0^{2\pi} S(e^{i\theta}) \overline{T(e^{i\theta})}\, d\mu(\theta); \tag{8}$$

conversely, if such a representation exists with some $d\mu \geqslant 0$, then by (3b) the choice $S = z^j$, $T = z^k$ shows the moment problem to have a solution. Thus we can view that problem as asking for the relationship between the scalar product of polynomials and their values. To pursue this matter, we focus on the operation which assigns to $S_n(z)$ its value at a specific point $z = \zeta$. As this is a bounded linear functional defined in Π_n, it can be represented by

$$S_n(\zeta) = [S_n, E_n^\zeta], \tag{9}$$

with $E_n^\zeta(z)$ a unique element of Π_n, which we call the *evaluation polynomial* for ζ; E_n^ζ is sometimes also referred to as the *reproducing kernel.*

PROPOSITION 2.

$$E_n^\zeta(\mu) = \overline{E_n^\mu(\zeta)} \quad \text{and} \quad E_n^\zeta(\zeta) = \|E_n^\zeta\|^2; \tag{10a}$$

$$E_n^\zeta(z) = \sum_{k=0}^{n} \overline{P_k(\zeta)} P_k(z)/t_k; \tag{10b}$$

$$\|E_n^\zeta\|^{-2} = \inf_{S_n(\zeta)=1} \|S_n\|^2. \tag{10c}$$

PROOF. From the defining property (9),

$$E_n^\zeta(\mu) = [E_n^\zeta, E_n^\mu] = \overline{[E_n^\mu, E_n^\zeta]} = \overline{E_n^\mu(\zeta)},$$

$$\|E_n^\zeta\|^2 = [E_n^\zeta, E_n^\zeta] = E_n^\zeta(\zeta).$$

If $E_n^\zeta = \sum_{k=0}^n a_k P_k$ is the decomposition of E_n^ζ in the orthogonal basis formed by $\{P_k\}$, then, as in any orthogonal expansion, by forming the scalar product of both sides with P_j we find, from (6), (3b), and (9),

$$a_j \|P_j\|^2 = a_j t_j = [E_n^\zeta, P_j] = \overline{[P_j, E_n^\zeta]} = \overline{P_j(\zeta)}.$$

Finally, if $1 = S_n(\zeta) = [S_n, E_n^\zeta]$ then by Schwarz's inequality $1 = |[S_n, E_n^\zeta]|^2 \leqslant \|S_n\|^2 \|E_n^\zeta\|^2$, and equality is attained only if S_n is proportional to E_n^ζ, i.e. for $S_n(z) = E_n^\zeta(z)/E_n^\zeta(\zeta)$.

We now single out $E_n^0(z)$, the evaluation polynomial at $z = 0$, which has a special connection to $P_n(z)$. For since $S(0)$ is the constant term of S, (9) and (3a) show that the coefficient vector ε_n of $E_n^0(z)$ must satisfy

$$\mathbf{C}_n\varepsilon_n = (1, 0, \ldots, 0). \tag{11a}$$

Since $\mathbf{C}_n$ is Toeplitz and Hermitian, we see on comparing with (5) that to obtain ε_n we need only write the components of $\bar{\tau}_n$ in reverse order. Thus $P_n(z)$ determines $E_n^0(z)$, and conversely. We can express this compactly by the equivalent formulae

$$E_n^0(z) = z^n\bar{P}_n\left(\frac{1}{z}\right),$$

$$P_n(z) = z^n\bar{E}_n^0\left(\frac{1}{z}\right). \tag{11b}$$

By construction, (10a), and (6), we have

$$\|E_0\|^2 = E_n^0(0) = t_n = \|P_n\|^2, \tag{12a}$$

and either from (11b) or from an explicit expansion,

$$\left|E_n^0(e^{i\theta})\right|^2 = \left|P_n(e^{i\theta})\right|^2. \tag{12b}$$

Of course, $E_n^0(z)$ has been singled out in previous discussions of the subject; it is often denoted by $P_n^*(z)$. The advantage of our treatment is that we have characterized it by its action in the scalar product, rather than by its analytic form.

We conclude by showing that the map from $(c_0, \ldots, c_n)$ to $E_n^0(z)$ is one-to-one.

PROPOSITION 3. *$E_n^0(z)$ determines $(c_0, \ldots, c_n)$.*

PROOF. We have seen in (11b) that $E_n^0(z)$ determines $P_n(z)$, and so also $E_n^0(z) - \overline{P_n(0)}P_n(z)/t_n = E_{n-1}^0(z)$ by (10b). Continuing in this way, we see that $E_n^0(z)$ determines all the $P_k(z)$, $k \leqslant n$. As the leading coefficient of every P_k is nonzero, this in turn defines, for each $j \leqslant n$, the coefficients in the expansion $z^j = \sum_{k=0}^{j} a_{j,k}P_k(z)$, from which c_j is determined by $c_j = [z^j, 1] = [\sum a_{j,k}P_k, P_0] = a_{j,0}$.

Spectral estimation. Suppose that $J + 1$ autocorrelation coefficients $1 = c_0, \ldots, c_J$ of a stationary process $\{X_k\}_{-\infty}^{\infty}$ (or, time series) are given. From the definition of c_k we see that

$$\sum_{j,k=0}^{J} a_j\bar{a}_k c_{j-k} = \sum_{j,k=0}^{J} a_j\bar{a}_k\mathscr{E}\left(X_j\bar{X}_k\right) = \mathscr{E}\left(\left|\sum_{k=0}^{J} a_kX_k\right|^2\right) > 0;$$

hence such coefficients always generate a sequence satisfying (2). Moreover, from (3b)

$$[A_J, B_J] = \mathscr{E}\left(\sum_{k=0}^{J} a_kX_k\,\overline{\sum_{m=0}^{J} b_mX_m}\right), \tag{13a}$$

so that the associated scalar product of two polynomials now finds a concrete interpretation as the correlation between the corresponding linear combinations of successive elements of the time series. In this identification, orthogonal polynomials represent uncorrelated random variables.

We call any nonnegative measure $d\mu(\theta)$ for which

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} e^{ik\theta}\, d\mu(\theta), \qquad 0 \leqslant k \leqslant J,$$

a *spectrum consistent with the data*, and denote by $\mathscr{M}_J$ the collection of all such measures. In other terminology, $\mathscr{M}_J$ consists of all solutions to the *truncated moment problem*, in which the representation (1) holds only for $c_0, \ldots, c_J$.

Important for us will be that each measure of $\mathscr{M}_J$ expresses the scalar product in Π_J by the formula (8), and, conversely, each positive measure so doing is a member of $\mathscr{M}_J$.

In the stochastic context, one frequently aims to describe the process at hand so as to clarify its character, behavior, or provenance. To this end, one generally tries, insofar as is practicable, to extract from it components that can be viewed as predictable, or coherent, leaving as remainder an uncorrelated part that is considered noise. There are no fixed rules. One can proceed in the time domain, by examining the interrelationships among the variables X_k for different instants $k_1, \ldots, k_m$; here, if merely $c_0, \ldots, c_J$ are known, such comparisons can be made over a span no wider than J. Alternatively, time-invariance being akin to periodicity, one can apply a suitable Fourier transformation to decompose the process as a whole into a sum of random but uncorrelated frequency components [**2**, p. 268]. Again, when only some of the correlations are specified this representation is not unique, providing, rather, only a range of possibilities consistent with the available information.

The family $\mathscr{M}_J$ plays an important role in each of these descriptions. A natural gauge for the closeness of X_k to a linear combination of some of the other X_m is the expected value of the squared deviation. Using (13a) and (8), this is given explicitly by the measures of $\mathscr{M}_J$ as

$$\mathscr{E}\left(\left|X_k - \sum_{|m-k|\leqslant J} a_m X_m\right|^2\right) = \frac{1}{2\pi}\int_0^{2\pi}\left|z^k - \sum_{|m-k|\leqslant J} a_m z^m\right|^2 d\mu(\theta), \tag{13b}$$

hence analysis over time corresponds to approximation in $L^2(d\mu)$. More directly still, when the process is expressed as a sum of uncorrelated frequency components, the mass of $d\mu(\theta)$ in an interval also measures the expected energy contributed by frequencies in that interval. Thus, as their name suggests, the spectra of $\mathscr{M}_J$ can be viewed as a family of snapshots of permitted frequency profiles for the process, often a vital aid to qualitative understanding. For these reasons, spectral estimation seeks information about $\mathscr{M}_J$.

Clearly, $\mathscr{M}_J$ is a convex set of measures, and it is not surprising that it contains a subfamily, whose members consist of $(J+1)$ point masses, which generates extreme points of $\mathscr{M}_J$. These have been called maximum-likelihood estimates [**33**], although that is something of a misnomer; we will describe a

construction below. More interesting, however, is that another measure in $\mathscr{M}_J$ is given by

$$d\nu_J \equiv \frac{E_J^0(0)\,d\theta}{\left|E_J^0(e^{i\theta})\right|^2} = \frac{t_J\,d\theta}{\left|P_J(e^{i\theta})\right|^2}, \tag{14}$$

the equality here stemming from (12).

PROPOSITION 4.
(a) $d\nu_J \in \mathscr{M}_J$;
(b) *Let* $\{s_k\}$ *be the correlation coefficients corresponding to* $d\mu_J$, *viz*,

$$s_k \equiv \frac{1}{2\pi}\int_0^{2\pi} e^{ik\theta}\,d\nu_J(\theta), \qquad k \geqslant 0.$$

Then for each $n \geqslant J$, *in the scalar product defined by* $\{s_k\}$ *in* (3),

$$E_n^0(z) = E_J^0(z).$$

PROOF. To show (a), we must verify that $s_k = c_k$, $0 \leqslant k \leqslant J$. But this follows from (b), for if $s_0, \ldots, s_J$ and $c_0, \ldots, c_J$ give rise to the same $E_J^0(z)$, they coincide by Proposition 3. To prove (b), we use the equivalent representation of the scalar product in question,

$$\left[S_n, E_J^0\right] = \frac{1}{2\pi}\int_0^{2\pi} S_n(e^{i\theta})\,\overline{E_J^0(e^{i\theta})}\,d\nu_n(\theta),$$

whence, by definition of $d\nu_J$,

$$\begin{aligned}\left[S_n, E_J^0\right] &= \frac{1}{2\pi}\int_0^{2\pi} \frac{S_n(e^{i\theta})}{E_J^0(e^{i\theta})} E_J^0(0)\,d\theta \\ &= \frac{1}{2\pi i}\int_{|z|=1} \frac{S_n(z)E_J^0(0)}{E_J^0(z)}\,\frac{dz}{z}.\end{aligned}$$

Now by (11), the zeros of $E_J^0(z)$ are given by $1/\bar{\zeta}_i$, with ζ_i the zeros of $P_J(z)$, therefore all lie in $|z| > 1$ by Proposition 1. Thus the only singularity of the integrand is the pole at $z = 0$, hence by Cauchy's formula $[S_n, E_J^0] = S_n(0)$ for each n, so that $E_n^0(z)$ coincides with $E_J^0(z)$, as was to be shown.

We observe that the construction of P_n from E_n^0 now shows that $P_n(z) = z^{n-J}P_J(z)$. Moreover, any polynomial S_k having all its zeros in $|z| < 1$ will be the kth orthogonal polynomial in some scalar product of the form (3), definable explicitly by the measure $d\theta/|S_k(e^{i\theta})|^2$. These facts were noted in [**17**, p. 43]. Finally, we remark that the appeal to the Cauchy formula in the preceding argument is the only point at which analysis enters our discussion. Even here it is inessential, since the rational integrand above can be decomposed into partial fractions and integrated explicitly.

