Maximum-Likelihood Estimation of Complex Sinusoids and Toeplitz Covariances

Michael J. Turmon and Michael I. Miller

***Abstract*—In an extension of previous methods for maximum-likelihood (ML) Toeplitz covariance estimation, new iterative algorithms for computing joint ML estimates of complex sinusoids in unknown stationary Gaussian noise are proposed. The number of sinusoids is assumed known, but their frequencies and amplitudes are not. The iterative algorithm, an adaptation of the expectation-maximization (EM) technique, proceeds from an initial estimate of the mean and Toeplitz covariance, and iterates between estimating the mean given the current covariance and vice versa, with likelihood increasing at each step. The resulting ML covariance estimates are compared to conventional estimators and Cramér-Rao bounds. An analysis of the Kay and Marple data set is also presented. The effectiveness of the new algorithm for estimating means in unknown noise is investigated, and the usefulness of simultaneously estimating the covariance and the mean is demonstrated.**

***Keywords*— spectrum estimation, covariance estimation, Toeplitz, circulant, maximum-likelihood.**

I. Introduction

In the problems we study here wide-sense stationary noise processes having a superimposed time-varying mean are observed over finite time intervals, and the goal is to estimate the frequency components of the mean as well as the spectral density of the stationary process. For this a non-zero mean complex Gaussian model is assumed, with the mean known to be a superposition of complex sinusoids having unknown deterministic frequencies and amplitudes. The additive noise is assumed stationary implying the covariance matrix of the random observation vectors is Toeplitz, i.e. constant along the diagonals. We adopt the method of constrained maximum likelihood: Given one or more vector samples from the normal density, find the maximum-likelihood estimate of the mean and covariance matrix, subect to the constraints that the mean is a superposition of a known number of sinusoids and the covariance is Toeplitz.

Our interest in the joint estimation of sinusoids and Toeplitz covariances is motivated via several problems in which we are currently involved. Estimating sinusoids in noise is the core of the nuclear magnetic resonance spectroscopy problem, for which we have already described iterative maximum-likelihood algorithms under the assumption that the additive noise covariance is known [2; 14]. However, no such method exists for the joint estimation problem, in which the noise covariance is unknown.

We are also involved in Toeplitz covariance estimation [5; 13; 15; 18; 20], our original motivation being for delay-Doppler radar imaging of diffuse targets in which the target reflectance at each point is modelled as a sample of a Gaussian process with a spectrum determined by the power at each Doppler frequency in the return signal. The imaging problem for such diffuse targets is essentially a Toeplitz covariance or spectrum estimation problem. The importance of a model also incorporating a deterministic mean component has been indicated by Haykin [9] in which specular and diffuse reflectors may be present in the environment resulting in a need for estimation of the complex sinusoids and the noise. Adding such glint or specular components to the radar imaging model immediately requires solution of the joint estimation problem.

The new contributions of this paper are threefold. Extending previous results [11; 15; 19], a new maximum-likelihood method for solving the joint mean and covariance estimation problem described above is presented. Secondly, the performance of this algorithm for estimating covariances when the mean is known, i.e. pure spectrum estimation, is explored via comparisons to the Cramér-Rao bound and to other estimators, in particular those analyzed by Kay and Marple [10]. (In this case the new algorithm for joint estimation reduces to the one of [15], but significant studies of this algorithm's performance have not to date been presented.) Finally the performance of the new algorithm for estimating means when the covariance is unknown is explored, and the benefits of simultaneously estimating the mean and covariance are demonstrated.

A. Problem Statement

Let $\mathcal{Y}_\mathcal{G} = \{\mathbf{y}_\mathrm{G}^{(1)}, \mathbf{y}_\mathrm{G}^{(2)}, \ldots, \mathbf{y}_\mathrm{G}^{(P)}\}$ be P independent $G \times 1$ random vectors from a multivariate complex Gaussian distribution of the "circular" class considered in [7]. The Gaussian has unknown Toeplitz covariance $\mathbf{K}_\mathrm{G}$ and mean $\boldsymbol{\mu}_\mathbf{G} = \sum_{i=1}^{M} m(i)\mathbf{w}(-\omega_i)$, a superposition of a known number M of sinusoids with unknown frequencies and amplitudes where $\mathbf{w}(\omega) = (1/\sqrt{N})[1\ e^{-j\omega}\ \ldots\ e^{-j\omega(G-1)}]^T$ (the constant N will be defined shortly). The log-likelihood is

$$L_i(\boldsymbol{\mu}_\mathbf{G}, \mathbf{K}_\mathrm{G}) = -P \log |\mathbf{K}_\mathrm{G}| - \sum_{r=1}^{P} (\mathbf{y}_\mathrm{G}^{(r)} - \boldsymbol{\mu}_\mathbf{G})^\dagger \mathbf{K}_\mathrm{G}^{-1} (\mathbf{y}_\mathrm{G}^{(r)} - \boldsymbol{\mu}_\mathbf{G}) \quad (1)$$

Michael Turmon was supported by NSF grant EID-8710984, a Creativity Award in Engineering, while at Washington University. He is currently in the Department of Electrical Engineering at Cornell University, Ithaca, NY 14853.

Michael Miller was supported by grants NSF PYI ECE-8552518, ARO P-29349-SDI, and ONR 59922, and is with the Electronic Signals and Systems Research Laboratory of the Department of Electrical Engineering at Washington University, St. Louis, MO 63130.

and the constrained MLE's are given by

$$(\hat{\boldsymbol{\mu}}_{\rm G},\, \hat{\mathbf{K}}_{\rm G}) = \underset{\substack{\boldsymbol{\mu}_{\rm G}\in\mathcal{U}_{\mathcal{G}} \\ \mathbf{K}_{\rm G}\in\mathcal{K}_{\mathcal{G}}}}{\arg\max}\; L_i(\boldsymbol{\mu}_{\mathbf{G}},\, \mathbf{K}_{\rm G})\;, \tag{2}$$

with constraint space

$$\mathcal{U}_{\mathcal{G}} := \Big\{\, \boldsymbol{\mu}_{\mathbf{G}} \in \mathbb{C}^G \;\Big|\; \boldsymbol{\mu}_{\mathbf{G}} = \sum_{i=1}^{M} m(i)\mathbf{w}(-\omega_i) \;\Big\} \tag{3}$$

where $\mathbb{C}$ is the field of complex numbers and $\mathcal{K}_{\mathcal{G}}$ is the set of $G\times G$ Hermitian positive-definite Toeplitz matrices. Taking the variation of the objective function yields the coupled equations which the maximizer must satisfy:

$$\begin{aligned}\operatorname{tr}\Bigg[\Bigg(\mathbf{K}_{\rm G}^{-1}\sum_{r=1}^{P}(\mathbf{y}_{\rm G}^{(r)}-\boldsymbol{\mu}_{\mathbf{G}})(\mathbf{y}_{\rm G}^{(r)}-\boldsymbol{\mu}_{\mathbf{G}})^{\dagger}\mathbf{K}_{\rm G}^{-1} \\ - P\mathbf{K}_{\rm G}^{-1}\Bigg)\delta\mathbf{K}_{\rm G}\Bigg] &= 0 \quad (4)\end{aligned}$$

$$\operatorname{Re}\left((\bar{\mathbf{y}}_{\rm G}-\boldsymbol{\mu}_{\mathbf{G}})^{\dagger}\mathbf{K}_{\rm G}^{-1}\,\delta\boldsymbol{\mu}_{\mathbf{G}}\right) = 0\,, \tag{5}$$

where $\bar{\mathbf{y}}_{\rm G}$ is the sample mean. For $(\boldsymbol{\mu}_{\mathbf{G}},\, \mathbf{K}_{\rm G})$ to solve (2), both conditions above must hold for all feasible variations $\delta\mathbf{K}_{\rm G}$ of $\mathbf{K}_{\rm G}$ and $\delta\boldsymbol{\mu}_{\mathbf{G}}$ of $\boldsymbol{\mu}_{\mathbf{G}}$. If $\boldsymbol{\mu}_{\mathbf{G}}$ is taken to be zero, (4) reduces to the trace condition of Burg et al. [1].

B. Constrained Solution

In subsequent sections we explore ways of solving the maximum-likelihood equations (4, 5). We extend our previous work [13; 15; 23] to include the case where not only must the Toeplitz covariance be estimated, but also the complex amplitudes and frequencies of the sinusoids comprising the mean. To avoid singularity of the Toeplitz-constrained ML estimate of (2), we constrain the solution along lines proposed previously by forcing the covariance to be not only Toeplitz, but also to have an embedding in a larger $N\times N$ circulant matrix, ensuring the existence of a positive-definite MLE [5]. To generate the ML estimates we will use numerical techniques based on the expectation-maximization algorithm of Dempster et al. [4].

If $\mathbf{K}_{\rm G}$ has a circulant embedding to length N, it must be of the form

$$\begin{aligned}\mathbf{K}_{\rm G} &= \sum_{i=0}^{N-1}\sigma(i)\,\mathbf{w}(2\pi i/N)\,\mathbf{w}(2\pi i/N)^{\dagger} \\ &= \mathbf{W}_{\rm G}^{\dagger}\boldsymbol{\Sigma}\mathbf{W}_{\rm G}\;,\end{aligned} \tag{6}$$

where $\mathbf{W}_{\rm G} = [\mathbf{w}(0)\,\mathbf{w}(2\pi/N)\cdots\mathbf{w}(2\pi(N-1)/N)]^T$ is $N\times G$ and $\boldsymbol{\Sigma} = \operatorname{diag}[\sigma(0)\cdots\sigma(N-1)]$. We denote as $\mathcal{K}_{\mathcal{G}}^N$ the set of $G\times G$ positive-definite Toeplitz matrices that can be generated by (6). $\mathcal{K}_{\mathcal{G}}^N$ is a strict subset of $\mathcal{K}_{\mathcal{G}}$, but as N increases, elements of $\mathcal{K}_{\mathcal{G}}^N$ are better able to approximate covariances in $\mathcal{K}_{\mathcal{G}}$. (The computational burden of increasing N is low if the FFT is used to perform multiplication by $\mathbf{W}_{\rm G}$ [6]. The circulant approximation to Toeplitz covariances is further discussed in [3; 12; 16].)

