Estimation of Structured Covariance Matrices

JOHN PARKER BURG, SENIOR MEMBER, IEEE, DAVID G. LUENBERGER, FELLOW, IEEE, AND DANIEL L. WENGER

Invited Paper

Abstract—Covariance matrices from stationary time series are Toeplitz. Multichannel and multidimensional processes have covariance matrices of block Toeplitz form. In these cases and many other situations, one knows that the actual covariance matrix belongs to a particular subclass of covariance matrices. This paper discusses a method for estimating a covariance matrix of specified structure from vector samples of the random process. The theoretical foundation of the method is to assume that the random process is zero-mean multivariate Gaussian, and to find the maximum-likelihood covariance matrix that has the specified structure. An existence proof is given and the solution is interpreted in terms of a minimum-entropy principle. The necessary gradient conditions that must be satisfied by the maximum-likelihood solution are derived and unique and nonunique analytic solutions for some simple problems are presented.

A major contribution of this paper is an iterative algorithm that solves the necessary gradient equations for moderate-sized problems with reasonable computational ease. Theoretical convergence properties of the basic algorithm are investigated and robust modifications discussed. In doing maximum-entropy spectral analysis of a sine wave in white noise from a single vector sample, this new estimation procedure causes no splitting of the spectral line in contrast to the Burg technique.

I. Introduction

IN DOING spectral analysis of a stationary time series, one modern approach is to use the "Burg technique" to estimate second-order statistics from the raw time series data and then to use the maximum-entropy method to generate an estimate of the power density spectrum [1], [2]. These two steps are independent in that one can use the Burg technique to estimate the autocorrelation function out to lag N and then use a conventional Fourier transformation with a window function to get the spectral estimate, or, one can use the conventional lag product method of estimating the autocorrelation function followed by use of the maximum-entropy method of spectral estimation. The Burg technique and the maximum-entropy method solve two separate but related problems.

The maximum-entropy method of spectral estimation can be considered to be a generalization of the autoregressive method of spectral estimation. That is, if the second-order statistics that are known about the spectrum consists of the first $N+1$ lags of the autocorrelation function and if the entropy of the time series is given by the integral of the logarithm of the spectrum, then the maximum-entropy estimation procedure generates an Nth order all-pole model as the functional form satisfying the variational extremum. Demonstrating its more fundamental nature, the maximum-entropy principle also tells us precisely how to do multichannel and multidimensional spectral estimation using correlation information about the spectrum. Actually, in a much broader sense, the maximum-entropy principle supplies us with a general approach to estimation theory in which one combines information with an extremal principle to select a possible solution to a problem.

In this paper, we shall be concerned with developing a similar generalization of the Burg technique. The approach again uses a variational principle combined with information to produce a feasible solution. The particular problem that we attack is simply stated. Given a set of vector samples from a random process, we wish to select a covariance matrix of specified structure that corresponds in a reasonable way to the given data. The solution formulation is to assume that the random process is zero-mean multivariate Gaussian, that the vector samples are independent, and to take as our solution the covariance matrix of specified structure that maximizes the probability of occurrance of our vector samples. As we shall see, it is easy to write down the probability of the vector samples given that the covariance matrix is R. In fact, the information in the vector samples is neatly compressed into the sample covariance matrix S, so we just end up with a function $p(S, R)$ in the two matrices S and R. R is constrained to be a covariance matrix of the proper structure while S is a random-sample covariance matrix without any special structure.

If we momentarily disregard statistical considerations, the $p(S, R)$ function gives us the desired variational formulation for the problem. That is, given the vector samples, we calculate the sample covariance matrix S and then solve for the constrained covariance matrix R that maximizes $p(S, R)$. Aside from the many technical questions that one might ask, there is the subjective question, namely, why is the R that maximizes this function a "good" solution to the problem? The response to this question is that $p(S, R)$ really comes from maximum-likelihood considerations and thus should, in some sense, give us a reasonable answer, even if the process is not Gaussian and the vector samples are not independent. After all, the process might be Gaussian and the samples independent. In the final analysis, however, the $p(S, R)$ variational principle will survive only if it works well in practice. And a practical principle must meet two criteria. It must work well on a large majority of meaningful situations and it must not be too difficult to compute. We will try to show that the $p(S, R)$ principle has the first of these attributes to a high degree and that the algorithm presented herein helps greatly to reduce the problem of numerical computation.

II. Derivation of the Variational Principle

Suppose a column vector x is drawn from an N-dimensional Gaussian distribution with zero mean and covariance matrix R. Using a superscript T for the matrix transpose, the corresponding probability density function is

Manuscript received May 11, 1982; revised June 22, 1982. This is an extended version of a paper presented at the 1982 NATO Advanced Study Institute on Nonlinear Stochastic Problems to be published by D. Reidel Publishing Company in a volume edited by J.M.F. Moura and R. C. Bucy.

J. P. Burg is with Time and Space Processing, Inc., Santa Clara, CA 95051.

D. G. Luenberger is with Stanford University, Stanford, CA 94305.

D. L. Wenger is at P.O. Box 221, Soquel, CA 95073.

0018-9219/82/0900-0963$00.75

$$p(x) = (2\pi)^{-N/2} |R|^{-1/2} \exp(-x^T R^{-1} x/2). \quad (1)$$

Now, instead of a single vector sample, suppose that we have M independent vector samples, x_m, $m = 1$ to M. The probability density for this set of vectors follows from (1) as

$$p(x_1, x_2, \cdots, x_M) = (2\pi)^{-MN/2} |R|^{-M/2} \cdot \exp\left(-\sum_{m=1}^{M} x_m^T R^{-1} x_m/2\right). \quad (2)$$

We consider the situation where R is unknown except that it is a member of a certain family $\mathfrak{R}$ of feasible covariances. This family is determined by the structure of the underlying source of the data vectors. For example, an important case is where $\mathfrak{R}$ is the collection of all positive definite symmetric Toeplitz matrices, corresponding to a vector sample being N consecutive values from a sampled stationary time series.

Given the set of vector samples, x_m, $m = 1$ to M, the R that belongs to $\mathfrak{R}$ and which maximizes (2) is the "maximum-likelihood" estimate of the covariance matrix. Since we are using (2) only as a function to be maximized, we do not change the problem if we maximize a strictly monotonic function of (2), for example, the natural logarithm of (2). Thus taking the logarithm of (2), we get

$$-(MN/2)\log(2\pi) - (M/2)\log|R| - (1/2)\sum_{m=1}^{M} x_m^T R^{-1} x_m.$$

Dropping the leading constant term and dividing through by $M/2$, we define our objective function $g(S, R)$ to be

$$g(S, R) = -\log|R| - (1/M)\sum_{m=1}^{M} x_m^T R^{-1} x_m. \quad (3)$$

Maximizing $g(S, R)$ is clearly equivalent to maximizing (2). To simplify (3), we employ a standard result from matrix theory.

The trace of a square matrix is defined to be the sum of the elements along the main diagonal of the matrix. Now if A is an r by s matrix and B is an s by r matrix, then both AB and BA are square matrices and thus their traces are defined. A well-known matrix theorem is that their traces are equal, even if they are different sized matrices. We now note that because it is a scalar, i.e., a one by one matrix,

$$x^T R^{-1} x = \mathrm{tr}(x^T R^{-1} x) = \mathrm{tr}(AB)$$

where $A = x^T$, a one by N matrix and $B = R^{-1}x$, an N by one matrix. The matrix theorem says that

$$x^T R^{-1} x = \mathrm{tr}(AB) = \mathrm{tr}(BA) = \mathrm{tr}(R^{-1} x x^T).$$

Note that $R^{-1}xx^T$ is an N by N matrix. Using this result, (3) can be written as

$$g(S, R) = -\log|R| - \mathrm{tr}\left(R^{-1}(1/M)\sum_{m=1}^{M} x_m x_m^T\right).$$

Defining the sample covariance matrix S to be

$$S = (1/M)\sum_{m=1}^{M} x_m x_m^T$$

we arrive at the compact equation for $g(S, R)$ of

$$g(S, R) = -\log|R| - \mathrm{tr}(R^{-1}S). \quad (4)$$

This is our basic objective function. We wish to find the R that maximizes this function, given the sample covariance matrix S and given that R is constrained to have a particular structure.

