The Role of Likelihood and Entropy in Incomplete-Data Problems: Applications to Estimating Point-Process Intensities and Toeplitz Constrained Covariances

MICHAEL I. MILLER AND DONALD L. SNYDER, FELLOW, IEEE

*The principle of maximum entropy has played an important role in the solution of problems in which the measurements correspond to moment constraints on some many-to-one mapping **h(x)**. In this paper we explore its role in estimation problems in which the measured data are statistical observations and moment constraints on the observation function **h(x)** do not exist. We conclude that:*

*1) For the class of likelihood problems arising in a complete-incomplete data context in which the complete data **x** are nonuniquely determined by the measured incomplete data **y** via the many-to-one mapping **y** = **h(x)**, the density maximizing entropy is identical to the conditional density of the complete data given the incomplete data. This equivalence results by viewing the measurements as specifying the domain over which the density is defined, rather than as a moment constraint on **h(x)**.*

2) The identity between the maximum entropy and the conditional density results in the fact that maximum-likelihood estimates may be obtained via a joint maximization (minimization) of the entropy function (Kullback-Liebler divergence). This provides the basis for the iterative algorithm of Dempster, Laird, and Rubin [1] for the maximization of likelihood functions.

3) This iterative method is used for maximum-likelihood estimation of image parameters in emission tomography and gamma-ray astronomy. We demonstrate that unconstrained likelihood estimation of image intensities from finite data sets yields unstable estimates. We show how Grenander's method of sieves can be used with the iterative algorithm to remove the instability. A bandwidth sieve is introduced resulting in an estimator which is smoothed via exponential splines.

4) We also derive a recursive algorithm for the generation of Toeplitz constrained maximum-likelihood estimators which at each iteration evaluates conditional mean estimates of the lag products based on the previous estimate of the covariance, from which the updated Toeplitz covariance is generated. We prove that the sequence of Toeplitz estimators has the property that they increase in likelihood, remain in the set of positive-definite Toeplitz covariances, and has all of its limit points stable and satisfying the necessary conditions for maximizing the likelihood.

INTRODUCTION

There has recently been a tremendous increase in the application of maximum-entropy techniques to constraint problems with nonunique solutions [2]–[6]. The rational, as first proposed by Jaynes [7] is that of all candidates consistent with a set of constraints the maximum-entropy (*maxent*) solution is the one which occurs with greatest multiplicity. The success of the entropy function is due to the property that the candidate solutions are concentrated strongly near the maxent one; solutions with appreciably lower entropy are atypical of those specified by the data [8]. The fact that entropy methods have been successful for the solution of underdetermined inference problems suggests that these methods may play an important role in the solution of maximum-likelihood (ML) parameter estimation problems. In particular, problems encompassed by a *complete–incomplete* data specification, in which the measured data specify many possible complete data sets over which the estimates may be obtained via maximization of the likelihood seem particularly well-suited for entropy techniques. For the problems examined in this paper, a function is estimated which parameterizes a known probability density; the actual data (denoted as the *complete data*) described by the density are not observed. Rather, observations consist of data (denoted as *incomplete data*) which nonuniquely specify the complete data via some set of many-to-one mappings.

Our motivation is that we have been working on problems in image reconstruction and spectrum estimation in which parameters are estimated from measurements which are both noisy, i.e., samples of a stochastic process, as well as incomplete [9]–[14]. For the former, images are reconstructed which are the intensity of a Poisson process; due to errors introduced by the measurement device, the data do not uniquely specify the point process. In the spectrum estimation problem, the Toeplitz covariance of a Gaussian random process is estimated from finite measurements of a stationary process. For both problems, ML techniques are an obvious choice for generating the parameter estimates; however, the fact that the measured data do not uniquely determine the underlying stochastic processes suggests that entropy may play a key role. In fact, both entropy and likelihood approaches have been applied for the solution of these problems [2]–[6], [9], [15], [16].

We shall explore the role that the entropy function plays in the generation of maximum-likelihood estimates (MLEs) when the data in the likelihood function are nonuniquely determined. Given a prior density f, the maxent density to which we refer is the density q maximizing the entropy

$$E(q, f) = -\int_D q(x) \log \left[\frac{q(x)}{f(x)}\right] dx \tag{1a}$$

Manuscript received February 21, 1986; revised April 7, 1987. This work was supported in part by the National Science Foundation under a PYIA Grant ECE-8552518 and under Grant ECS8407910, and the National Institutes of Health under Grant PR001380. This paper was presented in part at the International Society for Photo-Optical Engineering Conference, San Diego, CA, July 1986 and at the Information Theory Symposium, Ann Arbor, MI, October 1986.

The authors are with the Electronic Systems and Signals Research Laboratory, Department of Electrical Engineering, and the Institute for Biomedical Computing, Washington University, St. Louis, MO 63130, USA.

IEEE Log Number 8715352.

0018-9219/87/0700-0892$01.00 © 1987 IEEE

subject to the constraints $\boldsymbol{H}$ fixing mean values of the observation function given by

$$\boldsymbol{H} = \int_D q(x)\boldsymbol{h}(x)\, dx. \tag{1b}$$

The maxent density $\hat{q}$ is

$$\hat{q}(x) = \frac{\exp(\boldsymbol{v}^t\boldsymbol{h}(x)) f(x)}{Z(\boldsymbol{v})}, \qquad \text{for } x \in D. \tag{1c}$$

The Lagrange multiplier vector $\boldsymbol{v}$ is chosen so that $\hat{q}$ satisfies the constraints of (1b) over its support set D, with t denoting matrix transpose. Alternatively $-E(q, f)$ has been called the I-divergence, K-L number, or cross-entropy between q and f [17], [18].

The problem we set up is similar to that which generates $\hat{q}$ of (1c) in that we assume a prior density $f(x; \phi)$ describing the complete data x, parameterized by some function ϕ, and observations $y = \boldsymbol{h}(x)$ where $\boldsymbol{h}(x)$ is a many-to-one vector mapping from the complete data observations. We are interested in finding the maxent density $\hat{q}$ although we do not assume that the observations y provide moment constraints on $\boldsymbol{h}$. For the estimation problems in which we are involved, the set of observations $\boldsymbol{h}(x_i)$ for $i = 1, \cdots, N$ is small so that the average of $\boldsymbol{h}(x_i)$ may not be close to its expectation. This is in contrast to the results of Van Campenhout and Cover [19] who proved that for large N, the conditional distribution of a random variable x_1 given the empirical observation vector

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{h}(x_i), \qquad \text{where } x_1, x_2, \cdots$$

are i.i.d. random variables with common prior f, converges to the maxent distribution given in (1c). That is, for N large so that

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{h}(x_i)$$

is close to the expectation of $\boldsymbol{h}$, the exponential density of (1c) is identical to that based on rules of conditional probability.

To anticipate the results for finite observations, it is shown in Section I that by viewing the incomplete data y as restricting the domain over which the maxent density is defined, rather than as a moment constraint on $\boldsymbol{h}(x)$, the density maximizing $E(q, f)$ is identical to the conditional density derived via formal rules of conditional probability. This equivalence results in the fact that a large class of ML problems may be posed as a joint maximization of the entropy function. This joint-maximization view is then related to the *expectation-maximization* (EM) algorithm of Dempster, Laird, and Rubin [1] and the alternating minimization of Csiszar and Tusnady [20], and then applied to the iterative generation of MLEs in image reconstruction and Toeplitz constrained covariance estimation.

I. The Equivalence of the Conditional and Maxent Density

Our strategy in this section is to set up the *complete-incomplete* data models, from which it follows that the conditional and maxent densities are identical. This, in turn, results in the solution of classical likelihood problems via a maximization of the entropy function of (1).

A. Generation of the Complete-Incomplete Data Model

We begin by defining the underlying probability space $(\Omega, \sigma(\Omega), P)$, with sample points ω in Ω, events in the sigma field $\sigma(\Omega)$ of subsets of Ω, and probability measure P. Then we define the *complete-data* random variable X as a measurable function so that $X:(\Omega, \sigma(\Omega)) \rightarrow (\chi, \sigma(\chi))$, with the probability of an event $B \in \sigma(\chi)$ given by $P_X(B) = P\{\omega : X(\omega) \in B\}$. We shall for the entire paper *assume* that $P_X(\,)$ is absolutely continuous with density $f(x; \phi)$. The family of densities $f(x; \phi)$ parameterized by ϕ we term the *complete-data* densities. We say that we are given *incomplete data* if instead of observing x in χ, only the sample y is available, where $y = \boldsymbol{h}(x)$ for some measurable m-dimensional vector mapping $\boldsymbol{h}$; this mapping is, in general, many to one, so x is not uniquely specified by y. Thus the incomplete data y result from the existence of $m + 1$ sample spaces, the complete data space χ, and the m incomplete data spaces $Y_1, Y_2, \cdots, Y_m$. We denote the product space describing the incomplete data vector $Y = Y_1 \times Y_2 \times \cdots \times Y_m$.

The complete data x are a particular realization from χ, and the incomplete observed data y are a particular realization from Y. Therefore, the many-to-one mapping $\boldsymbol{h}(x)$ taking χ to Y specifies the subset $\chi(y) \subset \chi$ in which the complete data x is an element, with $\chi(y)$ given by the following relation:

$$\chi(y) = \{x : \boldsymbol{h}(x) = y\}. \tag{2}$$

The family of densities $g(y; \phi)$ describing the incomplete data are derived according to the following relation:

$$g(y; \phi) = \int_{\chi(y)} f(x; \phi)\, dx. \tag{3a}$$

The conditional density of x given y is then

$$\begin{aligned} k(x|x \in \chi(y), \phi) &= \frac{f(x; \phi)}{\int_{\chi(y)} f(x; \phi)\, dx}, && \text{for } x \in \chi(y) \\ &= 0, && \text{for } x \notin \chi(y). \end{aligned} \tag{3b}$$

B. Equivalence of Maxent and Conditional Density

By incorporating the complete-incomplete model of the measurements $y = \boldsymbol{h}(x)$ into the entropy formalism of (1) it follows directly that the density maximizing entropy $E(q, f)$ is identical to the conditional density $k(x|x \in \chi(y), \phi)$. Stated in another way, the density q closest to f in the cross-entropy sense becomes the conditional density.

Following the maximum-entropy approach given by (1a, b) we find the density q maximizing $E(q, f)$ subject to the constraints determined by the data y? Since moment values on $\boldsymbol{h}$ do not exist, the solution of the incomplete data problem is different then that stated in (1). The data y determine the domain $\chi(y)$ as given by (2) over which the complete data are defined. Therefore, rather than specifying moment constraints on $\boldsymbol{h}(x)$ the data y specify the domain D over which the maxent density has support, with the constraint given by

$$\int_{\chi(y)} q(x)\, dx = 1.$$

From Jensen's inequality, the density q maximizing entropy $E(q, f)$ subject to the support constraint

$$\int_{\chi(y)} q(x)\, dx = 1 \tag{4}$$

becomes

$$\hat{q}(x) = \frac{f(x;\phi)}{\displaystyle\int_{\chi(y)} f(x;\phi)\, dx}, \quad \text{for } x \in \chi(y). \tag{5}$$

This is precisely the conditional density denoted as $k(x|x \in \chi(y), \phi)$ in (3).

Remark 1: Relationship to the Maxent Density of (1c): By viewing the measurements y as determining the domain over which the density is defined, the density closest to the prior f in the cross-entropy sense is the conditional density of x given y. The fact that the conditional $k(x|x \in \chi(y), \phi)$ and the maxent density $\hat{q}(x)$ are identical is consistent with numerous results (see Jaynes [21] collected works for various examples), demonstrating that the set of maxent density are equivalent to those generated with formal rules of conditional probability. It is not of the same exponential form as in (1c) because moment constraints on $\boldsymbol{h}$ are not assumed.

The density of (5) may be related to the *maxent* density of (1c) via the results of Van Campenhout and Cover. For purpose of illustration, we choose a particularly simple many-to-one mapping for the incomplete data and apply their Theorems I and II. Assume $\{x_1, \cdots, x_N\}$ are the complete data which are i.i.d. discrete random variables with mass function $f(x)$ on the range $x \in \{1, 2, \cdots, m\}$. Given the incomplete data

$$y = \sum_{i=1}^{N} x_i$$

the conditional probability of $\{x_1, x_2, \cdots, x_N\}$ given y is from (5)

$$\begin{aligned} &k(x_1 = j_1, \cdots, x_N = j_N | y) \\ &\quad = \frac{f(x_1 = j_1, \cdots, x_N = j_N)}{\displaystyle\sum_{\{(j_1, \cdots, j_N): j_1 + \cdots + j_N = y\}} f(x_1 = j_1, \cdots, x_N = j_N)} \end{aligned}$$

and from the independence of the x_i's it follows that

$$\begin{aligned} &k(x_1 = j_1 | y) = f(x_1 = j_1) \\ &\quad \cdot \left[\frac{\displaystyle\sum_{\{(j_2, \cdots, j_N): j_2 + \cdots + j_N = y - j_1\}} \prod_{i=2}^{N} f(x_i = j_i)}{\displaystyle\sum_{\{(j_1, \cdots, j_N): j_i + \cdots + j_N = y\}} \prod_{i=1}^{N} f(x_i = j_i)} \right]. \end{aligned}$$

For $N \to \infty$, by Theorems I and II of Van Campenhout and Cover, the above density converges to the *maxent* one of (1c), with $h(x)$ replaced by x.