The mean $\boldsymbol{\mu}_{\mathbf{G}}$ of $\mathbf{y}_{\rm G}^{(r)}$ is similarly constrained according to

$$\boldsymbol{\mu}_{\mathbf{G}} = \sum_{i=1}^{M} m(k_i)\mathbf{w}(-2\pi k_i/N) \quad,\quad k_i\in\{0,\ldots,N-1\}\,, \tag{7}$$

with N chosen large enough to give the desired frequency resolution. The set of $G\times 1$ vectors that can be generated by such a superposition will be denoted $\mathcal{U}_{\mathcal{G}}^N$. It is convenient to write $\boldsymbol{\mu}_{\mathbf{G}} = \mathbf{W}_{\rm G}^{\dagger}\mathbf{m}_{\rm N}$ where $\mathbf{m}_{\rm N}\in\mathcal{M}_{\mathcal{N}}$, the set of complex N-vectors having at most M nonzero entries.

In summary, the constraint sets are formally defined as

$$\begin{aligned}\mathcal{U}_{\mathcal{G}}^N &:= \Big\{\, \boldsymbol{\mu}_{\mathbf{G}}\in\mathbb{C}^G \;\Big|\; \\ &\qquad \boldsymbol{\mu}_{\mathbf{G}} = \sum_{i=1}^{M} m(k_i)\mathbf{w}(\frac{2\pi k_i}{N})\;,\; 0\le k_i < N\,\Big\} \quad (8a)\\ \mathcal{K}_{\mathcal{G}}^N &:= \Big\{\, \mathbf{K}_{\rm G}\in\mathbb{C}^{G\times G} \;\Big|\; \\ &\qquad \mathbf{K}_{\rm G} = \sum_{i=0}^{N-1}\sigma(i)\,\mathbf{w}(\frac{2\pi i}{N})\mathbf{w}(\frac{2\pi i}{N})^{\dagger} > 0\,\Big\}, \quad (8b)\end{aligned}$$

where the notation $\mathbf{K}_{\rm G} > 0$ means $\mathbf{K}_{\rm G}$ is positive-definite. The constrained maximum-likelihood estimation problem is to choose $\boldsymbol{\mu}_{\mathbf{G}}$, $\mathbf{K}_{\rm G}$ maximizing the log-likelihood (1) subject to $\boldsymbol{\mu}_{\mathbf{G}}\in\mathcal{U}_{\mathcal{G}}^N$, $\mathbf{K}_{\rm G}\in\mathcal{K}_{\mathcal{G}}^N$.

Remark: These expressions for the mean and covariance have a physical interpretation apart from being representations of the process parameters. If, instead of the truncated DFT bases $\mathbf{w}(2\pi i/N)\,(0\le i<N)$, the entire N-length vectors are used to construct a larger $N\times N$ covariance matrix from $\boldsymbol{\Sigma}$, we have $\mathbf{K}_{\rm N} = \mathbf{W}_{\rm N}^{\dagger}\boldsymbol{\Sigma}\mathbf{W}_{\rm N}$, where $\mathbf{K}_{\rm G}$ appears in the upper left $G\times G$ corner of $\mathbf{K}_{\rm N}$. An augmented mean $\boldsymbol{\mu}_{\mathbf{N}} = \mathbf{W}_{\rm N}^{\dagger}\mathbf{m}_{\rm N}$ can be similarly constructed. Each $G\times 1$ sample $\mathbf{y}_{\rm G}^{(r)}$ can thus be viewed as the initial segment of an incompletely observed N-periodic process $\mathbf{y}_{\rm N}^{(r)}$ having mean $\boldsymbol{\mu}_{\mathbf{N}}$ and circulant covariance $\mathbf{K}_{\rm N}$. These augmented samples are the complete-data of the EM algorithms presented in the next chapters. It becomes useful to define the frequency-domain random vectors $\mathbf{c}_{\rm N}^{(r)} = \mathbf{W}_{\rm N}\mathbf{y}_{\rm N}^{(r)}$ having a normal density with log-likelihood

$$\begin{aligned}L_c(\mathbf{m}_{\rm N},\,\boldsymbol{\Sigma}) = \\ -P\sum_{i=0}^{N-1}\log\sigma(i) - \sum_{i=0}^{N-1}\frac{1}{\sigma(i)}\sum_{r=1}^{P}|c_{\rm N}^{(r)}(i)-m_{\rm N}(i)|^2 \quad (9)\end{aligned}$$

The sample mean of the $\mathbf{c}_{\rm N}^{(r)}$ is denoted $\bar{\mathbf{c}}_{\rm N}$.

II. Estimation of Constrained Means

For clarity of exposition we first develop a procedure for estimating the constrained mean alone. Of limited interest by itself, this procedure is of most use when incorporated into the joint mean/covariance estimation method in the next section. If the covariance $\mathbf{K}_{\rm G}$ is assumed known, the

MLE constrained to be a superposition of M sinusoids becomes

$$\hat{\boldsymbol{\mu}}_{\mathrm{G}} = \arg\max_{\mu_{\mathrm{G}} \in \mathcal{U}_{\mathcal{G}}^N} -\sum_{r=1}^{P} (\mathbf{y}_{\mathrm{G}}^{(r)} - \boldsymbol{\mu}_{\mathbf{G}})^\dagger \mathbf{K}_{\mathrm{G}}^{-1} (\mathbf{y}_{\mathrm{G}}^{(r)} - \boldsymbol{\mu}_{\mathbf{G}}) \,. \quad (10)$$

Two difficulties with this constrained maximization are illustrated by writing the objective function in terms of the orthogonal decomposition $\mathbf{K}_{\mathrm{G}} = \mathbf{Q}^\dagger \boldsymbol{\Lambda} \mathbf{Q}$:

$$\begin{aligned} -\sum_{r=1}^{P} (\mathbf{y}_{\mathrm{G}}^{(r)} - \boldsymbol{\mu}_{\mathbf{G}})^\dagger \mathbf{Q}^\dagger \boldsymbol{\Lambda}^{-1} \mathbf{Q} (\mathbf{y}_{\mathrm{G}}^{(r)} - \boldsymbol{\mu}_{\mathbf{G}}) & \\ &= -\sum_{r=1}^{P} (\mathbf{y}_{\mathrm{G}}^{(r)\prime} - \boldsymbol{\mu}_{\mathbf{G}}{}')^\dagger \boldsymbol{\Lambda}^{-1} (\mathbf{y}_{\mathrm{G}}^{(r)\prime} - \boldsymbol{\mu}_{\mathbf{G}}{}') \\ &= -\sum_{r=1}^{P} \sum_{i=0}^{G-1} \frac{1}{\lambda_i} |y_{\mathrm{G}}^{(r)\prime}(i) - \mu_{\mathrm{G}}{}'(i)|^2 \,, \quad (11) \end{aligned}$$

where the whitened data $\mathbf{y}_{\mathrm{G}}^{(r)\prime} = \mathbf{Q}\mathbf{y}_{\mathrm{G}}^{(r)}$ and mean $\boldsymbol{\mu}_{\mathbf{G}}{}' = \mathbf{Q}\boldsymbol{\mu}_{\mathbf{G}}$ have been used. If it were known which M sinusoids make up $\boldsymbol{\mu}_{\mathbf{G}}$ the maximization of (11) would be a linear least squares problem; however, the nonzero components are unknown, implying the solution is substantially more difficult. And if $\mathbf{K}_{\mathrm{G}}$ is unknown and must be estimated (as in the joint estimation problem) then for each candidate covariance there is a different orthogonal decomposition $\mathbf{Q}$.

We address these difficulties by assuming that, for large enough N, $\mathbf{K}_{\mathrm{G}} \in \mathcal{K}_{\mathcal{G}}^N$, so that the covariance can be represented using the basis vectors making up the mean. To see the simplification this allows, suppose for the moment that $G = N$. Then (10) may be rewritten in terms of the spectral decomposition $\mathbf{K}_{\mathrm{N}} = \mathbf{W}_{\mathrm{N}}^\dagger \boldsymbol{\Sigma} \mathbf{W}_{\mathrm{N}}$:

$$\hat{\mathbf{m}}_{\mathrm{N}} = \arg\max_{\mathbf{m}_{\mathrm{N}} \in \mathcal{M}_{\mathcal{N}}} -\sum_{i=0}^{N-1} \sum_{r=1}^{P} \frac{1}{\sigma_i} |c_{\mathrm{N}}^{(r)}(i) - m_{\mathrm{N}}(i))|^2 \,. \quad (12)$$

The problem now decouples across frequency bins, and the solution is obtained by setting $m_{\mathrm{N}}(i) = \bar{\mathbf{c}}_{\mathrm{N}}$ in the M bins where $|\bar{c}_{\mathrm{N}}(i)|^2/\sigma(i)$ is largest, and to zero elsewhere.

This indicates the maximization of (10) can be solved via the EM algorithm, with the complete-data space being the extension of the available G-length records to length N. The EM algorithm for constrained mean estimation is derived accordingly as follows. The expectation of the complete-data log-likelihood (9) becomes, after discarding terms not a function of $\mathbf{m}_{\mathrm{N}}$,

$$\begin{aligned} E\left[L_c(\mathbf{m}_{\mathrm{N}}) \,\middle|\, \mathcal{Y}_{\mathcal{G}}, \mathbf{m}_{\mathrm{N}}^{(p)} \right] = & \\ -\sum_{i=0}^{N-1} \frac{1}{\sigma(i)} \sum_{r=1}^{P} E\left[|c_{\mathrm{N}}^{(r)}(i) - m_{\mathrm{N}}(i)|^2 \,\middle|\, \mathcal{Y}_{\mathcal{G}}, \mathbf{m}_{\mathrm{N}}^{(p)} \right] \,, & \end{aligned}$$

from which the unconstrained mean estimate may be written

$$\begin{aligned} \tilde{m}_{\mathrm{N}}(i) &= \frac{1}{P} \sum_{r=1}^{P} E\left[c_{\mathrm{N}}^{(r)}(i) \,\middle|\, \mathcal{Y}_{\mathcal{G}}, \mathbf{m}_{\mathrm{N}}^{(p)} \right] \,, \text{ or} \\ \tilde{\mathbf{m}}_{\mathrm{N}} &= \mathbf{m}_{\mathrm{N}}^{(p)} + \boldsymbol{\Sigma} \mathbf{W}_{\mathrm{G}} \mathbf{K}_{\mathrm{G}}^{-1} \left(\bar{\mathbf{y}}_{\mathrm{G}} - \boldsymbol{\mu}_{\mathbf{G}}^{(\boldsymbol{p})} \right) \,. \end{aligned}$$

This estimate is constrained to have only M nonzero components by examining the penalty in likelihood paid by zeroing out the i^{th} bin of $\tilde{\mathbf{m}}_{\mathrm{N}}$:

$$\begin{aligned} \gamma_i &= \frac{1}{\sigma(i)} \left(\sum_{r=1}^{P} E\left[|c_{\mathrm{N}}^{(r)}(i)|^2 \,\middle|\, \mathcal{Y}_{\mathcal{G}}, \mathbf{m}_{\mathrm{N}}^{(p)} \right] - \right. \\ &\qquad \left. \sum_{r=1}^{P} E\left[|c_{\mathrm{N}}^{(r)}(i) - \tilde{m}_{\mathrm{N}}(i)|^2 \,\middle|\, \mathcal{Y}_{\mathcal{G}}, \mathbf{m}_{\mathrm{N}}^{(p)} \right] \right) \\ &= |\tilde{m}_{\mathrm{N}}(i)|^2 / P\sigma(i) \,. \end{aligned}$$