In the next section, we shall derive necessary conditions for a maximum in terms of the gradient of the objective function. Before doing that, however, we investigate here some general considerations about the existence of a maximum for a nonnegative definite R matrix. When S is singular, the general case is rather involved. Thus to simplify our discussion, we shall assume that the S matrix is positive definite. It is also understood that throughout this paper, S and R are always assumed to be symmetric matrices.

We shall first derive an inequality relation involving trace $(R^{-1}S)$. Given S and R, there always exists a nonsingular congruence transformation that simultaneously diagonalizes both S and R. That is,

$$A^T S A = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ & & \\ 0 & 0 & s_N \end{bmatrix} \quad \text{and} \quad A^T R A = \begin{bmatrix} r_1 & 0 & 0 \\ 0 & r_2 & 0 \\ & & \\ 0 & 0 & r_N \end{bmatrix}.$$

We shall normalize A so that its determinant is unity and thus

$$|S| = \prod_{n=1}^{N} s_n \quad \text{and} \quad |R| = \prod_{n=1}^{N} r_n.$$

The s_n are all positive. The r_n are nonnegative with the number of positive terms equal to the rank of R. Let us assume that R is nonsingular and so

$$A^{-1} R^{-1} A^{-T} = \begin{bmatrix} 1/r_1 & 0 & 0 \\ 0 & 1/r_2 & 0 \\ & & \\ 0 & 0 & 1/r_N \end{bmatrix}.$$

Then

$$\mathrm{tr}(R^{-1}S) = \mathrm{tr}(A^{-1}R^{-1}A^{-T}A^T S A) = \sum_{n=1}^{N} (s_n/r_n).$$

Now let us minimize trace$(R^{-1}S)$, keeping the determinant of R constant. Using a Lagrange multiplier, we need

$$\delta\left[\sum_{n=1}^{N} (s_n/r_n) + \lambda \prod_{n=1}^{N} r_n\right] = \sum_{m=1}^{N} \left[-(s_m/r_m^2) + (\lambda/r_m)\prod_{n=1}^{N} r_n\right] \delta r_m = 0$$

or

$$s_n = \lambda |R| r_n.$$

We solve for λ by

$$|S| = \prod_{n=1}^{N} s_n = \lambda^N |R|^N \prod_{n=1}^{N} r_n = [\lambda|R|]^N |R|$$

giving $s_n/r_n = |R^{-1}S|^{1/N}$ for all n. Thus holding $|R|$ constant, we have our desired inequality relation

$$\mathrm{tr}(R^{-1}S) \geq N|R^{-1}S|^{1/N} = N|S|^{1/N}/|R|^{1/N}.$$

We note that because this relationship is not changed by scaling R, we can drop the statement about holding $|R|$ constant. Furthermore, equality is achieved if and only if R is a multiple of S.

Using this relation, we now note that

$$g(S, R) \leqslant -\log |R| - N|S|^{1/N}/|R|^{1/N}.$$

Thus if S is positive definite, as $|R|$ goes to zero, $g(S, R)$ goes to minus infinity and our probability density goes to zero.

We now make the following observation for positive definite S. Let us assume that the region $\mathfrak{R}$ of allowable R matrices is simply connected and contains at least one positive definite matrix R_0. The value of $g(S, R_0)$ is some finite value. Now it is clear that as R is varied continuously within the region $\mathfrak{R}$, the value of $g(S, R)$ remains finite as long as R remains positive definite. As R approaches singularity, however, the value of $g(S, R)$ approaches minus infinity. It then follows that continuous variation of R in order to find a maximum will never lead to crossing of the boundary of singularity, because points near the boundary are worse than the original point R_0. This shows that, at least in terms of seeking a local maximum, attention can be focused on positive definite matrices. Thus in an iterative algorithm, excursions of R outside of the positive definite region should not be allowed.

We now introduce a metric on the space of N by N matrices by considering the N^2 elements of the matrix to be components in an N^2-dimensional Euclidean space. This metric is consistent with the inner product introduced in the next section. Let $\mathfrak{D}$ be the set of nonnegative definite symmetric matrices. Then $\mathfrak{D}$ is a closed, convex, connected, and unbounded subset of our N^2-dimensional vector space. The boundary between $\mathfrak{D}$ and its complement is the set of singular nonnegative definite symmetric matrices. Next, let $\mathfrak{B}_b$ be the set of matrices whose elements are less than or equal to b in magnitude. Thus $\mathfrak{B}_b$ is compact, convex, and connected. Let $\mathfrak{C}_b$ be the intersection of $\mathfrak{D}$ and $\mathfrak{B}_b$. Then $\mathfrak{C}_b$ is compact, convex, and connected. Its boundary with its complement consists of singular nonnegative definite symmetric matrices or positive definite symmetric matrices with some main diagonal element equal to b. We now prove that if S is positive definite and if $\mathfrak{R}$ is a closed subset of the class of nonnegative definite symmetric matrices, then our probability density has a maximum in $\mathfrak{R}$.

First, if $\mathfrak{R}$ contains only singular matrices, our probability density is zero over $\mathfrak{R}$ and we are finished. Thus assume that $\mathfrak{R}$ contains a positive definite matrix R_0 and so our probability density at R_0 has a finite positive value. Next, suppose b is larger than any element of R_0 and thus R_0 belongs to $\mathfrak{R} \cap \mathfrak{C}_b$. Now $\mathfrak{R} \cap \mathfrak{C}_b$ is compact and our probability density is continuous. Thus our probability density has a maximum in $\mathfrak{R} \cap \mathfrak{C}_b$. This maximum is equal to or greater than the value at R_0. Our proof will be complete if we show that for a large enough b, the probability density is less than it is at R_0 for all positive definite symmetric matrices with a diagonal element larger than b. Equivalently, we need to show that $g(S, R)$ goes to minus infinity as the maximum element of R goes to infinity.

Let us begin our proof by doing the orthogonal diagonalization of R so that

$$M^T R M = \begin{bmatrix} r_1 & 0 & 0 \\ 0 & r_2 & 0 \\ & & \\ 0 & 0 & r_N \end{bmatrix}$$

with $M^T M = I$ and with r_1 being the largest of the eigenvalues. Then we have

$$\operatorname{tr}(R^{-1} S) = \operatorname{tr}(M^T R^{-1} M M^T S M) = \sum_{n=1}^{N} (q_n / r_n)$$

where the q_n are the diagonal terms of $M^T S M$. If s is the minimum eigenvalue of S, then $q_n \geqslant s$ for all n. Then

$$g(S, R) \leqslant -\sum_{n=1}^{N} [\log (r_n) + (s/r_n)].$$

Since for $0 < x < \infty$, $-\log (x) - s/x \leqslant -\log (s) - 1$, we have

$$g(S, R) \leqslant -\log (r_1) - s/r_1 - (N - 1)[\log (s) + 1].$$

Now, if b is the maximum element of R, then

$$b < \operatorname{tr}(R) = \sum_{n=1}^{N} r_n \leqslant N r_1$$

and thus $r_1 > b/N$. Therefore, as b goes to infinity, r_1 goes to infinity and $g(S, R)$ goes to minus infinity and our proof is finished. Thus if $\mathfrak{R}$ is a closed set of nonnegative definite symmetric matrices, then there is a maximum value for $g(S, R)$ in $\mathfrak{R}$.

If $\mathfrak{R}$ is the space of nonnegative definite matrices, then we shall show later that there is only one maximum to our probability density and it occurs when $R = S$. If $\mathfrak{R}$ is more restricted, then there may be multiple maxima. One simple example is to suppose that $\mathfrak{R}$ is a line weaving through our nonnegative definite space and that it passes by S quite closely several times. Then, we would have multiple maxima in such an $\mathfrak{R}$ space. One can hope that there might be only one maximum for subsets that one finds in practice such as Toeplitz matrices. As a beginning study of multiple maxima, we give in Section IV two examples, one a linear variety and the other a linear manifold, in which there are two equal valued, symmetrically placed maxima.

III. The Necessary Conditions on the Gradient

The problem is to maximize $g(S, R)$ over the matrices R belonging to a class $\mathfrak{R}$. We shall assume that the class $\mathfrak{R}$ is defined by a linear variety and is a subset of the class of symmetric matrices. A good example is the subset of Toeplitz matrices. Since a linear variety is closed, its intersection with the set of nonnegative definite symmetric matrices is closed and thus, if S is positive definite, a maximum for $g(S, R)$ exists in the interior of this intersection. Note that we are not restricting $\mathfrak{R}$ to be nonnegative definite, but that we will be looking for a maximum in the positive definite region of $\mathfrak{R}$.