C. Maximum-Likelihood Via Joint-Maximization of the Entropy Function

Now it follows directly that the MLE $\hat{\phi}$ may be posed as a joint-entropy maximization. The MLE is obtained by maximizing the log likelihood of the incomplete data log $g(y; \phi)$. Applying (3a, b) the log likelihood of y is given by

$$\log g(y;\phi) = \log f(x;\phi) - \log k(x|x \in \chi(y), \phi) \tag{6}$$

and evaluating the expectation of $\log g(\cdot)$ in (6) with respect to the conditional density $k(x|x \in \chi(y), \phi)$ yields the following log likelihood to be maximized:

$$\begin{aligned} \log g(y;\phi) = &\int_{\chi(y)} k(x|x \in \chi(y), \phi) \log f(x;\phi)\, dx \\ &- \int_{\chi(y)} k(x|x \in \chi(y), \phi) \log k(x|x \in \chi(y), \phi)\, dx. \end{aligned} \tag{7}$$

Since the density maximizing $E(q, f)$ is precisely the conditional $k(x|x \in \chi(y), \phi)$, the log likelihood to be maximized becomes the following:

$$\log g(y;\phi) = \max_{\{q\}} \left\{ -\int_{\chi(y)} q(x) \log \left[\frac{q(x)}{f(x;\phi)} \right] dx \right\}$$

with q a density over $\chi(y)$; that is

$$\int_{\chi(y)} q(x)\, dx = 1.$$

The MLE is then simply

$$\hat{\phi} \leftarrow \underset{\{\phi\}}{\operatorname{argmax}} \left\{ \max_{\{q\}} E(q, f(\phi)) \right\}.$$

It follows that the MLE $\hat{\phi}$ is given by the following joint maximization:

$$\hat{\phi} \leftarrow \underset{\{\phi\}}{\operatorname{argmax}} \left\{ \int_{\chi(y)} \hat{q}(x) \log f(x;\phi)\, dx \right\} \tag{8a}$$

$$\hat{q} \leftarrow \underset{\{q\}}{\operatorname{denmax}} \left\{ -\int_{\chi(y)} q(x) \log \left[\frac{q(x)}{f(x;\hat{\phi})} \right] dx \right\}. \tag{8b}$$

The notation "argmax" means the MLE $\hat{\phi}$ is the argument which maximizes the expectation of $\log f(x;\phi)$; the notation "denmax" means the density $\hat{q}$ maximizes $E(q, f)$.

Because of the equivalence between the conditional and maxent densities, the incomplete data log likelihood is simply the joint maximum with respect to q, ϕ of the entropy function $E(q, f(\phi))$. This results in the estimation problem being expanded to what appears to be a larger problem in which both the parameters ϕ as well as density q must be estimated, which in turn implies the following iterative algorithms.

D. Iterative Joint Maximization of the Entropy Function

The Expectation-Maximization Algorithm of Dempster, Laird, and Rubin: The fact that maximum-likelihood problems may be viewed as a joint maximization of the entropy $E(q, f)$ results in the EM algorithm of Dempster *et al.* [1] for the iterative solution of maximum-likelihood problems. The EM algorithm yields the sequence of iterates $\{\phi^{(p)}; p = 0, 1, \cdots\}$ defined via the recursive maximization

$$\phi^{(p+1)} \leftarrow \underset{\{\phi\}}{\operatorname{argmax}} \{Q(\phi|\phi^{(p)})\} \tag{9a}$$

where $Q(\phi|\phi^{(p)})$ is the expectation of the complete data log likelihood given the incomplete data and the pth iterate $\phi^{(p)}$:

$$Q(\phi|\phi^{(p)}) = \int_{\chi(y)} k(x|x \in \chi(y), \phi^{(p)}) \log f(x; \phi)\, dx. \quad (9b)$$

Note, this is precisely (8a) with $\hat{q}(x) = k(x|x \in \chi(y), \phi^{(p)})$ for $\phi^{(p)}$ the pth iterate. The iterates of (8a) have the property that $\log g(y; \phi^{(0)}), \log g(y; \phi^{(1)}), \cdots$ is a monotonic nondecreasing sequence. This may be seen by defining the function

$$H(\phi|\phi^{(p)}) = \int_{\chi(y)} k(x|x \in \chi(y), \phi^{(p)}) \log k(x|x \in \chi(y), \phi)\, dx$$

and from (7) noting that

$$\log g(y; \phi^{(p+1)}) - \log g(y; \phi^{(p)}) = Q(\phi^{(p+1)}|\phi^{(p)}) - Q(\phi^{(p)}|\phi^{(p)}) + H(\phi^{(p)}|\phi^{(p)}) - H(\phi^{(p+1)}|\phi^{(p)}).$$

Jensen's inequality yields $H(\phi^{(p)}|\phi^{(p)}) - H(\phi^{(p+1)}|\phi^{(p)}) \geq 0$, and since the function $Q(\phi|\phi^{(p)})$ is maximized at stage $p + 1$, it follows that $\log g(y; \phi^{(p+1)}) \geq \log g(y; \phi^{(p)})$.

The Alternating Minimization of the K-L Divergence: In the derivation of (8), the joint maximization of $E(q, f)$ may be viewed as a joint minimization of the K-L divergence $-E(q, f)$ between the densities q and f, where f is varied via the parameter ϕ. This leads to the elegant results of Csiszar and Tusnady [20] and Musicus [22] demonstrating that the EM sequence of (9) is a particular example of an alternating minimization of (8). The iteration sequence becomes $\{q^{(1)}, f^{(1)}, q^{(2)}, f^{(2)}, \cdots\}$, where $q^{(p+1)}$ and $f^{(p+1)}$ are given by

$$q^{(p+1)} = \underset{\{q\}}{\text{denmin}} \left\{ \int_{\chi(y)} q(x) \log \left[\frac{q(x)}{f^{(p)}(x; \phi)} \right] dx \right\} \quad (10a)$$

and

$$f^{(p+1)} = \underset{\{f\}}{\text{denmin}} \left\{ \int_{\chi(y)} q^{(p+1)}(x) \log \left[\frac{q^{(p+1)}(x)}{f(x; \phi)} \right] dx \right\}. \quad (10b)$$

Note, "denmin" is a minimization over the densities q and f, where for the MLE setting f is varied via the parameters ϕ.

II. ML Estimation of Point-Process Intensities for Imaging

A. Imaging Model

The class of imaging problems which we are studying involve the estimation of spatial distributions of radioactivity from the measurement of discrete radioactive emissions [10], [11], [14]. These emissions are modeled as a Poisson process [23] with a spatially dependent intensity $\lambda(z)$. The image reconstruction takes into account two fundamental components characteristic of the imaging systems. i) The number of measurement points are low and therefore dominated by Poisson statistics; and ii) Due to the physics of the measurement systems, errors are introduced in the creation of the observed data. The basis for the ML solution is a model which hypothesizes the existence of two point processes. The first is the "emission process" denoted as $N(dz)$, which corresponds to the number of emission points having positions z in $[z, z + dz) \in Z$ for Z the space over which the radioactive tracer is reconstructed. The emissions $N(dz)$ are a spatial Poisson process with an unknown intensity $\{\lambda(z); z \in Z\}$. The second point process is the "measurement process" denoted as $N(dy)$, corresponding to the number of measured points in $[y, y + dy) \in Y$. The relationship between the radioactive emissions occurring at z and a measurement formed at y is given by

$$y = z + \epsilon \quad (11)$$

where ϵ is a random measurement-error vector with density $p(\cdot)$; since the errors are due to the measurement system, the density $p(\cdot)$ will depend on the imaging modality. For example, in time-of-flight positron-emission tomography, the error density is an elliptically shaped Gaussian function [24]; in electron-microscopic autoradiography it is "Cauchy-like" with long tails [14]; in single-photon tomography it is a symmetric one-dimensional Gaussian density [11].

The error vectors are assumed to be independent of the creation of radioactive events. It follows that the measurement process $N(dy)$ is Poisson with an intensity $\Theta(y)$, resulting from the convolution of the radioactivity distribution with the point-spread function [24], [25] given by

$$\Theta(y) = \int_Z p(y - z)\lambda(z)\, dz, \qquad \text{for } y \in Y. \quad (12)$$

Taking a direct assault on the estimation of λ via ML techniques yields the following log likelihood to be maximized:

$$\log g(y; \lambda) = -\int_Y \int_Z p(y - z)\lambda(z)\, dz dy + \int_Y \log \left[\int_Z p(y - z)\lambda(z)\, dz \right] N(dy). \quad (13)$$

The data vector y denotes the measured incomplete data $N(dy)$. Maximizing with the calculus of variations yields the following nonlinear integral equation which the MLE $\hat{\lambda}$ must satisfy:

$$1 = \int_Y \frac{p(y - z)N(dy)}{\int_Z p(y - r)\hat{\lambda}(r)\, dr}. \quad (14)$$

The integral equation of (14) has not been, to date, explicitly solved.

The fundamental difficulty with a direct maximization of the likelihood of (13) resulting in (14) is that the measurement point process does not uniquely specify the underlying emission point process. Recognizing this results in the introduction of the complete–incomplete data model and its iterative maximization.

B. Complete-Data Model

The complete data are defined as follows. Suppose that each point of the complete data is formed by labeling the emission points with a mark ϵ indicating the error associated with its measurement. The result is a marked point process [25], wherein each event in the complete data (z_j, ϵ_j) identifies the location z_j of the jth emission as well as the error ϵ_j associated with its measurement. The vector function $\boldsymbol{h}(\cdot)$ mapping points from the complete data space χ to the incomplete data space Y is defined by the component maps

$$h(z_j, \epsilon_j) = z_j + \epsilon_j = y_j, \qquad \text{for } j = 1, 2, \cdots, N_T$$

with N_T the total number of measurements. The complete data x are the set of emission points and error vectors

$\{(z_1, \epsilon_1), (z_2, \epsilon_2), \cdots, (z_{N_T}, \epsilon_{N_T})\}$, with the incomplete data y being the set of measurements $\{y_1, y_2, \cdots, y_{N_T}\}$. Note, given the measurement vector y, the emission locations as well as the error vectors are nonuniquely determined. It follows [25] that the log likelihood of the complete data is given by

$$\log f(x; \lambda) = -\int_Z \lambda(z)\, dz + \int_Z \log [\lambda(z)] N(dz) + \sum_{j=1}^{N_T} \log [p(\epsilon_j)]. \tag{15}$$

Having established the log likelihood of the complete data, the joint-maximizer conditions of (8) generate the following necessary conditions for $\hat{\lambda}$ an MLE:

$$\hat{\lambda} \leftarrow \underset{\{\lambda\}}{\operatorname{argmax}} \left\{ -\int_Z \lambda(z)\, dz + \int_Z \log [\lambda(z)] E_{\hat{q}}\{N(dz)\} \right\} \tag{16}$$

where $E_{\hat{q}}\{\cdot\}$ is a conditional expectation with respect to the maxent density of (8b). Evaluating the expectation with respect to $\hat{q}(x) = k(x|x \in \chi(y), \hat{\lambda})$ yields

$$E_{\hat{q}}\{N(dz)\} = \hat{\lambda}(z)\, dz \int_Y \frac{p(y - z) N(dy)}{\int_Z p(y - r) \hat{\lambda}(r)\, dr}$$

and performing the variation of (16) over λ yields the necessary condition for an interior point maximizer given by

$$\int_Y \frac{p(y - z) N(dy)}{\int_Z p(y - r) \hat{\lambda}(r)\, dr} = 1.$$

This is precisely the ML condition of (14) illustrating the fact that maximizing $\log g(y; \lambda)$ in the incomplete-data space is equivalent to maximizing

$$\int_{\chi(x)} \hat{q}(x) \log f(x; \lambda)\, dx$$

for $\hat{q}(x)$ the maxent density of (8b).

