A Method of Sieves for Multiresolution Spectrum Estimation and Radar Imaging

Pierre Moulin, *Member, IEEE,* Joseph A. O'Sullivan, *Member, IEEE,* and Donald L. Snyder, *Fellow, IEEE*

Abstract—**A method of sieves using splines is proposed for regularizing maximum-likelihood estimates of power spectra. This method has several important properties, including the flexibility to be used at multiple resolution levels. The resolution level is defined in terms of the support of the polynomial B-splines used. Using a discrepancy measure derived from the Kullback–Leibler divergence of parameterized density functions, an expression for the optimal rate of growth of the sieve is derived. While the sieves may be defined on nonuniform grids, in the case of uniform grids the optimal sieve size corresponds to an optimal resolution. Iterative algorithms for obtaining the maximum-likelihood sieve estimates are derived. Applications to spectrum estimation and radar imaging are proposed.**

Index Terms—**Maximum-likelihood, EM algorithm, sieves, Kullback–Leibler information, splines, spectrum estimation, radar imaging.**

I. Introduction

RADAR images can be obtained by producing an estimate of the reflectivity of the target. Under certain physical assumptions, there exists a linear transformation from the reflectivity, expressed in delay and Doppler coordinates, to the observations. When the target surface is rough, the reflectivity can be modeled as an uncorrelated Gaussian random process [26]. The second-order statistics of the reflectivity are known as the *scattering function* of the target; the scattering function can be displayed as an image of the target. The imaging problem is to estimate the scattering function from the observations. In a one-dimensional version of the problem, the spectral density of a wide-sense stationary Gaussian random process is to be estimated from (noisy) observations.

For both the radar imaging and the spectrum estimation problems, the objective is to estimate the second-order statistics $\{S(u),\ u \in U \subset R^d\}$ of a complex, uncorrelated Gaussian random process $\{z(u),\ u \in U\}$ from a finite set of complex-valued observations $\{r(t),\ t \in T\}$. The function S is in $L^1(U)$ and is nonnegative and real. The observations are described by a linear transformation of $\{z(u),\ u \in U\}$, corrupted by an additive noise $\{n(t),\ t \in T\}$ with known statistics,

$$r(t) = \int_U G(t,u)\,z(u)\,du + n(t), \qquad t \in T, \quad (1.1a)$$

$$E\left[z(u)\,z^*(u')\right] = S(u)\,\delta(u-u'), \qquad u \in U. \quad (1.1b)$$

We view both the radar imaging and the spectrum estimation problems as statistical inference problems, and we use methods of statistical estimation theory to solve them, based on the model (1.1). Although taking advantage of the information available in the form of a statistical model appears to be a natural strategy, this approach does not appear frequently in the radar literature. Statistical methods in spectrum estimation are much more common, but they generally assume more information about the process than we do here (e.g., superposition of sinusoids, MA, AR and ARMA models [19], known correlation coefficients [5]).

The function S is the unknown parameter in the model (1.1) and can be estimated via the method of maximum-likelihood (ML). So far, this approach has received only marginal attention in the radar and spectrum estimation literatures. Capon's method for spectrum estimation is sometimes called an ML method, but this has been recognized to be a misnomer [19]. Other existing ML methods assume a special form for the spectral density [4], [19]. To our knowledge, the first use of ML for spectrum estimation in the general sense studied here is due to Chow and Grenander [7]. In radar imaging, research using this approach has been reported by Snyder *et al.* [25].

Since ML estimation of a whole function such as S from finite data is an ill-posed problem, regularization of the estimates is needed. We use Grenander's *method of sieves* [13], which offers a convenient, general framework for solving such estimation problems. With this method, estimates are sought over a subset of the parameter set. The size of the restricted set results from a tradeoff between stability of the estimates and accuracy of the representation. The sieve is a

Manuscript received February 19, 1991; revised August 22, 1991. This work was supported by Contract No. N00014-86-K-0370 from the Office of Naval Research and by the National Science Foundation Grant No. MIP-8722463. This work was presented in part at the 28th Allerton Conference, Urbana-Champaign, IL, September 1989; in part at the IEEE International Symposium on Information Theory, San Diego, CA, January 14–19, 1990; and in part at the IEEE International Symposium on Information Theory, Budapest, Hungary, June 24–28, 1991.