The particularly simple form of the spectrum $d\nu_J$ suggests that it ought to have some other distinguishing property among the spectra of $\mathscr{M}_J$. We can describe its special role in terms of prediction.

Linear prediction. The problem of linear prediction from the past for a stationary time series asks how well the value of the present random variable X_0 can, on the average, be approximated by some linear combination of the previously observed X_k, $k < 0$. In view of (13b), the equivalent question is to approximate 1 by a linear combination of the exponentials $\{e^{ik\theta}\}$, $k < 0$, in the metric of $L^2(d\mu)$, with $d\mu$ the spectrum of the entire process; that is, we seek

$$I_\infty(d\mu) = \inf_{a_k} \frac{1}{2\pi} \int_0^{2\pi} \left|1 - \sum_{k\geqslant 1} a_k e^{-ik\theta}\right|^2 d\mu(\theta)$$

$$= \inf_{a_k} \frac{1}{2\pi} \int_0^{2\pi} \left|1 - \sum_{k\geqslant 1} a_k e^{ik\theta}\right|^2 d\mu(\theta).$$

We interpret $I_\infty(d\mu)$ as measuring how accurately the process can be predicted at $t = 0$ from knowledge of its entire past.

If we restrict the approximation to only the first J exponentials, defining

$$I_J(d\mu) \equiv \inf_{a_k} \frac{1}{2\pi} \int_0^{2\pi} \left|1 - \sum_{k=1}^{J} a_k e^{ik\theta}\right|^2 d\mu(\theta), \tag{15}$$

then $\{I_J\}$ is a decreasing sequence and $I_\infty = \lim I_J$. Suppose $d\mu \in \mathscr{M}_J$. On expanding the integrand of (15) we see that I_J depends only on $c_0, \ldots, c_J$, and will therefore be unchanged should $d\mu$ be replaced by any other spectrum of $\mathscr{M}_J$. It is easy to determine I_J for these spectra. For, letting $S_J(z) = 1 - \sum_{k=1}^{J} a_k z^k$, we have $S_J \in \Pi_J$, with $S_J(0) = 1$; conversely, any $S_J(z)$ with $S_J(0) = 1$ can be written in this form. Thus (15) becomes

$$I_J = \inf_{S_J(0)=1} \|S_J\|^2, \tag{16}$$

hence $I_J = 1/E_J^0(0)$ by (10c) and (10a). Finally, on applying Cramer's rule to (5), and using (12a), we find

$$|\mathbf{C}_{n-1}|/|\mathbf{C}_n| = t_n = \|P_n\|^2, \tag{17}$$

with $|\mathbf{C}_k|$ the determinant of $\mathbf{C}_k$, so that by (12a),

$$I_J = |\mathbf{C}_J|/|\mathbf{C}_{J-1}|. \tag{18}$$

Suppose now that the correlations $c_0, \ldots, c_J$ are given. Each spectrum $d\mu$ of $\mathscr{M}_J$ defines a corresponding $I_\infty(d\mu)$. By monotonicity of the $I_k(d\mu)$,

$$0 \leqslant I_\infty(d\mu) \leqslant 1/E_J^0(0) = I_J.$$

The extreme measures of $\mathscr{M}_J$ given by $(J + 1)$ mass points yield $I_{J+1}(d\mu) = 0$, corresponding to perfect predictability. On the other hand, with the measure $d\nu_J$ of (14), for $n > J$ we have by Proposition 4

$$I_n(d\nu_J) = 1/E_n^0(0) = 1/E_J^0(0),$$

hence also $I_\infty(d\nu_J) = I_J$. Thus $d\nu_J$ enjoys the opposite property: it produces the poorest prediction from the past, consistent with the given correlations. Stated equivalently, all the extensions of the given $c_0, \ldots, c_J$ to a complete

sequence of correlation coefficients are produced by the measures of $\mathcal{M}_J$, and it is $d\nu_J$ which yields that extension for which the process is least predictable from its past. This happens because, by Proposition 4b, the spectrum $d\nu_J$ produces a *Markovian* process: that is, one in which the entire past influences the present through only the preceding J lags.

Maximum entropy. So far in the discussion entropy has played no role, nor have logarithms even made an appearance. They enter from the information-theoretic formulation which, for a stochastic process $F = \{X_k\}_{k=-\infty}^{\infty}$ with marginal probability densities $p(x_{-n}, \ldots, x_n)$, defines

$$h_{2n+1}(F) = -\int \cdots \int p(x_{-n}, \ldots, x_n) \log p(x_{-n}, \ldots, x_n)\, dx_{-n} \cdots dx_n,$$

and seeks to maximize the *entropy rate*

$$H(F) = \limsup_{n \to \infty} h_n(F)/n,$$

subject to prescribed values of the $J + 1$ consecutive autocorrelation coefficients

$$c_k \equiv \mathscr{E}\left(X_j \bar{X}_{j+k}\right) \equiv \int \cdots \int x_j \bar{x}_{j+k} p(x_{-n}, \ldots, x_n)\, dx_{-n} \cdots dx_n,$$

$$0 \leqslant k \leqslant J, 0 \leqslant j.$$

Here, as the penetrating discussion by B. S. Choi and T. M. Cover [**11**] makes clear, it is assumed that $\mathscr{E}(X_j \bar{X}_{j+k})$ is independent of j for only the first $J + 1$ lags, so that the process need not be wide-sense stationary.

A zero-mean Gaussian process $G = \{Y_i\}_{i=-\infty}^{\infty}$ is one for which all the joint densities are given by

$$p_G(y_m, y_{m+1}, \ldots, y_{m+n}) = |M_n|^{-1/2} (2\pi)^{-(n+1)/2} e^{-(1/2) y \mathbf{M}_n^{-1} y}, \qquad (19)$$

with $y = (y_m, \ldots, y_{m+n})$, $\mathbf{M}_n$ the $(n + 1) \times (n + 1)$ matrix of corresponding correlation coefficients $\mathscr{E}(y_{m+i} \bar{y}_{m+j})$, and $|\mathbf{M}_n|$ its determinant. G is stationary if and only if all the matrices $\mathbf{M}_k$ are Toeplitz. Gaussian processes play a major role in stochastic analysis, by providing an accurate and often analytically tractable model for a host of physical phenomena. Here, too, they figure centrally, for, as is easy to see, they maximize the entropy integrals h_k when the correlation coefficients are fixed. Indeed, suppose that G is the Gaussian process having the same autocorrelation coefficients as F. Since $\log p_G$ is quadratic, $\int p_F \log p_G\, dx$ depends only on the correlations, so that $\int p_G \log p_G\, dx = \int p_F \log p_G\, dx$, and the basic property

$$\int p_F \log \frac{p_G}{p_F}\, dx \leqslant 0,$$

valid for distributions by convexity of the logarithm, now implies that, for each k, $h_k(F) \leqslant h_k(G)$.

It is interesting that at this point there are two distinct paths to follow. On the one hand, the elegant argument of [**11**] uses elementary information-theoretic inequalities, also based only on convexity of the logarithm, to conclude that $h_{n+1}(G) \leqslant h_{n+1}(G_\nu)$, with G_ν the stationary Gauss-Markov

process of minimum order having the prescribed correlations—that is, the process whose spectrum is dv_J of (14). On the other hand, by a direct computation from (19),

$$h_{n+1}(G) = \frac{n+1}{2}\log\left(2\pi e|C_n|^{1/(n+1)}\right), \tag{20}$$

with $\mathbf{C}_n$ the matrix of correlation coefficients of F, and on writing

$$|\mathbf{C}_n| = \frac{|\mathbf{C}_n|}{|\mathbf{C}_{n-1}|}\frac{|\mathbf{C}_{n-1}|}{|\mathbf{C}_{n-2}|}\cdots\frac{|\mathbf{C}_1|}{|\mathbf{C}_0|}|\mathbf{C}_0|, \tag{21}$$

we obtain an immediate connection with the approximation problem addressed by prediction theory. For suppose, to begin with, that F is wide-sense stationary, so that all the matrices $\mathbf{C}_k$ are Toeplitz. Then from (18) and the monotonicity of I_k,

$$|\mathbf{C}_k|/|\mathbf{C}_{k-1}| \leqslant I_J \quad \text{for } k \geqslant J, \tag{22}$$

so that $|\mathbf{C}_n| \leqslant I_J^{n-J}|\mathbf{C}_J|$ for $n \geqslant J$, whence by (20)

$$h_{n+1}(G) \leqslant h_{n+1}(G_v) \tag{23}$$

and therefore $H(F) \leqslant H(G) \leqslant H(G_v)$. Moreover, by (21), the same monotonicity implies also that $\lim_{n\to\infty}|\mathbf{C}_n|^{1/(n+1)}$ exists, hence so does $\lim_{n\to\infty} h_n(G)/n$. A minor modification suffices when only $\mathbf{C}_J$ is Toeplitz. Here (3) still defines a scalar product for polynomials, but now (4) holds only when $S = z^k S_J(z)$ and $T = z^k T_J(z)$, with S_J and $T_J \in \Pi_J$ and $k \geqslant 0$, for then (3) involves solely the Toeplitz submatrix $\mathbf{C}_J$. From (17), as before, $|\mathbf{C}_k|/|\mathbf{C}_{k-1}| = \|P_k\|^2$, and $[S_k, P_k] = s_k$, the leading coefficient of S_k, so that, on applying Schwarz's inequality as in the proof of (10c),

$$|\mathbf{C}_k|/|\mathbf{C}_{k-1}| = \min_{S_k \text{ monic}} \|S_k\|^2. \tag{24}$$

As, in general, $\|P_k\|^2 \neq \|E_k^0\|^2$ for $k > J$, the successive minima in (24) are no longer necessarily monotonic as k increases. Nevertheless, by considering the trial polynomial $z^{k-J}P_J(z)/t_J$ in (24), we see by (4) that, for $k \geqslant J$, $|\mathbf{C}_k|/|\mathbf{C}_{k-1}| \leqslant I_J$, whence (23) follows as earlier, although here $\lim_{n\to\infty} h_n(G)/n$ need not exist.

Thus an analytic extremal problem involving the combinatorial notion of entropy, and a geometric one concerned with best approximation in a Euclidean space, lead to the same solution. Although the corresponding proofs—the first, based on convexity of the logarithm; the second, on orthogonal projection—appear entirely different, a similar decomposition seems to underlie them both.

Finally, we note that the connection between prediction and entropy is usually established by considerably deeper limiting results: the theorem of Szegö-Kolmogorov-Krein [**19**, p. 49] which asserts that

$$\log I_\infty(d\mu) = \frac{1}{2\pi}\int_0^{2\pi} \log S(\theta)\, d\theta, \tag{25}$$

with $S(\theta)\,d\theta$ the absolutely continuous part of $d\mu$; and Szegö's eigenvalue distribution theorem [17, p. 65], which shows that for stationary processes the right-hand side of (25) also represents $\lim_{n\to\infty}(n+1)^{-1}\log|\mathbf{C}_n|$. In our discussion we link these quantities already in finite dimensions. Moreover, in our case, where $I_J = 1/E_J^0(0)$ and $d\nu = E_J^0(0)|E_J(e^{i\theta})|^{-2}\,d\theta$, the equality (25) follows trivially on applying the mean-value theorem to the harmonic function $\log(E_J^0(0)|E_J^0(z)|^{-2})$.