Iterative Solution: The maximization sequence is straightforwardly derived. The maximizer $\lambda^{(p+1)}$ at iteration $p + 1$ is generated by taking the conditional expectation of $N(dz)$ in (16) with respect to the density $\hat{q}^{(p)}$ from the pth iteration, and then maximizing with respect to λ to determine λ^{p+1}. This yields the sequence $\{\lambda^{(p)}; p = 1, \cdots\}$ defined via the iteration

$$\lambda^{(p+1)}(z) = \lambda^{(p)}(z) \int_Y \frac{p(y - z) N(dy)}{\int_Z p(y - r)\, \lambda^{(p)}(r)\, dr}, \quad \text{for } z \in Z. \tag{17}$$

The maxent density $q^{(p+1)}$ becomes

$$q^{(p+1)}(x) = \underset{\{q\}}{\operatorname{denmax}} \left\{ -\int_{\chi(y)} q(x) \log \frac{q(x)}{f(x; \lambda^{(p+1)})}\, dx \right\} = k(x|x \in \chi(y), \lambda^{(p+1)}). \tag{18}$$

The maximization of the discrete likelihood via (17) was first derived and implemented for positron-emission tomography by Shepp and Vardi [16], and later by Lange and Carson [26] for transmission tomography and Snyder and Politte [9] for tomography systems with time-of-flight. We have derived similar solutions for single-photon tomography and electron-microscopic autoradiography, with the appropriate imaging models chosen in each [11], [14].

Remark 2: Relationship of the Poisson Process Additive Error Model to Gamma-Ray Astronomy: The Poisson model with intensity given by the convolution of the image with a known point-spread function is precisely the model proposed by both Frieden and Wells [27], [28] and Skilling, Strong, and Bennett [29] for reconstructions in gamma-ray astronomy. For gamma-ray detection, the function $p(\cdot)$ denotes the point spread of the detector array and the function λ corresponds to the two-dimensional intensity map of gamma rays. The basic departure between the imaging model described by Skilling *et al.* [29] and the one described here is, as Skilling points out, the point-spread function representing the operating characteristics of the telescope (Scarsi *et al.*, 1977 [30]) vary as a function of the energy of the detected photons. The single intensity of the measurement process $\Theta(y)$ in (12) is not adequate. The model is modified to accommodate this as follows. Imagine that each measurement point is marked with the value corresponding to its energy, resulting in the multi-energy measurement processes $\{N^1(dy), N^2(dy), \cdots, N^K(dy)\}$, for K the number of energy states. Assuming the measurements $N^k(dy)$ are mutually independent with the kth intensity given by the convolution of $\lambda^k(z)$ with the density $p_k(\cdot)$, then the intensity $\lambda^k(z)$ describes the rate at which photons of energy k are created. The log likelihood becomes

$$\log g(y; \lambda) = -\sum_{k=1}^{K} \int_Y \int_Z p_k(y - z) \lambda^k(z)\, dz dy + \sum_{k=1}^{K} \int_Y \log \left[\int_Z p_k(y - z) \lambda^k(z)\, dz \right] N^k(dy).$$

Assuming, as did Skilling *et al.* [29] that the total rate of production of gamma rays, and not the intensity $\lambda^k(z)$ for each energy state is the desired image, then the sequence becomes

$$\lambda^{(p+1)}(z) = \lambda^{(p)}(z) \sum_{k=1}^{K} \int_Y \frac{p_k(y - z) N^k(dy)}{\int_Z p_k(y - r) \lambda^{(p)}(r)\, dr}. \tag{19}$$

C. Convergence of (17) to the MLE

Vardi, Shepp, and Kaufman [31] proved that a discrete implementation of (17) has global convergence properties; the initial estimate $\lambda^{(0)}(z)$ can be any positive bounded function with the sequence converging to an MLE $\hat{\lambda}(z)$ satisfying the necessary and sufficient maximizer conditions. For purpose of discussion, we outline here their proof. We emphasize that their proof is for a discrete implementation of the image model and iteration sequence. As we will show in the next section, the nondiscretized log-likelihood function of (13) has no maximum when the space of parameters is not restricted, and the iteration will not converge.

The neat proof of Vardi *et al.* breaks into two parts: i) Showing that if the iteration of (17) converges, the Kuhn–Tucker conditions are satisfied and therefore the convergence point of the algorithm maximizes the log likelihood; and ii) Showing that every sequence converges. Proof of i) follows since the log likelihood is concave and Shepp and Vardi [16] showed that the convergence point satisfies the necessary and sufficient Kuhn–Tucker conditions. Proof of ii) is much more subtle, and it is here that Vardi *et al.* invoke the results of Csiszar and Tusnady [20]. They show that for the particular alternating maximization of (17), the K-L divergence between any limit point of the sequence and successive iterates of the algorithm decreases. This coupled with the fact that every sequence has a set of subsequential limit points due to the compactness of the iteration set and the fact that the limit points are stable, implies global converge for the full sequence of iterates.

D. Inconsistency of Unconstrained-Likelihood Estimation of Image Intensities

Estimates of radioactivity distributions derived via the ML method perform better, as measured by signal-to-noise ratio and resolution metrics, than non-likelihood-based imaging methods [14], [32]–[34]. However, images produced via the algorithm of (17) exhibit noise-like artifacts in the form of sharp peaks and valleys located randomly throughout the image field, with these artifacts worsening as the algorithm climbs the "likelihood hill" towards the MLE [10], [35]. This degradation has been observed by other investigators [36], [37], and we argue that the noise artifact is fundamental to any algorithm generating unconstrained likelihood estimates.

To illustrate the fundamental problem, assume that the measurements $N(dz)$ are from a Poisson process with intensity $\lambda(z)$. Then the log likelihood becomes

$$-\int_Z \lambda(z)\, dz + \int_Z \log\, [\lambda(z)] N(dz). \qquad (20)$$

Direct maximization of (20) for the estimation of the image intensity $\lambda(z)$ yields a set of Dirac delta functions, centered at the points of the N observations; that is

$$\hat{\lambda}(z) = \sum_{i=1}^{N} \delta(z - z_i)$$

where the N data points occur at $z_1, z_2, \cdots, z_N$. This solution is obviously unacceptable because it contradicts our *a priori* knowledge about the image. It is expected that $\lambda(z)$ is bounded, and at least piecewise-continuous. An unconstrained maximization of (20) fails to produce meaningful estimates. The iterative algorithm described for the imaging problems will suffer in precisely the same manner, whether measurement errors exist or not. The fundamental problem is that the likelihood is unbounded above over the unconstrained set of measurable functions.

We have described the imaging problem via a continuous model to illustrate precisely these difficulties. We emphasize that the discretization required for performing the implementation does not help matters. Discretizing into pixels of width Δ results in an MLE which is a series of pulses with heights proportional to the number of observations in each pixel; that is

$$\hat{\Lambda}(i) = \int_{i\Delta}^{(i+1)\Delta} N(dz).$$

The unconstrained maximization will result in an extremely rough estimate of the radioactivity distribution. The implementation also has the undesirable property that as it becomes finer with decreasing pixel size, the problem becomes worse; a situation termed "dimensional instability" by Tapia and Thompson [38]. The discrete version of the problem does not remove the fundamental difficulty. For finite data sets, unconstrained maximum likelihood may yield estimates which are unacceptable.

Constrained Estimation Via Penalty Methods: We now discuss the application of the *method of sieves*, first developed by Grenander [39], for the generation of consistent estimates. Grenander notes that in ML problems such as these the parameter space (positive measurable functions of finite measure) is too large. He proposes maximizing the likelihood over a constrained subspace S_m, and then relaxing the constraint with sample size by allowing the subspace to grow. Under the condition that the sieve grows sufficiently slowly this produces consistent estimates. For the conventional tomographic imaging problem, bandwidth constraints are introduced via the choice of reconstruction filters which have high-frequency rolloffs. This leads to a particularly attractive sieve for the ML problem given by the sequence of subsets S_m, for $m = 1, 2, \cdots$, specified by

$$S_m = \left\{ \lambda : \int \left| \frac{d\sqrt{\lambda(z)}}{dz} \right|^2 dz \le m \right\}. \qquad (21)$$

If m grows sufficiently slowly with sample size, the MLE $\hat{\lambda} \in S_m$ is consistent [40].

We implement this sieve using the *penalty method* of Good and Gaskins [41], where the MLE is generated by performing the maximization of the likelihood with a penalty $\Phi(\lambda)$ expressing the smoothness constraint of the sieve on the intensity λ. The penalty constrained MLE is then given by the parameters $\hat{\lambda}$ which maximize the sum of the log likelihood and penalty given by

$$\log g(y; \lambda) - \Phi(\lambda). \qquad (22)$$

As pointed out by Dempster *et al.*, the EM algorithm is perfectly suited for the maximization of the expression in (22). The sequence $\{\lambda^{(p)}; p = 1, 2, \cdots\}$ is given by the following recursion:

$$\lambda^{(p+1)} = \underset{\{\lambda\}}{\operatorname{argmax}} \left\{ \int_{X(y)} q^{(p)}(\boldsymbol{x}) \log f(\boldsymbol{x}; \lambda)\, d\boldsymbol{x} - \Phi(\lambda) \right\} \qquad (23)$$

with $q^{(p)}$ generated in the previous iteration. Maximizing (23) at the $p + 1$st stage results in the sum of (22) increasing at each iteration of the algorithm as well and maximizing (23) is simpler than maximizing the sum in (22) directly.

The sieve constraint of (21) is implemented via the penalty

$$\Phi(\lambda) = \int \left| \frac{d\sqrt{\lambda(z)}}{dz} \right|^2 dz - (2\pi B)^2 \int \lambda(z)\, dz.$$

Note, the constant m determining the size of the sieve simply corresponds to the second term in $\Phi(\lambda)$. At iteration $p + 1$ the quantity to be maximized becomes

$$-\int \lambda(z)\, dz + \int \log\, [\lambda(z)\, h^{(p)}(z)\, dz - \nu \cdot \left[\int \left| \frac{d\sqrt{\lambda(z)}}{dz} \right|^2 dz - (2\pi B)^2 \int \lambda(z)\, dz \right].$$

The constant B corresponds to the second central moment of the energy spectrum of $\sqrt{\lambda}$ and is determined by the choice of m in the sieve; ν corresponds to the scalar Lagrange multiplier weighting the penalty function; and $h^{(p)} = E_{\hat{q}}\{N(dz)\}$ for $\hat{q}$ the maxent density from the pth iteration. Performing the variation over λ yields the following equation to be solved on the p + 1st iteration:

$$\sqrt{\lambda^{(p+1)}(z)} = \int \frac{K(z-u)h^{(p)}(u)}{\sqrt{\lambda^{(p+1)}(u)}}\, du. \qquad (24a)$$

The kernel $K(\cdot)$ is given by

$$K(z) = \frac{1}{2\sqrt{\nu\beta}} \exp\left[-\sqrt{\beta/\nu}\,|z|\right] \qquad (24b)$$

with $\beta = 1 - \nu(2\pi B)^2$. When there are no measurement errors, the MLE becomes a sum of exponential splines with knots at the data points. We have solved the integral equation of (24) recursively, thereby requiring an iterative procedure at each stage of the maximization. (See Snyder and Miller [10] for other details.)

Remark 3: Entropy Penalty Constraint: Various investigators have maximized likelihood subject to an entropy penalty on the image. For example, Frieden [27], Daniell and Gull [3], and Skilling *et al.* [29] have maximized the sum of the log likelihood with a Bayesian prior, where the negative of the penalty function $\Phi(\lambda)$ corresponds to the Bayes prior describing the image statistics. They have suggested using an entropy prior, denoted as $\Phi_e(\lambda)$, which was first proposed by Frieden [27] to describe the image statistics. The entropy prior $\Phi_e(\lambda)$ is proportional to the function

$$-\int \lambda(z) \log \lambda(z)\, dz.$$

They argue that for a large number of counts, the entropy prior is proportional to the number of ways of generating a particular image field. Thus they recommend maximizing $\log g(y; \lambda) + \nu\Phi_e(\lambda)$ with respect to λ, where $\log g(y; \lambda)$ is the log likelihood of the data. The Lagrange multiplier ν is chosen so that the log likelihood and entropy function are weighted various amounts. For example, Frieden chooses $\nu = 1$, thereby maximizing the total sum. Skilling chooses the Lagrange multiplier so that the log likelihood attains a certain value.

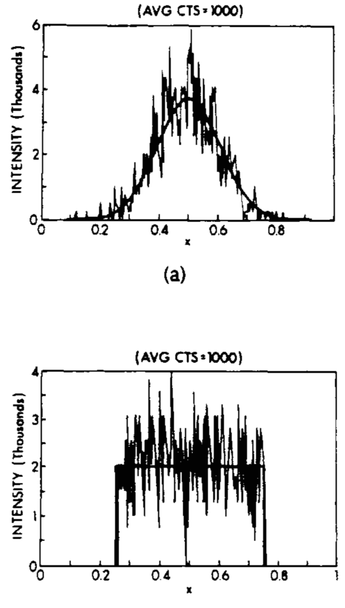
As Frieden and Skilling point out, maximization of the sum of the log likelihood and entropy prior is very difficult (see, for example, Skilling and Bryan [4]). Our results on the penalty-constrained problem pertains, with the functional to be maximized given by

$$-\int_E \lambda(z)\, dz + \int_E \log\left[\lambda(z)\right] h^{(p)}(z)\, dz + \nu\Phi_e(\lambda).$$

The maximization with respect to λ is much simpler because $h^{(p)}(z)$ is not varied. In fact, since the entropy is a strictly concave function, unique maxima will exist as the incomplete data log likelihood of (13) is concave.