P. Moulin is with Bell Communications Research, MRE 2M-393, 445 South Street, Morristown, NJ 07960.

J. A. O'Sullivan and D. L. Snyder are with the Electronic Systems and Signals Research Laboratory, Department of Electrical Engineering, Washington University, St. Louis, MO 63130.

IEEE Log Number 9104352.

collection of all such subsets, indexed by a positive parameter μ called mesh size.

Some important problems arise regarding the design of the sieve. The estimation procedure should be tractable. It is also desired that the estimates be consistent, in the sense that a certain measure $d(S:\hat{S}^{(N)})$ of the estimation error for the estimator $\hat{S}^{(N)}$ tends to zero as the sample size N increases. The estimates obtained with the spectrum estimation method reported in [7] are consistent, but computational issues have not received detailed attention. Furthermore, the extension to more general problems of the type (1.1) is not considered. On the other hand, the estimation method studied by Snyder *et al.* [25] is applicable to problems of the class (1.1), but regularization issues are only partially addressed.

In this paper, we construct and analyze a sieve for problems of the class (1.1). Our method is based on a representation for the parameter-function S in terms of known, elementary functions $\{\psi_m(u),\ u \in U\}$ indexed by m in a countable set Λ. Each function in a sieve subset is a linear combination of Q elementary functions indexed by a subset Λ_Q of Λ.

$$S(u) = \sum_{m \in \Lambda_Q} a(m)\psi_m(u), \qquad u \in U. \tag{1.2}$$

The coefficients $\{a(m),\ m \in \Lambda_Q\}$ represent the function and are viewed as parameters for which ML estimates are sought, subject to nonnegativity constraints on S. If Q grows with N at an appropriate rate, consistent estimates can be obtained. A method for estimating the parameters is developed, based on an alternating maximization algorithm.

There is great flexibility in the design of the elementary functions, under the constraint that the functions satisfy certain technical conditions for convergence of the estimates in the parameter set. A case can be made for elementary functions that have local support. In this instance, the parameters describe the local behavior of S, and the complexity of the estimation algorithm is reduced significantly. The particular representation studied here is given in terms of polynomial B-splines. Since B-splines are nonnegative, the nonnegativity constraint on S can be enforced by simply requiring that the parameters $\{a(m),\ m \in \Lambda_Q\}$ be nonnegative. The latter constraint is easily incorporated in our estimation procedure.

The B-splines are defined over a partition $\{\boldsymbol{P}_m,\ m \in \Lambda_Q\}$ of $\boldsymbol{U}$. The resolution at which the estimate $\hat{S}$ is represented at a given $u \in \boldsymbol{U}$ is the size of $\boldsymbol{P}_m$ for $u \in \boldsymbol{P}_m$. In particular, for uniform partitions, the resolution is also uniform in $\boldsymbol{U}$ and is $O(Q^{-1})$. The mesh size of the sieve, defined as Q^{-1}, is a measure of resolution. Thus, ML estimates at different resolution levels can be produced by performing an estimation within our sieve for different mesh sizes.

We study convergence of the estimates as a function of the rate of growth of the sieve $Q(N)$. Convergence is defined with respect to a discrepancy measure derived from the Kullback–Leibler divergence of parameterized density functions. We obtain consistency results and show how these results can be interpreted in terms of the smoothness properties of S. The optimal rate of growth of the sieve is also determined.

Finally, we discuss certain algorithmic issues. In general, the numerical operations required for solving the ML equations are computationally expensive. However, under a large-sample approximation to the model, the complexity of the processing is greatly reduced, at insignificant cost for the statistical performance.