This cost is a normalized measure of the energy in the ith frequency bin of the mean. The constraint is enforced by setting $m_{\mathrm{N}}^{(p+1)}(i) = \tilde{m}_{\mathrm{N}}(i)$ if γ_i is among the M largest costs, and to zero otherwise. Formally, define indices such that $\gamma_{k_0} \geq \gamma_{k_1} \geq \cdots \geq \gamma_{k_{N-1}}$, and let $\Gamma_M = \{k_0, \ldots, k_{M-1}\}$. The above procedure can then be summarized as

Algorithm 1 (EM Constrained Mean Estimation) *Given a set of observations $\mathcal{Y}_{\mathcal{G}}$ from a stationary Gaussian process with an unknown mean and known Toeplitz covariance $\mathbf{K}_{\mathrm{G}} \in \mathcal{K}_{\mathcal{G}}^N$, an EM algorithm for computing the maximum-likelihood estimate of its constrained mean is defined by the following iteration:*

- *Find the unconstrained mean estimate* $\tilde{\mathbf{m}}_{\mathrm{N}} = \mathbf{m}_{\mathrm{N}}^{(p)} + \boldsymbol{\Sigma} \mathbf{W}_{\mathrm{G}} \mathbf{K}_{\mathrm{G}}^{-1} (\bar{\mathbf{y}}_{\mathrm{G}} - \boldsymbol{\mu}_{\mathbf{G}}^{(\boldsymbol{p})})$
- *Calculate the costs* $\gamma_i = |\tilde{m}_{\mathrm{N}}(i)|^2 / P\sigma(i)$, *for each bin* $i = 0 \cdots N-1$.
- *Set* $m_{\mathrm{N}}^{(p+1)}(i) = \begin{cases} \tilde{m}_{\mathrm{N}}(i) & i \in \Gamma_M \\ 0 & i \notin \Gamma_M \end{cases}$.

Remark: If $M \geq G$, there is no unique MLE $\hat{\mathbf{m}}_{\mathrm{N}}$. Indeed, since any set of M distinct direction vectors is linearly independent, there are $\binom{N}{M}$ equivalent solutions if $M = G$, and infinitely many if $M > G$, so we require $M < G$ as a necessary condition for a unique estimate to exist.

III. Estimation of Constrained Means and Toeplitz Covariances

Although the mean arises from point sources, suppose that the interfering noise is diffuse and its strength does not vary wildly with frequency. Less resolution is therefore needed to estimate the spectrum than the mean. This suggests representing the spectrum with a superposition of basis functions; accordingly, suppose $N = JQ$ and write

$$\sigma(i) = \sum_{m=0}^{J-1} s(m) \psi_m(i), \qquad s(m) \geq 0 \quad (13a)$$

$$\psi_m(i) = \begin{cases} 1 & mQ \leq i < (m+1)Q \\ 0 & \text{else} \end{cases} \,. \quad (13b)$$

Each of the J basis functions has a disjoint support of Q integers. This reparameterizes the spectrum in terms of $\mathbf{S} = [s(1) \cdots s(J-1)]^T$. We define

$$\mathcal{K}_{\mathcal{G}}^{\mathcal{S}} = \{ \mathbf{K}_{\mathrm{G}} = \mathbf{W}_{\mathrm{G}}^\dagger \boldsymbol{\Sigma} \mathbf{W}_{\mathrm{G}} \in \mathcal{K}_{\mathcal{G}}^N \mid \sigma(i) = \sum_{m=0}^{J-1} s(m) \psi_m(i) \} \,, \quad (14)$$

the set of covariance matrices that can be generated via spectra of the form (13). The complete-data log-likelihood of (9) becomes

$$L_c(\mathbf{m}_{\mathrm{N}}, \mathbf{S}) = -PQ \sum_{m=0}^{J-1} \log s(m) - \sum_{m=0}^{J-1} \frac{1}{s(m)} \sum_{i=mQ}^{(m+1)Q-1} \sum_{r=1}^{P} |c_{\mathrm{N}}^{(r)}(i) - m_{\mathrm{N}}(i)|^2 \,. \quad (15)$$

It is the expectation of this re-parameterized complete-data log-likelihood function that is maximized in the M-step of the EM algorithm.

Remark: If $J = N$, the spectrum is composed of N basis functions and no smoothing is done. As J approaches unity, the spectrum becomes white. If $J = G$ the number of spectrum parameters equals the number of covariance parameters determinable from the G-length observation. More elaborate schemes have been explored using polynomial spline representations of the spectrum [17].

A direct EM algorithm solution to the joint mean and covariance estimation problem requires an exhaustive search over all the allowable $\binom{N}{M}$ configurations of mean energy. As an alternative to a strict EM solution, recall that the most important properties of the EM algorithm stem from its nondecreasing likelihood [24]. Below we craft a *generalized EM* (GEM) algorithm having this property.[1]

Let each iteration consist of two stages named M and K for mean and covariance estimation respectively. On iteration $p+1$, stage M uses $(\boldsymbol{\mu}_{\mathbf{G}}^{(\boldsymbol{p})}, \mathbf{K}_{\mathrm{G}}^{(p)})$ to perform one step of the EM algorithm for computing a constrained mean when the process covariance is known. The covariance is taken to be $\mathbf{K}_{\mathrm{G}}^{(p)}$, and the resulting mean estimate is $\boldsymbol{\mu}_{\mathbf{G}}^{(\boldsymbol{p+1})}$. Then stage K uses $(\boldsymbol{\mu}_{\mathbf{G}}^{(\boldsymbol{p+1})}, \mathbf{K}_{\mathrm{G}}^{(p)})$ to find a new covariance estimate $\mathbf{K}_{\mathrm{G}}^{(p+1)}$ by performing one step of the EM algorithm for covariance estimation when the mean is known to be $\boldsymbol{\mu}_{\mathbf{G}}^{(\boldsymbol{p+1})}$, and the iteration is complete. Since both stages are EM steps,

$$L_i(\boldsymbol{\mu}_{\mathbf{G}}^{(\boldsymbol{p})}, \mathbf{K}_{\mathrm{G}}^{(p)}) \le L_i(\boldsymbol{\mu}_{\mathbf{G}}^{(\boldsymbol{p+1})}, \mathbf{K}_{\mathrm{G}}^{(p)}) \le L_i(\boldsymbol{\mu}_{\mathbf{G}}^{(\boldsymbol{p+1})}, \mathbf{K}_{\mathrm{G}}^{(p+1)}) \,,$$

and the likelihood is nondecreasing.

Separating the mean and covariance estimation phases this way allows the spectrum constraint and mean constraint to be easily introduced. The potential disadvantage of the GEM solution is that its two-stage maximization may reach a fixed point not otherwise attained by a joint maximization algorithm (e.g., the joint likelihood function has a ridge in the plane defined by one covariance parameter and one mean parameter). This has not been observed in practice.

Now a method for jointly estimating the constrained mean and Toeplitz covariance follows easily. The procedure performed by stage M is precisely algorithm 1. It will be interlaced with the following method of constrained covariance estimation. The maximization problem associated with stage K is

$$\hat{\mathbf{K}}_{\mathrm{G}} = \arg\max_{\mathbf{K}_{\mathrm{G}} \in \mathcal{K}_{\mathcal{G}}^{\mathcal{S}}} -P \log |\mathbf{K}_{\mathrm{G}}| - \sum_{r=1}^{P} (\mathbf{y}_{\mathrm{G}}^{(r)} - \boldsymbol{\mu}_{\mathbf{G}})^{\dagger} \mathbf{K}_{\mathrm{G}}^{-1} (\mathbf{y}_{\mathrm{G}}^{(r)} - \boldsymbol{\mu}_{\mathbf{G}}) \quad (16)$$

where the mean $\boldsymbol{\mu}_{\mathbf{G}}$ is considered known. To find the EM iteration, the expectation of the complete-data log-likelihood of (15) is differentiated to obtain the M-step:

$$s^{(p+1)}(m) = \frac{1}{PQ} \sum_{i=mQ}^{(m+1)Q-1} \sum_{r=1}^{P} E\left[|c_{\mathrm{N}}^{(r)}(i) - m_{\mathrm{N}}(i)|^2 \,\middle|\, \mathcal{Y}_{\mathcal{G}}, \boldsymbol{\Sigma}^{(p)} \right] \quad (17)$$

At each iteration, the expected variances of $c_{\mathrm{N}}^{(r)}(i)$ given the current parameter estimate are found and averaged over realizations and the support sets of the spectrum basis functions. The explicit form for the processing of stage K is easily derived, and is here expressed using the intermediate quantity $\tilde{\boldsymbol{\Sigma}}$:

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{(p)} \mathbf{W}_{\mathrm{G}} \mathbf{K}_{\mathrm{G}}^{(p)-1} \times \left[\frac{1}{P} \sum_{r=1}^{P} \left(\mathbf{y}_{\mathrm{G}}^{(r)} - \boldsymbol{\mu}_{\mathbf{G}} \right) \left(\mathbf{y}_{\mathrm{G}}^{(r)} - \boldsymbol{\mu}_{\mathbf{G}} \right)^{\dagger} \right] \mathbf{K}_{\mathrm{G}}^{(p)-\dagger} \mathbf{W}_{\mathrm{G}}^{\dagger} \boldsymbol{\Sigma}^{(p)} + \boldsymbol{\Sigma}^{(p)} - \boldsymbol{\Sigma}^{(p)} \mathbf{W}_{\mathrm{G}} \mathbf{K}_{\mathrm{G}}^{(p)-1} \mathbf{W}_{\mathrm{G}}^{\dagger} \boldsymbol{\Sigma}^{(p)} \quad (18)$$

$$s^{(p+1)}(m) = \frac{1}{Q} \sum_{i=mQ}^{(m+1)Q-1} \tilde{\boldsymbol{\Sigma}}(i,i) \,, \quad (19)$$

and the updated spectrum estimate $\boldsymbol{\Sigma}^{(p+1)}$ is obtained from $\mathbf{S}^{(p+1)}$ via (13). In summary we have

Algorithm 2 (Mean and Covariance Estimation) *Given a set of observations $\mathcal{Y}_{\mathcal{G}}$ from a stationary Gaussian process having an unknown mean and covariance, a GEM algorithm for computing the maximum-likelihood estimate of its constrained mean and Toeplitz covariance is defined by the following iteration:*

- *Find the unconstrained mean estimate* $\tilde{\mathbf{m}}_{\mathrm{N}} = \mathbf{m}_{\mathrm{N}}^{(p)} + \boldsymbol{\Sigma}^{(p)} \mathbf{W}_{\mathrm{G}} \mathbf{K}_{\mathrm{G}}^{(p)-1} (\bar{\mathbf{y}}_{\mathrm{G}} - \boldsymbol{\mu}_{\mathbf{G}}^{(\boldsymbol{p})})$
- *Calculate the costs* $\gamma_i = |\tilde{m}_{\mathrm{N}}(i)|^2 / P\sigma^{(p)}(i)$, *for each bin* $i = 0 \cdots N-1$.
- *Set* $m_{\mathrm{N}}^{(p+1)}(i) = \begin{cases} \tilde{m}_{\mathrm{N}}(i) & i \in \Gamma_M \\ 0 & i \notin \Gamma_M \end{cases}$.
- *Compute* $\mathbf{S}^{(p+1)}$ *and* $\boldsymbol{\Sigma}^{(p+1)}$ *with (19) and (13), respectively.*

Remark: By setting $M = 0$ and $J = N$ the mean estimation stage vanishes and the spectrum is not smoothed, so the covariance estimator of [15] is obtained.