E. Simulations Using the Bandwidth Constraint

In this section we describe the result of applying the "bandwidth-penalty" constraint to the derivation of MLEs when there are no measurement errors. These simulations are based on two one-dimensional distributions, one being a smooth Gaussian profile and one a rectangular profile. A Poisson process was generated with a mean in each pixel of $\Lambda(i)$, where $\Lambda(i)$ is the integral over one pixel of the Gaussian and rectangular distributions. Fig. 1 shows the MLEs of the Gaussian (a) and rectangular profiles (b) based on a Poisson simulation containing an average of 1000 counts in the 512-bin simulation. The histogram shown in Fig. 1 results from a direct maximization of the discrete likelihood of (20) and is therefore the unconstrained MLE.



(a)

(b)

Fig. 1. The graphs show the maximum-likelihood estimates of the one-dimensional Poisson process, generated with a Gaussian (a) and rectangular (b) mean rate of discharge. An average of 1000 measurement points were simulated in the 512-bin histogram.

The one-dimensional histograms of Fig. 1 demonstrate the "dimensional instability" that the unconstrained MLEs exhibit. Notice the occurrence of large variations between adjacent pixel estimates of Λ; this effect gets worse if the pixel size or the number of measurement points are decreased.

Plotted in Fig. 2 are the results of applying the bandwidth penalty to the Poisson simulations. The estimates of Fig. 2 were obtained using the bandwidth penalty resulting in the exponential spline estimate of (24). For the simulations there were no measurement errors so that $h(u)$ in (24a) becomes simply a sum of delta Dirac functions centered at the measurement locations. Since the nonlinear equation (24) has no obvious analytic solution, we solved it recursively. Denoting the iterates by $\gamma^{(0)}(x), \gamma^{(1)}(x), \cdots$, then the iteration we define is

$$\lambda^{(k+1)}(x) = \gamma^{(k)}(x) \sum_{i=1}^{N} \frac{K(x-x_i)}{\gamma^{(k)}(x_i)}, \qquad k = 0, 1, 2, \cdots$$

where $K(x)$ is given in (24) and

$$\gamma^{(k+1)}(x) = \sqrt{\lambda^{(k+1)}(x)}.$$

For the initial estimate $\gamma^{(0)}(x)$, we used the square root of the histogram estimate of $\lambda(x)$. While we have no mathematical

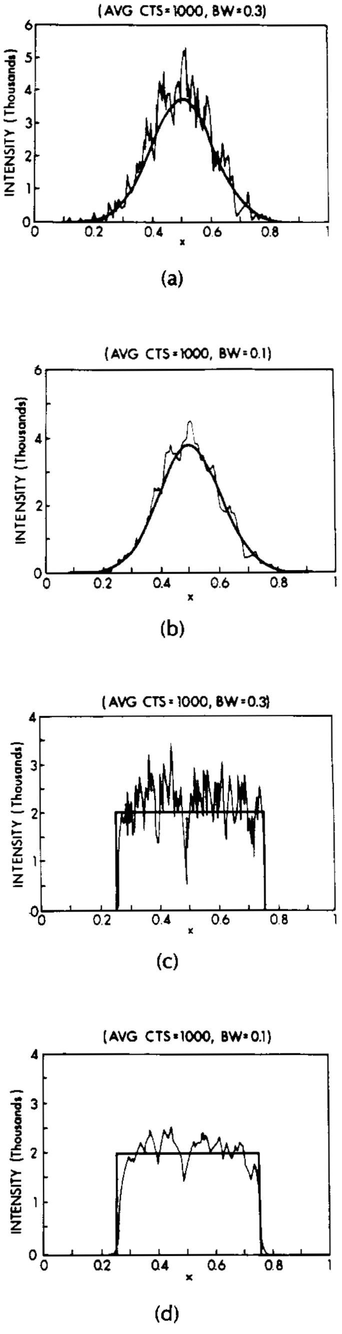


Fig. 2. Plots of the maximum-likelihood estimates generated with the bandwidth penalty of equations (24a, b) for different values of the bandwidth BW. Both rows show the estimates derived from the simulation data of Fig. 1, generated with BW = 0.3 (a), (c) and BW = 0.1 (b), (d).

proof of the convergence of the iterates to the solution of (24), we did observe convergence for all the simulation experiments we attempted. Each panel in Fig. 2 shows the estimates derived with a different bandwidth constraint as specified by the value B corresponding to the second central moment of the energy spectrum. For this simulation, we selected $\nu(2\pi B)^2 = 0.5$ and $B = (BW)(0.5/\Delta x)$, where $\Delta x = 1/512$ is the interval width and BW was either 0.1 or 0.3.

As demonstrated by the results in Fig. 2 as the bandwidth B of the MLE is decreased, the effect is to smooth the variations between adjacent estimates in the histogram.

III. ML Estimation of Toeplitz Constrained Covariances

As first shown by Burg, when given exact values of some finite set of the autocovariances of a stationary series, the maxent spectrum is of the same analytical form as that resulting from an autoregressive model [5], [8], [42]. The maxent density becomes a multivariate Gaussian, constructed by maximizing entropy subject to covariance constraints in the form of (1b). For the covariance estimation problem we now address, samples from the actual series do exist, whereas the autocovariance values themselves are unknown. In order to apply the entropy theory of (1a, b, c), investigators have generated second-order statistics from the series, from which moment constraints in the form of (1b) are assumed [5], [43]. As just one example, a sum of lag products smoothed with a triangular window to reduce the variance at the edge of the data sequence has been used [5]. For this choice of covariance constraints, the maxent method generates a spectrum consistent with the autoregressive model with coefficients determined by the triangularly smoothed lag products.

We instead derive MLEs of the Toeplitz constrained covariances under the Gaussian model. As shown by Burg *et al.* [15], the MLE problem involves the solution of a difficult set of matrix equations, for which no simple closed-form expression has yet been found. The solution we propose is an iterative one in which the measured G-length "incomplete" data series y_G with corresponding Toeplitz covariance K_G is embedded in a larger N-periodic "complete" data series y_N having circulant covariance K_N. By so doing we generate a constrained MLE, where the constraint set is the set of matrices $\mathbf{K}_G$ with the property that $K_G \in \mathbf{K}_G$ has an N-periodic positive-definite extension, and the MLE maximizes likelihood over the constrained set $\mathbf{K}_G$. The constrained maximization over the sieve has the advantage that i) for all observation vectors y_G the MLE exists and the estimator is assured to be nonsingular, and ii) the difficult maximization with respect to K_G is solved via simpler iterative ones over the set of N-periodic extensions. We see i) as an extremely important property as the basic assumption upon which the entire ML procedure is based is the notion that the underlying process is full in the sense that its covariance is nonsingular.

This section proceeds as follows. First we define the sieve of positive semi-definite matrices $\mathbf{K}_G$ having an N-periodic positive semi-definite extension. Then we set up the constrained MLE problem over $\mathbf{K}_G$, derive the complete-data model and iterative algorithm, and discuss the existence of the MLE within the sieve and the convergence properties of the algorithm.

A. Definition of Constraint Set $\mathbf{K}_G$

The sieve set $\mathbf{K}_G$ corresponding to the set of all $G \times G$ Toeplitz matrices K_G having N-periodic positive semi-definite extensions is defined via the set of $N \times N$ circulant positive semi-definite matrices $\mathbf{K}_N$. The set $\mathbf{K}_N$ is given by

$$\mathbf{K}_N = \{K_N : K_N = W^{+}\Sigma W\} \qquad (25a)$$

for Σ a diagonal matrix diag $[\sigma_0, \cdots, \sigma_{N-1}]$ with $\sigma_i \geq 0$ and W the $N \times N$ matrix of normalized orthogonal discrete Fourier transform columns. The notation $[\ ]^\dagger$ denotes Hermitian transpose. Then any $K_G \in \mathbf{K}_G$ is the upper left $G \times G$ matrix of some $K_N \in \mathbf{K}_N$ given by

$$K_G = W_G^\dagger \Sigma W_G \tag{25b}$$

where W_G is the $N \times G$ submatrix made up of the columns of W, and the sieve set $\mathbf{K}_G$ simply becomes

$$\mathbf{K}_G = \{K_G : K_G = W_G^\dagger \Sigma W_G\}. \tag{25c}$$

The constrained MLE of K_G is found by maximizing the likelihood over all $K_G \in \mathbf{K}_G$.

B. MLE of Toeplitz Covariance $K_G \in \mathbf{K}_G$

We now find the MLE of $K_G \in \mathbf{K}_G$ by embedding the measured $\mathbf{y}_G$ in an N-periodic process with covariance $K_N \in \mathbf{K}_N$. We do this in two steps. First we assume that $\mathbf{y}_G = \mathbf{y}_N$; that is, the observations are a full period of the process. Given the MLE for that problem, it becomes clear how to set up the algorithm for the problem in which the data $\mathbf{y}_G$ are of length $G < N$, the period of the process.

1) MLE of K_N Given $\mathbf{y}_N$: Given the series $\{y_0, y_1, \cdots, y_{N-1}\}$ of length N from a stationary, zero-mean Gaussian N-periodic process for which $y_0 = y_N, y_1 = y_{N+1}, \cdots$, we want to find the MLE of $K_N \in \mathbf{K}_N$. The Gaussian density describing $\mathbf{y}_N = [y_0 \cdots y_{N-1}]^t$ becomes

$$f(\mathbf{y}_N; K_N) = (2\pi)^{-N/2} \det{}^{-1/2} K_N \exp\left(-\frac{1}{2}\mathbf{y}_N^\dagger K_N^{-1} \mathbf{y}_N\right) \tag{26}$$

where det denotes matrix determinant. The necessary condition for the MLE is given by the following trace expression (Burg *et al.*) [15]:

$$\operatorname{tr}\left[(K_N^{-1}\mathbf{y}_N\mathbf{y}_N^\dagger K_N^{-1} - K_N^{-1})\,\delta K_N\right] = 0 \tag{27}$$

for all allowable variations of $K_N \in \mathbf{K}_N$. We now derive an explicit form for the MLE $\hat{K}_N$ in terms of the lag products which are the sufficient statistics for the circulant positive-definite covariances. We also show below that the likelihood is strictly concave, implying that the MLE is unique and the trace condition of (27) is a sufficient condition.

Using the Fourier transform matrix W from the orthogonal decomposition of K_N yields the data in the rotated coordinates $\mathbf{c}_N = W\mathbf{y}_N$ with $\mathbf{c}_N = [c_0 \cdots c_{N-1}]^t$. Viewing the problem in the rotated data converts it into one of estimating the eigenvalues σ_m corresponding to the spectral power at discrete frequencies $[2\pi(m)]/N$, for $m = 0, \cdots, N-1$. Rewriting the density in the rotated coordinates, taking natural logarithm, and discarding terms which are not a function of the parameters yields the following expression to be maximized with respect to Σ:

$$-\log\det\Sigma - \mathbf{c}_N^\dagger\Sigma^{-1}\mathbf{c}_N. \tag{28}$$

With Σ diagonal, the MLE of the spectral coefficients becomes

$$\hat{\sigma}_m = |c_m|^2, \qquad \text{for } m = 0, \cdots, N-1 \tag{29}$$

and $\hat{\Sigma} = \operatorname{diag}[|c_0|^2, \cdots, |c_{N-1}|^2]$. From (27), the MLE $\hat{K}_N(k, \ell)$, for $k, \ell = 1, \cdots, N$ becomes

$$\hat{K}_N(k, \ell) = \frac{1}{N}\sum_{m=0}^{N-1}\hat{\sigma}_m \exp\left(-j2\frac{\pi}{N}(\ell - k)m\right). \tag{30}$$

Note in (30) that the covariance is Toeplitz. Substituting (29) into (30) and using the fact that $c_m = \mathbf{w}_m^t\mathbf{y}_N$ for $\mathbf{w}_m$ the mth column of W yields the following satisfying relation for the MLE:

$$\hat{K}_N(k, \ell) = \frac{1}{N}\sum_{m=0}^{N-1} y(m)y^*(\langle m + \ell - k\rangle_N). \tag{31}$$

The notation $\langle\ \rangle_N$ denotes modulo N and y^* denotes complex conjugate. For the Gaussian stationary N-periodic process, the ML estimates of the covariances are just a linear sum of lag products. In Appendix I we show that the estimates of (31) satisfy the trace conditions of (27).