A recursion for $\{P_n\}$; Schur parameters. The following supplementary information stems from only the basic property (4) of the scalar product, and is independent of Proposition 4.

$P_k(z)/t_k$ has leading coefficient 1 and is orthogonal to Π_{k-1}; thus $\|P_k/t_k\| = t_k^{-1/2}$ measures the distance of z^k from Π_{k-1}, and is positive if and only if $\mathbf{C}_k$ is positive definite. In light of the correspondence (13a), the square of this distance is also the residue left by the best linear forward predictor of length k, whose coefficients therefore coincide with those of $z^k - P_k(z)/t_k$; analogously, those of $1 - E_k^0(z)/t_k$ give the best backward predictor.

By (4), $zP_k(z)/t_k \in \Pi_{k+1}$ is already orthogonal to all polynomials of the form zS_{k-1}, so it needs only a one-dimensional adjustment to be orthogonal to all of Π_k, thereby producing P_{k+1}. Specifically, $Q(z) \equiv P_{k+1}(z)/t_{k+1} - zP_k(z)/t_k \in \Pi_k$, since the leading terms cancel, and $[zS_{k-1}, Q] = 0$. Thus Q and E_k^0 are orthogonal to the same $(k-1)$-dimensional subspace of Π_k, hence are proportional, so that

$$\frac{P_{k+1}(z)}{t_{k+1}} = \frac{zP_k(z)}{t_k} + \gamma_k \frac{E_k^0(z)}{t_k}. \tag{26}$$

This is the *Levinson algorithm* for generating $\{P_k\}$ efficiently, which leads to a corresponding computational improvement in the inversion of Toeplitz matrices. The mirror symmetry of P_k and E_k^0 here suggests that the subspaces we consider should grow, not at one end only, like our polynomials, but simultaneously at both ends—a hint we will follow later.

Evaluating (26) at $z = 0$ and recalling (12a),

$$\gamma_k = \frac{P_{k+1}(0)}{t_{k+1}} = \frac{P_{k+1}(0)}{\|E_{k+1}^0\|^2} = \left[\frac{P_{k+1}}{\|P_{k+1}\|}, \frac{E_{k+1}^0}{\|E_{k+1}^0\|}\right],$$

so that γ_k is interpretable as the correlation between the normalized forward and backward prediction errors of length $k+1$. Alternatively, forming the scalar product of (26) with E_k^0, and recalling that, by (4), $\|zP_k\|^2 = \|P_k\|^2 = t_k$, gives

$$\gamma_k = -\left[\frac{zP_k(z)}{t_k}, E_k^0\right] = -\left[\frac{zP_k(z)}{\|zP_k\|}, \frac{E_k^0}{\|E_k^0\|}\right], \tag{27}$$

so that $-\gamma_k$ can likewise be thought of as the correlation between the normalized backward prediction error of length k and the forward prediction error of length k, advanced by one step. This formula follows immediately also from the fact that, by (26) and the definition of P_{k+1}, $\|zP_k + \gamma E_k^0\|^2$, quadratic in γ,

must be minimized at $\gamma = \gamma_k$. On applying the transformation (11) between P_k and E_k^0, formula (26) can be written equivalently as

$$\frac{E_{k+1}^0(z)}{t_{k+1}} = \frac{E_k^0(z)}{t_k} + \bar{\gamma}_k \frac{zP_k(z)}{t_k}. \tag{28}$$

Thus (26) and its alternate (28) represent a way of generating the successive forward and backward predictors by means of the constants γ_k. These equations can also be viewed as modeling wave propagation in a stratified medium, in which the wave is partially transmitted and partially reflected at each layer boundary. In this interpretation, they have been applied to geophysical exploration [**35**] and to inverse scattering [**3, 4**]. An acoustical version has been used to model the vocal tract [**16, 31**], while an electrical analogue, given by a lattice filter network, allows an efficient implementation [**22, 23**].

To investigate the $\{\gamma_k\}$, let us rewrite (26) in the form

$$\frac{P_{k+1}(z)}{t_{k+1}} - \gamma_k \frac{E_k^0(z)}{t_k} = \frac{zP_k(z)}{t_k},$$

in which the components on the left are orthogonal, by definition of P_{k+1}. On taking norms, and recalling (4) and (12a), we find

$$\frac{1}{t_{k+1}^2} + \frac{|\gamma_k|^2}{t_k^2} = \frac{1}{t_k^2},$$

whence

$$t_{k+1}^2 = \frac{t_k^2}{1 - |\gamma_k|^2}. \tag{29}$$

Suppose now that $\mathbf{C}_k$ is positive definite. If $\mathbf{C}_{k+1}$ is likewise positive definite, so that $P_{k+1}(z)$ is well defined, then $|\gamma_k| < 1$ by (29). Conversely, for each γ_k with $|\gamma_k| < 1$, we can define t_{k+1} by (29) and $P_{k+1}(z)$ by (26). Thereupon P_{k+1} is automatically orthogonal to all polynomials of the form $zS_{k-1}(z)$; to be orthogonal to all of Π_k it need only be orthogonal to 1, a condition which determines c_{k+1}. With c_{k+1} at this value, the matrix $\mathbf{C}_{k+1}$ will be positive definite and will generate $P_{k+1}(z)$ as the $(k+1)$-st orthogonal polynomial. Thus we see that the choice of $\{\gamma_0, \ldots, \gamma_k\}$ with $|\gamma_i| < 1$ corresponds in a one-to-one way to positive definite Toeplitz matrices $\{\mathbf{C}_{k+1}\}$.

This fact, combined with (27), or the equivalent

$$\gamma_k = \frac{-2[zP_k, E_k^0]}{\|zP_k\|^2 + \|E_k^0\|^2}, \tag{30}$$

forms the basis of *Burg's procedure* for generating the orthogonal polynomials directly from sample data when the autocorrelation coefficients are not available and cannot be reliably computed. Here if $\gamma_0, \ldots, \gamma_{k-1}$ are presumed known, P_k and E_k^0 are determined from (26) and (28), and as the scalar products defining γ_k by (30) represent expected values, they are estimated directly by averaging over all the consecutive segments of length $k+1$ which the data contains [**7, 9**].

The numbers $\{\gamma_k\}$ are sometimes referred to as *partial correlation coefficients* or, in view of their physical interpretation, as *reflection coefficients*. They are also the *Schur parameters*, whose fundamental importance has been exposed in the far-ranging papers [**3, 21, 22, 32**]. To see this, we form the quotient $B_k(z) \equiv P_k(z)/E_k^0(z)$. We have $B_0(z) = 1$ and, by (26) and (28),

$$B_{k+1}(z) = \frac{zB_k(z) + \gamma_k}{1 + \bar{\gamma}_k z B_k(z)},$$

or

$$B_k(z) = \frac{B_{k+1}(z) - B_{k+1}(0)}{z\left(1 - B_{k+1}(z)\,\overline{B_{k+1}(0)}\right)},$$

the Schur algorithm [**3, 5**]. This relationship is one link between the scalar product structure and complex-variable methods. As (26) and (28) are equivalent, it is curious that the simple expedient of considering them jointly, in the form of $B_k(z)$, rather than individually, should form this highly productive connection.

Extremal spectra in M_n; maximum likelihood estimation. Suppose that $\{E_n^{\alpha_i}(z)\}$, $i = 1, \ldots, n+1$, with $|\alpha_i| = 1$, is a set of $n+1$ evaluations which are mutually orthogonal as elements in Π_n. Using $\{E_n^{\alpha_i}(z)/\|E_n^{\alpha_i}\|\}$ as an orthonormal basis, we can write for any S_n, $T_n \in \Pi_n$,

$$S_n(z) = \sum_i a_i E_n^{\alpha_i}(z)/\|E_n^{\alpha_i}\|, \tag{31}$$

with

$$a_i = \left[S_n, E_n^{\alpha_i}/\|E_n^{\alpha_i}\|\right] = S_n(\alpha_i)/\|E_n^{\alpha_i}\|,$$

by definition of $E_n^{\alpha_i}$. Expanding T_n similarly, we find from (31)

$$[S_n, T_n] = \sum_i S_n(\alpha_i)\,\overline{T_n(\alpha_i)}\,/\|E_n^{\alpha_i}\|^2, \tag{32}$$

which is a representation having the desired form (8), in which the measure $d\mu(\theta)$ consists of masses $\|E_n^{\alpha_i}\|^{-2}$ at $z = \alpha_i$, $1 \leqslant i \leqslant n+1$.

The evaluations E_n^{α} and E_n^{β} are orthogonal provided $0 = [E_n^{\alpha}, E_n^{\beta}] = E_n^{\alpha}(\beta)$. Thus the mutual orthogonality from which (32) springs requires that, for each of the α_i, $E_n^{\alpha_i}$ vanish at the remaining α_j, $j \neq i$. We therefore consider the zeros of E_n^{α}. To this end, we derive a compact expression for $E_n^{\zeta}(z)$ which is of interest in its own right.

By definition of E_n^{ζ} and by (4) we have, for each $S_{n-1} \in \Pi_{n-1}$,

$$\begin{aligned} 0 &= \left[(z - \zeta)S_{n-1}, E_n^{\zeta}\right] = \left[zS_{n-1}, E_n^{\zeta}\right] - \left[S_{n-1}, \bar{\zeta}E_n^{\zeta}\right] \\ &= \left[zS_{n-1}, (1 - \bar{\zeta}z)E_n^{\zeta}\right]. \end{aligned}$$

Consequently, $(1 - \bar{\zeta}z)E_n^{\zeta} \in \Pi_{n+1}$ lies in the orthogonal complement of the $(n-1)$-dimensional subspace of Π_{n+1} generated by $\{zS_{n-1}\}$. Again by definition and (4), this two-dimensional complementary subspace contains $E_n^0(z)$ and $zP_n(z)$—elements which, being polynomials of different degrees,

are linearly independent and therefore span. Thus

$$(1 - \bar{\zeta}z)E_n^{\zeta}(z) = aE_n^0(z) + bzP_n(z). \tag{33}$$

By (10b), the leading coefficient of $-\bar{\zeta}zE_n^{\zeta}(z)$ coincides with that of $-\bar{\zeta}\overline{P_n(\zeta)}zP_n(z)/t_n$, so that $b = -\bar{\zeta}\overline{P_n(\zeta)}/t_n$, and evaluation of (24) at $z = 1/\bar{\zeta}$ then determines a. Applying (11) to simplify the resulting expression, we obtain

$$E_n^{\zeta}(z) = \frac{E_n^0(z)\,\overline{E_n^0(\zeta)} - zP_n(z)\,\overline{\zeta P_n(\zeta)}}{t_n(1 - \bar{\zeta}z)}, \tag{34}$$

the *Christoffel-Darboux formula*. By the same argument, if a linear combination of $E_n^0(z)$ and $zP_n(z)$ vanishes at $z = 1/\bar{\alpha}$, then on writing