With these assumptions, it is easy to characterize the solution of the problem in terms of the gradient of the objective function. Specifically, the gradient must be orthogonal to variations in $\mathfrak{R}$. Geometrically, that is all there is to it. Of course, to make this more concrete, it is necessary to define an inner product on the space of matrices so that the notions of gradient and orthogonality have specific meanings. We shall define the inner product of two matrices C and D to be given by the trace of $C^T D$. Note that the inner product is symmetric, bilinear, and that the inner product of a nonzero matrix with itself is positive.

With this definition, we now need to find the gradient of $g(S, R)$ with respect to R. We shall do this by deriving the variation of g in terms of the variation of R. We note that if R

is an N by N symmetric matrix, we may have up to $N(N-1)/2$ independent variables! One is thus faced with the thought of having to deal with very large matrices for moderate sized problems. Fortunately, the large number of equations can still be treated in terms of just N by N matrices.

To derive the necessary conditions, we begin with some definitions and two matrix theorems. First, we define the variation of R to be

$$\delta R = \begin{bmatrix} \delta R(1,1) & \delta R(1,2) & \cdots & \delta R(1,N) \\ \delta R(2,1) & \delta R(2,2) & \cdots & \delta R(2,N) \\ \cdots\cdots\cdots & \cdots & \cdots & \cdots \\ \delta R(N,1) & \delta R(N,2) & \cdots & \delta R(N,N) \end{bmatrix}$$

where $\delta R(i,j)$ is the variation of the i,jth element of R.

Our first matrix theorem gives us an expression for the variation of the determinant of R in terms of the variation of R. If $|R|$ is not zero, then

$$\delta|R| = |R| \operatorname{tr}(R^{-1}\delta R).$$

One derives this equation by noting that if the determinant of R is explicitly written out in terms of its elements, then the coefficient of $R(i,j)$ in this expansion is the cofactor of $R(i,j)$. In the inverse of R, the j,ith element is equal to the cofactor of the i,jth element of R divided by the determinant of R. In our above equation, we see that this is the coefficient of the variation of the i,j element of R. Now, noting that $\delta \log |R| = \delta|R|/|R|$, we have the important corollary that

$$\delta \log |R| = \operatorname{tr}(R^{-1}\delta R).$$

Our second useful matrix theorem gives us the variation of the inverse of R in terms of the variation of R. We first express the relation between the elements of R and R inverse by the matrix identity

$$RR^{-1} = I.$$

Taking the variation of this identity, we have

$$\delta R R^{-1} + R\delta(R^{-1}) = \delta I = 0, \text{ the null matrix.}$$

Our result is then

$$\delta(R^{-1}) = -R^{-1}\delta R R^{-1}.$$

Now we can derive the variation of $g(S,R)$ easily as

$$\begin{aligned} \delta g(S,R) &= -\delta \log |R| - \delta \operatorname{tr}(R^{-1}S) \\ &= -\operatorname{tr}(R^{-1}\delta R) - \operatorname{tr}[\delta(R^{-1})S] \\ &= -\operatorname{tr}(R^{-1}\delta R - R^{-1}\delta R R^{-1}S) \\ &= \operatorname{tr}(R^{-1}SR^{-1}\delta R - R^{-1}\delta R). \end{aligned}$$

Our expression for the variation of $g(S,R)$ is thus neatly written as

$$\delta g(S,R) = \operatorname{tr}[(R^{-1}SR^{-1} - R^{-1})\,\delta R].$$

The condition for maximization is that the gradient $R^{-1}SR^{-1} - R^{-1}$ is orthogonal to changes in $\mathfrak{R}$ space. That is, the variation of g is zero for any feasible variation of R. Thus the equation we shall solve is

$$\operatorname{tr}[(R^{-1}SR^{-1} - R^{-1})\,\delta R] = 0. \tag{5}$$

It often happens that the structural constraint on the variation of R is satisfied by R itself. The Toeplitz constraint is one case of this. We can then replace δR in (5) by R itself and the equation remains true. Our equation then says that

$$\operatorname{tr}[R^{-1}S] = N. \tag{6}$$

If we substitute this into $g(S,R)$, we have

$$g(S,R) = -\log |R| - N.$$

From this, we see that if R itself satisfies the structural constraints on the variation of R, then we can restate our variational principle as:

> Minimize the determinant of R under the constraints that R belongs to $\mathfrak{R}$ and that the trace of $(R^{-1}S)$ equals N.

This is a very interesting and intuitive way of stating our variational principle. In Section II, we derived the inequality relation

$$\operatorname{tr}(R^{-1}S) \geq N|R^{-1}S|^{1/N} = N|S|^{1/N}/|R|^{1/N}.$$

Using this, we see that (6) places a scale factor constraint on R so that $|R|$ is equal to or greater than $|S|$, with equality occurring only if R can be equal to S. We shall show later that (6) gives us the minimum variance estimate of the optimum scale factor for R when we are dealing with a Gaussian process. In the Gaussian situation, we also note that the entropy of the random process is given by $\log |R|$, so this special case of our general variational principle is saying:

> Choose the R that corresponds to the minimum-entropy process under the auxiliary constraint that R is normalized by the minimum variance scale factor. The entropy of the estimated process is equal to or greater than, but as close as possible to, the entropy of the process corresponding to the sample covariance matrix.

To show that this alternative principle is consistent with the more general principle, let us minimize $\log |R|$ under the constraints that (6) holds and that δR belongs to $\mathfrak{R}$. Using a Lagrange multiplier λ, we can write

$$\delta[-\log |R| - \lambda \operatorname{tr}(R^{-1}S)] = -\operatorname{tr}[(R^{-1} - \lambda R^{-1}SR^{-1})\,\delta R] = 0.$$

Since both R and δR belong to $\mathfrak{R}$, we can set $\delta R = R$. Then we have

$$\operatorname{tr}(I - \lambda R^{-1}S) = 0 \quad \text{or} \quad \lambda \operatorname{tr}(R^{-1}S) = N.$$

Thus (6) tells us that $\lambda = 1$. Therefore, if δR belongs to $\mathfrak{R}$, minimizing $\log |R|$ under (6) is equivalent to maximizing $g(S,R)$.

IV. Some Simple Cases of Covariance Estimation and their Solutions

In this section, we shall formulate some relatively simple cases of covariance estimation and solve them using our variational principle. We assume that our data consisted of a set of N-dimensional vector samples, and that we have already formed the sample covariance matrix S. To simplify our discussions, we shall assume here that the sample covariance matrix is positive definite. If S is singular, then in some cases, our solution is also singular, which requires a more careful consideration of our basic equations. We point out here, however, that S being singular does not mean in general that R is singular. For example, for Toeplitz structures, we normally obtain a nonsingular matrix for our estimate even if S is formed from a single vector sample.

With S positive definite, our solution is interior to the space

of nonnegative definite matrices and a maximum must satisfy the necessary gradient conditions. Thus if the gradient equations have only one solution, then the corresponding maximum is the only maximum. In most of the examples in this section, the solution is proven to be unique. Uniqueness if $\mathfrak{R}$ is Toeplitz is not known, but two simple examples are given showing nonuniqueness for linear variety and linear manifold cases.

In some of the chosen examples below, we already "know" what the best answer should be. In other cases, one may not be so certain that the derived answer is the best. In this latter case, use of the variational principle may end up changing one's intuition about what answers do make the most sense.

The Unconstrained Case

The simplest problem to solve is when the R matrix is not constrained. Then the variation of R is arbitrary and the gradient must be identically zero. Then we have from (5) that

$$R^{-1}SR^{-1} - R^{-1} = 0$$

which gives us immediately the unique solution of $R = S$. Of course, we have already covered this case in Section III since the variation of R belongs to $\mathfrak{R}$ and R can be equal to S.