This paper is organized as follows. In Section II, we present the statistical model and introduce our notation. In Section III, the ML estimation problem is stated and the method of sieves introduced. In Section IV, we construct a sieve for the problem (1.1). In Section V, the relationship between the selected resolution level and the statistical performance of the estimator is examined and the optimal performance is derived. In Section VI, an expectation-maximization algorithm is given for computing the estimates. Simulations using this algorithm are presented in Section VII.

II. Model

Under physical assumptions detailed in [25], the radar echo is a linear superposition of the reflections off the target:

$$s_R(t) = \int_{-f_{\max}}^{f_{\max}} \int_0^{\tau_{\max}} e^{j2\pi f(t-\tau/2)} s_T(t-\tau)\, z(f,\tau)\, df\, d\tau, \qquad 0 \le t < T, \tag{2.1}$$

where $z(f, \tau)$ is the reflectivity viewed in delay (τ) and Doppler (f) coordinates and $s_T(t)$ is the complex envelope of the radar transmitted signal. In order to set up the model in a more general perspective, we define u as the pair of coordinates (f, τ), $\boldsymbol{U}$ as the domain of z, and $G(t, u)$ as the kernel of the linear transformation (2.1). The observations are a set of samples of the echo (2.1), corrupted by an additive noise w:

$$r(n\Delta t) = \int_U G(n\Delta t, u) z(u)\, du + w(n\Delta t), \qquad n = 0, \cdots, N-1, \tag{2.2}$$

where Δt is the sampling interval and N the number of samples. We adopt a *diffuse-target* model for the reflectivity. In this instance, z is modeled by a zero-mean, complex Gaussian random process with orthogonal random variables [25], [26], and its covariance takes the diagonal form

$$E\left[z(u) z^*(u') \right] = S(u)\delta(u - u'), \qquad u, u' \in \boldsymbol{U}. \tag{2.3}$$

In the radar context, the function S is known as the *scattering function* of the target and may be viewed as an image of the target. The additive noise $w(n\Delta t)$ is modeled as an independent Gaussian discrete-time random process, with zero-mean and known variance N_0. The discrete-time covariance matrix for the data is assumed to be (strictly) positive definite and is derived from (2.2) and (2.3):

$$K_r := \left[\int_U G(n\Delta t, u) G^*(m\Delta t, u) \cdot S(u)\, du + N_0 \delta_{nm} \right]_{0 \le n, m < N}, \tag{2.4}$$

where δ_{nm} is the Kronecker delta.

The model (2.2) and (2.4) also applies to spectrum estimation. In this instance, u is the frequency and z and S are known as the spectral process and the spectral density, respectively. In fact, this spectrum estimation problem is a special case of the radar imaging problem, where the target is concentrated in one dimension (f), and the transmitted signal is a constant equal to one.

We estimate S in $\boldsymbol{L}^1(\boldsymbol{U})$ or a subspace thereof from the model (2.2) and (2.4), under a nonnegativity constraint on S. $\boldsymbol{U}$ is a subset of $\boldsymbol{R}^d$ of the form $\boldsymbol{U}_1 \times \cdots \times \boldsymbol{U}_d$, where $\{\boldsymbol{U}_i = [u_{\min}(i), u_{\max}(i)],\ i = 1, \cdots, d\}$ are intervals on the real line. In spectrum estimation, the assumption of boundedness implies that the process is bandlimited; in radar imaging, that the target has finite extent. We call z the *process*, and S its *intensity function*.

Two quantities of interest in the following are the *ambiguity function*,

$$A_N(u, v) := (1/E_G) \sum_{n=0}^{N-1} G(n\Delta t, u) G^*(n\Delta t, v), \quad (2.5)$$

and the squared Hilbert–Schmidt norm of G,

$$E_G := \int_U (1/N) \sum_{n=0}^{N-1} |G(n\Delta t, u)|^2 \, du. \quad (2.6)$$

The technical results presented in this paper all rely on the following assumptions.