$$(1 - \bar{\alpha}z)T_n(z) = aE_n^0(z) + bzP_n(z)$$

we see that $T_n \in \Pi_n$ is orthogonal to $\{(z - \alpha)S_{n-1}\}$, so that

$$T_n = cE_n^{\alpha}(z). \tag{35}$$

We now specialize to $|\zeta| = 1$. By (10b), the leading coefficient of E_n^{ζ} is $\overline{P_n(\zeta)}/t_n$, which does not vanish by Proposition 1, so that $E_n^{\zeta}(z)$ has n zeros; (34) and (12b) show that these are given by the solutions of

$$\frac{zP_n(z)}{E_n^0(z)} = \frac{\overline{E_n^0(\zeta)}}{\bar{\zeta}\overline{P_n(\zeta)}} = \frac{\zeta P_n(\zeta)}{E_n^0(\zeta)} \equiv \gamma, \tag{36}$$

distinct from $z = \zeta$. By (12b), $|\gamma| = 1$. Now if α is a solution of (36), then so is $1/\bar{\alpha}$, for by (11b) the value of the left-hand side of (36) at $z = 1/\bar{\alpha}$ is $1/\bar{\gamma} = \gamma$. Consequently, applying (35) to the linear combination $\gamma E_n^0(z) - zP_n(z)$ which vanishes at $z = 1/\bar{\alpha}$, we find

$$c(1 - \bar{\alpha}z)E_n^{\alpha}(z) = \gamma E_n^0(z) - zP_n(z),$$

and as the right-hand side also vanishes at $z = \alpha$, while $E_n^{\alpha}(\alpha) > 0$, we see that $|\alpha| = 1$; the same argument shows also that all the zeros of $\gamma E_n^0(z) - zP_n(z)$ are distinct. We conclude that the map $zP_n(z)/E_n^0(z)$ winds $|z| = 1$ onto the unit circumference n times—a fact that also follows easily from the maximum principle for analytic functions. Evidently, the set of solutions of (36) is unchanged if ζ is replaced by any one of them. In sum, $E_n^{\zeta}(z)$ has n distinct zeros $\{\alpha_i\}$, $1 \leqslant i \leqslant n$, of unit modulus, and adding to E_n^{ζ} the evaluations $\{E_n^{\alpha_i}\}$ produces a mutually orthogonal set and justifies (32).

To see how the α_i change as a function of ζ, we begin with $\zeta = 1$ and denote the zeros of E_n^1 in counterclockwise order by $\{\beta_i\}$. Then as ζ moves monotonically on the unit circumference to β_1, the motion of the $\alpha_i = \alpha_i(\zeta)$ produces a monotone displacement of each β_i to β_{i+1}, and β_n to 1. In turn, this generates a one-parameter family of measures $d\mu_{\zeta}$, $0 \leqslant \arg\zeta < \arg\beta_1$, each consisting of $n + 1$ mass points; every point α on $|z| = 1$ appears once in this family, carrying the mass $\|E_n^{\alpha}\|^{-2}$. By (32), these measures all belong to $\mathscr{M}_n$. They, too, have an extremal characterization.

PROPOSITION 5. *$\|E_n^\alpha\|^{-2}$ is the largest mass which a measure of $\mathcal{M}_n$ can carry at the point $z = \alpha$, $|\alpha| = 1$, and any measure which does so coincides with $d\mu_\zeta$, with α a zero of E_n^ζ.*

PROOF. If $\alpha = e^{i\phi}$ is a point of mass m for a measure $d\mu(\theta) \in \mathcal{M}_n$, then from (8)

$$\|E_n^\alpha\|^2 = \frac{1}{2\pi}\int_0^{2\pi} |E_n^\alpha(e^{i\theta})|^2 \, d\mu(\theta) \geqslant m|E_n^\alpha(e^{i\phi})|^2,$$

so that $m \leqslant \|E_n^\alpha\|^2/|E_n^\alpha(\alpha)|^2 = \|E_n^\alpha\|^{-2}$. Equality occurs here only if $E_n^\alpha(e^{i\theta})$ vanishes on the support of $d\mu(\theta)$, excepting $\theta = \phi$. Thus $d\mu(\theta)$ consists of masses at α and at the zeros $\{\alpha_i\}$, $1 \leqslant i \leqslant n$, of $E_n^\alpha(z)$. Since $E_n^{\alpha_i}$ vanishes at all of these points except $z = \alpha_i$, formula (8) for $\|E_n^{\alpha_i}\|^2$ shows the mass of $d\mu(\theta)$ at $z = \alpha_i$ to be $\|E_n^{\alpha_i}\|^{-2}$, as required.

In view of Proposition 5, the measures $d\mu_\zeta$ have been called *maximum likelihood spectra* [**33**]; we note that in spectral estimation this term has a special sense, different from its usual meaning [**30**]. By (3a), the coefficient vector ε_n^ζ of E_n^ζ, which defines $d\mu_\zeta$, is given by the solution of

$$\mathbf{C}_n\varepsilon_n^\zeta = (1, \zeta, \ldots, \zeta^n).$$

From (10b),

$$\|E_n^\zeta\|^2 = \sum_{k=0}^{n} |P_k(\zeta)|^2/t_k,$$

and comparison with (13) shows immediately that—as observed by Burg [**8**]—the reciprocal of the maximum-likelihood estimate is the sum of the reciprocals of the maximum entropy estimates for all lower orders.

Matrix factorization. We now show how the polynomials $P_k(z)$ and $E_k^0(z)$ can be used to invert the matrix $\mathbf{C}_n$.

Any orthonormal basis in Π_n generates at once a factorization of $\mathbf{C}_n^{-1}$ and of $\mathbf{C}_n$. For suppose that $\{Q_k\}$, $0 \leqslant k \leqslant n$, constitutes such a basis, so that $R = \Sigma[R, Q_k]Q_k$ for any $R \in \Pi_n$, whence

$$[R, S] = \sum_{k=0}^{n} [R, Q_k]\overline{[S, Q_k]} = \sum_{k=0}^{n} [R, Q_k][Q_k, S]. \tag{37}$$

Let us henceforth extend the polynomial representation of a vector also to the scalar product $(\cdot, \cdot)$; thus if $A(z)$ and $B(z)$ are polynomials having respective coefficient vectors a and b, and $\mathbf{M}$ is a matrix, we set $(A(z), B(z)) \equiv (a, b)$, and denote by $\mathbf{M}A$ the polynomial with coefficient vector $\mathbf{M}a$. Further, we will say that a row, or column, of $\mathbf{M}$ is given by $A(z)$ if it consists of the sequence $a_0, a_1, \ldots, a_n$ of coefficients of A. With this notation (37) becomes, by (3a),

$$(\mathbf{C}_n^{-1}(\mathbf{C}_n R), \mathbf{C}_n S) = \sum_{k=0}^{n} (\mathbf{C}_n R, Q_k)(Q_k, \mathbf{C}_n S). \tag{38}$$

Since $\mathbf{C}_n$ is invertible, any nth degree polynomials X and Y can be represented as $\mathbf{C}_n R$ and $\mathbf{C}_n S$, respectively, so that (38) is equivalent to

$$(\mathbf{C}_n^{-1}X, Y) = \sum_{k=0}^{n} (X, Q_k)(Q_k, Y),$$

expressing the fact that

$$\mathbf{C}_n^{-1} = \mathbf{M}^*\mathbf{M}, \tag{39}$$

with $\mathbf{M}$ the matrix whose $(k+1)$st row consists of $\overline{Q}_k$, $0 \leqslant k \leqslant n$. On forming inverses, this yields

$$\mathbf{C}_n = \mathbf{M}^{-1}(\mathbf{M}^*)^{-1}.$$

Since, by orthonormality of the Q_k,

$$\delta_{ij} = [Q_i, Q_j] = (\mathbf{C}_n Q_i, Q_j),$$

$\mathbf{M}^{-1}$ is the matrix whose $(k+1)$st column is given by $\mathbf{C}_n Q_k$, $0 \leqslant k \leqslant n$, so that $(\mathbf{M}^*)^{-1}$ has rows $\overline{\mathbf{C}_n Q_k}$.

By applying this to $Q_k = P_k(z)/\sqrt{t_k}$, we obtain the *Cholesky decomposition*

$$\mathbf{C}_n^{-1} = \mathbf{L}_1^*\mathbf{L}_1,$$

with $\mathbf{L}_1$ lower-triangular, having $\overline{P}_k(z)/\sqrt{t_k}$ as $(k+1)$st row, and

$$\mathbf{C}_n = \mathbf{L}_1^{-1}(\mathbf{L}_1^*)^{-1} = \mathbf{U}_1^*\mathbf{U}_1,$$

with $\mathbf{U}_1$ upper-triangular, having $\overline{\mathbf{C}_n P_k}/\sqrt{t_k}$ as $(k+1)$st row, $0 \leqslant k \leqslant n$ [**24**]. Since for $j > k$

$$\left[z^j E_{n-j}^0, z^k E_{n-k}^0\right] = \left[z^{j-k} E_{n-j}^0, E_{n-k}^0\right] = 0, \tag{40}$$

the last equality holding because $z^{j-k}E_{n-j}^0$, of degree $n-k$, is evaluated at the origin by E_{n-k}^0, it follows that $\{z^k E_{n-k}/\sqrt{t_{n-k}}\}$ is likewise an orthonormal set —indeed, by (11a), the mirror image of $\{P_k/\sqrt{t_k}\}$. This in turn yields a factorization

$$\mathbf{C}_n^{-1} = \mathbf{U}_2^*\mathbf{U}_2,$$

with $\mathbf{U}_2$ upper-triangular, having $z^k\overline{E}_{n-k}/\sqrt{t_{n-k}}$ as its $(k+1)$st row, and

$$\mathbf{C}_n = \mathbf{L}_2^*\mathbf{L}_2,$$

with $\mathbf{L}_2$ lower-triangular, having $\overline{C}_n z^k \overline{E}_{n-k}/\sqrt{t_{n-k}}$, a polynomial of degree k, as its $(k+1)$st row, $0 \leqslant k \leqslant n$.

In using such decompositions, it is a considerable computational and conceptual advantage for the factors to be Toeplitz as well as triangular, because the matrix multiplications can then be interpreted as convolutions. Now if $\mathbf{U}$ is an upper-triangular Toeplitz matrix having the vector $\bar{v}$ as its first row, and $V_n(z)$ is the polynomial with coefficient vector v, then $\mathbf{U}a$ has components given by

$$\left(A_n(z), z^k V(z)\right) = \left(A_n(z), z^k W_{n-k}(z)\right),$$

$0 \leqslant k \leqslant n$, where $W_{n-k}(z) \equiv \sum_{j=0}^{n-k} v_j z^j$ consists of V_n truncated to degree $n-k$. Thereupon,

$$\mathbf{U}A_n(\zeta) = \sum_{k=0}^{n} \zeta^k \left(A_n(z), z^k W_{n-k}(z)\right) = \left(A_n(z), \sum_{k=0}^{n} \bar{\zeta}^k z^k W_{n-k}(z)\right),$$

$$= \left(A_n(z), \frac{V_n(z) - cz^{n+1}}{1 - \bar{\zeta}z}\right), \tag{41}$$

where $c = \bar{\zeta}^{n+1}V_n(1/\bar{\zeta})$ causes the numerator to vanish at $z = 1/\bar{\zeta}$, as it must for the quotient to be a polynomial. Thus the appearance of a function of this form in a scalar product signals the presence of an upper-triangular Toeplitz matrix with top row consisting of $\overline{V}_n(z)$. In particular, the Christoffel-Darboux formula suggests itself as a source of such factors. We give two examples.