IV. Covariance Estimation Performance

We have noted that with $M = 0$ and $J = N$, algorithm 2 is equivalent to the covariance estimator proposed

[1] The proposed algorithm is not a GEM algorithm in the sense originally set forth by Dempster et al. in [4].

in [15]. In this section we examine the performance of these equivalent estimators by comparing the error moments of the MLE to those of simple lag product methods and the Cramér-Rao bound. We also compare the ML covariance estimator to the group of estimators presented by Kay and Marple [10]. It will become convenient to specify a covariance matrix $\mathbf{K}_{\mathrm{G}}$ by its covariance sequence $K(\tau)$, $0 \leq \tau < G$, which is the first row of the matrix. To avoid wraparound effects, we will assume $G \leq N/2$ throughout this section.

A. Performance Benchmarks

The three simple lag-product estimators in table 1 will be used as benchmarks for evaluating the performance of the MLE. The first is CD-ML, the MLE of the covariance when the full process period (the complete data) is given. CD-ML is the minimum-variance unbiased estimator of $\mathbf{K}_{\mathrm{N}}$. When $G < N$ pieces of data are available, two estimators are commonly used: the unbiased lag-product estimator UBLP and the biased lag-product estimator BLP, also known as the periodogram. The variance of $\hat{K}_{\mathrm{ublp}}(\tau)$ generally increases as τ approaches G; BLP lowers this variance by scaling the estimate down for large τ, at the expense of some bias in this region.

To find the Cramér-Rao variance bound, parameterize a covariance matrix $\mathbf{K}_{\mathrm{G}} \in \mathcal{K}_{\mathcal{G}}$ via the $2G-1$ real numbers $k_{-G+1}, \ldots, k_{G-1}$ defined by $\mathbf{K}_{\mathrm{G}}(0,0) = k_0$ and $\mathbf{K}_{\mathrm{G}}(0,\tau) = k_\tau + jk_{-\tau}$ for $0 < \tau < G$. For an estimate $\hat{\mathbf{K}}_{\mathrm{G}} \in \mathcal{K}_{\mathcal{G}}$, we therefore have

$$Var\{\hat{\mathbf{K}}_{\mathrm{G}}(0,\tau)\} = \begin{cases} Var\{\hat{k}_\tau\} & \tau = 0 \\ Var\{\hat{k}_\tau\} + Var\{\hat{k}_{-\tau}\} & 0 < \tau < G \end{cases} . \tag{20}$$

The $(2G-1)\times(2G-1)$ Fisher information matrix $\mathbf{J}$ of these scalars, indexed from $-G+1$ to $G-1$, has elements (see [21])

$$\begin{aligned} \mathbf{J}(m,n) &= -E\left[\frac{\partial^2 L_i(\mathbf{K}_{\mathrm{G}})}{\partial k_m \partial k_n}\right] \\ &= P \operatorname{tr}\left(\mathbf{K}_{\mathrm{G}}^{-1}\frac{\partial \mathbf{K}_{\mathrm{G}}}{\partial k_m}\mathbf{K}_{\mathrm{G}}^{-1}\frac{\partial \mathbf{K}_{\mathrm{G}}}{\partial k_n}\right) . \end{aligned} \tag{21}$$

Letting $\mathbf{B}_m \in \mathbb{C}^{G\times G}$ have ones on its m^{th} superdiagonal and zeros elsewhere allows us to write

$$\frac{\partial \mathbf{K}_{\mathrm{G}}}{\partial k_m} = \begin{cases} j(\mathbf{B}_m - \mathbf{B}_m^T) & -G < m < 0 \\ \mathbf{I}_{\mathrm{G}} & m = 0 \\ \mathbf{B}_m + \mathbf{B}_m^T & 0 < m < G \end{cases} . \tag{22}$$

The bound is obtained by inverting $\mathbf{J}$ and applying (20). This formulation expresses the Hermitian and Toeplitz constraints, but not the circulant extension one; the bound is thus applicable to estimation over the constraint set $\mathcal{K}_{\mathcal{G}}$. (See [8] for ways to incorporate the circulant extension constraint.)

B. Performance versus Lag Product Methods

The following performance study compares the first two error moments of ID-ML to CD-ML, UBLP, BLP and the Cramér-Rao bound. Two spectra were examined: one white, the other consisting of several impulses (figure 1). To evaluate the ID-ML estimates, we generated 500 independent sets of data $\mathcal{Y}_{\mathcal{G}}$, each consisting of one data record (i.e. $P = 1$), having one of the sample spectra. Algorithm 2 with $M = 0$ and $J = N$ was applied to each realization, resulting in 500 independent ID-ML covariance estimates from which the sample means and variances of ID-ML were calculated. The circulant extension in the EM algorithm was chosen as $N = 32$ to agree with the length of the process spectra. Several values of G were studied for each process.

Figure 2 shows the experimental results for the white process. The leftmost two panels show the components of root mean-square error (RMSE), bias and standard deviation, when $G = 16$. Panel *(a)*, plotted on an expanded scale, shows that all four estimators are essentially unbiased for this process, with the bias of BLP hidden because $K(\tau) = 0$ for $\tau \neq 0$. Comparing panels *(a)* and *(d)*, we see that estimator standard deviation dominates the RMSE error measure for this covariance. BLP fares well for such processes. In fact, its standard deviation falls below that of CD-ML, which is an efficient estimator given N pieces of data. UBLP, because of its rise in variance at the edge of the window, does poorly by comparison with BLP even though the former is efficient for a white process. ID-ML suffers no rise in variance as τ approaches G. For low values of τ it does not attain the level of UBLP because ID-ML is based on being able to fill in missing lag products with conditional expectations. The predictions fail when samples are independent.

The rightmost four panels show the RMSE of the three covariance estimators and the Cramér-Rao variance bound, plotted versus lag number τ. Each curve corresponds to a different data record length G. Also on these four panels is the RMSE of CD-ML, which equals the Cramér-Rao bound when given the full $N = 32$ length data record. Let us examine the panels one by one.

The RMSE of UBLP changes in a regular way. The curves are shifted versions of each other, each reaching a maximum value of $K(0)$ at $\tau = G-1$. As G decreases, average RMSE across all lags increases rapidly. For a white covariance, the Cramér-Rao bound (panel *(e)*) is the same as the RMSE of UBLP.

For several values of G and τ, the RMSE of BLP is actually lower than the Cramér-Rao lower bound given the full process period. Acceptance of some bias allows BLP to reduce its variance, and since the variance component dominates RMSE for this process, the net effect is a drop in RMSE.

Finally, the RMSE curves for ID-ML start out (at $G = 16$) almost flat, and, as G is decreased, become more peaked, though not to the extent of UBLP. This is a benefit of the circulant extension constraint. For example, at $G = 16$, the first 16 covariances can be immediately estimated from

the available lag products. But because $\mathbf{K}_G \in \mathcal{K}_G^N$, symmetry completely determines the extension of $\mathbf{K}_G$ to $\mathbf{K}_N$, save for one free covariance. The conditional expectations used by ID-ML to fill in the unavailable lag products then work with essentially complete knowledge of the process. But as G decreases, more parameters in $\mathbf{K}_N$ are free and the problem approaches the unconstrained one, with the performance of ID-ML becoming closer to UBLP, which is the unconstrained MLE. At any rate, the MLE suffers no catastrophic rise in variance at the edge of the lag window, a benefit of the circulant extension constraint.

Figure 3 shows the corresponding information for the impulsive process. Let us first examine the bias and variance when $G = 16$. The bias of BLP is evident. ID-ML, however, is essentially unbiased. In panel *(d)*, UBLP exhibits its characteristic variance rise for large lags, and BLP its decrease in variance. ID-ML essentially attains the standard deviation of CD-ML. *This is because this process, having broad covariance support, can be predicted by the MLE.*

Moving to the plots of RMSE for various G, the UBLP curves are again shifted versions of each other, though they are now closer to the RMSE of CD-ML. Performance in relation to the white process is improved due to the decrease in the variance of the process samples in relation to the overall magnitude of the covariances being estimated. The effect of bias on BLP is evident. While remaining low, the RMSE of BLP varies in proportion to the magnitude of the process covariance.

The G-data CRLB is quite close to the N-data CRLB except at $G = 4$, where by numerical accident the covariance $\mathbf{K}_G$ is virtually white (see figure 1*d*). The lower bound is insensitive to changes in G since for this process, each piece of data does not carry independent information about the covariances.

The striking observation (Fig. 3 *f*) is that the maximum-likelihood method essentially attains the Cramér-Rao lower bound, except for the anomaly at $G = 4$. The broad covariance support of the process allows the MLE to accurately predict the unknown lags.

C. Performance versus Other Spectrum Estimators

Now we compare the performance of the MLE to the group of estimators examined by Kay and Marple [10]. The process spectrum used in their study is shown in figure 4*a*. The frequency axis ranges from zero to 0.5 and represents the fraction of the sampling frequency: for a real-valued process such as this one the spectrum is even. The PSD consists of two components clustered at 0.20 and 0.21 having equal energy, a third smaller spike at 0.10, and broadband Gaussian noise. The spikes, which are due to deterministic sinusoids, are intended as a frequency resolution test, and the lower-amplitude spike to see if weak signals are masked by strong ones.