Strict concavity: Demonstrating strict concavity of the likelihood and therefore uniqueness of the MLE of (31) is straightforward. Rewriting the complete-data Gaussian likelihood of (28) in the rotated coordinates using $V = \Sigma^{-1}$ where

$$V = \operatorname{diag}\left[\frac{1}{\sigma_0}, \cdots, \frac{1}{\sigma_{N-1}}\right]$$

yields the following function to be maximized:

$$\log f(\mathbf{c}_N; V) = -\frac{N}{2}\log 2\pi + \frac{1}{2}\sum_{i=0}^{N-1}\log\nu_i - \frac{1}{2}\sum_{i=0}^{N-1}\nu_i|c_i|^2.$$

Demonstrating a unique maximizer with respect to K_N amounts to showing strict concavity with respect to V. The second variation of the log likelihood becomes

$$\begin{aligned}\frac{\partial^2\log f(\mathbf{y}_N; V)}{\partial\nu_j\,\partial\nu_k} &= \frac{1}{2\nu_j^2}, \qquad j = k\\ &= 0, \qquad j \neq k.\end{aligned}$$

Since the Hessian matrix of second derivatives is negative-definite, the log likelihood $\log f(\mathbf{y}_N; V)$ is strictly concave. Now we generate the MLE of $K_G \in \mathbf{K}_G$ given $G < N$ pieces of data.

2) MLE of Toeplitz $K_G \in \mathbf{K}_G$ Given $\mathbf{y}_G$: Given the series $\{y_0, y_1, \cdots, y_{G-1}\}$ from a stationary, zero-mean Gaussian process of period $N > G$ the likelihood of $\mathbf{y}_G$ becomes

$$g(\mathbf{y}_G; K_G) = (2\pi)^{-G/2}\det{}^{-1/2}K_G\exp\left(\frac{-1}{2}\mathbf{y}_G^\dagger\mathbf{K}_G^{-1}\mathbf{y}_G\right). \tag{32}$$

The necessary interior point condition for the MLE is as follows:

$$\operatorname{tr}\left[(K_G^{-1}\mathbf{y}_G\mathbf{y}_G^\dagger K_G^{-1} - K_G^{-1})\delta K_G\right] = 0 \tag{33}$$

for all variations of $K_G \in \mathbf{K}_G$. The fundamental difference between the trace condition of (33) and that in (27) is that for $\mathbf{y}_G \neq \mathbf{y}_N$ all of the lag products of the full period are not available. Recognizing this results in the following complete-data model and iteration sequence.

Complete-data model: The proper choice for the *complete data* becomes the N-dimensional vector $\mathbf{y}_N$ consisting of the given $\mathbf{y}_G$ augmented by the $N - G$-dimensional vector $\mathbf{y}_A = [y_G \cdots y_{N-1}]^t$, with the *incomplete data* the observed vector $\mathbf{y}_G$. The many-to-one function $\mathbf{h}$ mapping $\mathbf{y}_N$ to $\mathbf{y}_G$ ignores all points corresponding to the augmented vector of length $N - G$; that is, $\mathbf{y}_G = \mathbf{h}(\mathbf{y}_N)$. The complete-data likelihood is given by (26), and transforming into the rotated coordinates yields the following function to be maximized:

$$-\sum_{m=0}^{N-1}\log\sigma_m - \sum_{m=0}^{N-1}\frac{|c_m|^2}{\sigma_m}.$$

From the complete-data model we apply the joint-entropy maximization of (8) to derive the MLE conditions. Performing the variation over the eigenvalues σ_m, and taking the expectation with respect to the maxent density specified by (8b) yields

$$\hat{\sigma}_m = E\{|c_m|^2|\hat{\Sigma}, y_G\}, \qquad \text{for } m = 0, \cdots, N-1 \tag{34}$$

where $\hat{\Sigma} = \text{diag}\,[\hat{\sigma}_0, \cdots, \hat{\sigma}_{N-1}]$. In the original coordinates this becomes

$$\hat{K}_N(k, \ell) = \frac{1}{N}\sum_{m=0}^{N-1} E\{y(m)\, y^*(\langle m + \ell - k\rangle_N)|\, y_G, \hat{K}_N\} \tag{35}$$

for $k, \ell = 1, \cdots, N$. Below we demonstrate that $\hat{K}_G \in \boldsymbol{K}_G$ the first $G \times G$ submatrix of $\hat{K}_N$ satisfying the lag-product condition of (35) satisfies the trace condition of (33). This illustrates what we find to be most exciting about posing the estimation problem via the joint maximization of (8). The relatively simple lag-product conditions of (35) resulting from a direct maximization of the expectation of the complete data log likelihood has significant intuitive appeal to the trace condition of (33) derived by maximizing the incomplete data log likelihood directly.

Iterative algorithm: The iteration is straightforwardly derived. The maximizer $K_N^{(p+1)}$ at iteration $p+1$ is obtained by evaluating the conditional expectation of (34) with respect to the eigenvalues from the previous iteration, yielding the following estimates:

$$\sigma_m^{(p+1)} = E\{|c_m|^2|\Sigma^{(p)}, y_G\}, \qquad \text{for } m = 0, 1, \cdots, N-1 \tag{36a}$$

with

$$\Sigma^{(p+1)} = \text{diag}\,[\sigma_0^{(p+1)}, \cdots, \sigma_{N-1}^{(p+1)}].$$

The covariances at iteration $p+1$ are obtained by simply transforming back to the original coordinates

$$K_N^{(p+1)}(k, \ell) = \frac{1}{N}\sum_{m=0}^{N-1} E\{y(m)\, y^*(\langle m + \ell + k\rangle_N)|\, y_G, K_N^{(p)}\}. \tag{36b}$$

For the problem in which the observations y_G are embedded in an N-periodic process the constrained ML procedure over the sieve set $\boldsymbol{K}_G$ requires the augmentation of the lag products via generation of conditional mean and mean-square estimates of the missing lags. At the convergence point, the covariance estimates are those values which equal the sum of the conditional mean of the lag products.

The iteration defined by (36) is generated using the standard formulas for the conditional mean and variance of a Gaussian process (see Rhodes [44], for example). We do this by noting from (36a) that the eigenvalue matrix $\Sigma^{(p+1)}$ is simply made up of the diagonal elements of the conditional correlation matrix $E\{\mathbf{c}_N\mathbf{c}_N^\dagger|\Sigma^{(p)}, y_G\}$, from which (36b) is obtained via the discrete Fourier transform operation. Denoting the cross covariance of $\mathbf{c}_N$ and y_G as K_{cy} the conditional mean matrix $E\{\mathbf{c}_N\mathbf{c}_N^\dagger|\Sigma, y_G\}$ becomes

$$E\{\mathbf{c}_N\mathbf{c}_N^\dagger|\Sigma, y_G\} = K_{cy}K_G^{-1}y_G y_G^\dagger K_G^{-\dagger}K_{cy}^\dagger + \Sigma - K_{cy}K_G^{-1}K_{cy}^\dagger. \tag{37}$$

Using the fact that $y_G = W_G^\dagger \mathbf{c}_N$ implies $K_{cy} = \Sigma W_G$ from which the conditional mean on iteration $p+1$ becomes

$$E\{\mathbf{c}_N\mathbf{c}_N^\dagger|\Sigma^{(p)}, y_G\} = \Sigma^{(p)}W_G K_G^{(p)-1}y_G y_G^\dagger K_G^{(p)-\dagger}W_G^\dagger\Sigma^{(p)\dagger} + \Sigma^{(p)} - \Sigma^{(p)}W_G K_G^{(p)-1}W_G^\dagger\Sigma^{(p)\dagger}. \tag{38}$$

The estimate $\Sigma^{(p+1)}$ becomes the diagonal entries of the conditional correlation matrix $E\{\mathbf{c}_N\mathbf{c}_N^\dagger|\Sigma^{(p)}, y_G\}$ of (38).

Stable points satisfying the necessary maximizer conditions: The stable points of (38) are shown to satisfy the necessary maximizer conditions of (33) as follows. If $K_G^{(\ell)}$ is a stable point then $\Sigma^{(\ell+1)} = \Sigma^{(\ell)}$. Defining $D[A]$ to be the diagonal matrix with entries defined by the diagonal elements of an arbitrary square matrix A, then since $\Sigma^{(\ell)}$ is stable we have

$$D[E\{\mathbf{c}_N\mathbf{c}_N^\dagger|\Sigma^{(\ell)}, y_G\} - \Sigma^{(\ell)}] = 0$$

with 0 the null matrix. From (38) it follows that

$$D[\Sigma^{(\ell)}W_G K_G^{(\ell)-1}y_G y_G^\dagger K_G^{(\ell)-\dagger}W_G^\dagger\Sigma^{(\ell)\dagger} - \Sigma^{(\ell)}W_G K_G^{(\ell)-1}W_G^\dagger\Sigma^{(\ell)\dagger}] = 0.$$

Pre and post multiplying by $\delta\Sigma\Sigma^{(\ell)-1}$ and $\Sigma^{(\ell)-\dagger}$, respectively, yields

$$D[\delta\Sigma W_G(K_G^{(\ell)-1}y_G y_G^\dagger K_G^{(\ell)-\dagger}W_G^\dagger - K_G^{(\ell)-1}W_G^\dagger)] = 0$$

for all $\delta\Sigma$. Taking the trace and rearranging terms yields

$$\text{tr}\,[(K_G^{(\ell)-1}y_G y_G^\dagger K_G^{(\ell)-\dagger} - K_G^{(\ell)-1})W_G^\dagger\delta\Sigma W_G] = 0.$$

Since $K_G \in \boldsymbol{K}_G$ it follows that $\delta K_G = W_G^\dagger\delta\Sigma W_G$. Therefore, we have proven that for all $\delta\Sigma$ the above trace is zero, implying the trace is zero for all variations δK_G in the class of feasible ones, and therefore the stable points of (36) and (38) satisfy the trace condition of (33).

C. Existence and Convergence of the MLE Over Constraint Set $\boldsymbol{K}_G$

The iterative procedure of (36) and (38) maximizes the likelihood $g(y_G; K_G)$ over all covariances $K_G \in \boldsymbol{K}_G$. Clearly, we have constrained the ML problem by introducing the maximization over the constrained set $\boldsymbol{K}_G$. We may expect that the MLE produced over the unconstrained set of Toeplitz matrices would be different than the constrained MLE generated via (36) and (38). The constraint on $\boldsymbol{K}_G$ has been introduced for the following reasons. For the Toeplitz covariance estimation problem only one observation vector is available. However, it has been shown (Fuhrmann and Miller [45]) that given a single vector observation y_G there is a nonzero probability that the MLE over the unconstrained set of Toeplitz matrices will not exist, that is, the likelihood may have no maximizer over the set of positive-definite Toeplitz matrices and any algorithm maximizing likelihood will generate a singular estimator. The necessary and sufficient condition from [45] for the failure of the MLE procedure is that there exists a singular matrix in the set of Toeplitz matrices which has the observation vector y_G in its range. The argument of [45] is restated as follows. Assume K_G^s is singular and in the feasible region of Toeplitz matrices and construct matrix $K_G^s + \epsilon I$ also Toeplitz with $\epsilon > 0$. Then, y_G in the range of K_G^s implies

$$y_G = \sum_{i=1}^{M} \alpha_i \gamma_i$$

for γ_i the nonzero eigenvectors of K_G^s with $M < G$. Evaluating the log likelihood yields

$$\log g(y_G; K_G^s + \epsilon I) = -\log \det (K_G^s + \epsilon I) - \sum_{i=1}^{M} \frac{\alpha_i^2}{\sigma_i + \epsilon}$$

with σ_i corresponding to the nonzero eigenvectors of K_G^s. It is clear, that as $\epsilon \rightarrow 0$ so that $K_G^s + \epsilon I$ converges to singularity the likelihood increases without bound. The necessary part of the argument follows from the fact that if there is no K_G^s singular with y_G in its range, then the log likelihood converges to minus infinity for any sequence converging to singularity. Thus the MLE is guaranteed to be nonsingular if there exists no K_G^s in the feasible region with y_G in its range.

The difficulty of maximizing the likelihood over the unconstrained set of Toeplitz matrices is analogous to the one we faced in the imaging problem in which the set of functions containing λ were too large. By restricting the maximization to the constraint set the estimator does exist. This is precisely what we recommend here, in which the smaller set $\boldsymbol{K}_G$ over which the maximization is performed forces the existence of a well-defined nonsingular positive-definite Toeplitz MLE. The existence of the MLE over $\boldsymbol{K}_G$ follows (see Fuhrmann and Miller [45] for a more general description) simply from the fact that the probability that a data vector y_G is in the range space of any singular $K_G^s \in \boldsymbol{K}_G$ is zero. This is because y_G being in the range of a singular $K_G^s \in \boldsymbol{K}_G$ implies that it lies in some $M \leq G - 1$-dimensional space spanned by the M discrete Fourier transform columns of W_G. However, since the actual covariance of the process is assumed nonsingular, the Gaussian measure of the M-dimensional subspace spanned by M columns of W_G is zero. Therefore, with probability one there exists no data vector y_G in the range of a singular $K_G^s \in \boldsymbol{K}_G$ and the MLE over $\boldsymbol{K}_G$ is well defined.