From (34) we have

$$R_n(\zeta) = [R_n, E_n^{\zeta}]$$
$$= t_n^{-1}\left[R_n, \overline{E_n^0(\zeta)}\,\frac{E_n^0(z) - d_1 z^{n+1}}{1 - \bar{\zeta}z} - \bar{\zeta}\,\overline{P_n(\zeta)}\,\frac{(zP_n(z) - t_n z^{n+1}) - d_2 z^{n+1}}{1 - \bar{\zeta}z}\right],$$

where d_1 and d_2 divide the component of z^{n+1} in the numerator of (34), so that each of the above numerators separately vanishes at $z = 1/\bar{\zeta}$. As we have seen, the two quotients correspond to upper-triangular Toeplitz matrices $\mathbf{U}$ and $\mathbf{V}$ generated by $\overline{E_n^0}$ and $\overline{P_n^*}$ respectively, where $P_n^*(z) \equiv zP_n(z) - t_n z^{n+1}$ is the truncation of $zP_n(z)$ to degree n. Thus by (41)

$$R_n(\zeta) = t_n^{-1}\sum_{k=0}^{n} \zeta^k E_n^0(\zeta)[R_n, z^k E_n^0(z)] - \zeta P_n(\zeta)[R_n, z^k P_n^*(z)],$$

whence

$$[R_n, S_n] = t_n^{-1}\sum [R_n, z^k E_n^0(z)][\zeta^k E_n^0(\zeta), S_n] - [R_n, z^k P_n^*(z)][\zeta^k P_n^*(\zeta), S],$$

so that, as in (39),

$$\mathbf{C}_n^{-1} = t_n^{-1}(\mathbf{U}^*\mathbf{U} - \mathbf{V}^*\mathbf{V}).$$

This formula is due to Gohberg-Semencul [**24**; **15**, p. 86].

Finally, we turn to the set of orthonormal evaluations figuring in (32). As we have seen, the set $\{\alpha_k\}$ is parameterized by a point $\gamma = e^{i\theta}$ on the unit circumference. By (39), we have

$$\mathbf{C}_n^{-1} = (\mathbf{M}_n^{\theta})^*\mathbf{M}_n^{\theta}, \tag{42}$$

where the kth row of $\mathbf{M}_n^{\theta}$ is given by $\overline{E}_n^{\alpha_k}/\|E_n^{\alpha_k}\|$, $1 \leqslant k \leqslant n+1$. The strong kinship among these rows enables us further to factor $\mathbf{M}_n^{\theta}$. For by (36), $\{\alpha_k\}$ consists of all the solutions of $\zeta P_n(\zeta)/E_n^0(\zeta) = e^{i\theta}$, so that, from (34),

$$E_n^{\alpha_k}(z) = \overline{E_n^0(\alpha_k)}\,\frac{E_n^0(z) - e^{-i\theta}zP_n(z)}{t_n(1 - \bar{\alpha}_k z)}.$$

Consequently, for any polynomial $A_n(z)$, the components of $\mathbf{M}_n^{\theta}A_n$ consist of

$$\frac{E_n^0(\alpha_k)}{\|E_n^{\alpha_k}\|}\left(A_n(z), \frac{E_n^0(z) - e^{-i\theta}zP_n(z)}{t_n(1 - \bar{\alpha}_k z)}\right).$$

By (41), these coincide with $\|E_n^{\alpha_k}\|^{-1}E_n^0(\alpha_k)\mathbf{U}_n^{\theta}A(\zeta)$ evaluated at $\zeta = \alpha_k$, where $\mathbf{U}_n^{\theta}$ is the upper-triangular Toeplitz matrix with top row given by $t_n^{-1}\{\overline{E}_n^0 - e^{i\theta}\overline{P}_n^*\}$. We therefore have

$$\mathbf{M}_n^{\theta} = \mathbf{D}_1^{\theta}\mathbf{R}_n^{\theta}\mathbf{U}_n^{\theta},$$

with $\mathbf{R}_n^\theta$ the matrix whose kth row is $(1, \alpha_k, \ldots, \alpha_k^n)$, and $\mathbf{D}_1^\theta$ diagonal, with entries $\|E_n^{\alpha_k}\|^{-1} E_n^0(\alpha_k)$, $1 \leqslant k \leqslant n+1$. On substituting into (42), and letting $\mathbf{D}^\theta = (\mathbf{D}_1^\theta)^* \mathbf{D}_1^\theta$, we find

$$\mathbf{C}_n^{-1} = (\mathbf{U}_n^\theta)^* \mathbf{T}_n^\theta \mathbf{U}_n^\theta, \tag{43}$$

where $\mathbf{T}_n^\theta = (\mathbf{R}_n^\theta)^* \mathbf{D}^\theta \mathbf{R}^\theta$. Since each $|\alpha_k| = 1$, $\mathbf{T}_n^\theta$ is Toeplitz. Thus $\mathbf{C}_n^{-1}$, although not itself Toeplitz, can be associated with a family of Toeplitz matrices $\mathbf{T}_n^\theta$ by the simple coordinate transformation of (43).

The orthogonal polynomials of $\mathbf{T}_n^\theta$ can now be readily determined. For we have already noted that $Q_k(z) \equiv \mathbf{C}_n z^k E_{n-k}^0(z)$, $0 \leqslant k \leqslant n$, are mutually orthogonal in the scalar product defined by $\mathbf{C}_n^{-1}$, since

$$\left(\mathbf{C}_n^{-1} Q_j, Q_k\right) = \left(z^j E_{n-j}^0(z), \mathbf{C}_n z^k E_{n-k}^0(z)\right) = \left[z^j E_{n-j}^0, z_k E_{n-k}^0\right]$$

and the last quantity vanishes by (39) when $j \neq k$. On applying (43) we therefore find, for $j \neq k$,

$$\begin{aligned} 0 = \left(\mathbf{C}_n^{-1} Q_j, Q_k\right) &= \left((\mathbf{U}_n^\theta)^* \mathbf{T}_n^\theta \mathbf{U}_n^\theta Q_j, Q_k\right) \\ &= \left(\mathbf{T}_n^\theta \mathbf{U}_n^\theta Q_j, \mathbf{U}_n^\theta Q_k\right). \end{aligned}$$

Now by (11a), $Q_k(z)$ has degree k and is monic. Since $\mathbf{U}_n^\theta$ is upper-triangular with 1 on the main diagonal, the same is true for $\mathbf{U}_n^\theta Q_k$; these are therefore the monic orthogonal polynomials associated with $\mathbf{T}_n^\theta$. By (26), the corresponding reflection coefficients δ_k^θ are given as

$$\delta_k^\theta = \mathbf{U}_n^\theta Q_{k+1}(0),$$

so that, from the definition of $\mathbf{U}_n^\theta$, for $0 \leqslant k \leqslant n-1$,

$$\begin{aligned} \delta_k^\theta &= \left(Q_{k+1}, t_n^{-1}\{E_n^0 - e^{-i\theta} P_n^*\}\right) \\ &= \left(\mathbf{C}_n z^{k+1} E_{n-k-1}^0, t_n^{-1}\{E_n^0(z) - e^{-i\theta} P_n^*(z)\}\right) \\ &= t_n^{-1}\left[z^{k+1} E_{n-k-1}^0, E_n^0 - e^{-i\theta} P_n^*\right]. \end{aligned}$$

Now on writing

$$P_n^* \equiv z P_n(z) - t_n z^{n+1} = z P_n(z) - z^{k+1} t_n \left\{\left(z^{n-k} - \frac{P_{n-k}}{t_{n-k}}\right) + \frac{P_{n-k}}{t_{n-k}}\right\},$$

the decomposition arranged so that the first bracketed term has degree $n-k-1$, and repeatedly applying (4), the evaluation property of E_n^0, and the definition of P_j, we find

$$\delta_k^\theta = -e^{i\theta} \frac{\overline{P}_{n-k}(0)}{t_{n-k}} = -e^{i\theta} \overline{\gamma}_{n-k-1}.$$

Thus the reflection coefficients associated with $\mathbf{T}_n^\theta$, the Toeplitz factor of $\mathbf{C}_n^{-1}$, are those of $\mathbf{C}_n$, taken in reverse order, followed by a suitable rotation and

complex conjugation. The representation (43), as well as this curious fact, were described and discussed with extensive applications in [23, 22].

In sum, orthogonal polynomials on $|z| = 1$ can be equivalently defined from a positive measure $d\mu$ by using the Gram-Schmidt process in $L^2(d\mu)$, from the Fourier coefficients of such a measure by solving (5), or from the partial correlation coefficients by applying the recursion (26). These elements are therefore interrelated, and the preceding sections have described some of the connections among them, useful for both theory and application. All of these ideas have continuous analogues, which we sketch in the Appendix.

Conclusion. The notion of entropy is introduced as a measure of randomness, and in that way, intuitively, larger entropy corresponds to less information. This feeling was formulated with greater precision by the entropy concentration theorem [20], which shows that among all distributions satisfying a given set of linear constraints, that could be observed in a repeated experiment, the entropy is very heavily concentrated in a neighborhood of its maximum. This stems from the combinatorial fact that the entropy of a distribution measures the number of ways in which it can be realized. Thus the entropy-maximizing distribution occurs by far the most frequently among the available candidates, and is thereby viewed as the least informative choice.

However, another sense in which a process may be informative is that of its own coherence, measured by its predictability from the past. This in turn is controlled by the autocorrelation coefficients, and if the first n of these are prescribed, it is the remaining ones which determine how much information about the current outcome is available from previous observations. In these terms, the least informative process consistent with the data is that for which prediction ahead is poorest.

That it is surprising for these two descriptions to characterize the same spectrum was observed originally by Burg. Our elementary constructions show that already in finite dimension the combinatorial notion of entropy, and orthogonal decomposition, are closely related. What does this connection imply, and how is the structure of each problem reflected in the other? In particular, is some equivalent of the phenomenon of entropy concentration to be found in the geometry of prediction? This question, and similar others, seem interesting and remain open.

Acknowledgment. I am very grateful to J. J. Benedetto and D. Slepian for stimulating conversations.

Appendix: the continuous limit. In this section we explain how our present considerations apply to second-order Sturm-Liouville differential equations—an important class, governing a variety of mechanical and electrical inverse problems. That such equations are related to the moment problem was pointed out by M. G. Krein [27], but the precise connection is far from obvious. Here we trace it explicitly, stressing a geometric interpretation. Our object is not to give proofs, which are often more efficiently obtainable by other methods, but to guide intuition. We focus on the following remarkable result of [28].

THEOREM (M. G. KREIN). *Suppose that* $D(t)$, $0 \leqslant t < 2R$, *is a continuous function such that, for each* $0 \leqslant r < R$, *the equation*

$$q(t;r) + \int_{-r}^{r} D(|t-s|)q(s;r)\,ds = 1, \qquad 0 \leqslant t \leqslant r, \tag{A1}$$

has a unique continuous solution $q(t;r)$. *Then this function generates a differential equation*

$$\frac{d}{dr}\left(p(r)\frac{dy}{dr}\right) + \lambda^2 p(r)y(r) = 0, \qquad 0 \leqslant r < R, \tag{A2}$$

with $p(r) = q^2(r;r)$, *having solutions*

$$\phi(r,\lambda^2) = \left(\frac{d}{dr}\int_0^r q(s;r)\cos\lambda s\,ds\right)/p(r), \tag{A3}$$

$$\psi(r,\lambda^2) = \left(\frac{d}{dr}\int_0^r q(s;r)\omega(s,\lambda)\,ds\right)/p(r), \tag{A4}$$

where

$$\lambda\omega(t,\lambda) = \sin\lambda t + 2\int_0^t D(t-s)\sin\lambda s\,ds; \qquad 0 \leqslant t < R,$$

which satisfy the initial conditions $\phi(0,\lambda^2) = 1$, $\phi'(0,\lambda^2) = 0$; $\psi(0,\lambda^2) = 0$, $\psi'(0,\lambda^2) = 1$.