Algorithm 2 with $N = J = 200$ and $M = 0$ was applied to the single $G = 64$ length data record provided in [10]. The resulting spectrum estimate $\hat{\sigma}_{\text{id-ml}}(i)$ is shown in figure 4*b* with the broadband part of the process spectrum superimposed. The ID-ML estimates of all three spikes agree well in both location and in magnitude with the process spectrum. Further, the maximum-likelihood algorithm infers the shape of the broadband noise, although the estimate is extremely rough and shows substantial deviations from the true spectrum. The tendency of the maximum-likelihood approach seems to be to concentrate spectral energy at several frequencies, even though the true spectrum may be broadband.

An extensive comparison, adapted directly from [10], is shown in the remainder of the figure. The purpose of the comparison is to illustrate the properties of each technique rather than to make detailed performance comparisons since only one data set was analyzed. The estimators shown include lag product methods (*c*), autoregressive models (*d, e, f, h*), moving average models (*g*), and line spectrum models (*i, k*). The ML method is able to resolve components arising from deterministic sources without producing spurious components, and provides quantitative estimates of the amplitudes of the spectral lines. Only one other method (the Prony method in (*k*)) quantitatively estimated the power at these frequencies.

V. Joint Mean and Covariance Estimation Performance

Now the performance of the joint mean and covariance estimator of algorithm 2 is examined. Table 2 introduces some benchmarks which provide a basis for evaluation of the performance of the new ML estimator. For each estimator costs are given which are to be computed across each frequency index. Just as in solving (12), the M frequencies having the largest costs are allowed to be nonzero. (Since we are primarily concerned with estimating the frequency correctly, it is not important what value the nonzero bins take on.)

The first three estimators are based on the full N samples of the process, which aren't available in practice. First is CD-ML, the joint MLE of the mean and covariance, which is obtained by applying algorithm 2 with $G = N$. Next are two Fourier-based complete-data processing methods. If the spectrum is known, the ML estimate of the constrained mean is obtained just as in solving (12). This is called CD-FS to indicate that it uses the spectrum. In practice, the process spectrum is unknown, but one can nevertheless find an estimator by analogy with CD-FS by computing the costs assuming a white spectrum; this is CD-F.

If $G < N$ samples are known, we have first the estimate of algorithm 2, called ID-ML. Fourier estimators are formulated using the zero-padded N-point Fourier transform $\mathbf{W}_G \bar{\mathbf{y}}_G = \bar{\mathbf{c}}_{N|G}$. Costs can be computed assuming the spectrum is known (ID-FS) or not (ID-F), just as in the complete-data case.

A. Performance versus Fourier Methods

First let us examine the ability of the estimators to discriminate between the deterministic mean and noise in the form of covariance or spectrum energy. Define the signal-to-noise ratio *(SNR)* as

$$SNR \;=\; 10\log_{10}\frac{|E\{\bar{y}_{\mathrm{G}}(0)\}|^2}{Var\{\bar{y}_{\mathrm{G}}(i)\}} \;=\; 10\log_{10}\frac{P|\boldsymbol{\mu}_{\mathbf{G}}(0)|^2}{\mathbf{K}_{\mathrm{G}}(0,0)}\,, \tag{23}$$

a loose measure of the estimability of the mean. The procedure is to vary SNR and determine the point at which mean estimates degrade. The parameters used for the simulation study are $N = 32$, $G = J = 16$, $M = 1$, and $P = 4$. The two processes studied are shown in figure 5. The process in the upper two panels is a single sinusoid in a field of wideband nonoverlapping noise; the second is the same mean in narrowband overlapping noise. SNR is varied by adjusting the scale parameter α. We shall examine the mean and variance of the single element of Γ_M, the integer representing frequency of the complex sinusoid. As previously, 500 independent realizations of data were generated from which sample means and variances were calculated.

Results for the nonoverlapping process are shown in figure 6. Panel *(c)* shows the negative of the RMSE of frequency estimates[2] versus SNR. At high SNR, all estimators perform equally well. At about 0dB, the incomplete-data estimators' performance begins to degrade rapidly, but in different ways for each. The ID-F variance and bias increase rapidly, with the variance eventually becoming small. At first ID-F chooses predominantly the correct frequency, with the incorrect estimates falling in the range 16–31, making the variance large. As SNR decreases further, all estimates concentrate in the range of the noise, the variance drops, and the bias increases. The histograms in panels *(d)* and *(e)* illustrate this: at -3dB, the open bars showing individual ID-F frequency estimates are concentrated at the correct value with a few outliers. At -12dB, ID-F generates no correct estimates, but they are more concentrated, lowering the variance.

ID-ML maintains a low bias over the range of SNR examined here. It has approximately the same threshold of 0dB as does ID-F; however, the effects of higher noise are not as pronounced: ID-ML never reaches the high bias characteristic of ID-F (see panels *(d)* and *(e)*). Panel *(e)* illustrates this clearly: 78 of 100 of the ML frequency estimates are correct while none of the ID-F estimates are close to the true value. In fact, referring to the chart of RMSE in panel *(c)*, the performance of ID-ML apparently eventually exceeds that of ID-FS.

Error statistics for the narrowband, interfering noise process are shown in figure 7. This time ID-F fares well because the noise and the deterministic signal lie at approximately the same frequency, and choosing either yields an acceptable estimate. ID-ML is the next best performer, followed by ID-FS.

A histogram of ID-ML and ID-F estimates before performance degrades is shown in panel *(d)*. ID-ML again outperforms the Fourier approach. Even after RMSE performance of ID-ML starts to degrade (panel *(e)*), its estimates are qualitatively more accurate than ID-F, with more estimates concentrated at the true value.

B. Benefits of Joint Estimation

We conclude the simulation results by demonstrating that, for the ML estimator, estimates of sinusoid frequencies increase in accuracy as spectrum estimates improve. The mean and spectrum used for the experiment are the same as shown in the upper panels of figure 5. The parameter choices are $N = 32$, $G = J = 16$. The following procedure was used:

1. The process mean $\mathbf{m}_{\mathrm{N}}$ and signal-to-noise ratio were held constant.
2. The number of realizations P and the spectrum scale constant α were varied to maintain constant SNR while changing the accuracy with which the spectrum can be estimated.

This is done as follows. For constant SNR (see (23)) the ratio $P|\boldsymbol{\mu}_{\mathbf{G}}(0)|^2/\mathbf{K}_{\mathrm{G}}(0,0)$ must be constant, which implies $\mathbf{K}_{\mathrm{G}}(0,0) \propto P$ or $\mathbf{K}_{\mathrm{G}} \propto P$ since all of $\mathbf{K}_{\mathrm{G}}$ is scaled uniformly. If this is done the density of $\bar{\mathbf{y}}_{\mathrm{G}}$ remains constant as P varies.[3] Since CD-F, ID-F, CD-FS, and ID-FS are calculated using only the statistic $\bar{\mathbf{y}}_{\mathrm{G}}$, the performance of the Fourier estimators does not vary as P changes.

However, the estimability of the spectrum *does* change. Standard deviation of spectrum estimates tends to increase linearly with P because $\mathbf{K}_{\mathrm{G}}$ is growing. At the same time, the increase in the size of the dataset pushes the standard deviation down by a factor of $\sqrt{P}$. (Strictly true for an efficient unbiased estimator, this is appropriate as a heuristic for the MLE as well.) The combined effect is that the standard deviation, and also the RMSE (since bias makes a negligible contribution to RMSE) go up as $\sqrt{P}$ when P increases. Looking at this another way, *as P increases, the RMSE of the spectrum estimates relative to their expected value goes down as $\sqrt{P}$.*

This is shown in figure 8 for $SNR = -3$dB. The RMSE of a sample element of the spectrum estimate, $\hat{\sigma}_{\text{id-ml}}(16)$, is shown across a range of P. The solid line is the RMSE normalized by the process spectrum $\sigma(16)$, and the dashed line is the same RMSE normalized by $\sqrt{P}$. The solid line confirms that the RMSE of the ML spectrum estimate relative to the process spectrum decreases as P increases. The dashed one indicates that the dependence is indeed ap-

[2] The negative of RMSE, standard deviation, and bias are plotted to give the impression of performance "falling" as error increases.

[3] Using superscripts to indicate dependence on P we have $\mathbf{y}_{\mathrm{G}}^{(r)(P)} \sim N(\mu_{\mathrm{G}},\, \mathbf{K}_{\mathrm{G}}^{(P)})$, hence $\bar{\mathbf{y}}_{\mathrm{G}}^{(P)} \sim N(\mu_{\mathrm{G}},\, (1/P)\mathbf{K}_{\mathrm{G}}^{(P)})$. Since $\mathbf{K}_{\mathrm{G}}^{(P)} \propto P$, write $\mathbf{K}_{\mathrm{G}}^{(P)} = P\mathbf{K}_{\mathrm{G}}^{(1)}$, yielding $\bar{\mathbf{y}}_{\mathrm{G}}^{(P)} \sim N(\mu_{\mathrm{G}},\, \mathbf{K}_{\mathrm{G}}^{(1)})$.

proximately $1/\sqrt{P}$. Similar effects occur across the entire spectrum.

The effect of such improving spectrum estimates on mean estimates is shown in figure 9. Panel *(c)* summarizes the result by showing the RMSE of the frequency estimates. At $P = 2$, the RMSE's of CD-ML, ID-ML, CD-F, and ID-F are quite high: about 14 for a quantity that ranges between 0 and 31. As P increases, the RMSE of the joint ML estimators decreases rapidly, until finally at $P = 16$ there is no longer any error in the ML frequency estimate. By this point ID-ML outperforms ID-FS as well as ID-F. The RMSE of the Fourier-based approaches (including that of ID-FS) remains constant, since the statistics of $\bar{\mathbf{y}}_G$ are constant, while that of the MLE improves due to better spectrum estimates.

VI. Conclusions

From simulation studies we conclude that the ML estimates of stationary covariances appear to be essentially unbiased provided that the underlying process is modeled correctly. Although the maximum-likelihood estimator offers lower variance than the UBLP estimator in almost all situations, the effectiveness of the MLE is greatest when the underlying process has broad covariance support. In such cases the ID-ML estimator can give performance approaching the complete-data maximum-likelihood estimator. Furthermore, the maximum-likelihood method offers better frequency resolution than many other contemporary model-based methods. It quantitatively estimates signal power.

A new algorithm has been developed to estimate the process mean consisting of complex sinusoids of unknown frequencies and the Toeplitz covariance jointly. Experiments indicate that, because the ML method takes the spectrum into account, estimates of the sinusoidal frequencies present can be greatly improved. An example is a process where there is substantial noise away from the process mean. However, if the noise is in a narrow band coincident with the mean, then improvements in terms of RMSE are more modest. In terms of the proportion of correct frequency estimates rather than their RMSE, though, there are improvements even for these processes. The benefit of estimating the spectrum simultaneously with the mean was demonstrated. As spectrum estimates improve, the estimates of the process mean also improve.