Remark 4: Relationship to the Full Rank Outer Product Condition of Burg et al. The fact that with probability one there exists no singular $K_G^s \in \boldsymbol{K}_G$ having y_G in its range implies that covariance sequences converging to any singular $K_G^s \in \boldsymbol{K}_G$ have log likelihoods which converge to $-\infty$. This follows from precisely the argument of Burg *et al.* [15] where they demonstrated that under the assumption that the outer products of multiple independent data vectors were full-rank implied that algorithms with increasing likelihood will not converge to singular estimates in the constraint region. The essence of the proof is in Appendix II as part of Theorem 1. We emphasize, that for the Toeplitz problem it seems unnatural to us to introduce multiple independent vector copies into the problem since a single observation vector supports longer lags. By performing the maximization over the set $\boldsymbol{K}_G$ with circulant positive-definite extension we have assured the existence of the MLE without assuming multiple vector copies.

Convergence to the MLE: Proving convergence of (36) and (38) to the set of MLEs involves three parts. i) Showing that a maximizer of the likelihood is bounded over $\boldsymbol{K}_G$ and attains its bound in the set. ii) Showing that all limit points of the sequences of (36) and (38) are stable which implies the set of limit points satisfy the necessary maximizer conditions. iii) Showing that every sequence of (36) and (38) converges. We have proven i) and ii) in Appendix II. Our proof of ii) basically follows that of Cover [46]. We sketch briefly here the ideas in the proof of the Appendix.

Part i) relies on the fact that over the constrained set the Gaussian density has value zero for all singular matrices $K_G \in \boldsymbol{K}_G$. This then assures that algorithms with increasing likelihood such as (36) and (38) will not converge to singularity, the iteration sequence is contained in a compact set in $\boldsymbol{K}_G$, and that the log likelihood is upper semicontinuous and finite from above over the compact set. Then it follows that the likelihood is both bounded above and attains its maximum.

Part ii) is proven by showing that the algorithm produces a monotonic sequence of likelihoods, the map defined by (38) is continuous over the set of positive-definite iterates, and the set of limit points of the sequences of the algorithm are all stable and therefore satisfy the maximizer conditions.

For part iii) we must show that the full sequence converges. This still remains an open research question.

D. Simulations Comparing MLE and Conventional Lag Product Estimators

Now we show simulation results from two different experiments in which an N-periodic Gaussian process was simulated with period $N = 32$. We have examined the performance of the iterative ML estimator of (35) and (36) by choosing $G = 16$ pieces of the full period of the process and generating estimates of the first 15 covariances. For the performance comparison we have made we have generated four different estimators. These four are as follows:

1) *N-Lag ML Estimator:* The first is the lag-product estimator given the complete period of the N-length process, which is just the MLE given the complete data as demonstrated in Section III-B1. It is given by (31) as

$$\hat{K}(\tau) = \frac{1}{N} \sum_{i=0}^{N-1} y(i)\, y^*(\langle i + \tau \rangle_N).$$

2) *G-Lag ML Estimator:* This is the ML estimator given the incomplete data corresponding to the first $G < N$ pieces of the process, and is given by the convergence point of the iterative algorithm of (35) and (36). For the simulations we show, we have found 10 iterations of the algorithm to be sufficient for convergence.

3) *G-Lag Biased Estimator:* This estimator is the conventional biased estimator given by

$$K_b(\tau) = \frac{1}{G} \sum_{i=0}^{G-1-\tau} y(i)\, y^*(i + \tau).$$

4) *G-Lag Unbiased Estimator:* This estimator is the conventional unbiased estimator given by

$$K_u(\tau) = \frac{1}{G - \tau} \sum_{i=0}^{G-1-\tau} y(i)\, y^*(i + \tau).$$

In Fig. 3 we show the covariance and spectra from both experiments in which the stationary zero-mean Gaussian processes were generated. The left column, Fig. 3(a), shows the covariance for both processes; the right (Fig. 3(b)) shows the spectra given by the DFT of the covariances. We have chosen these two covariances as they illustrate two extreme examples. The first process has a relatively broad low-frequency spectrum, with the second having a sharp high-frequency spectrum.

With the specified covariances shown in Fig. 3, a total of 300 independent realizations of the stationary periodic pro-

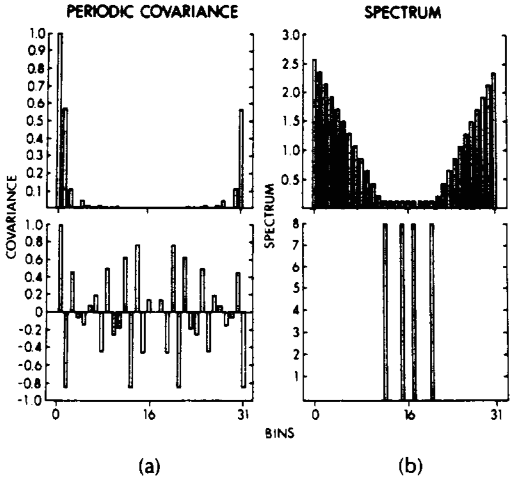


Fig. 3. Plots of the covariance (a) and spectra (b) of the zero-mean Gaussian N-periodic processes used in the simulations. The period of the process was $N = 32$.

cesses were generated. From each realization, the four different estimators of the covariance sequence were generated, and the sample bias and standard deviation were calculated. Shown in Figs. 4 and 5 are the four estimators

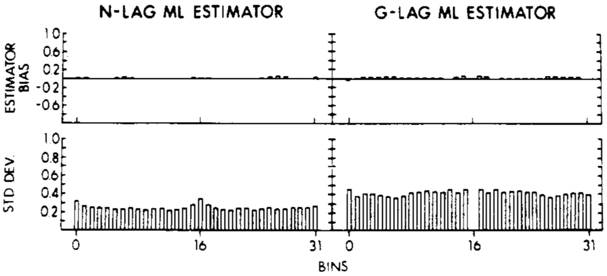


Fig. 4. Plots of the bias (top row) and standard deviation (bottom row) of the N-lag MLE (left column) and the G-lag MLE (right column) given by the iteration of (36). Sample statistics were generated from 300 realizations of the process with covariance and spectra given in the top row of Fig. 3.

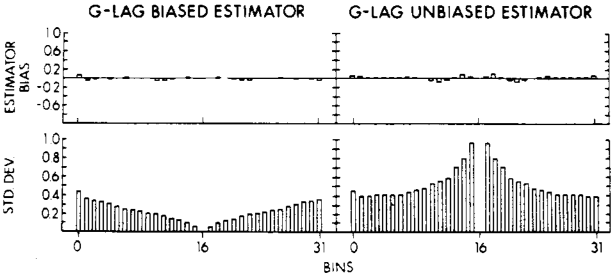


Fig. 5. Plots of the bias (top row) and standard deviation (bottom row) of the G-lag biased (left column) and the G-lag unbiased (right column) estimators. Sample statistics were generated from 300 realizations of the process with covariance and spectra given in the top row of Fig. 3.

generated from the low-frequency process having covariance given in the top row of Fig. 3. In the left column of Fig. 4 is the N-lag MLE. Plotted in the right column is the G-lag MLE. Plotted in Fig. 5 are the G-lag biased and G-lag unbiased estimators.

The results of Figs. 4 and 5 show that for this process, the estimators have extremely small biases. We should expect this, as near the edge of the data collection window the true covariances are virtually zero so that the actual value of the biases are extremely small. The major result illustrated by this experiment is that the G-lag unbiased estimator has a variance which grows exponentially at the edge of the data collection window. The G-lag MLE shows a standard deviation which is roughly constant and only slightly larger than the N-lag MLE.

Plotted in Figs. 6 and 7 are the results from the second experiment, in which we simulated the high-frequency process with covariances given in the bottom row of Fig. 3. The

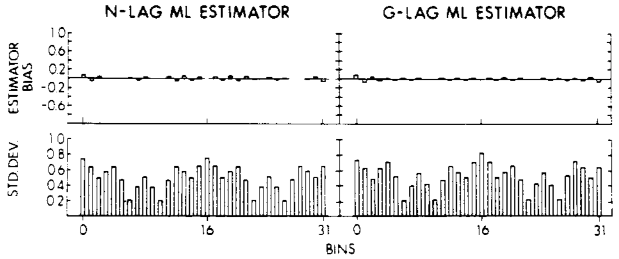


Fig. 6. Plots of the bias (top row) and standard deviation (bottom row) of the N-lag MLE (left column) and the G-lag MLE given by the iteration of (36). Sample statistics were generated from 300 realizations of the process with covariance and spectra given in the bottom row of Fig. 3.

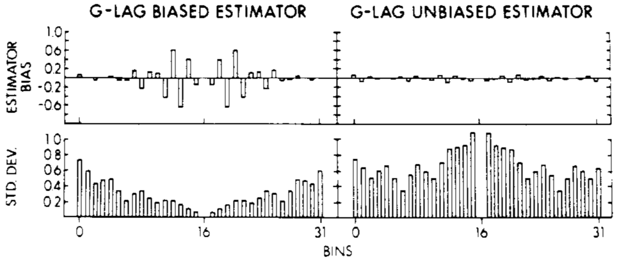


Fig. 7. Plots of the bias (top row) and standard deviation (bottom row) of the G-lag biased (left column) and the G-lag unbiased (right column) estimators. Sample statistics were generated from 300 realizations of the process with covariance and spectra given in the bottom row of Fig. 3.

left column of Fig. 6 shows that the N-lag MLE is unbiased. The G-lag MLE is only slightly more biased for the longer lags where there is little data. Comparing the actual size of the bias of the G-lag MLE to the value of the underlying covariances we see that the bias is a small percentage of the true value. For this process, the standard deviation is virtually identical for both estimation algorithms.

These results change sharply for the biased and unbiased estimators. As seen in the left column of Fig. 7, the G-lag biased estimator has a large amount of bias, far more than the G-lag MLE. The G-lag unbiased estimator, although slightly less biased than the G-lag MLE, has a standard deviation which is much larger than the MLE.

These results are summarized in the plot of Fig. 8 showing the mean-squared error (MSE) of the 17 unique covariance lags $K(\tau)$, $0 \leq \tau \leq 16$. The MSE was generated by summing

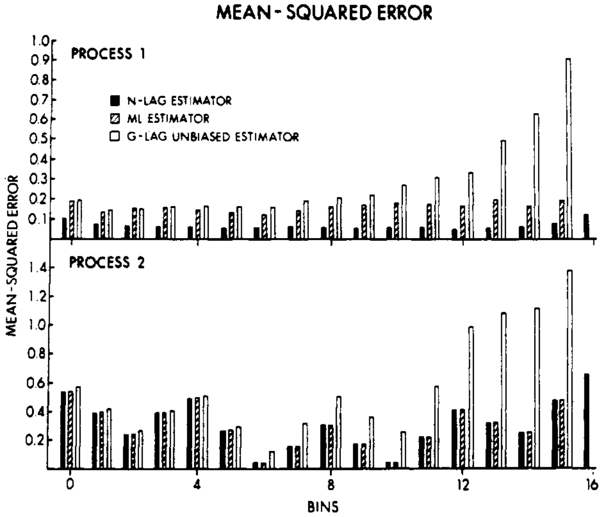


Fig. 8. Plots of the mean-squared error of the N-lag MLE (solid bars), G-lag MLE (dashed bars), and G-lag unbiased (open bars) estimators.

the square of the bias and the variance of the estimators in each bin. Plotted in the figure are the N-lag MLE (solid bars), the G-lag MLE (dashed bars), and the G-lag unbiased estimator (open bars). Since the estimators based on $G = 16$ data points cannot estimate $K(16)$, only the MSE for the N-lag estimator is plotted for lag 16 as seen by the last solid single bar.

The plot derived from process 1 (top row) shows that the MSE of the unbiased estimator grows towards the edge of the data window, and the N-lag estimator has approximately one half the MSE of the G-lag MLE. This results from the fact that process 1 is "relatively white" thereby implying that each piece of data carries independent information. Since the N-lag MLE is based on twice the number of lags as the G-lag MLE ($N = 32$, $G = 16$) the variance and resulting MSE is approximately one half for the N-lag MLE.

For process 2 (bottom row), we see the major result of the new maximum-likelihood method. Here, the spectrum is fairly concentrated so that the missing data can be fairly well estimated via the conditional expectations, resulting in the fact that the MSE of the G, N-lag MLEs are identical for the lags which the data support (up to $\tau = 15$ for $G = 16$). We conclude from these two extremely simple preliminary simulations what we find to be most encouraging about the ML method for estimating Toeplitz covariances. In an attempt to decrease the bias of the estimator the MLE fills in the missing lags with conditional mean estimates. The resulting decrease in bias would imply an increase in spectral resolution. This increase in spectral resolution does not seem to accompany a large increase in variance as exemplified by the mean-squared error plots.