C1) $(1/N)|A_N(u, v)|^2$ is a resolution of the identity in $\boldsymbol{L}^1$, i.e., for all $f \in \boldsymbol{L}^1$, $\lim_{N\to\infty} \| f(\cdot) - \int_U f(u)(1/N)|A_N(u, \cdot)|^2 \, du \|_{L^1} = 0$.

C2) $A_N(u_k, u_l) = (N/|\boldsymbol{U}|)\delta_{kl}$ on a uniform grid of N cells in $\boldsymbol{U}$, where δ_{kl} is the Kronecker delta.

C3) E_G is finite and independent of N.

C4) If $N_0 = 0$, there exists $\epsilon > 0$ such that $\inf_U S \geq \epsilon$.

The first two assumptions imply that the ambiguity function is arbitrarily narrow for N large enough and are satisfied in spectrum estimation as well as in radar imaging, with practical transmitted signals [23, Section 3.4.1]. The fourth assumption establishes a restriction on the class of intensity functions, in the case of noise-free observations. Assumptions C1)–C4) are referred to as Condition C).

We use the following notation. Sets and spaces are in boldface. We denote by $\boldsymbol{S}_+$ the set of nonnegative functions in a space of functions $\boldsymbol{S}$. Let $u = (u(1), \cdots, u(d))$ and $v = (v(1), \cdots, v(d))$ be two elements of $\boldsymbol{R}^d$ or $\boldsymbol{Z}^d$. Then, we define:

$u \leq v$, if and only if $u(i) \leq v(i)$, $i = 1, \cdots, d$,
$u < v$, if and only if $u(i) < v(i)$, $i = 1, \cdots, d$,
$u + v = (u(1) + v(1), \cdots, u(d) + v(d))$,
$u \cdot v = (u(1)v(1), \cdots, u(d)v(d))$,
$|u| = \Pi_{i=1}^d |u(i)|$,
$0 = (0, \cdots, 0)$, $1 = (1, \cdots, 1)$and $\infty = (\infty, \cdots, \infty)$.

III. Maximum-Likelihood Estimation

A. ML Formulation

The probability density for the observations r is a complex, Gaussian density [12] parameterized by the covariance matrix K_r: $f(r:K_r) = \pi^{-N}(\det K_r)^{-1}\exp(-r^\dagger K_r^{-1} r)$ in which the super † denotes the Hermitian-transpose operator on matrices and vectors. The log likelihood functional for the intensity function S is defined as

$$l(S) := -\ln \det K_r - r^\dagger K_r^{-1} r, \quad (3.1)$$

where K_r is given in terms of S by (2.4), and the term not dependent on S has been discarded. Any maximizer of this functional is a ML estimate for S.

The solution to (3.1) involves the maximization of the likelihood with respect to a structure-constrained covariance matrix. Problems of this class have been studied by Burg *et al.* [6] at a certain level of generality. Applications have been found in the context of Toeplitz and circulant Toeplitz structure constraints [22] and in radar imaging [25].

B. Sieves

The estimation problem described above is ill posed, since the parameter-function S belongs to an infinite-dimensional space. Stable solutions to ill-posed estimation problems can be obtained using the *method of sieves* due to Grenander [13]. A *sieve in a parameter space* $\boldsymbol{A}$ is a family of subsets $\boldsymbol{S}(\mu)$ of $\boldsymbol{A}$ indexed by a positive parameter μ called the *mesh size*, chosen as a function of the sample size. The likelihood functional is maximized over the restricted set $\boldsymbol{S}(\mu)$. A sieve must satisfy two conditions.

S1) A restricted ML estimate exists over each set $\boldsymbol{S}(\mu)$.

S2) Any element of $\boldsymbol{A}$ can be approximated with arbitrary accuracy by an element of $\boldsymbol{S}(\mu)$, for μ small enough.