Unknown Scale Factor Case

Suppose that we know the covariance matrix up to an unknown scale factor. An example of this is if one knows the shape of a spectrum as a function of frequency, but does not know the average power. This occurs if one passes white noise of unknown power through a known filter. For our general problem, we let $R = aW$, where W is a given positive definite symmetric matrix that is known to be proportional to the true covariance matrix and "a" is the unknown scale factor. In the white-noise example, we note that the elements of W would be obtained from the autocorrelation of the impulse response of the filter out to lag $N-1$. Then, with $R = aW$, $\delta R = (\delta a)\, W$ and we have from (5) that

$$\operatorname{tr}[(aW)^{-1}S(aW)^{-1} - (aW)^{-1}]\, W = 0$$

or

$$\operatorname{tr}[(aW)^{-1}S - I] = 0$$

giving

$$a = (1/N)\operatorname{tr}[W^{-1}S]. \tag{7}$$

One might not recognize that this unique solution is indeed the best answer since in our above example of white noise of unknown power passing through a known filter, the output power is normally estimated by a direct power average over our data samples. We shall now show in general that using the information provided by knowing W, (7) gives us the minimum variance estimate of "a."

Let us express the unique Cholesky decomposition of W in the form

$$W = G^{-1}G^{-T}$$

where G is lower triangular and its main diagonal consists of positive terms. Let us assume that x is one of our column vector samples and let us create the vector sample y by the linear transformation

$$y = Gx.$$

We can now write

$$aI = aGWG^T = \text{average value of } Gxx^TG^T$$

$$= \text{average value of } yy^T.$$

Thus the vector sample y is made up of N independent random variables of uniform variance "a." If x is multivariant Gaussian, so is y; and in this case, the best estimate of the variance of the y variables is simply the average square value over all elements in all vector samples. Weighting each independent sample equally gives us the minimum variance estimate. We now derive our estimate for "a" in terms of S and W as

$$aN = \text{sample average of the tr of } yy^T$$

$$= \text{sample average of the tr of } Gxx^TG^T$$

$$= \operatorname{tr}(GSG^T) = \operatorname{tr}(G^TGS)$$

$$= \operatorname{tr}(W^{-1}S), \quad \text{since } W^{-1} = G^TG.$$

We now see that our variational principle has indeed led us to the best estimate of "a."

If W were already equal to our solution matrix R, then we see that "a" = 1 and (7) becomes (6). Thus the optimally scaled R matrix does indeed satisfy (6) as discussed at the end of Section III.

The Burg Technique Case

One of the main features of the Burg technique is that the problem of estimating the reflection coefficients of a stationary time series is turned into one of estimating the covariance matrix of a pair of random variables whose individual variances are known to be equal. In this case, the structure of the two by two covariance matrix is of the form

$$R = \begin{bmatrix} a & b \\ b & a \end{bmatrix}, \quad \text{with } \delta R = \begin{bmatrix} \delta a & \delta b \\ \delta b & \delta a \end{bmatrix}.$$

Equation (5) now tells us that the sum of the two diagonal elements of the gradient must be zero. Also, since the gradient matrix is always symmetric, we see that the gradient matrix must actually be diagonal in this present case. Thus the gradient is of the form

$$R^{-1}SR^{-1} - R^{-1} = \begin{bmatrix} c & 0 \\ 0 & -c \end{bmatrix}.$$

We now invoke a matrix theorem concerning transposes about the minor diagonal, i.e., flipping the matrix about the diagonal that runs upward to the right at 45°. If we denote the transpose about the minor diagonal by a pre-superscript T, then it is easy to prove that if $AB = C$, then

$$({}^TB)({}^TA) = {}^TC.$$

Then we see that since $R^{-1} = {}^TR^{-1}$, we have

$$R^{-1}SR^{-1} - R^{-1} + R^{-1}({}^TS)\,R^{-1} - R^{-1} = 0$$

the null matrix, or

$$R^{-1}(S + {}^TS)\,R^{-1} = 2R^{-1}$$

or that finally

$$R = (1/2)(S + {}^TS).$$

Thus the estimated covariance matrix R is the minor diagonal symmetrized version of S, as given by the Burg technique.

The above use of the transpose about the minor diagonal gives us the following interesting and useful theorem. If the structure of R is such that R is equal to its minor diagonal transpose, then the inverse of R and the variation of R have this same property. Now, since transposing a matrix around either diagonal does not change its trace, we see that if R is also minor diagonal symmetric, then averaging the minor diagonal transposed form of (5) with itself, we have that R also satisfies

$$\operatorname{tr}\left[R^{-1}((S + {}^{T}S)/2)\,R^{-1} - R^{-1}\right]\delta R = 0.$$

Thus we can average S with its minor diagonal transpose and use this average in (5) to get the same answer for R. In fact, one can now note that if $\mathfrak{R}$ is minor diagonal symmetric, then S can be replaced by the minor diagonal symmetrized matrix without any change to our objective function (4). Thus we can start with this replacement and not change the functional form or numerical value of any of our equations. This is of more than passing significance.

One could hope that, when there are three or more variables whose variances are known to be equal, the optimum estimate of their variance is also the average of the sample average. Unfortunately, this simple property does not extend beyond two variables. The reader can investigate for himself why this is so from a mathematical point of view. An intuitive feeling for this fact can be developed if one supposes that the S matrix happens to indicate that the middle variable is almost independent of the other variables, but that the other variables are strongly dependent. Then, weighting all the variables equally to estimate the variance would not seem to be the right thing to do. This line of reasoning only says that a straight average is not optimum. The variational principle gives us a solution to this problem. Unfortunately, it cannot be written down explicitly.

Prediction Error Filter Interpretation

The Burg technique is based on the properties of prediction error filters. It is interesting that the maximum-likelihood procedure also can be interpreted in terms of prediction error filters, but in a more indirect manner. This interpretation arises if one considers that the mth column of R^{-1} is proportional to the prediction error filter that predicts the mth random variable from the rest of the variables. The diagonal terms of R^{-1} are the reciprocal values of the mean-square errors of the corresponding prediction error filters.

The congruence transformation of S by R^{-1}, that is $R^{-1}SR^{-1}$, gives us the covariance matrix of the prediction error filter variables, scaled by the reciprocals of their mean-square errors. Now, the necessary gradient condition (5) can be rewritten as

$$\operatorname{tr}\left[(R^{-1}SR^{-1} - R^{-1}RR^{-1})\,\delta R\right] = 0.$$

If $S' = R^{-1}SR^{-1}$ and $R' = R^{-1}RR^{-1} = R$ are the transformed covariance matrices, then we can write

$$\operatorname{tr}\left[(S' - R')\,\delta R\right] = 0.$$

Thus the gradient condition is more directly related to the scaled prediction error filter variables than they are to the untransformed variables. We shall use this observation in the sequel.

Two Variables with Fixed Value Constraints

In addition to the above two variable case in which the two variances are known to be equal, there is a class of two variable problems in which the precise value of one or more of the second-order statistics is known. These problems arise in practical situations and are also interesting from a philosophical point of view. There are basically four such problems, two of which have closed-form solutions and the other two require the solution of a cubic equation. Note that in these problems, the variation of R does not belong to $\mathfrak{R}$.

One of the Variances is Known: Let us assume that the variance of the first variable is known to be unity. Choosing unity is clearly as general as any other constant. So let

$$R = \begin{bmatrix} 1 & c \\ c & b \end{bmatrix} \quad \text{and} \quad S = \begin{bmatrix} A & C \\ C & B \end{bmatrix}.$$

Now since

$$\delta R = \begin{bmatrix} 0 & \delta c \\ \delta c & \delta b \end{bmatrix}$$

we see that

$$R^{-1}SR^{-1} - R^{-1} = \begin{bmatrix} \alpha & 0 \\ 0 & 0 \end{bmatrix}$$

which leads to

$$\begin{bmatrix} A-1 & C-c \\ C-c & B-b \end{bmatrix} = \begin{bmatrix} 1 & c \\ c & c^2 \end{bmatrix}.$$

Thus $\alpha = A - 1$ and we have $c = C/A$ and $b = B + (1 - A)(C/A)^2$. Our theoretical development has shown that if S is positive definite, then R is positive definite. We verify this and see what happens in the singular case by checking $b \geqslant 0$ and $b - c^2 \geqslant 0$. Clearly, the first inequality is true if the second is true, and the second is true if $b = B - C^2/A = (AB - C^2)/A \geqslant 0$. If B is zero, then b is zero and R is singular.