References

[1] J. P. Burg, D. G. Luenberger, and D. L. Wenger. Estimation of structured covariance matrices. *Proc. IEEE*, 70(9):963–74, September 1982.

[2] S. C. Chen, T. J. Schaewe, R. S. Teichman, and M. I. Miller. Parallel algorithms for maximum likelihood nuclear magnetic resonance spectroscopy. *Jour. Magnetic Resonance*, March 1993.

[3] A. Dembo, C. L. Mallows, and L. A. Shepp. Embedding nonnegative-definite Toeplitz matrices in nonnegative-definite circulant matrices, with application to covariance estimation. *IEEE Trans. Infor. Theory*, 35(6):1206–12, November 1989.

[4] A. D. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Jour. Roy. Stat. Soc. Ser. B*, 39(1):1–37, 1977.

[5] D. R. Fuhrmann and M. I. Miller. On the existence of positive-definite maximum-likelihood estimates of structured covariance matrices. *IEEE Trans. Infor. Theory*, 34(4):722–9, July 1988.

[6] D. R. Fuhrmann, M. J. Turmon, and M. I. Miller. Efficient implementation of the EM algorithm for Toeplitz covariance estimation. In *Proc. Ann. Conf. Infor. Sci. Sys.*, Princeton, N. J., March 1988. Princeton Univ.

[7] N. R. Goodman. Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). *Ann. Math. Stat.*, 34(1):152, March 1963.

[8] J. D. Gorman and A. O. Hero. Lower bounds for parametric estimation with constraints. *IEEE Trans. Infor. Theory*, 36(6):1285–1301, November 1990.

[9] S. Haykin, V. Kezys, and E. Vertatschitsch. Maximum likelihood for angle-of-arrival estimation in multipath. In Simon Haykin, editor, *Advances in Spectrum Analysis and Array Processing*, volume II, page 143. Prentice-Hall, 1991. (conclusion).

[10] S. M. Kay and S. L. Marple Jr. Spectrum analysis—A modern perspective. *Proc. IEEE*, 69(11):1380–1418, November 1981.

[11] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.

[12] J. Malley. *Statistical Applications of Jordan Algebras*. Lecture Notes in Statistics. Springer. Accepted for publication.

[13] M. I. Miller, D. R. Fuhrmann, J. A. O'Sullivan, and D. L. Snyder. Maximum-likelihood methods for Toeplitz covariance estimation and radar imaging. In Simon Haykin, editor, *Advances in Spectrum Analysis and Array Processing*, volume II, pages 145–172. Prentice-Hall, 1991.

[14] M. I. Miller and A. Greene. Maximum-likelihood estimation for nuclear magnetic resonance spectroscopy. *Jour. Magnetic Resonance*, 83:525–548, 1989.

[15] M. I. Miller and D. L. Snyder. The role of likelihood and entropy in incomplete-data problems. *Proc. IEEE*, 75(7):892–907, July 1987.

[16] S. Morgera. The role of abstract algebra in structured estimation theory. *IEEE Trans. Infor. Theory*, 38(3):1053–65, May 1992.

[17] P. Moulin. *A Method of Sieves for Radar Imaging and Spectrum Estimation*. PhD thesis, Washington University, St. Louis, May 1990.

[18] P. Moulin, J. A. O'Sullivan, and D. L. Snyder. A method of sieves for multiresolution spectrum estimation and radar imaging. *IEEE Trans. Infor. Theory*, 38(2 pt. II):801–13, March 1992.

[19] D. B. Rubin and T. H. Szatrowski. Finding maximum likelihood estimates of patterned covariance ma-

trices by the EM algorithm. *Biometrika*, 69(3):657–60, March 1982.

[20] D. L. Snyder, J. A. O'Sullivan, and M. I. Miller. The use of maximum-likelihood estimation for forming images of diffuse radar-targets from delay-doppler data. *IEEE Trans. Infor. Theory*, 35(3):536–548, May 1989.

[21] P. Stoica and A. Nehorai. Performance study of conditional and unconditional direction-of-arrival estimation. *IEEE Trans. Sig. Proc.*, 38(10):1783–1795, October 1990.

[22] M. J. Turmon. Maximum-likelihood estimation of constrained means and Toeplitz covariances with application to direction-finding. Master's thesis, Washington University, St. Louis, August 1990.

[23] M. J. Turmon, M. I. Miller, D. L. Snyder, and J. A. O'Sullivan. Performance evaluation of maximum-likelihood Toeplitz covariance estimates generated using the EM algorithm. In *Proc. Fourth ASSP Workshop on Spect. Est. and Modeling*, pages 182–5, August 1988.

[24] C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, 11(1):95–103, 1983.

Michael J. Turmon was born in Kansas City, Missouri in 1964. He received bachelors degrees in Computer Science and Electrical Engineering from Washington University in 1987. Under the guidance of Michael I. Miller, he got the M.S. in Electrical Engineering from that institution in 1990. An NSF Creativity Award in Engineering supported his research on spectrum estimation in those years.

Since 1990 Michael has been in the Department of Electrical Engineering at Cornell University, where he is studying the statistical properties of neural network pattern classifiers with Terrence Fine. At issue is the amount of training data required to ensure that the ultimate performance of networks in a given family is accurately reflected in their performance on the training data.

Michael I. Miller received the B.S. degree in electrical engineering from the S.U.N.Y. at Stony Brook, 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 Electical Engineering and Institute for Biomedical Computing Faculty of Washington University in St. Louis, MO, where he is currently 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. Dr. Miller is a recipient of the Presidential Young Investigator Award.

Tables and Figures Follow

Covariance Estimators and Their Error Moments

Name	Estimator	Bias	Variance
CD-ML	$\hat{K}_{\text{cd-ml}}(\tau)=\frac{1}{PN}\sum_{r=1}^{P}\sum_{i=0}^{N-1} y_{\text{N}}^{(r)}(i)\, y_{\text{N}}^{(r)*}(i+\tau)_N$	0	$\frac{1}{PN}\sum_{i=0}^{N-1}\lvert\mathbf{K}_{\text{N}}(0,i)\rvert^2$
UBLP	$\hat{K}_{\text{ublp}}(\tau)=\frac{1}{P(G-\tau)}\sum_{r=1}^{P}\sum_{i=0}^{G-\tau-1} y_{\text{G}}^{(r)}(i)\, y_{\text{G}}^{(r)*}(i+\tau)$	0	$\frac{1}{P(G-\tau)^2}\sum_{i=0}^{G-\tau-1}\sum_{j=0}^{G-\tau-1}\lvert\mathbf{K}_{\text{G}}(i,j)\rvert^2$
BLP	$\hat{K}_{\text{blp}}(\tau)=\frac{1}{PG}\sum_{r=1}^{P}\sum_{i=0}^{G-\tau-1} y_{\text{G}}^{(r)}(i)\, y_{\text{G}}^{(r)*}(i+\tau)$	$-\frac{\tau}{G}K(\tau)$	$\frac{1}{PG^2}\sum_{i=0}^{G-\tau-1}\sum_{j=0}^{G-\tau-1}\lvert\mathbf{K}_{\text{G}}(i,j)\rvert^2$

Table 1: Lag product covariance estimators and their first two error moments. The notation $(i)_N$ denotes $i \bmod N$.

Mean Estimators

Name	Notation	Costs
CD-ML	$\hat{m}_{\text{cd-ml}}(i)$	$\lvert\bar{c}_{\text{N}}(i)\rvert^2/\hat{\sigma}_{\text{cd-ml}}(i)$
CD-FS	$\hat{m}_{\text{cd-fs}}(i)$	$\lvert\bar{c}_{\text{N}}(i)\rvert^2/\sigma(i)$
CD-F	$\hat{m}_{\text{cd-f}}(i)$	$\lvert\bar{c}_{\text{N}}(i)\rvert^2$
ID-ML	$\hat{m}_{\text{id-ml}}(i)$	$\frac{\lvert E\{\bar{c}_{\text{N}}(i) \mid \hat{\mathbf{m}}_{\text{id-ml}},\, \hat{\mathbf{\Sigma}}_{\text{id-ml}}\}\rvert^2}{\hat{\sigma}_{\text{id-ml}}(i)}$
ID-FS	$\hat{m}_{\text{id-fs}}(i)$	$\lvert\bar{c}_{\text{N}\mid\text{G}}(i)\rvert^2/\sigma(i)$
ID-F	$\hat{m}_{\text{id-f}}(i)$	$\lvert\bar{c}_{\text{N}\mid\text{G}}(i)\rvert^2$

Table 2: Maximum-likelihood and Fourier-based constrained mean estimators for complete-data (CD) and incomplete-data (ID) cases. Only the costs used to discriminate among the N frequency bins of the mean are shown.

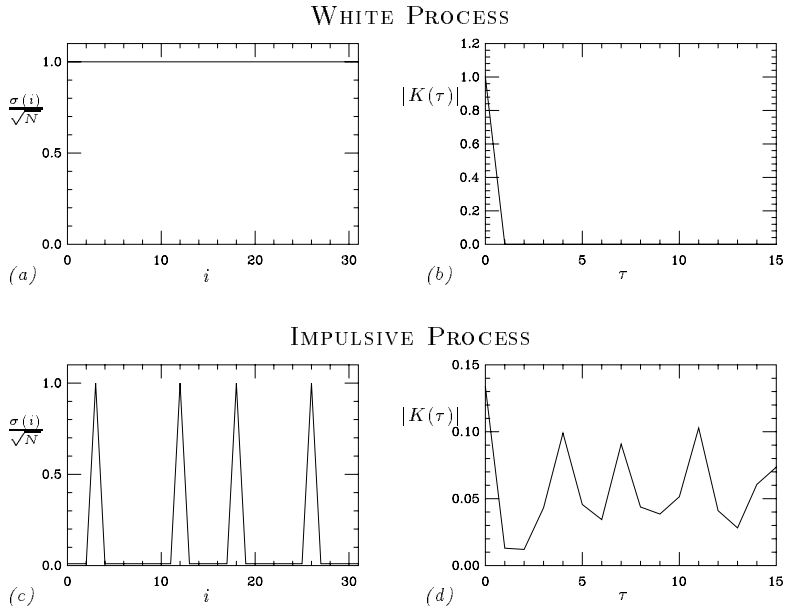

Figure 1: Noise processes used in simulation studies of estimator performance. (The triangular shape of the discrete impulses is a result of connecting discrete points with lines; while technically more correct, bar plots can overwhelm the eye.)