We will show that this theorem can be interpreted as a continuous version of some of the preceding relationships. Since the kernel $D(|t-s|)$ is akin to a Toeplitz matrix, we might expect it to generate a first-order equation analogous to (26). We aim to explain, firstly, how the second-order equation (A2) enters the problem, and, secondly, what the solutions (A3) and (A4) represent in terms of our earlier constructions. We note from the outset that the hypotheses of the theorem require neither a positive-definite, nor even a Hermitian kernel, and are therefore weaker than ours; we propose our point of view less to recover all of these results than to understand them better.

If in the trigonometric polynomial $A_n(z) = \sum_{k=0}^{n} a_k z^k$ we let $z = e^{i\lambda/M}$ and correspondingly rescale the polynomial by the factor $1/M$, we obtain

$$\frac{A_n(z)}{M} = \frac{1}{M}\sum_{k=0}^{n} a_k e^{i\lambda k/M},$$

a Riemann sum approximation to an integral of the form

$$\int_0^r a(t)e^{i\lambda t}\,dt, \tag{A5}$$

in which $r = n/M$, and $a(t)$ has the value a_k at $t = k/M$. Thus we can think of Fourier transforms of the form (A5) as continuous analogues (obtained as $M \to \infty$) of trigonometric polynomials, in which r corresponds to the degree, and $a(t)$ to the coefficient vector. The Toeplitz matrix $\mathbf{C}_n$, which defined the scalar product for polynomials, similarly passes into a positive-definite difference kernel $C(s-t)$, defined on a finite interval. To parallel the earlier discussion, we should next construct continuous versions of orthogonal polynomials, but for this purpose it is awkward to use a limiting form of (5), since

its right-hand side represents a function vanishing except for $(n-1)/M \leqslant s \leqslant n/M$ where the value is 1, and this does not have a well-defined limit as $M \to \infty$ (nevertheless, see [29] for a construction along these lines). We therefore take a readily available alternative approach. For from (10b),

$$\overline{P_n(1)}\, P_n(z)/t_n = E_n^1(z) - E_{n-1}^1(z), \tag{A6}$$

and, since the value of a polynomial at $z = 1$ is the sum of its coefficients, (3a) shows that the coefficient vector η_n of $E_n^1(z)$ is given by the solution of

$$\mathbf{C}_n \eta_n = (1, 1, \ldots, 1),$$

which evidently has as limit

$$\int_0^r C(s-t)e(s;r)\,ds = 1, \qquad 0 \leqslant t \leqslant r. \tag{A7}$$

As this equation does not generally have a solution for smooth C, we incorporate into the kernel a component of $\delta(0)$, representing (A7) by

$$e(t;r) + \int_0^r D(s-t)e(s;r)\,ds = 1, \qquad 0 \leqslant t \leqslant r. \tag{A8}$$

Another effect of this modification is that in the discrete approximation, which now has a component of the identity, $t_{rM} \to 1$ in the continuous limit. In sum, we conclude that when the covariance function

$$\delta(s-t) + D(s-t), \qquad 0 \leqslant s, t \leqslant r \tag{A9}$$

replaces the covariance matrix $\mathbf{C}_n$, functions (A5) replace polynomials $A_k(z)$, with r corresponding to k; the scalar product, paralleling (3a), is given by

$$[A(\lambda), B(\lambda)] \equiv \int_0^r a(t)\left[\overline{b(t)} + \overline{\int_0^r D(s-t)b(s)\,ds}\right] dt,$$

and the solution of (A8) defines the evaluation

$$E_r(\lambda) \equiv \int_0^r e(t;r)e^{it\lambda}\,dt,$$

analogous to $E_k^1(z)$; here we note that $\lambda = 0$ represents $z = 1$. Continuing, $e(r;r)$ corresponds to the leading coefficient of $E_k^1(z)$ which, according to (3a) and (5), is $[E_k^1, P_k] = \overline{P_k(1)}$, and, from (A6), the equivalent of $\overline{P_k(1)}P_k(z)/t_k$ is $dE_r(\lambda)/dr$, so that

$$\pi_r(\lambda) \equiv e(r;r)^{-1}\frac{dE_r(\lambda)}{dr}$$

replaces $P_k(z)/t_k$. For these relationships, as for our subsequent remarks, a good illustration is a constant kernel D, for which all the quantities, discrete and continuous, are easy to find explicitly.

Now we turn to (26), which we rewrite as

$$\frac{P_{k+1}(z)}{t_{k+1}} - \frac{P_k(z)}{t_k} = (z-1)\frac{P_k(z)}{t_k} + \gamma_k \frac{E_k^0(z)}{E_k^0(0)}.$$

Recalling the relationship (11b) between E_k^0 and P_k, we see that, since $z - 1 = e^{i\lambda/M} - 1 \to i\lambda/M$, in the continuous limit we obtain

$$\frac{d\pi_r(\lambda)}{dr} = i\lambda\pi_r(\lambda) + \gamma(r)e^{ir\lambda}\overline{\pi_r(\lambda)}\,. \tag{A10}$$

This equation is simplified by the substitution

$$\pi_r(\lambda) = e^{ir\lambda/2}F_r(\lambda), \tag{A11}$$

which replaces $\pi_r(\lambda)$ by the Fourier transform of a function supported on the symmetric interval $|t| \leqslant r/2$. Indeed, such a modification was already prefigured earlier in our constructions, when the symmetry between $P_k(z)$ and $E_k^0(z)$ suggested that the matrices $\{\mathbf{C}^n\}$ be used to define the scalar product for two-sided trigonometric expressions of the form $\sum_{k=-[n/2]}^{[n/2]} a_k z^k$, rather than for trigonometric polynomials. Before doing this, however, we note that (A11) converts (A10) into

$$\frac{dF_r(\lambda)}{dr} - \frac{i\lambda}{2}F_r(\lambda) = \gamma(r)\overline{F_r(\lambda)}\,,$$

and, although this equation is of only first order, its real and imaginary components combine to produce two different second-order equations, for $\mathrm{Re}\{F_r(\lambda)\}$ and $\mathrm{Im}\{F_r(\lambda)\}$, respectively [**29**].

We now return to the discrete problem. recasting it in terms of two-sided expressions. We will follow the discussion of [**31**] with only minor modifications. For simplicity, suppose that $\mathbf{C}_m$ is real. If m is even, (2) will also define a scalar product in the space of functions of the form $\sum_{k=-n}^{n} a_k z^k$, with $m = 2n$. Since $\mathbf{C}_m$ is symmetric, functions generated by even and odd coefficient sequences are orthogonal with respect to the scalar product, and we concentrate on the former, denoting by $\mathscr{T}_n$ the $(n+1)$-dimensional space of functions of the form $\sum_{-n}^{n} a_k z^k$, $a_{-k} = a_k$; as before, we will call $(a_n, \ldots, a_1, a_0, a_1, \ldots, a_n)$ the *coefficient vector* of an element of $\mathscr{T}_n$, and denote it by α_n, despite its having $2n+1$ components. For odd m, the scalar product is defined on a different space of functions, but as the distinction between odd and even disappears in the continuous limit, we restrict attention to $\mathscr{T}_n$. Now if $A \in \mathscr{T}_{n-1}$, then $(z + z^{-1})A$ also has even coefficients, hence belongs to $\mathscr{T}_n$, and from (4) we obtain

$$\left[(z + z^{-1})A, B\right] = \left[A, (z + z^{-1})B\right],$$

whenever either scalar product is defined, in particular if A or $B \in \mathscr{T}_{n-1}$. Consequently, the operation propelling an element from one of our subspaces to the next, which had previously been the unitary multiplication of a trigonometric polynomial by z, is now self-adjoint. This accounts for the fact that, despite starting with a Toeplitz matrix, our present formulation leads to results which more closely resemble the power moment problem, generated by Hankel matrices; we can see this directly by noting that members of $\mathscr{T}_n$ are polynomials of degree n in $\cos\theta$. In particular, there will be a three-term recursion for the analogues of orthogonal polynomials, and it is plausible that this should yield a second-order differential equation in the continuous limit.

Following the previous development, we introduce an orthogonal basis $\{T_k(z)\}$ which spans the successive subspaces $\{\mathscr{T}_k\}$, defined by coefficient vectors $\{\tau_k\}$ which satisfy

$$\mathbf{C}_{2k}\tau_k = (1,0,\ldots,0,1). \tag{A12}$$

Evidently, τ_k is real, hence T_k is real-valued; in terms of the earlier constructions, $T_k(z) = z^{-k}[P_{2k} + E_{2k}^0(z)]$. Letting

$$t_0' = 1;\ t_k' = t_{2k} + P_{2k}(0), \qquad k \geqslant 1,$$

denote the leading coefficient of $T_k(z)$, we see by (3a) that

$$\|T_0\|^2 = t_0';\quad \|T_k\|^2 = [T_k, T_k] = 2t_k', \qquad k \geqslant 1. \tag{A13}$$

Similarly, we introduce the evaluations $G_n^\zeta(z)$, defined by

$$[A_n, G_n^\zeta] = A_n(\zeta), \qquad A_n \in \mathscr{T}_n. \tag{A14}$$

As before,

$$G_n^\zeta(z) = \sum_{k=0}^{n} T_k(z)\,\overline{T_k(\zeta)}\,/\|T_k\|^2, \tag{A15}$$

so that, by (A13), the leading coefficient g_n of G_n^1 is

$$g_n = t_n'\,\overline{T_n(1)}\,/\|T_n\|^2 = T_n(1)/2, \qquad n \geqslant 1, \tag{A16}$$

and the coefficient vector γ_n of G_n^1 satisfies

$$\mathbf{C}_{2n}\gamma_n = (1,1,\ldots,1). \tag{A17}$$

Again, following the proof of (34), we obtain, for $n \geqslant 1$,

$$\{(z + z^{-1}) - (\zeta + \zeta^{-1})\}G_n^\zeta(z) = \frac{T_{n+1}(z)T_n(\zeta) - T_n(z)T_{n+1}(\zeta)}{2t_n'}, \tag{A18}$$

so that, letting $z = 1$,

$$(\zeta + \zeta^{-1} - 2)G_n^\zeta(1) = \frac{T_n(1)T_{n+1}(1)}{2t_{n+1}'}\left\{\frac{T_{n+1}(\zeta)}{T_{n+1}(1)} - \frac{T_n(\zeta)}{T_n(1)}\right\}. \tag{A19}$$

This is clearly a discrete version of (A2) and (A3). Indeed, let Δ be the difference operator, defined for a sequence $\{A_j\}$ by

$$\Delta A_j \equiv A_j - A_{j-1}.$$

We then find from (A16), (A13), and (A15) that, setting

$$\nu_k \equiv g_k/(t_k')^{1/2}, \qquad k \geqslant 1, \tag{A20}$$

we have

$$\frac{\Delta G_n^\zeta(1)}{\nu_n^2} = 2\frac{T_n(\zeta)}{T_n(1)}, \tag{A21}$$

so that, on applying Δ to (A19), we obtain, for $n \geqslant 1$,

$$(\zeta + \zeta^{-1} - 2)\nu_n^2\frac{T_n(\zeta)}{T_n(1)} = \Delta(t_n'/t_{n+1}')^{1/2}\nu_n\nu_{n+1}\Delta\frac{T_{n+1}(\zeta)}{T_{n+1}(1)}. \tag{A22}$$

Now as we have seen, in the limiting process ζ is replaced by $e^{i\lambda/M}$ so that $(\zeta + \zeta^{-1} - 2) \to -\lambda^2/M^2$, while $M\Delta \to d/dr$ and, by (A17),

$$G_n^1(\zeta)/M \to \int_{-r}^{r} q(s;r)e^{i\lambda s}\,ds,$$

with q defined by (A1). Moreover, $t_n \to 1$ for $n = Mr$, so that $\nu_n \to q(r;r)$, and

$$G_n^\zeta(1)/M = G_n^1(\zeta)/M \to \int_{-r}^{r} q(s;r)e^{is\lambda}\,ds = 2\int_0^r q(s;r)\cos s\lambda\,ds,$$

so that (A21) represents the solution (A3).