Having $c = C/A$ is clearly reasonable and perhaps intuitive. To see that the solution for b is also reasonable, one can note that to do the linear least mean square prediction of the second variable from the first involves merely multiplying the first variable by c. One could estimate the resulting mean-square prediction error from either S or R. Actually, b is such that both of these estimates are the same, that is,

$$[-c \quad 1]\begin{bmatrix} A & C \\ C & B \end{bmatrix}\begin{bmatrix} -c \\ 1 \end{bmatrix} = [-c \quad 1]\begin{bmatrix} 1 & c \\ c & b \end{bmatrix}\begin{bmatrix} -c \\ 1 \end{bmatrix} = b - c^2. \tag{8}$$

This interpretation, of course, follows directly from the previous observations about transforming into the prediction error variables and from noting that the lower right term of the gradient is zero.

One Variance and the Cross Variance are Known: Suppose in addition to knowing that the first variable has unity variance, it is also known that the cross variance is c. Using the above expressions for R and S, we note that

$$R^{-1} = 1/(b - c^2) \begin{bmatrix} b & -c \\ -c & 1 \end{bmatrix} \quad \text{and} \quad R^{-1}SR^{-1} - R^{-1} = \begin{bmatrix} \alpha & \gamma \\ \gamma & 0 \end{bmatrix}.$$

Only the bottom right-hand equation needs to be solved and it is

$$b = c^2 + c^2A - 2cC + B.$$

We note again that this b is such as to make the two prediction error estimates as shown in (8) equal.

The Cross Variance is Known: Here we start with

$$R = \begin{vmatrix} a & c \\ c & b \end{vmatrix} \quad \text{and} \quad R^{-1}SR^{-1} - R^{-1} = \begin{vmatrix} 0 & \gamma \\ \gamma & 0 \end{vmatrix}.$$

Pre- and post-multiplying this last equation by R, we have

$$S - R = \gamma R \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} R = \gamma \begin{bmatrix} a & c \\ c & b \end{bmatrix} \begin{bmatrix} c & b \\ a & c \end{bmatrix}$$

or

$$\begin{bmatrix} A - a & C - c \\ C - c & B - b \end{bmatrix} = \gamma \begin{bmatrix} 2ac & ab + c^2 \\ ab + c^2 & 2bc \end{bmatrix}.$$

From the two equations on the main diagonal, we see that $A = a(1 + 2\gamma c)$ and $B = b(1 + 2\gamma c)$. Thus a has the same proportion to A as b has to B. To actually solve for a, we end up with a cubic equation in a that says

$$[-c/a \quad 1] \begin{bmatrix} a & c \\ c & b \end{bmatrix} \begin{bmatrix} -c/a \\ 1 \end{bmatrix} = [-c/a \quad 1] \begin{bmatrix} A & C \\ C & B \end{bmatrix} \begin{bmatrix} -c/a \\ 1 \end{bmatrix}.$$

Thus we see again that the solution says that the prediction error is the same whethe we apply the optimum estimated filter to the sample covariance matrix or to the estimated covariance matrix.

Both Variances are Known: We can assume without loss of generality that both variances are unity. In working out the equation for c, we end up with a cubic that again says that the sample and estimated least mean square error in predicting one variable from the other are equal.

An Example of Nonuniqueness of Solution

Suppose we consider the two matrices

$$Y = \begin{bmatrix} 1 + a & 0 \\ 0 & 1 - a \end{bmatrix} \quad \text{and} \quad Z = \begin{bmatrix} 1 - a & 0 \\ 0 & 1 + a \end{bmatrix}$$

and let $R(x) = Y(1 + x)/2 + Z(1 - x)/2$. Thus $R(x)$ defines a one-dimensional linear variety in our four-dimensional vector space. Defining

$$A(x) = (1 + a)(1 + x)/2 + (1 - a)(1 - x)/2 = 1 + ax$$

and

$$B(x) = (1 - a)(1 + x)/2 + (1 + a)(1 - x)/2 = 1 - ax$$

we have

$$R(x) = \begin{bmatrix} A(x) & 0 \\ 0 & B(x) \end{bmatrix}.$$

We note that $R(x)$ is characterized by being diagonal and having its trace equal to 2.

Suppose now that our sample covariance matrix is $(1 - a^2)/2$ times the identity matrix. Then our objective function $g(S, R)$ is

$$\begin{aligned} g[S, R(x)] &= -\log [A(x)B(x)] - [(1 - a^2)/2][1/A(x) + 1/B(x)] \\ &= -\log [1 - (ax)^2] - (1 - a^2)/[1 - (ax)^2] \\ &= \log \{(1 - a^2)/[1 - (ax)^2]\} - (1 - a^2)/[1 - (ax)^2] \\ &\quad + 1 + [-\log (1 - a^2) - 1]. \end{aligned}$$

Letting $y = (1 - a^2)/[1 - (ax)^2]$, our objective function can be written as

$$[\log (y) - y + 1] + [-\log (1 - a^2) - 1].$$

Since $\log(y) - y + 1 \leqslant 0$ for $0 < y < \infty$, and is equal to zero only when $y = 1$, we see that the objective function has its only maxima at $x = \pm 1$ and that these two maxima are both equal to $-\log (1 - a^2) - 1$. The minimum between these two maxima is at $x = 0$ and has the value $-(1 - a^2)$.

As a verification, we check the necessary conditions by noting that the gradient is

$$\begin{aligned} R^{-1}SR^{-1} - R^{-1} &= [(1 - a^2)/2] \begin{bmatrix} 1/(1 + ax)^2 & 0 \\ 0 & 1/(1 - ax)^2 \end{bmatrix} \\ &\quad - \begin{bmatrix} 1/(1 + ax) & 0 \\ 0 & 1/(1 - ax) \end{bmatrix} \\ &= (-1/2) \begin{bmatrix} (1 + 2ax + a^2)/(1 + ax)^2 & 0 \\ 0 & (1 - 2ax + a^2)/(1 - ax)^2 \end{bmatrix}. \end{aligned}$$

We note that

$$\delta R(x) = a \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \delta x.$$

Thus the condition on the gradient that we are looking for is that its two diagonal terms be equal. This occurs only when $x = \pm 1$ or when $x = 0$, verifying our necessary conditions.

If one forms a linear manifold from two diagonal matrices, the first with diagonal values of $[2/(1 + a), 1/2, 2/(1 - a)]$ and the second with $[2/(1 - a), 1/2, 2/(1 + a)]$, with the magnitude of "a" less than one, then if the sample covariance matrix is the identity matrix, one finds that there are two maxima for

$g(S, R)$, occurring when R is equal to one of the generating matrices. Thus even when one can scale the R matrix with an arbitrary multiplier, there can be more than one solution to the maximum-likelihood estimation procedure.

V. The Inverse Iteration Algorithm

We shall now develop a general iterative process for determining a solution to the necessary conditions as given by (5). The approach used here is somewhat abstract, but it greatly simplifies the development since concepts are isolated from the intricate computations that may be required to express the concepts in concrete form. To avoid additional complexity in our derivation, we shall consider here only the common situation where the variation of R belongs to $\mathfrak{R}$. The case in which elements of R are constrained to have particular numerical values can be treated by similar methods by using a reduced dimensionality.

The Algorithm

The iterative process that we consider here shall be termed *inverse iteration*. At any stage, we begin with an approximation R_k and a new approximation R_{k+1} is determined as follows:

1) Find D_k belonging to $\mathfrak{R}$ so that $g(S - D_k, R_k)$ satisfies the necessary conditions.

2) Put $R_{k+1} = R_k + D_k$.

Note that in step 1, we find the change (belonging to $\mathfrak{R}$) in our data S that makes the current approximation optimal. Then, in step 2 the approximation is updated by the negative of this virtual change in the data. That is why we have chosen to name the method inverse iteration. The main reason for using inverse iteration for this problem is that the gradient of the objective function is linear with respect to S, so the problem implied by step 1 is a linear problem. Note that at each step of our iteration, our approximation satisfies our linear constraints. Experience has shown, however, that the new approximation can jump out of the positive definite region. We shall discuss later a small modification of the basic algorithm that handles this possibility.

Improving Direction

There are some basic properties of the inverse iteration process that are extremely important. The first is that the direction of change D_k is an improving direction. That is, if R_k is changed by adding a small amount of D_k, the objective function will increase over what it was with R_k. We shall now prove this.