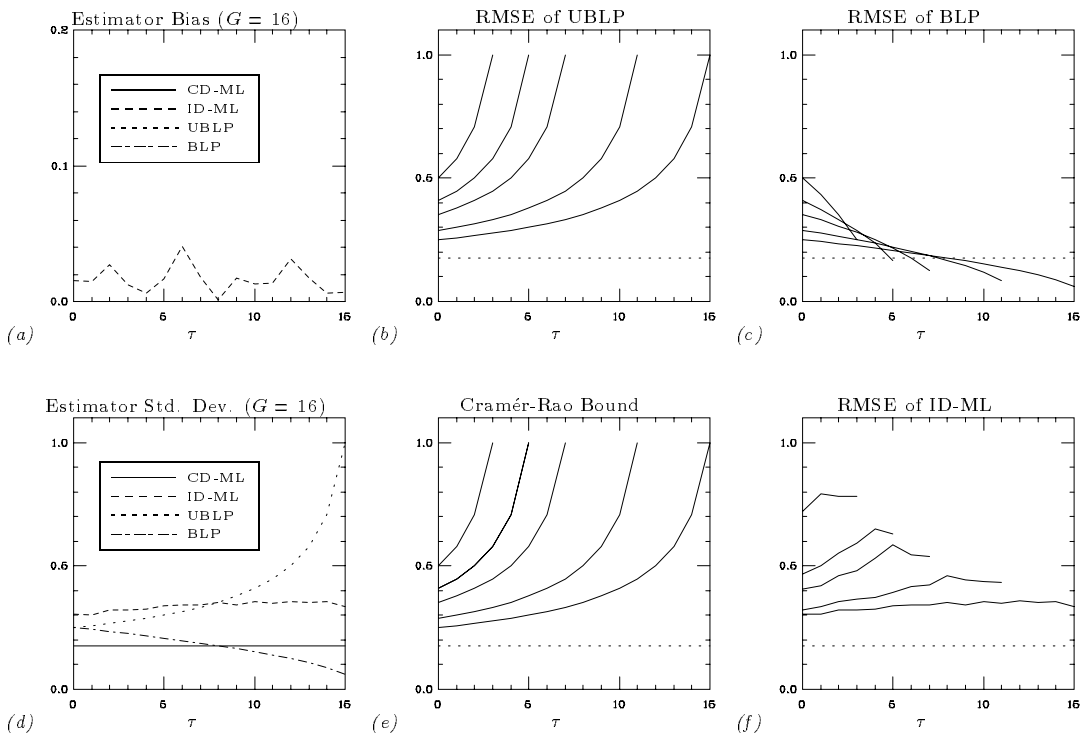

Figure 2: The leftmost two panels show the standard deviation and modulus of bias for the white process and $G = 16$ plotted versus lag number τ. The bias is shown in expanded scale. The rightmost four panels show RMS errors for the white process with $G = 4$, 6, 8, 12, and 16, plotted versus τ. The UBLP, BLP, and ID-ML estimators and the Cramèr-Rao bound are shown, together with the CD-ML estimator, which is in dotted lines on those four panels.

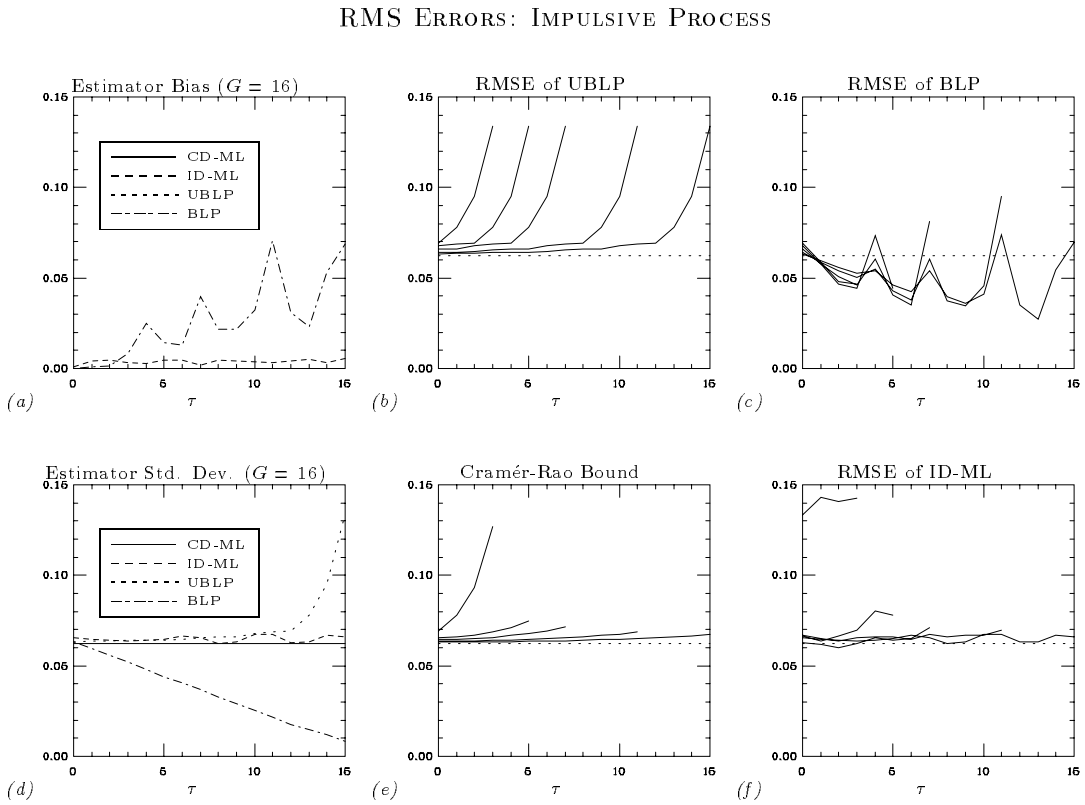

Figure 3: The leftmost two panels show the standard deviation and modulus of bias for the white process plotted versus lag number τ for $G = 16$. The rightmost four panels show RMS errors with $G = 4$, 6, 8, 12, and 16, plotted versus τ. The UBLP, BLP, and ID-ML estimators and the Cramèr-Rao bound are shown, together with the CD-ML estimator, which is in dotted lines on those four panels.

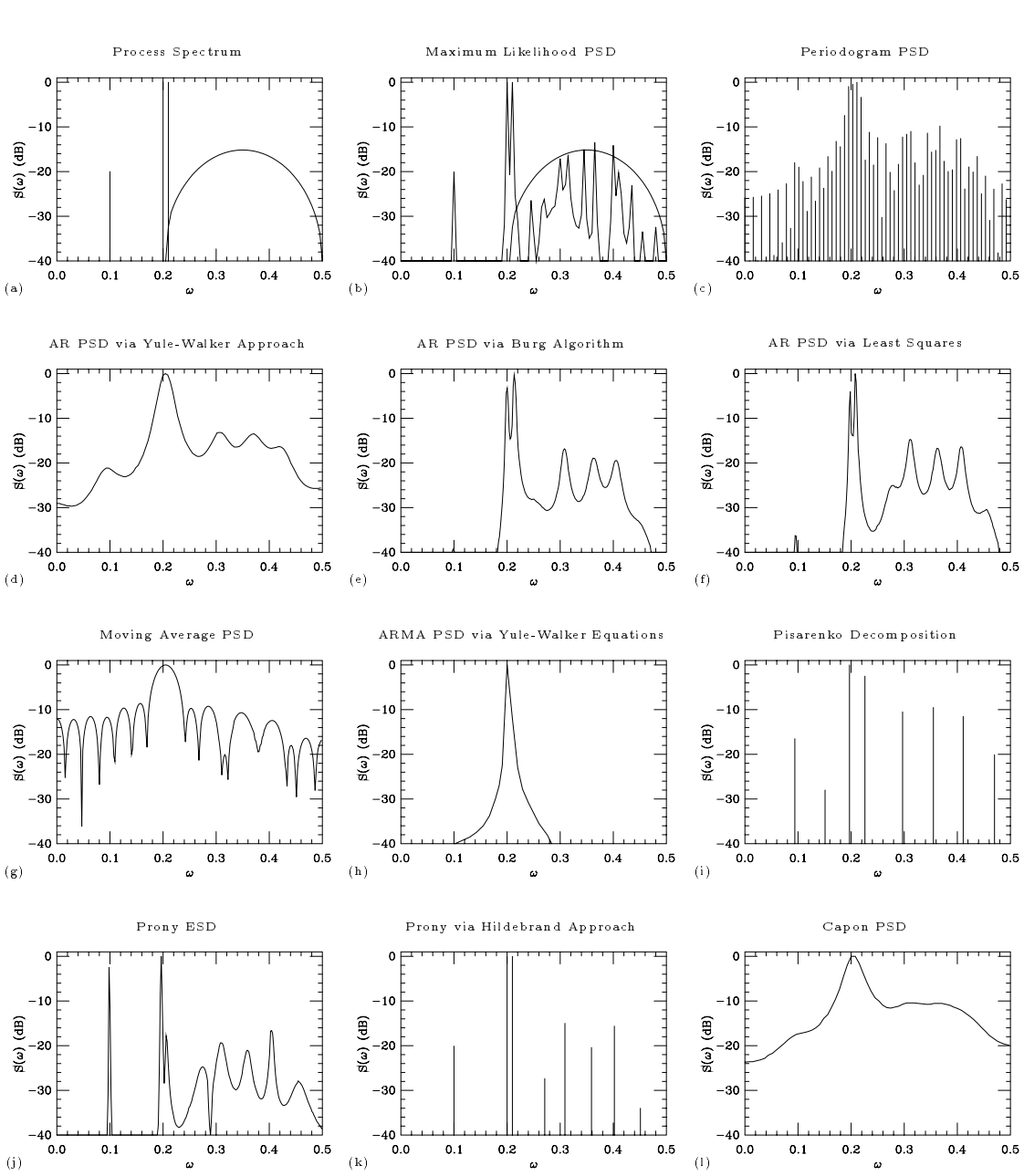

Figure 4: Comparison of spectrum estimators based on their spectrum estimates from a sample data record. The spectrum of the process is in panel *(a)*, and the MLE is in panel *(b)* with the broadband part of the process spectrum superimposed on it. These graphs are adapted from [10].