Under certain conditions, sieves can be designed to ensure consistency of the estimates. This generally requires that the mesh size μ of the sieve decrease at an adequate rate as the sample size increases. In this fashion, the sieve subsets are large enough to ensure a good representation of any function in $\boldsymbol{A}$, yet they are also small enough to guarantee stable estimates. In anticipation of the convergence analysis in Section V, we define a "distance" between elements of the parameter set.

C. Measure of the Estimation Error

Consider two probability density functions $f_\theta(x)$ and $f_{\theta'}(x)$ parameterized by θ and $\theta' \in \boldsymbol{A}$, $x \in \boldsymbol{R}^N$. A measure of the "distance" between θ and θ' is given by the average Kullback–Leibler information per sample:

$$\bar{d}^{(N)}(\theta : \theta') := (1/N) I(f_\theta : f_{\theta'}) = (1/N) \int f_\theta(x) \ln \frac{f_\theta(x)}{f_{\theta'}(x)} \, dx. \quad (3.2)$$

We assume that (3.2) converges to a limit $\bar{d}(\theta:\theta')$, at the rate $O(N^{-1})$, as N tends to infinity. The functional (3.2) is positive, nonsymmetric in its arguments, and does not satisfy the triangle inequality. It is thus not a distance. Functionals of this type are sometimes called *directed distances* [18]. In the context of estimation, θ' can be an estimator for the unknown parameter θ, based on the observation x, so the

left-hand side of (3.2) is also a random variable. We define a directed distance from θ to θ' which is independent of x,

$$d^{(N)}(\theta : \theta') := E\left[\bar{d}^{(N)}(\theta : \theta'(x))\right] = \int f_\theta(x)(1/N) I(f_\theta : f_{\theta'(x)})\, dx. \quad (3.3)$$

The following definition formalizes the concept of convergence under the information distance $\bar{d}$.

Definition 1: If for all $\epsilon > 0$ and $\theta \in A$ there exists $\theta^* \in \bigcup_\mu S(\mu)$ such that $\bar{d}(\theta : \theta^*) < \epsilon$, then the sieve is said to be dense in the information sense in A.

We use this definition by analogy to the classical definition of denseness in topological spaces, although here the directed distance does not induce a topology on A.

D. Application

Let the function S be the parameter θ in the previous discussion. In the following lemma, we give an expression for the distance $\bar{d}$ from S to another function S' in the parameter set.

Lemma 1: Under Condition C),

$$\bar{d}(S : S') = |U|^{-1} \int_U \left[-\ln \frac{S(u) + N_0/E_G}{S'(u) + N_0/E_G} - 1 + \frac{S(u) + N_0/E_G}{S'(u) + N_0/E_G} \right] du. \quad (3.4)$$

Outline of the proof: Denote by K'_r the covariance associated to S' by (2.4). The directed distance

$$\bar{d}^{(N)}(S : S') = \frac{1}{N} I(f_S : f_{S'}) = -\frac{1}{N} \ln \det K_r K_r'^{-1} - 1 + \frac{1}{N} \operatorname{tr}\left[K_r K_r'^{-1}\right],$$

sometimes called *likelihood loss* [3], is known to converge to (3.4) in the case of Toeplitz matrices, that is, when $G(t, u) = \exp(j2\pi tu)$ [14]. In [23, Section 6.4.1], these results have been extended to the larger class of functions G that satisfy Condition C). The rate of convergence is $O(N^{-1})$.

The information distance (3.4) is zero, if and only if $S' = S$ almost everywhere. In spectrum estimation, (3.4) is recognized as the *Itakura–Saito distance* between two spectral densities [16] when $N_0 = 0$. From a more general standpoint, the Itakura–Saito distance is also suitable for measuring the "distance" between two arbitrary positive functions, in a sense made precise by Csiszar [8]. Thus, (3.4) can be viewed as a *generalized Itakura–Saito distance* for measuring the distance between higher dimensional spectral densities in the presence of noise.