What we wish to show is that the gradient of $g(S, R)$ with respect to R has a positive inner product with the direction D determined by step 1. This means that small movement along D will yield a positive change in g. In mathematical terms, we wish to show that

$$\operatorname{tr}[(R^{-1}SR^{-1} - R^{-1})\,D] > 0.$$

The left side of this can be rewritten as

$$\operatorname{tr}[(R^{-1}(S - D)\,R^{-1} - R^{-1})\,D] + \operatorname{tr}[R^{-1}DR^{-1}D].$$

The first term is zero by construction since this is the requirement that R is optimal for the data $S - D$. We now prove that the second term is positive.

Since R is positive definite, we again do the Cholesky decomposition as we did in Section IV, this time in the form

$$R^{-1} = G^T G.$$

Then

$$\operatorname{tr}[R^{-1}DR^{-1}D] = \operatorname{tr}[G^T GDG^T GD] = \operatorname{tr}[(GDG^T)\,(GDG^T)].$$

Since D is a symmetric matrix, the matrix GDG^T is symmetric and we are looking at its inner product with itself. Since D is not zero, this inner product is positive and we have proven that D is an improving direction.

Quadratic Approximation

Although D defines an improving direction, it is not yet clear how far to move in that direction. The basic algorithm moves a distance qD, with $q = 1$. Computational experience has shown that the best value for q is indeed often equal to unity, but it would be best to move the distance that maximizes the objective function. To initiate an investigation of this general subject, we derive here a quadratic approximation in q for our objective function. It is interesting how easily this and higher order approximations can be generated by use of the trace function.

Using the matrix theorems developed in Section III, we derive the second-order expansion of $g(S, R + qD)$ as a function of q about $q = 0$ by first noting that

$$dg(S, R + qD)/d^2q|_{q=0} = \operatorname{tr}(R^{-1}DR^{-1}D) + \operatorname{tr}[(R + qD)^{-1}D(R + qD)^{-1}S].$$

The second derivative, evaluated at $q = 0$, can now be derived as

$$d^2g(S, R + qD)/d^2q\Big|_{q=0} = \operatorname{tr}(R^{-1}DR^{-1}D) - 2\operatorname{tr}(R^{-1}DR^{-1}DR^{-1}S). \quad (9)$$

Having already shown that the first derivative evaluated at $q = 0$ is just the first term in (9), we can now write down the Taylor expansion out to second order as

$$g(S, R + qD) = g(S, R) + (q + q^2/2)\operatorname{tr}(R^{-1}DR^{-1}D) - q^2\operatorname{tr}(R^{-1}DR^{-1}SR^{-1}D).$$

Differentiating with respect to q and setting the result to zero gives us the maximum of this quadratic expression as

$$q_{\max} = [\operatorname{tr}(R^{-1}DR^{-1}D)]/[2\operatorname{tr}(R^{-1}DR^{-1}SR^{-1}D) - \operatorname{tr}(R^{-1}DR^{-1}D)].$$

One of the unproven conjectures of this paper is that $q_{\max}$ approaches unity as the iterative algorithm converges.

VI. Computation of the Inverse Iteration

The iterative algorithm discussed in the previous section is based on repeatedly solving a linear problem in order to determine the new estimate of the covariance matrix R. The solution of this linear problem is, computationally, the most difficult part of the algorithm. By suitable formulation and use of available structure, however, a very efficient procedure can be developed.

Formulation and Duality

The necessary condition for a solution has been given by (5). It is repeated here (with slightly different notation) as that of finding an R that satisfies

$$\operatorname{tr}[(R^{-1}SR^{-1} - R^{-1})\,Q] = 0$$

for all Q in $\mathfrak{R}$. The iterative process is deduced by assuming a trial solution R is given. A new trial R' is sought that satisfies

$$\mathrm{tr}\,[(R^{-1}SR^{-1} - R^{-1}R'R^{-1})\,Q] = 0 \tag{10}$$

for all Q in $\mathfrak{R}$. This is linear in the unknown R'.

A problem of this type is a generalized projection problem. It can be solved by selecting a basis either in $\mathfrak{R}$ or in the orthogonal complement of $\mathfrak{R}$. We propose here to use a basis in $\mathfrak{R}$ itself since in most applications, such as the case where $\mathfrak{R}$ is the space of symmetric Toeplitz matrices, this will be of considerably less dimension than the dimension of the orthogonal complement.

Let the basis in $\mathfrak{R}$ be Q_m, $m = 1$ to M, where we are assuming that the space is of dimension M. We note that the Q_m are symmetric matrices. Then we write an expansion for the unknown R' as

$$R' = \sum_{m=1}^{M} x_m Q_m. \tag{11}$$

This expansion is substituted into the basic linear condition (10) which leads to

$$\sum_{m=1}^{M} \mathrm{tr}\,[R^{-1}Q_m R^{-1}Q_j]\; x_m = \mathrm{tr}\,[R^{-1}SR^{-1}Q_j]$$

for $j = 1$ to M.

This is a system of M equations in the M unknowns x_m, $m = 1$ to M. Solution of this system yields the next approximation R' by using (11). If we define the matrix A by

$$A_{ij} = \mathrm{tr}\,[R^{-1}Q_i R^{-1}Q_j]$$

and

$$c_j = \mathrm{tr}\,[R^{-1}SR^{-1}Q_j]$$

the system takes the standard form

$$Ax = c. \tag{12}$$

Proof that A is Positive Definite

We can show that the matrix A defined above is symmetric and positive definite. Actually, the symmetry is seen immediately. Being positive definite is quite important for implementation, since it means that the efficient algorithms for solution of symmetric positive definite systems can be employed.

Consider

$$\sum_{i,j=1}^{M} x_i A_{ij} x_j$$

$$= \sum_{i,j=1}^{M} x_i\,\mathrm{tr}\,[R^{-1}Q_i R^{-1}Q_j]\; x_j = \mathrm{tr}\,[R^{-1}BR^{-1}B]$$

where

$$B = \sum_{j=1}^{M} Q_j x_j.$$

We have shown before that when R is positive definite and B is not the null matrix, this expression is positive, showing that A is positive definite.

The Overall Procedure

The overall algorithm is outlined by the following steps:

1) Start with an initial positive definite approximation to R that lies in $\mathfrak{R}$. (We usually start with the identity matrix for the Toeplitz constrained problems.)

2) Using this R, calculate all the traces required to define the matrix A and solve the system (12). Then use (11) to obtain a tentative new approximation R'.

3) Evaluate the function $g(S, R')$. If this value is not larger than with the previous approximation or if the new approximation for R is not positive definite, then define $D = R' - R$ and try an approximation of the form $R + qD$ for $q = \frac{1}{2}$. Keep cutting q by a factor of two until an approximation satisfying the above requirements is found. Set the new R equal to this value and go back to step 2.

There are, of course, numerous variations possible, especially regarding the step-size determination.

VII. Sine Wave in White Noise Examples

In this section, we shall compare the maximum-likelihood procedure with the Burg technique for estimating a Toeplitz covariance matrix from a single vector sample. This vector sample is supposed to be eleven consecutive samples from a sinusoid in white-noise process. The estimated Toeplitz matrix provides us with an estimate of the autocorrelation of the process out to lag 10. Instead of trying to compare results by looking at the estimated autocorrelation values, we have plotted the corresponding maximum entropy spectra. This spectral domain presentation best illustrates a major problem with the Burg technique; namely, that of splitting a single spectral line into multiple lines. The vector samples have been chosen to produce line splitting in a controlled manner and thus clearly expose the cause of the phenomena. An outstanding property of the maximum-likelihood estimation procedure is that it is impervious to this line splitting problem, even for the worst case vectors.

Our sample covariance matrix S is very singular since it is of rank one. (Actually, since Toeplitz matrices are minor diagonal symmetric, as a first step we can take the sample covariance matrix that we get from time reversing our vector sample and average it with our initial matrix to get a rank two matrix for S.) In our theoretical development of the properties of the maximum-likelihood procedure, we have restricted S to being positive definite. This guarantees that the R matrix is positive definite. However, even with S being singular, the Toeplitz matrix constraint on R normally produces a positive definite answer. We shall not give a proof of this, but it takes special conditions on the vector sample for the estimated covariance matrix to be singular using either estimation procedure.