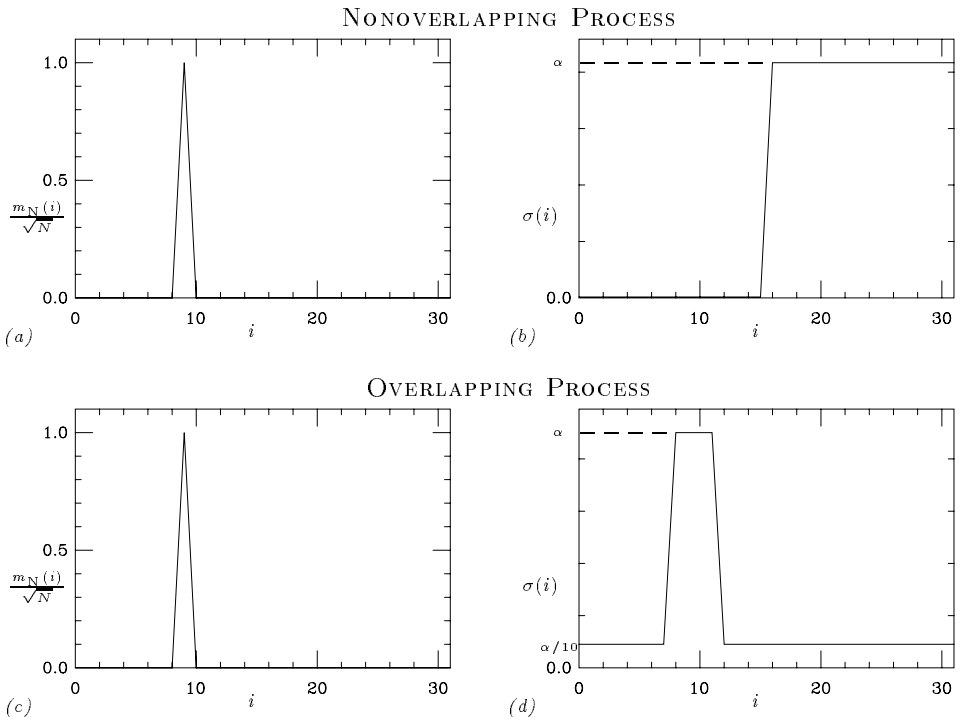

Figure 5: Frequency-domain means (left) and spectra (right) of the two processes examined in the simulation studies. SNR is controlled via the scaling parameter α.

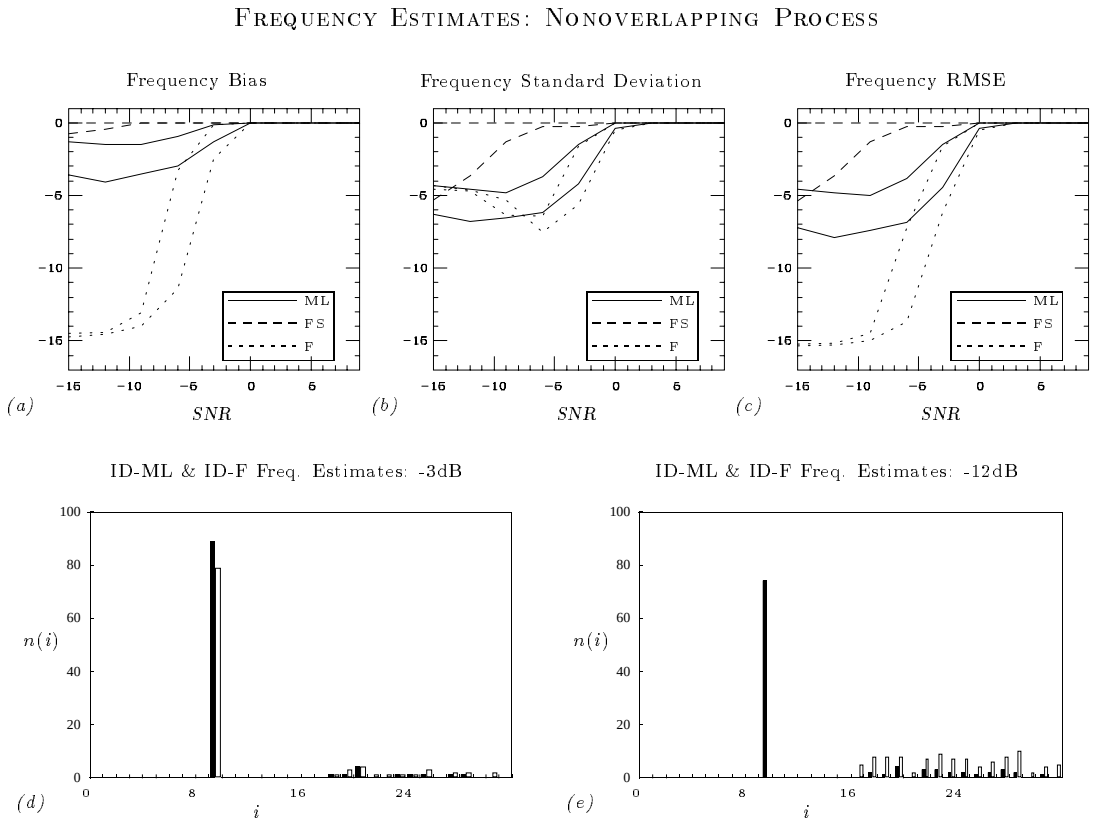

Figure 6: Nonoverlapping process frequency estimates. The upper panels show the negative of the bias, standard deviation, and RMSE of the frequency estimators over a range of SNR. Both complete- and incomplete-data estimators are plotted using the same line style; each CD estimator lies above the corresponding ID estimator. The lower panels show histograms based on 100 independent ID-ML (solid bars) and ID-F (open bars) estimates at -3dB and -12dB.

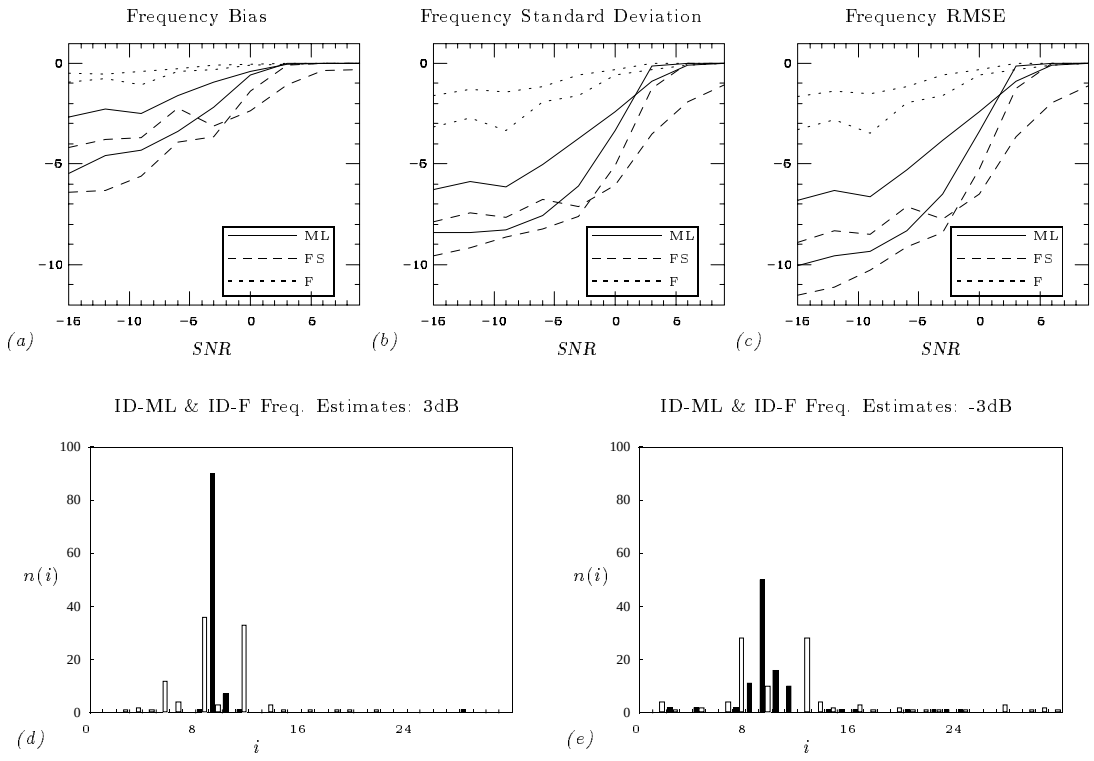

Figure 7: Overlapping process frequency estimates. The upper panels show the negative of the bias, standard deviation, and RMSE of the frequency estimators over a range of SNR. Both complete- and incomplete-data estimators are plotted using the same line style; each CD estimator lies above the corresponding ID estimator. The lower panels show histograms based on 100 independent ID-ML (solid bars) and ID-F (open bars) estimates at 3dB and -3dB.

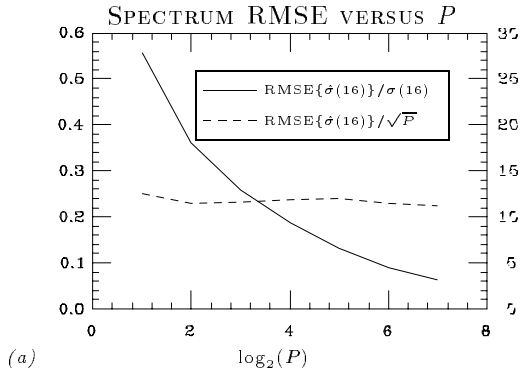

Figure 8: RMSE of $\hat{\sigma}_{\text{id-ml}}(16)$ versus P. The solid line (scale at left) is this quantity normalized by the process spectrum, and the dashed line (scale at right) is the RMSE divided $\sqrt{P}$.

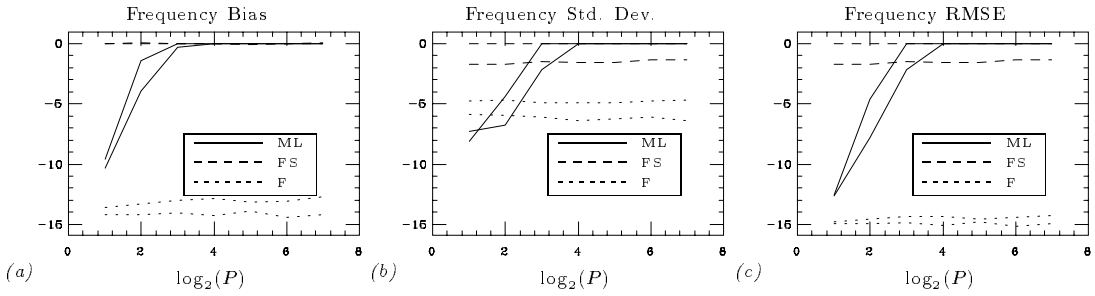

Figure 9: The negatives of bias, standard deviation, and RMSE of frequency estimates versus $\log_2(P)$. Both complete- and incomplete-data estimators are plotted using the same line style; each CD estimator lies above the corresponding ID estimator.