To find a second solution of (A2), we trace the source of (A22) to the fact (A18) that the projection of $\{(z + z^{-1}) - (\zeta + \zeta^{-1})\}G_n^\zeta(z)$ onto $\mathscr{T}_{n-1}$ is independent of n. Specifically, let $\mathscr{P}_n$ denote the orthogonal projection onto $\mathscr{T}_n$; then we can write (A18) as

$$\mathscr{P}_{n-1}\{(z + z^{-1}) - (\zeta + \zeta^{-1})\}G_n^\zeta = 0. \tag{A23}$$

With the added normalization

$$[1, G_n^\zeta] = 1, \tag{A24}$$

this characterizes G_n^ζ, for if some $G_n \in \mathscr{T}_n$ satisfies (A23) and (A24), then on writing

$$A_n(z) = A_n(\zeta) + \{(z - z^{-1}) - (\zeta + \zeta^{-1})\}F_{n-1}(z),$$

with

$$F_{n-1}(z) \equiv \frac{A_n(z) - A_n(\zeta)}{(z + z^{-1}) - (\zeta + \zeta^{-1})},$$

we have, for $|\zeta| = 1$,

$$\begin{aligned}[A_n, G_n] &= A_n(\zeta)[1, G_n] + [F_{n-1}, \{(z + z^{-1}) - (\zeta + \zeta^{-1})\}G_n] \\ &= A_n(\zeta),\end{aligned}$$

so that G_n coincides with the evaluation G_n^ζ. By the same argument, specifying $\mathscr{P}_{n-1}\{(z + z^{-1}) - (\zeta + \zeta^{-1})\}H_n(z)$ for an element $H_n \in \mathscr{T}_n$ is equivalent to prescribing the effect of H_n in the scalar product on $\mathscr{T}_n$. In particular, suppose that, for $|\zeta| = 1$, H_n^ζ satisfies

$$\mathscr{P}_{n-1}\{(z + z^{-1}) - (\zeta + \zeta^{-1})\}H_n^\zeta = 1, \tag{A25}$$

$$[1, H_n^\zeta] = 0; \tag{A26}$$

then on decomposing A_n as above, we find

$$[A_n, H_n^\zeta] = \left[\frac{A_n(z) - A_n(\zeta)}{(z + z^{-1}) - (\zeta + \zeta^{-1})}, 1\right], \qquad A_n \in \mathscr{T}_n. \tag{A27}$$

In order to express H_n^ζ in closed form, define for $k \leqslant n$

$$U_k(\zeta) \equiv [T_k, H_k^\zeta] = \left[\frac{T_k(z) - T_k(\zeta)}{(z + z^{-1}) - (\zeta - \zeta^{-1})}, 1\right]. \tag{A28}$$

Writing $H_n^\zeta = \Sigma_{k=0}^n \beta_k T_k$, we then find

$$\beta_j = \left[H_n^\zeta, T_j\right] / \|T_j\|^2 = \overline{U_j(\zeta)} / \|T_j\|^2,$$

so that, paralleling (A15),

$$H_n^\zeta(z) = \sum_{k=0}^{n} T_k(z)\, \overline{U_k(\zeta)} / \|T_k\|^2, \tag{A29}$$

while by (A25),

$$\left\{(z + z^{-1}) - (\zeta - \zeta^{-1})\right\} H_n^\zeta(z) = 1 + aT_{n+1}(z) + bT_n(z). \tag{A30}$$

To determine a and b we evaluate at $z = \zeta$, obtaining

$$aT_{n+1}(\zeta) + bT_n(\zeta) + 1 = 0,$$

and form the scalar product of H_{n+1}^ζ with (A30) and (A18) to find

$$aU_{n+1}(\zeta) + bU_n(\zeta) = 0,$$

$$1 = \frac{U_{n+1}(\zeta)T_n(\zeta) - U_n(\zeta)T_{n+1}(\zeta)}{2t_n'},$$

respectively. Thus

$$\left\{(z + z^{-1}) - (\zeta + \zeta^{-1})\right\} H_n^\zeta(z) = 1 + \frac{T_{n+1}(z)U_n(\zeta) - T_n(z)U_{n+1}(\zeta)}{2t_n'},$$

whence, proceeding as earlier with (A18), we see that

$$\frac{\Delta H_n^\zeta(1)}{\nu_n^2} = 2\frac{U_n(\zeta)}{T_n(1)} \tag{A31}$$

provides a second solution of (A22).

Another way of deriving these solutions is to interpret (A22) as the three-term recursion satisfied by the normalized $\{T_k(z)/\|T_k\|\}$. Specifically, if we write $(z + z^{-1})T_k(z)/\|T_k\| = \Sigma_{j=0}^{k+1} \beta_j T_j(z)/\|T_j\|$, then

$$\beta_j = \left[(z + z^{-1})T_k, T_j\right] / \|T_k\|\,\|T_j\| = \left[T_k, (z + z^{-1})T_j\right] / \|T_k\|\,\|T_j\|,$$

and as $(z + z^{-1})T_j \in \mathcal{T}_{j+1}$, β_j vanishes for $j + 1 < k$ by definition of T_k. Thus

$$(z + z^{-1})\frac{T_k(z)}{\|T_k\|} = \sigma_{k-1}\frac{T_{k-1}(z)}{\|T_{k-1}\|} + \rho_k\frac{T_k(z)}{\|T_k\|} + \sigma_k\frac{T_{k+1}(z)}{\|T_{k+1}\|}, \tag{A32}$$

with

$$\sigma_{-1} = 0$$

$$\rho_k = \left[(z + z^{-1})T_k, T_k\right] / \|T_k\|^2$$

$$\sigma_k = \left[(z + z^{-1})T_k, T_{k+1}\right] / \|T_k\|\,\|T_{k+1}\|, \qquad k \geqslant 0.$$

These coefficients define a recursion

$$(\zeta + \zeta^{-1})\, y_k = \sigma_{k-1}y_{k-1} + \rho_k y_k + \sigma_k y_{k+1} \tag{A33}$$

which, if applied starting with $k = 1$, generates for each ζ two independent solutions, determined by choice of initial values y_0 and y_1. We can verify directly that (A33) is equivalent to (A22). For since, by (A15), the leading term of $(z + z^{-1} - 2)G_k^1(z)$ is $(z + z^{-1})T_k(z)T_k(1)/\|T_k\|^2$, and the rest lie in $\mathscr{T}_k$, to which T_{k+1} is orthogonal, on forming the scalar product of (A18) with T_{k+1} we find, recalling (A13),

$$\sigma_k = (t_k'/t_{k+1}')^{1/2}, \qquad k \geqslant 1.$$

Then, setting $z = 1$ in (A32), we can solve for ρ_k to obtain, for $k \geqslant 2$,

$$\rho_k = 2 - T_{k-1}(1)/T_k(1) - t_k' T_{k+1}(1)/t_{k+1}' T_k(1),$$

and introducing these expressions into (A33) yields

$$(\zeta + \zeta^{-1} - 2)\nu_k^2 \frac{y_k(\zeta)(2t_k')^{1/2}}{T_k(1)} = \Delta(t_k'/t_{k+1}')^{1/2}\nu_k\nu_{k+1}\Delta\frac{y_{k+1}(\zeta)(2t_{k+1}')^{1/2}}{T_{k+1}(1)}.$$

Thus the recursion (A33) can be recast into (A22), with solutions given in terms of y_k by $y_k(\zeta)(2t_k')^{1/2}/T_k(1)$. To identify the $\{y_k\}$, let $\mathbf{D}$ be the tridiagonal symmetric matrix, with main diagonal consisting of $\{\rho_j\}$ and the two adjoining diagonals of $\{\sigma_j\}$, $j \geqslant 0$. By (A32), if

$$S_n(z) = \sum_{k=0}^{n} \beta_k \frac{T_k(z)}{\|T_k\|}, \tag{A34}$$

then $(z + z^{-1})S_n(z)$, when expanded in $\{T_k/\|T_k\|\}$, has as coefficients the components of $\mathbf{D}\beta$. Now let $y_k(\zeta)$, $k \geqslant 2$, be the solution of (A33) generated from initial values y_0 and y_1, and in (A34) set $\beta_k = y_k(\zeta)$, $0 \leqslant k \leqslant n$. Then $\mathbf{D}\beta$ has at most $n + 2$ nonzero components and, by (A33), in all but possibly the first and the last two components it coincides with $(\zeta + \zeta^{-1})\beta$. The action of $\mathscr{P}_j$ being to truncate an expansion of the form (A34) to $k \leqslant j$, we therefore have

$$\mathscr{P}_{n-1}(z + z^{-1}) \sum_{k=0}^{n} y_k(\zeta)\frac{T_k(z)}{\|T_k\|}$$
$$= \sum_{k=0}^{n-1} (\zeta + \zeta^{-1}) y_k(\zeta)\frac{T_k(z)}{\|T_k\|} + \{y_0\rho_0 + y_1\sigma_1 - (\zeta + \zeta^{-1})y_0\}\frac{T_0(z)}{\|T_0\|}$$

or

$$\mathscr{P}_{n-1}\{(z + z^{-1}) - (\zeta + \zeta^{-1})\} \sum_{k=0}^{n} y_k(\zeta)\frac{T_k(z)}{\|T_k\|} = y_0(\rho_0 - (\zeta + \zeta^{-1})) + y_1\sigma_1. \tag{A35}$$

We thus see that the three-term recursion is but a concrete implementation of an equation like (A23) or (A25). Specifically, on choosing $y_0 = 1$ and $y_1 = (\zeta + \zeta^{-1} - \rho_0)/\sigma_0$ in (A35), $\Sigma_0^n y_k(\zeta)T_k(z)/\|T_k\|$ satisfies (A23) and (A24), whence it coincides with $G_n^\zeta(z)$; the desired solution, $y_k(\zeta)(2t_k')^{1/2}/T_k(1)$, of (A22) is then $\Delta G_n^\zeta(1)/2\nu_n^2$, recovering (A21). By (A15), $y_k(\zeta) = T_k(\zeta)/\|T_k\|$.