Our comparison of the two estimation techniques is done using three different but similar vector samples. Each vector sample is eleven points long and consists of a unit amplitude cosine wave whose period is eight sample points long. Taking the sampling period to be 0.005 s, the foldover frequency is 100 Hz, and the frequency of the cosine wave is 25 Hz. The three cases differ in the beginning phase of the cosine wave. The first case starts the cosine function at 45°, the second at 90°, and the third at 135°. Fig. 1 shows the cosine wave. The points from 1 to 11 go into the first case vector, the points from 2 to 12, the second, and the points from 3 to 13, the third. To each of these cosine vectors, we have added the same vector (0.000562, −0.019127, 0.007377, −0.000149, −0.007479, −0.013960, 0.003510, 0.012380, 0.006979,

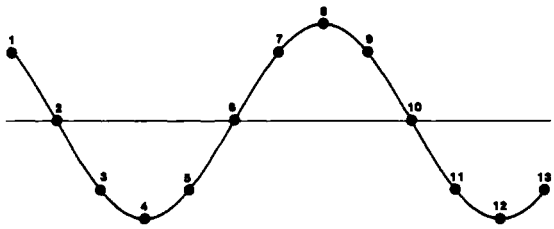

Fig. 1. Sampled cosine function used in the examples.

0.003092, 0.010053). This vector was generated from eleven randomly chosen numbers from a zero-mean Gaussian process with variance 0.0001. Thus our three cases have the same noise and differ only in the beginning phase of the cosine. We note that the average power of a unit amplitude sinusoid is 0.5 and thus the average power of our sinusoid in white noise process is 0.5001.

Before discussing the examples, we review how the Burg technique estimates the first two lag values of the autocorrelation function, $R(0)$ and $R(1)$. $R(0)$ is estimated by the sample average power. $R(1)$ is then estimated by first estimating the first reflection coefficient $C(1)$ and then using $R(1) = -C(1)R(0)$. For our vector sample, x_n, $n = 1$ to 11, the Burg formula for $C(1)$ is

$$C(1) = -2 \sum_{n=1}^{10} (x_n x_{n+1}) / \sum_{n=1}^{10} (x_n^2 + x_{n+1}^2).$$

If the summation were from $n = 1$ to 8, that is, over one period of the cosine, then if we disregarded the added noise vector, the numerator would be -5.656856 and the denominator would be 8.0. The $C(1)$ would be -0.707107, which is precisely the correct value. As we shall see, summing in the two extra terms will cause an error in the estimate of $C(1)$ and lead to line splitting.

45° Case

The first important feature of this vector sample (points 1 to 11 of Fig. 1) is that the average square value of the cosine alone over the eleven samples is $5/11 = 0.454545$. With the noise added, the sample average is 0.457055. This low sample average for the power is due to starting at 45°. One period of the cosine wave is eight samples and the three extra samples are 0.707107, 0.000000, and -0.707107. The average square value of these three extra samples pulls the average down. The second important fact is that in the calculation of $C(1)$, if we again neglect the low-level white noise, the $n = 9$ and $n = 10$ terms add zero to the numerator, while they add 1.000000 to the denominator. Thus our estimate for $C(1)$ would be -0.628540 instead of -0.707107. This value is too positive and forces the estimate to put more power into higher frequencies than it should. Fig. 2 shows the maximum-entropy spectrum corresponding to the autocorrelation values out to $R(10)$ as estimated by the Burg technique. The Burg technique has discovered that the time series is highly predictable, but because it did not get the correct value for $C(1)$, it ended up estimating multiple lines to account for the high predictability. It should be noted that once the Burg technique has decided on a value for $C(1)$, this value is unchanged as the higher order reflection coefficients are estimated. Thus although the higher order estimates are dependent on the estimated value of $C(1)$, they cannot change $C(1)$. Also, since there is a strict one-to-one correspondence between normalized power spectra and sequences of reflection coefficients, the estimation of the higher

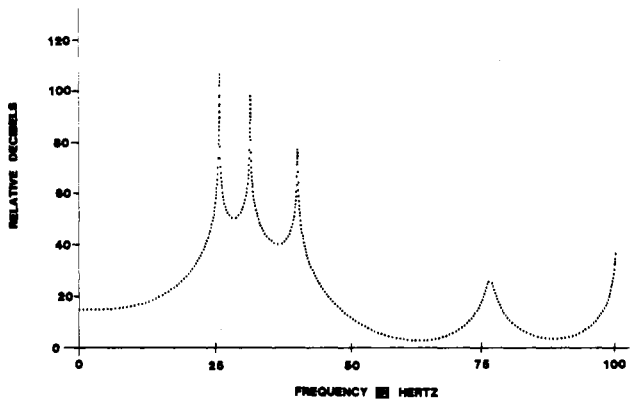

Fig. 2. 45° case, Burg technique maximum-entropy spectrum.

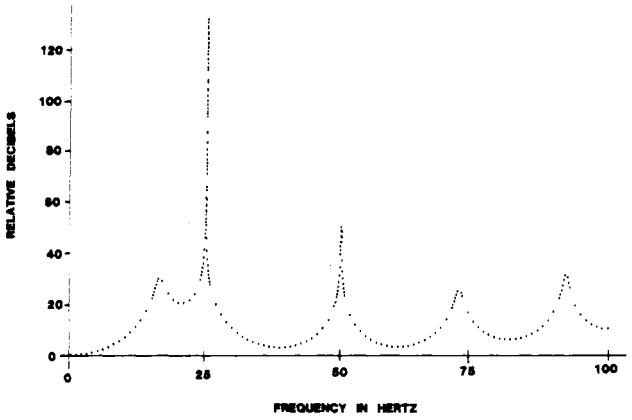

Fig. 3. 45° case, maximum-likelihood maximum-entropy spectrum.

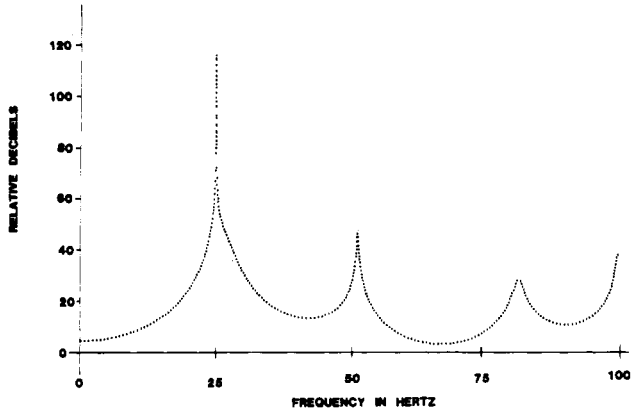

Fig. 4. 90° case, Burg technique maximum-entropy spectrum.

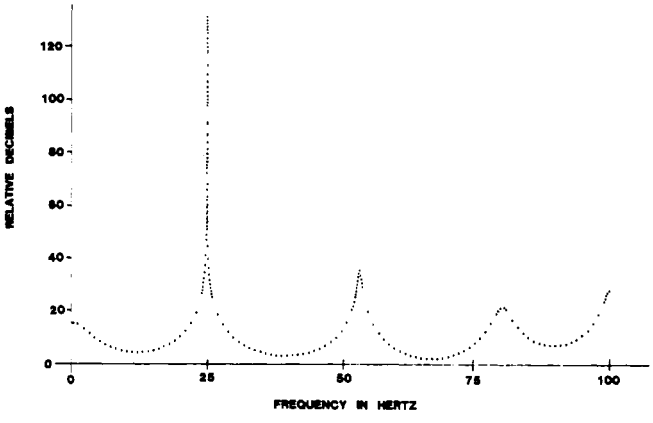

Fig. 5. 90° case, maximum-likelihood maximum-entropy spectrum.

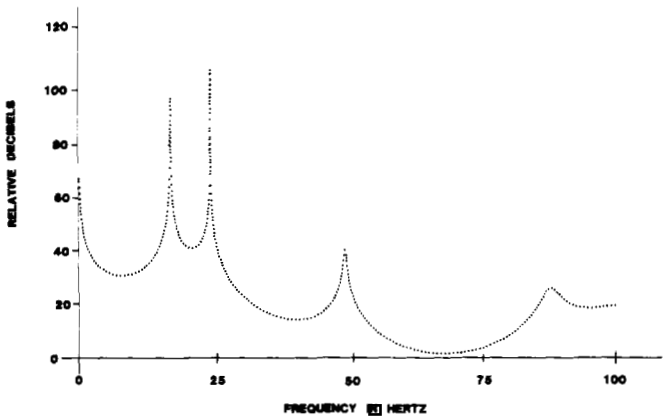

Fig. 6. 135° case, Burg technique maximum-entropy spectrum.