Conclusions

We have shown that when given finite observations $\boldsymbol{h}(\boldsymbol{x})$ of a random process $\boldsymbol{x}$ with density $f(\boldsymbol{x}; \phi)$, the set of maximum entropy densities are identical to those generated via rules of conditional probability. By viewing the measurements as specifying the domain over which the maxent density $q(\boldsymbol{x})$ is defined, rather than as a moment constraint on the observation function $\boldsymbol{h}$, the maxent density closest to the prior in the cross-entropy sense is just the conditional density $k(\boldsymbol{x}|\boldsymbol{x} \in \chi(\boldsymbol{y}), \phi)$. Because of this identity, maximum-likelihood parameter estimates may be obtained by solving a joint maximization (minimization) of the entropy function (K-L divergence). This reduces to finding the parameters ϕ which maximize the expected value of the log of the prior $E\{\log f(\boldsymbol{x}; \phi)\}$, where the expectation is taken with respect to the maxent density.

It is precisely this view which results in the application of the iterative methods of Dempster *et al.* and Csiszar and Tusnady for the generation of ML estimates in tomography and gamma-ray astronomy. By performing constrained maximum-likelihood estimation subject to the bandwidth constraints, the ML problem becomes well-posed. The iterative maximization is ideally suited for generating constrained MLEs, and we have demonstrated that the bandwidth constraint results in exponential spline smoothing which yields much improved estimates.

This constrained ML approach has also been adopted for the Toeplitz covariance estimation problem in which we find the MLE within the set of positive-definite Toeplitz matrices with periodic extensions. By imposing the periodic extension constraint on the maximization it insures the existence of a nonsingular estimator. We demonstrated that when given the complete data of a full period of the process the MLE of the Toeplitz constrained covariance involves the sum of the usual lag-product statistics. For the problem in which the observed data is of length $G < N$ the assumed period of the process, conditional mean estimates of the lag products which are outside of the data collection window must be generated, from which the Toeplitz covariances are obtained. We also show via simulations that estimates generated using the ML procedure have excellent bias and variance properties when compared with conventional lag-product estimators.

Appendices

I. The MLE of (31) Satisfies the Trace Condition of (27) Given y_N

Given the complete data y_N from a stationary zero-mean Gaussian series of period N, we show the MLE of (30)

$$\hat{K}_N(\tau) = \frac{1}{N}\sum_{m=0}^{N-1} y(m)\, y^*(\langle m + \tau\rangle_N)$$

satisfies the necessary trace conditions of (27) for an interior point maximizer.

Rewriting the Gaussian log likelihood of the complete data of (25) via the matrix trace yields

$$\log f(y_N; K_N) = -\frac{1}{2}\operatorname{tr}\,[K_N^{-1} y_N y_N^{\dagger}] - \log\left[\int \exp\left(-\frac{1}{2}\operatorname{tr}\,[K_N^{-1} y_N y_N^{\dagger}]\right) m(dy_N)\right] \tag{A1}$$

where $m(dy_N)$ is the N-dimensional Lebesgue measure. Denoting the sufficient statistics as $S = y_N y_N^{\dagger}$, and using the fact that $\delta K_N^{-1} = -K_N^{-1}\delta K_N K_N^{-1}$ yields the following variation equation:

$$\delta \log f(y_N; K_N) = -\frac{1}{2}(\operatorname{tr}\,[(K_N^{-1} S K_N^{-1} - K_N^{-1})\delta K_N]).$$

The necessary condition is the gradient is orthogonal to all variations δK_N in the class of feasible variations

$$\text{tr}\,[(K_N^{-1}SK_N^{-1} - K_N^{-1})\delta K_N] = 0. \qquad \text{(A2)}$$

Rewriting K_N^{-1} of (25) as $W^{-1}\Sigma^{-1}W^{-\dagger}$ yields

$$K_N^{-1}SK_N^{-1} = W^{-1}\Sigma^{-1}W^{-\dagger}SW^{-1}\Sigma^{-1}W^{-\dagger}.$$

Substituting this into (A2) for the first term, and noticing that $\delta K = W^{\dagger}\delta\Sigma W$ yields the following expression for the variation of the log likelihood:

$$\delta \log f(y_N; K_N) = \text{tr}\,[(\Sigma^{-1}W^{-\dagger}SW^{-1}\Sigma^{-1} - \Sigma^{-1})\delta\Sigma].$$

Using the fact that $\delta\Sigma^{-1} = \Sigma^{-1}\delta\Sigma\Sigma^{-1}$, and rearranging terms yields the following equation:

$$\delta \log f(y_N; K_N) = \text{tr}\,[(W^{-\dagger}SW^{-1} - \Sigma)\delta\Sigma^{-1}].$$

Substituting the estimates of (29) into Σ, and performing the trace operation results in the variation being zero as we had set out to prove.

II. Convergence of the Algorithm of (36) and (38) to the Set of Limit Points Satisfying the Necessary Maximizer Conditions

In this Appendix we prove that given the observation vector y_G the set of limit points of the iteration of (36) and (38) are all stable and satisfy the necessary maximizer conditions.

Definition: Define the set $\boldsymbol{K}_G^b$ as the set of all positive-definite Toeplitz matrices K_G of dimension G with entries bounded by some b with positive semidefinite extensions $K_N \in \boldsymbol{K}_N$, and $\overline{\boldsymbol{K}}_G^b$ its closure the set of all nonnegative definite bounded G-dimensional Toeplitz matrices.

Define the sequence of covariance estimates $K_G^{(p)}$ via the iteration of (36) mapping $K_G \to M(K_G)$ such that each step $K_G^{(p+1)} = M(K_G^{(p)})$ is given by the following equation for $k, \ell = 1, \cdots, G$:

$$K_G^{(p+1)}(k, \ell) = \frac{1}{N}\sum_{m=0}^{N-1} E\{y(m)\,y^*(\langle m + \ell - k\rangle_N)|y_G, K_G^{(p)}\}. \qquad \text{(A3)}$$

The full N-periodic covariance K_N is given by (A3) for $k, \ell = 1, \cdots, N$. The log likelihood $L(K_G) = \log g(y_G; K_G)$ is defined for $K_G \in \boldsymbol{K}_G^b$ to be

$$L(K_G) = -\frac{G}{2}\log(2\pi) - \frac{1}{2}\log\det K_G - \frac{1}{2}y_G^{\dagger}K_G^{-1}y_G \qquad \text{(A4a)}$$

and for singular $K_G^s \in \overline{\boldsymbol{K}}_G^b$ given by

$$L(K_G^s) = -\infty, \qquad \text{for } K_G^s \in \overline{\boldsymbol{K}}_G^b. \qquad \text{(A4b)}$$

Then with

$$L^{\max} = \max_{\{K_G : K_G \in \overline{\boldsymbol{K}}_G^b\}} L(K_G)$$

we will show that

1) $L(K_G)$ is both bounded above and attains its max $L^{\max}$ over the set $\overline{\boldsymbol{K}}_G^b$; and,
2) $L(K_G^{(p)})$, for $K_G^{(p)}$ defined via (A3), is a monotonic nondecreasing sequence; and,
3) all limit points of the sequence $K_G^{(p)}$ are stable and satisfy the necessary maximizer conditions of (33).

We first show that for all $K_G \in \boldsymbol{K}_G^b$ the log likelihood $L(K_G)$ is continuous. Over its closure $\overline{\boldsymbol{K}}_G^b$ it is upper semicontinuous. It is therefore bounded and has a max in $\overline{\boldsymbol{K}}_G^b$.

Theorem 1: Given the set $\overline{\boldsymbol{K}}_G^b$ defined above, then

a) $\overline{\boldsymbol{K}}_G^b$ is compact in G^2-Euclidean norm.
b) $L(K_G)$ given by (A4) is continuous over the interior of $\boldsymbol{K}_G^b$ and upper semicontinuous over its closure $\overline{\boldsymbol{K}}_G^b$.
c) $L^{\max} = \max\limits_{\{K_G \in \overline{\boldsymbol{K}}_G^b\}} L(K_G) < \infty.$

Proof: Clearly a) follows because $\overline{\boldsymbol{K}}_G^b$ is a set of covariances closed and bounded in G^2-dimensional Euclidean space. To prove upper semicontinuity we note that $L(K_G)$ is continuous over the entire interior $\boldsymbol{K}_G^b$; therefore, we must examine its behavior at the boundary, in particular those covariances $K_G^s \in \overline{\boldsymbol{K}}_G^b$ which are singular. We must show that for all sequences $\{K_G^{(p)}\} \subset \boldsymbol{K}_G^b$ converging to $K_G^s \in \overline{\boldsymbol{K}}_G^b$ singular that

$$\lim_{\{K_G^{(p)} \to K_G^s\}} L(K_G^{(p)}) = -\infty.$$

We do this by constructing a particular sequence $K_G^{(p')} \to K_G^s$ with $L(K_G^{(p')}) \to -\infty$. Then by the continuity of L over $\boldsymbol{K}_G^b$ it follows that all sequences converging to the singular covariance have likelihood which converge to minus infinity. We proceed by choosing $K_G^{(p')}$ as $K_G^s + \epsilon I$ with $\epsilon > 0$. Clearly $K_G^s + \epsilon I \in \boldsymbol{K}_G^b$ and can, therefore, be written as

$$K_G^s + \epsilon I = \sum_{i=1}^{M} (\sigma_i + \epsilon)\gamma_i\gamma_i^{\dagger} + \sum_{M+1}^{G} \epsilon\beta_i\beta_i^{\dagger}$$

where there are $M \leq G - 1$ eigenvectors γ_i of K_G^s corresponding to the nonzero eigenvalues σ_i, and $G - M$ eigenvectors β_i spanning the $G - M$-dimensional orthogonal complement of the γ_i's. Writing y_G via the eigenvectors as

$$y_G = \sum_{i=1}^{M} \alpha_i\gamma_i + \sum_{i=M+1}^{G} \alpha_i\beta_i$$

then the log likelihood becomes

$$L(K_G^s + \epsilon I) = \sum_{i=1}^{M}\left(-\log[\sigma_i + \epsilon] + \frac{\alpha_i^2}{\sigma_i + \epsilon}\right) + \sum_{i=M+1}^{G}\left(-\log[\epsilon] + \frac{\alpha_i^2}{\epsilon}\right).$$

Since $K_G^s \in \overline{\boldsymbol{K}}_G^b$ it can be written via (25c) as $K_G^s = W_G^{\dagger}\Sigma W_G$. Any G columns of W_G are linearly independent and therefore K_G^s lies in the space spanned by M of those columns which is, of course, the same space as is spanned by the set $\{\gamma_1, \cdots, \gamma_M\}$. Since the underlying covariance of the process is nonsingular, the data vector y_G will with probability one have a component in the orthogonal complement of the M-dimensional span of $\{\gamma_1, \cdots, \gamma_M\}$, implying $\alpha_i \neq 0$ for some $i \geq M + 1$. Therefore, with probability one

$$\lim_{\epsilon \to 0} L(K_G^s + \epsilon I) = -\infty.$$

It follows by continuity of $L(K_G)$ that for all sequences in $\boldsymbol{K}_G^b$, converging to K_G^s the log likelihood converges to minus infinity, and upper semicontinuity is proved. Part c) follows from the fact that every upper semicontinuous function over a compact set is bounded and achieves its maximum.

Next we show that each step of the algorithm yields an improved log likelihood.

Theorem 2: For the iterates defined in (A3) and the log likelihood of (A4)

$$L(K_G^{(p+1)}) - L(K_G^{(p)}) \geq Q(K_N^{(p+1)}|K_N^{(p)}) - Q(K_N^{(p)}|K_N^{(p)}) \geq 0$$

for $Q(K_N|K_N^{(p)})$ as defined via the iteration of (11a) given by

$$Q(K_N|K_N^{(p)}) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\det K_N - \frac{1}{2}E\{y_N^{\dagger}K_N^{-1}y_N|y_G, K_N^{(p)}\}.$$

Proof: This follows directly from the fact that the iteration is an instance of an EM algorithm as proved in Section III-B, and the properties of that sequence discussed in Section I-D.

Corollary: $L(K_G^{(p+1)}) = L(K_G^{(p)})$ if and only if $K_N^{(p+1)} = K_N^{(p)}$.

Proof: By definition of K_G as the first $G \times G$ submatrix of K_N, $K_N^{(p+1)} = K_N^{(p)} \rightarrow K_G^{(p+1)} = K_G^{(p)}$. Now if $L(K_G^{(p+1)}) = L(K_G^{(p)})$, then $Q(K_N^{(p+1)}|K_N^{(p)}) = Q(K_N^{(p)}|K_N^{(p)})$ by Theorem 2. From the strict concavity of the complete data log likelihood shown in Section III-B1, $K_N^{(p+1)}$ is the unique global maximizer of $Q(K_N|K_N^{(p)})$, implying that $K_N^{(p+1)} = K_N^{(p)}$.

Now we show that all the limit points are stable.