Similarly, if $y_0 = 0$ and $y_1 = 1/\sigma_1$ in (A35), $\Sigma_0^n y_k(\zeta)T_k(z)/\|T_k\|$ satisfies (A25) and (A26) so that it coincides with $H_n^\zeta(z)$, and the corresponding solution of (A22) is given by $\Delta H_n^\zeta(1)/2\nu_n^2$, recovering (A31). Here, by (A29), $y_k(\zeta) = U_k(\zeta)/\|T_k\|$. Thus both $\{T_k(\zeta)\}$ and $\{U_k(\zeta)\}$ can be efficiently generated by the recursion (A33), from appropriate initial values.

We have already seen that the solution obtained from (A21) corresponds to (A3); it remains to find a limiting expression for that of (A31). To this end, we represent $H_n^\zeta(1)$ as $[H_n^\zeta(z), G_n^1(z)] = [G_n^1, H_n^\zeta]$, and evaluate this scalar product from its definition (3a) by determining $\mathbf{C}_{2n}\eta_n$, with η_n the coefficient vector of $H_n^\zeta(z)$. In turn, to do this, we invoke (A27) to obtain

$$\left[z^k + z^{-k}, H_n^\zeta(z)\right] = \left[\frac{(z^k + z^{-k}) - (\zeta^k + \zeta^{-k})}{(z + z^{-1}) - (\zeta + \zeta^{-1})}, 1\right], \qquad k \leqslant n. \text{(A36)}$$

As the coefficient vector of $z^k + z^{-k}$ has entries 1 in the $\pm k$th position and 0 elsewhere, the left-hand side of (A36) yields twice the kth component of $\mathbf{C}_{2n}\eta_n$, by (3a). Since, explicitly,

$$\begin{aligned}
\frac{(z^k + z^{-k}) - (\zeta^k + \zeta^{-k})}{(z + z^{-1}) - (\zeta + \zeta^{-1})} &= \frac{\left((z\zeta)^k - 1\right)(z^k - \zeta^k)}{(z\zeta)^{k-1}(z\zeta - 1)(z - \zeta)} \\
&= \left\{1 + \frac{1}{z\zeta} + \cdots + \frac{1}{(z\zeta)^{k-1}}\right\}\left\{z^{k-1} + z^{k-2}\zeta + \cdots + \zeta^{k-1}\right\} \\
&= z^{k-1} + (\zeta + \zeta^{-1})z^{k-2} + (\zeta^2 + 1 + \zeta^{-2})z^{k-3} \\
&\quad + \cdots + (\zeta + \zeta^{-1})z^{-(k-2)} + z^{-(k-1)},
\end{aligned}$$

for the right-hand side of (A36) we obtain

$$\begin{aligned}
&c_{k-1} + c_{k-2}(\zeta + \zeta^{-1}) + c_{k-3}(\zeta^2 + 1 + \zeta^{-2}) \\
&\quad + \cdots + c_0(\zeta^{k-1} + \zeta^{k-3} + \cdots + \zeta^{-k+3} + \zeta^{-k+1}) \\
&\quad + \cdots + c_{-(k-2)}(\zeta + \zeta^{-1}) + c_{-(k-1)}. \qquad \text{(A37)}
\end{aligned}$$

In the discrete approximation to the integral operator of (A1), c_0 represents the component of the identity so, for $k = sM$, its coefficient converges to $\int_{-s}^{s} e^{i\lambda t}\,dt/2$, since only half of the powers of ζ are counted, while the remainder of (A37) has as its limit, correspondingly,

$$2\int_0^s D(s - x)\left\{\frac{1}{2}\int_{-x}^{x} e^{i\lambda t}\,dt\right\}dx, \qquad 0 \leqslant s \leqslant r.$$

It follows that the limiting value of $\mathbf{C}_{2n}\eta_n$ at $s > 0$ is given by

$$\frac{1}{2}\frac{\sin\lambda s}{\lambda} + \int_0^s D(s - x)\frac{\sin\lambda x}{\lambda}\,dx = \frac{1}{2}\omega(s, \lambda), \qquad 0 \leqslant s < r.$$

Since the coefficient vector of G_n^1 tends to $q(s;r)$, we see by (3a) that the continuous limit of $H_n^\zeta(1) = [G_n^1, H_n^\zeta]$ becomes

$$\int_{-r}^{r} q(s;r)\frac{1}{2}\omega(s,\lambda)\,ds = \int_0^r q(s;r)\omega(s,\lambda)\,ds,$$

so that (A31) corresponds to (A4), as was to be shown.

References

1. N. I. Akhiezer, *The classical moment problem*, Hafner, New York, 1965.
2. L. Breiman, *Probability and stochastic processes*, Houghton-Mifflin, Boston, 1969.
3. A. M. Bruckstein and T. Kailath, *Inverse scattering for discrete transmission-line models*, SIAM Review **7** (1986), 1332–1349.
4. A. M. Bruckstein, B. C. Levy, and T. Kailath, *Differential methods in inverse scattering*, SIAM J. Appl. Math. **45** (1985), 312–335.
5. A. Bultheel, *Error analysis of incoming and outgoing schemes for the trigonometric moment problem*, Padé Approximation and Applications, Lecture Notes in Math., vol. 888, Springer-Verlag, Berlin and New York, 1981, pp. 100–109.
6. J. P. Burg, *Maximum entropy spectral analysis*, Proc. 37th Meet. Soc. Exploration Geophysicists, 1967; reprinted in Modern Spectrum Analysis (D. G. Childers, ed.), IEEE Press, New York, 1978, pp. 34–39.
7. _____, *A new analysis technique for time series data*, NATO Adv. Study Inst. on Signal Processing, Enschede, Netherlands, 1968; reprinted in Modern Spectrum Analysis (D. G. Childers, ed.), IEEE Press, New York, 1978, 42–48.
8. _____, *The relationship between maximum entropy spectra and maximum likelihood spectra*, Geophysics, 1972; reprinted in Modern Spectrum Analysis (D. G. Childers, ed.), IEEE Press, New York, 1978, pp. 132–133.
9. _____, *Maximum entropy spectral analysis*, Ph.D. dissertation, Stanford University, Stanford, California, 1975.
10. D. G. Childers, ed., *Modern spectrum analysis*, IEEE Press, New York, 1978.
11. B. S. Choi and T. M. Cover, *An information-theoretic proof of Burg's maximum entropy spectrum*, Proc. IEEE **72** (1984), 1094–1095.
12. I. Csiszàr and G. Tusnàdy, *Information geometry and alternating minimization procedures*, Statist. Decisions (1984), Suppl. 1, 205–237.
13. A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Royal Stat. Soc. Ser. B **39** (1977), 1–38.
14. H. Dym and A. Iacob, *Applications of factorization and Toeplitz operators to inverse problems*, Toeplitz Centennial Memorial Conference (I. Gohberg, ed.), Birkhäuser, Basel-Boston-Stuttgart, 1982, pp. 233–260.
15. I. C. Gohberg and I. A. Fel'dman, *Convolution equations and projection methods for their solution*, Transl. Math. Monographs, vol. 41, Amer. Math. Soc., Providence, RI, 1974.
16. B. Gopinath and M. M. Sondhi, *Inversion of the telegraph equation and the synthesis of non-uniform lines*, Proc. IEEE **59** (1971), 383–392.
17. U. Grenander and G. Szegö, *Toeplitz forms and their applications*, Univ. of Calif. Press, Berkeley, 1957.
18. S. Haykin, ed., *Nonlinear methods of spectral analysis*, Springer-Verlag, New York, 1979.
19. K. Hoffman, *Banach spaces of analytic functions*, Prentice-Hall, New York, 1962.
20. E. T. Jaynes, *On the rationale of maximum entropy methods*, Proc. IEEE **70** (1982), 939–952.
21. T. Kailath, *A theorem of I. Schur and its impact on modern signal processing*, I. Schur Methods in Operator Theory and Signal Processing (I. Gohberg, ed.), Operator Theory: Advances and Applications, vol. 18, Birkhäuser, Basel-Boston-Stuttgart, 1986, pp. 9–30.
22. T. Kailath, A. Bruckstein, D. Morgan, *Fast matrix factorizations via discrete transmission lines*, Linear Algebra Appl. **75** (1986), 1–25.
23. T. Kailath and H. Lev-Ari, *On mappings between covariance matrices and physical systems*, Contemporary Mathematics (B. Datta, ed.), Amer. Math. Soc., vol. 47, Providence, R.I., 1985, pp. 241–252.

24. T. Kailath, A. Vieira, and M. Morf, *Inverses of Toeplitz operators, innovations, and orthogonal polynomials*, SIAM Review **20** (1978), 106–119.

25. S. J. Karlin and W. J. Studden, *Tchebycheff systems: With applications in analysis and statistics*, Interscience, New York, 1966.

26. M. G. Krein, *The ideas of P. L. Čebyšev and A. A. Markov in the theory of limiting values of integrals and their further development*, Uspehi Mat. Nauk (NS) **6** (1951), 3–120; Amer. Math. Soc. Transl. Ser. 2, **12** (1959), 1–122.

27. _____, *Solution of the inverse Sturm-Liouville problem*, Dokl. Akad. Nauk SSSR **76** (1951), 21–24. (Russian)

28. _____, *On integral equations which generate second-order differential equations*, Dokl. Akad. Nauk SSSR **97** (1954), 21–24. (Russian)

29. _____, *Continuous analogues of propositions on polynomials orthogonal on the unit circle*, Dokl. Akad. Nauk SSSR **105** (1955), 637–640. (Russian)

30. R. T. Lacoss, *Autoregressive and maximum likelihood spectral analysis methods*, Aspects of Signal Processing, Part 2, NATO Adv. Study Inst., La Spezia, Italy, 1976 (G. Tacconi, ed.), D. Reidel, Boston, pp. 591–615.

31. H. J. Landau, *The inverse problem for the vocal tract and the moment problem*, SIAM J. Math. Anal. **14** (1983), 1019–1035.

32. H. Lev-Ari and T. Kailath, *Lattice-filter parametrization and modeling of nonstationary processes*, IEEE Trans. Inf. Theory **IT-30** (1984), 2–16.

33. T. L. Marzetta and S. W. Lang, *Power spectral density bounds*, IEEE Trans. Inf. Theory **IT-30** (1984), 117–122.

34. A. Papoulis, *Maximum entropy and spectral estimation: a review*, IEEE Trans. Acoustics, Speech, and Signal Proc. **ASSP-29** (1981), 1176–1186.

35. E. A. Robinson, *Spectral approach to geophysical inversion by Lorentz, Fourier, and Radon transforms*, Proc. IEEE **70** (1982), 1039–1054.

36. J. E. Shore and R. W. Johnson, *Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy*, IEEE Trans. Inf. Theory **IT-26** (1980), 26–37; *Comments and corrections*, IEEE Trans. Inf. Theory **IT-29** (1983), 942–943.

37. _____, *Properties of cross-entropy minimization*, IEEE Trans. Inf. Theory **IT-27** (1981) 472–482.

38. J. M. Van Campenhout and T. M. Cover, *Maximum entropy and conditional probability*, IEEE Trans. Inf. Theory **IT-27** (1981), 483–489.

39. Y. Vardi, L. A. Shepp, and L. Kaufman, *A statistical model for positron emission tomography*, J. Amer. Statist. Assoc. **80** (1985), 8–37.

AT & T Bell Laboratories, Murray Hill, New Jersey 07974