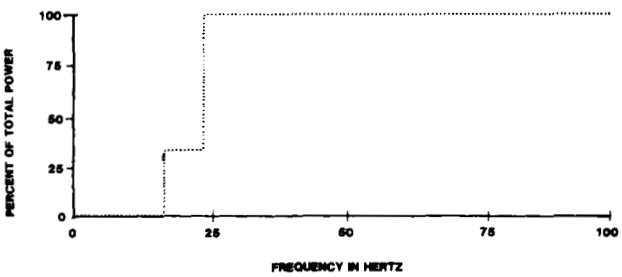

Fig. 7. Distribution function of spectrum in Fig. 6.

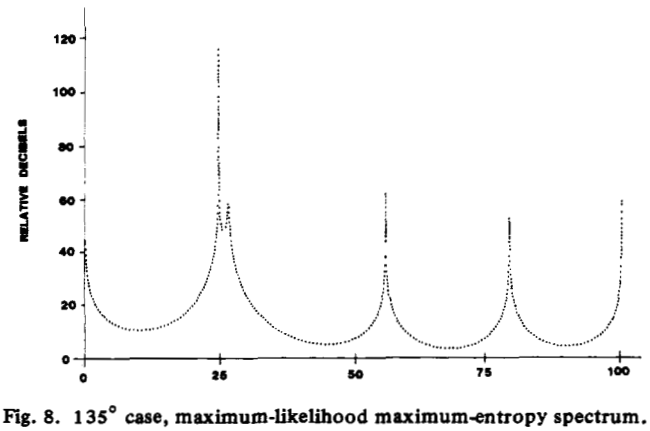

Fig. 8. 135° case, maximum-likelihood maximum-entropy spectrum.

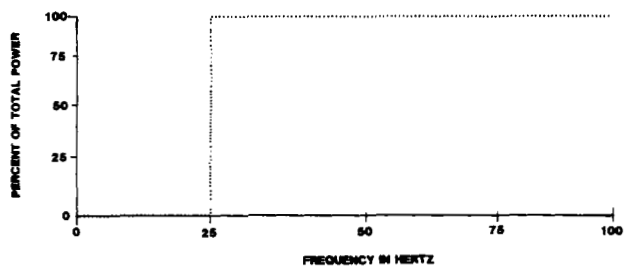

Fig. 9. Distribution function of spectrum in Fig. 8.

order coefficients cannot be influenced in such a way as to cover up an error in a lower order coefficient. Looking again at Fig. 2, the maximum power spectral line has a frequency of 25.870376 Hz which is not a particularly accurate estimate of the true frequency. Two other lines of considerable power have been split off to higher frequencies.

Fig. 3 shows the tenth-order maximum-entropy spectrum corresponding to the maximum-likelihood estimate of the autocorrelation function for this 45° case. Here we have one overwhelming peak at 24.948942 Hz, an error of 5 parts in 10 000. In addition, the estimate for the total average power is 0.503475! Remembering that the sample average was 0.457055, how can the maximum-likelihood procedure get an estimate that is so much closer to the "correct" answer of 0.5001? To try to find an explanation of this phenomena, one should ponder the Unknown Scale Factor Case of Section IV.

90° Case

This vector sample (points 2 to 12 of Fig. 1) has the cosine wave positioned so that the sample averaged square value has the correct value of 0.5. With the noise added, the sample average is 0.499437. More important, however, is the fact that the Burg technique calculation of $C(1)$ is correct for this positioning. Looking at Fig. 4, we see that there is no line splitting for this vector and the major peak frequency is an accurate 25.001776 Hz.

Fig. 5 is the spectrum for the maximum-likelihood estimated autocorrelation function. The peak frequency of 24.931945 Hz is not as good as the Burg technique answer, but one would believe that it is in the range of values produced by random-noise vectors. The power estimate is 0.500531, which is slightly closer to 0.5001 than the sample average power. The two techniques seem to be similar in performance for this vector sample. Even the two maximum-entropy spectra are similar.

135° Case

The three extra samples of the cosine (points 3 to 13 of Fig. 1) are −0.707107, −1.000000, and −0.707107. Thus the sample average of the cosine wave is 6/11 = 0.545455. With the noise added, the sample average is 0.542202. Again, for the cosine wave alone, the Burg technique calculation of $C(1)$ has a significant error, giving the too negative value of −0.771389. This causes too much power to be put into low frequencies. Indeed, Fig. 6 shows a second major peak at about 17 Hz with the primary peak having a frequency of 24.228206 Hz. Maximum entropy spectra such as show in Fig. 6 can be misleading in estimating the power under a peak. Fig. 7 is the normalized distribution function of the power density spectrum in Fig. 6. That is, it is the running integral starting at zero frequency, normalized by the total power so that it is read in percent power. This figure shows that about 35 percent of the power is in the split-off peak and 65 percent in the main peak. The power in the white noise is too small to see on this scale.

Fig. 8 is the maximum-likelihood estimated spectrum. Again, there is only one overwhelming peak whose frequency is 24.930899 Hz. Right next to it is a tiny peak but with a total power much too small to call it a split-off peak. It is interesting that the maximum-entropy assumption and the autocorrelation function values call for a peak to occur that close to the main peak. However, in Fig. 4 one can see the indication of a budding peak on the right side of the main peak. The maximum-likelihood estimate for the total power is 0.497147. Again, this estimate is remarkable in the face of a sample averaged estimate of 0.542202 for the total power.

Fig. 9 is the distribution function of the spectrum in Fig. 8. It simply shows that almost all of the power is in the 25-Hz line.

To sum up the results of this comparison, the maximum-

likelihood estimated covariance matrix gave us an estimated autocorrelation function for our sinusoid in white-noise process that is always equal to and usually much superior to the Burg technique estimate. Regardless of the beginning phase of the cosine, the maximum-likelihood technique gave us a single peak that engulfed the rest of the spectrum, a center frequency estimate within 7 parts per 10 000, and a power estimate well within a 1-percent error. Considering that the Burg technique has become the standard with which to compare new methods, the performance of the maximum-likelihood estimation procedure is indeed impressive.

VIII. Summary and Conclusions

The approach of maximum-likelihood estimation of a structured covariance matrix is both theoretically sound and computationally feasible. In evaluating the procedure with simulated random vector data from a stationary time series, we have found that the maximum-entropy spectra obtained from the autocorrelation functions estimated by this generalization of the Burg technique to be consistantly better and often much better than those obtained by the usual Burg technique. In particular, there is no evidence of line splitting with the maximum-likelihood procedure.

As discussed in the Introduction, the Burg technique and maximum-entropy spectral analysis are natural partners in doing spectral estimation of single-channel time series. However, the use of the new technique to estimate multichannel and multidimensional covariance matrices, perhaps followed by maximum-entropy spectral analysis, is of particular importance since there is only a moderately successful generalization of the Burg technique to the multichannel case and none at all to the multidimensional case. Multidimensional spectral estimation from data originating from an array of sensors lying in a stationary field of propagating plane waves is one of the most important challenges today. This new estimation technique provides a solution to the problem of analyzing the array data into a positive definite covariance matrix, from which one can then estimate the spectrum via the maximum-entropy method. Not having a good data analysis method for getting the multidimensional covariance matrix has been a drawback to the usefulness of high-resolution multidimensional spectral analysis as well as limiting its effectiveness for real data.

There is a great amount of research and testing to be done on this new estimation technique. Conditions for the uniqueness of the solution to the variational principle should be developed, especially for the general class of Toeplitz structured matrices. The properties of the iterative algorithm are mostly unknown and large improvements in its overall performance are likely. From a theoretical point of view, it appears that this maximum-likelihood procedure can be strongly related to general maximum-entropy theory as preached by Edwin T. Jaynes. In particular, in a private correspondence, John E. Shore has shown that this maximum-likelihood estimation theory is interrelated to his work with minimum cross-entropy [3]. The possibility of uniting these two estimation theories is indeed exciting.

References

[1] D. G. Childers, Ed., *Modern Spectrum Analysis*. New York: IEEE Press, 1978.

[2] J. P. Burg, "Maximum entropy spectral analysis," Ph.D. dissertation, Dep. Geophys., Stanford Univ., Stanford, CA, 123 pp., 1975.

[3] J. E. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 230-237, Apr. 1981.