Lemma 1: The iterates $\{K_G^{(p)}; p = 1, 2, \cdots\}$ are contained in $\boldsymbol{K}_G^b$ the interior of the bounded set of Toeplitz covariance matrices $\boldsymbol{K}_G^b$, and do not converge to some singular $K_G^s \in \bar{\boldsymbol{K}}_G^b$.

Proof: Clearly each iterate is Toeplitz, positive-definite and by the proof of Burg *et al.* [15] each entry is bounded. Suppose, however, that $K_G^{(p)} \rightarrow K_G^s$ for K_G^s singular. Then by the argument in Theorem 1, the log likelihood $L(K_G^{(p)}) \rightarrow -\infty$. This is impossible because we assumed that $L(K_G^{(0)}) > -\infty$, and by Theorem 2 the likelihood monotonically increases.

Lemma 2: The set of limit points A_{K_G} of the sequence $\{K_G^{(p)}; p = 0, 1, \cdots\}$ is nonempty.

Proof: By Lemma 1, the sequence is in the compact set $\bar{\boldsymbol{K}}_G^b$, and by the compactness of $\bar{\boldsymbol{K}}_G^b$ has a convergent subsequence.

Lemma 3: The map $M(K_G)$ is a continuous function for all $K_G \in \boldsymbol{K}_G^b$.

Proof: Continuity of the map over $\boldsymbol{K}_G^b$ follows directly from (38).

Lemma 4: All limit points $K_G^{(\ell)} \in A_{K_G}$ are stable; i.e., $M(K_G^{(\ell)}) = K_G^{(\ell)}$.

Proof: Define $\Delta(K_G^{(p)}) = L(K_G^{(p+1)}) - L(K_G^{(p)})$. By Theorem 2, $L(K_G^{(p)})$ is monotonically nondecreasing and from Theorem 1 has an upper bound $L^{\max}$, implying $\Delta(K_G^{(p)}) \rightarrow 0$.

Let $K_G^{(\ell)} \in A_{K_G}$ be any limit point of the sequence $\{K_G^{(p)}; p = 0, 1, \cdots\}$ defined by (A3), and let $K_G^{(p_j)}$ be a subsequence converging to $K_G^{(\ell)}$. Since $K_G^{(p_j)} \rightarrow K_G^{(\ell)}$ and by the continuity of the log likelihood, $L(K_G^{(p_j)}) \rightarrow L(K_G^{(\ell)})$. From Lemma 2, $M(K_G)$ is continuous with respect to K_G so that $K_G^{(p_j)} \rightarrow K_G^{(\ell)}$ implies

$$L(M(K_G^{(p_j)})) \rightarrow L(M(K_G^{(\ell)})).$$

Therefore, the convergence of the subsequence $K_G^{(p_j)}$ implies $\Delta(K_G^{(p_j)}) \rightarrow \Delta(K_G^{(\ell)})$. But $\Delta(K_G^{(p_j)}) \rightarrow 0$ by the boundedness of the likelihood, implying that $\Delta(K_G^{(\ell)}) = 0$.

From Corollary 1, $\Delta(K_G^{(\ell)}) = 0$ implies $M(K_N^{(\ell)}) = K_N^{(\ell)}$ and thus all the limit points are stable and therefore satisfy the necessary maximizer conditions.

Acknowledgment

Discussions with Dr. E. T. Jaynes have been extremely helpful in connecting the maximum-entropy techniques to maximum-likelihood estimation and are greatly appreciated. Various important suggestions were made by Dr. D. Fuhrmann and Dr. E. Chornoboy which helped focus the work on the Toeplitz covariance estimation algorithm, and M. Turmon contributed to the implementation results.

References

[1] A. D. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. B39, pp. 1-37, 1977.

[2] S. F. Gull and G. J. Daniell, "Image reconstruction from incomplete and noisy data," *Nature*, vol. 272, p. 686, 1978.

[3] S. F. Gull and J. Skilling, "The maximum entropy method," in *Indirect Imaging*, J. A. Roberts, Ed. London, UK: Cambridge Univ. Press, 1984.

[4] J. Skilling and R. K. Bryan, "Maximum entropy image reconstruction," *MNRAS*, submitted 1982.

[5] J. P. Burg, "Maximum entropy spectral analysis," Univ. Microfilms no. 75-25, Stanford Univ., Stanford, CA, 1975.

[6] J. E. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 2, pp. 230–237, Apr. 1981.

[7] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, pp. 620–630, 1957.

[8] ——, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, pp. 939–952, Sept. 1982.

[9] D. L. Snyder and D. G. Politte, "Image reconstruction from list-mode data in an emission tomography system having time-of-flight measurements," *IEEE Trans. Nucl. Sci.*, vol. NS-30, pp. 1843–1849, 1983.

[10] D. L. Snyder and M. I. Miller, "The use of sieves to stabilize images produced with the EM algorithm for emission tomography," *IEEE Trans. Nucl. Sci.*, vol. NS-32, pp. 3864–3872, Oct. 1985.

[11] M. I. Miller, D. L. Snyder, and T. Miller, "Maximum likelihood reconstruction for single photon emission computed tomography," *IEEE Trans. Nucl. Sci.*, vol. NS-32, no. 1, pp. 769–778, Feb. 1985.

[12] M. I. Miller, "Algorithms for removing recovery related distortion from auditory-nerve discharge patterns," *J. Acoust. Soc. Amer.*, vol. 77, pp. 1452–1464, 1985.

[13] M. I. Miller, N. Karamanos, and W. E. Bosch, "EM algorithms for estimating parameters from single-memory Markov point-processes having a multiplicative intensity," presented at the 23rd Annual Allerton Conf., Univ. of Illinois, Urbana, IL, Oct. 1985.

[14] M. I. Miller, K. B. Larson, J. E. Saffitz, D. L. Snyder, and L. J. Thomas, Jr., "Maximum-likelihood estimation applied to electron-microscope autoradiography," *J. Electron Microscopy Tech.*, 1985.

[15] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices," *Proc. IEEE*, vol. 70, no. 9, pp. 963–974, Sept. 1982.

[16] L. A. Shepp and Y. Vardi, "Maximum-likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imag.*, vol. MI-1, pp. 113–121, 1982.

[17] S. Kullback, in *Information Theory and Statistics.* New York, NY: Wiley, Dover, 1959, 1968.

[18] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *Ann. Prob.*, vol. 3, pp. 146–158, 1975.

[19] J. M. Van Campenhout and T. M. Cover, "Maximum entropy and conditional probability," *IEEE Trans. Informat. Theory*, vol. IT-27, no. 4, pp. 483–489, July 1981.

[20] I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedures," *Statistics and Decisions* (Supplement Issue no. 1). Munchen, West Germany: R. Oldenbourg Verlag, 1984, pp. 205–237.

[21] E. T. Jaynes, in *Papers on Probability, Statistics and Statistical*

Physics, R. D. Rosenkrantz, Ed. Dordrecht: The Netherlands/ Boston: USA/London, England: D. Reidel, 1983.

[22] B. R. Musicus, "Iterative algorithms for optimal signal reconstruction and parameter identification given noisy and incomplete data," MIT thesis, Cambridge, MA, 1982.

[23] R. D. Evans, in *The Atomic Nucleus*. New York, NY: McGraw-Hill, 1955.

[24] D. L. Snyder, L. J. Thomas, Jr., and M. M. Ter-Pogossian, "A mathematical model for positron emission tomography systems having time-of-flight measurements," *IEEE Trans. Nucl. Sci.*, vol. NS-28, pp. 3575-3583, 1981.

[25] D. L. Snyder, *Random Point Processes*. New York, NY: Wiley, 1975.

[26] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comput. Assisted Tomography*, vol. 8, no. 2. New York, NY: Raven Press, Apr. 1984, pp. 306-316.

[27] B. R. Frieden, "Restoring with maximum likelihood and maximum entropy," *J. Opt. Soc. Amer.*, vol. 62, no. 4, pp. 511-518, 1972.

[28] B. R. Frieden and D. C. Wells, "Restoring with maximum entropy. III. Poisson sources and backgrounds," *J. Opt. Soc. Amer.*, vol. 68, no. 1, pp. 93-103, 1978.

[29] J. Skilling, A. W. Strong, and K. Bennett, "Maximum-entropy image processing in gamma-ray astronomy," *Mon. Not. R. Astr. Soc.*, vol. 187, pp. 145-152, 1979.

[30] L. Scarsi, in *Proc. 12th ESLAB Symp.*, vol. ESA-124, p. 3, 1977.

[31] Y. Vardi, L. A. Shepp, and L. Kaufman, "A statistical model for positron emission tomography," *J. Amer. Statist. Assoc.*, vol. 80, pp. 8-35, Mar. 1985.

[32] D. Politte, "Reconstruction algorithms for time-of-flight assisted positron-emission tomographs," M.S. thesis, Sever Institute of Technology, Washington Univ., St. Louis, MO, Dec. 1983 (supervised by D. L. Snyder).

[33] M. Mintun, J. Gorman, and D. L. Snyder, "Evaluation of the maximum-likelihood method for reconstruction of images in positron emission tomography," in *Proc. Soc. Nuclear Medicine 32nd Annual Meeting* (Houston, TX, June 1985).

[34] J. M. Ollinger and D. L. Snyder, "An evaluation of an improved method for computing histograms in dynamic tracer studies using positron-emission tomography," *IEEE Trans. Nucl. Sci.*, vol. NS-33, pp. 435-438, Feb. 1986.

[35] M. I. Miller, D. L. Snyder, and S. M. Moore, "An evaluation of the use of sieves for producing estimates of radioactivity distributions with the EM algorithm for PET," *IEEE Trans. Nucl. Sci.*, vol. NS-33, pp. 492-495, Feb. 1986.

[36] L. A. Shepp, Y. Vardi, J. B. Ra, S. K. Hilal, and Z. H. Cho, "Maximum-likelihood with real data," *IEEE Trans. Nucl. Sci.*, vol. NS-31, pp. 910-913, 1984.

[37] R. Carson, personal communication, 1985.

[38] R. A. Tapia and J. R. Thompson, *Nonparametric Probability Density Estimation*. Baltimore, MD: Johns Hopkins Univ. Press, 1978.

[39] U. Grenander, *Abstract Inference*. New York, NY: Wiley, 1981.

[40] S. Geman, "Sieves for nonparametric estimation of densities and regressions," *Repts. Pattern Anal.*, vol. 99, D.A.M. Brown Univ., 1981.

[41] I. J. Good and R. A. Gaskins, "Nonparametric roughness penalties for probability densities," *Biometrika*, vol. 58, no. 2, pp. 255-277, 1971.

[42] A. van den Bos, *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 493-494, 1971.

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

[44] I. B. Rhodes, "A tutorial introduction to estimation and filtering," *IEEE Trans. Automat. Contr.*, vol. AC-16, no. 6, pp. 688-706, 1971.

[45] D. R. Fuhrmann and M. I. Miller, *On the Singularity of Maximum Likelihood Estimates of Structured Covariance Matrices*, Monograph of Electronic Systems and Signals Research Lab., Washington Univ., St. Louis, MO, 1986.

[46] T. M. Cover, "An algorithm for maximizing expected log investment return," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 369-373, Mar. 1984.

Michael I. Miller received the B.S. degree in electrical engineering from the S.U.N.Y. at Stony Brook, NY, in 1976, the M.S. degree in electrical engineering from the Johns Hopkins University, Baltimore, MD, in 1978, and the Ph.D. degree in biomedical engineering from the Johns Hopkins University in 1983.

Since 1984 he has been on the Electrical Engineering and Institute for Biomedical Computing Faculty of Washington University in St. Louis, MO, where he is currently an Associate Professor. His research interests include speech coding in the central nervous system, image processing, and digital signal processing. Most recently, his interests have been in the development of iterative algorithms on parallel computers for tomography and spectrum estimation in direction of arrival array processing. He is a recipient of the Presidential Young Investigator Award.

Donald L. Snyder (Fellow, IEEE) received the B.S. degree in electrical engineering from the University of Southern California, Los Angeles, in 1961 and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1963 and 1966, respectively.

From 1966 to 1969, he was on the faculty of the Massachusetts Institute of Technology. Since 1969, he has been on the Electrical Engineering Faculty of Washington University in St. Louis, MO. He served as Chairman of the Department of Electrical Engineering from 1976 to 1986 and as Associate Director of the Biomedical Computer Laboratory, Washington University School of Medicine. He is presently Director of the Electronic Systems and Signals Research Laboratory in the Department of Electrical Engineering. He is the author of papers on the theories of random processes, estimation, decision, and systems and the application of these theories to practical problems arising in communications, and to radar and biomedical imaging. Most recently, his interest has been in the development and application of random point process models in optical communication and radiology and in estimation-theoretic approaches to radar imaging. He is the author of the textbook *Random Point Processes* (New York, NY: Wiley, 1975), which develops point-process models with emphasis on applications.

Dr. Snyder served as Associate Editor for Random Processes for the IEEE TRANSACTIONS ON INFORMATION THEORY and was the 1981 President of the IEEE Information Theory Group.