IV. Spline Sieve

In this section, we propose a sieve based on multiresolution concepts. The mesh size of the sieve is a measure of the resolution of the estimates. ML estimates at a desired resolution level are obtained by selecting the appropriate mesh size. The sieve is based on a representation for the unknown function in terms of known elementary functions, a classical technique used in nonparametric statistics.

A. Definition of a Sieve

The intensity function S belongs to the subset B_+ of some linear space $B \subset L^1(U)$. Define a countable index set Λ and let $\{\psi_m,\ m \in \Lambda\}$ be a set of (nonnegative) elementary functions in B_+. It is not required that this set be orthogonal. Consider the set A of all functions in the span of $\{\psi_m,\ m \in \Lambda\}$, with nonnegative coefficients:

$$A := \left\{ S \in B \mid S(u) = \sum_{m \in \Lambda} a(m)\psi_m(u),\quad a \geq 0 \right\}. \quad (4.1)$$

We define our parameter set to be A. The functions in this set are represented by means of a countable number of coefficients $\{a(m),\ m \in \Lambda\}$. Clearly A is a subset of B_+. As we shall see, $\{\psi_m,\ m \in \Lambda\}$ can be designed so that any function in B_+ can be approximated with arbitrary accuracy by a function in A under the information distance (3.4). We define a sieve $S(\Lambda_Q)$ on S in A by truncating the representation (4.1) to a finite number of terms Q,

$$S(\Lambda_Q) := \left\{ S \in B \mid S(u) = \sum_{m \in \Lambda_Q} a(m)\psi_m(u),\quad a \geq 0 \right\}, \quad (4.2)$$

where the index set Λ_Q has cardinality Q, and the union of all sets $\{\Lambda_Q,\ Q \in N\}$ is equal to Λ. The functions in the subset $S(\Lambda_Q)$ of A are parameterized by a finite number Q of coefficients $\{a(m),\ m \in \Lambda_Q\}$. In Section VI, we show how estimates of these coefficients can be produced numerically using an EM algorithm.

B. ML Estimator

The loglikelihood function for the parameters a is given by

$$l(a) = -\ln \det K_r - r^\dagger K_r^{-1} r. \quad (4.3)$$

From (2.4) and (4.2), the dependence of K_r on a is as follows,

$$K_r = \sum_{m \in \Lambda_Q} a(m) K_m + N_0 I_N, \quad (4.4)$$

where we have defined the $N \times N$ *basis covariance matrix*

$$K_m = \left(\int_U G(n\Delta t, u) G^*(p\Delta t, u) \psi_m(u)\, du \right)_{0 \leq n,p < N}, \quad (4.5)$$

and the $N \times N$ identity matrix I_N.

Anderson [2] has studied a similar problem of covariance matrix estimation in which the unknown covariance matrix is a linear combination of given matrices. This problem can be viewed as a special case of ours where it is known that the covariance is in a given subset of the sieve.

Conditions for a Maximizer: The loglikelihood (4.3) is to be maximized subject to nonnegativity constraints on the components of a. The necessary and sufficient conditions for some $\hat{a}$ to be a local maximum of the loglikelihood can be derived from the Kuhn–Tucker conditions [20].

Proposition 1 (Necessary Condition for a Local Maximum): A necessary condition for K_r to be a local maximum of the loglikelihood is

$$\operatorname{tr}\left[K_m K_r^{-1}\left(rr^\dagger - K_r\right)K_r^{-1}\right] + \mu(m) = 0, \qquad m \in \Lambda_Q, \tag{4.6}$$

where $\mu(m) = 0$ if $a(m) > 0$, and $\mu(m) \geq 0$ if $a(m) = 0$.

Proof: Equation (4.6) is just the necessary Kuhn–Tucker condition for maximizing a function subject to the inequality constraints $\{a(m) \geq 0,\ m \in \Lambda_Q\}$ [20]. The first term in the left-hand side of (4.6) is the gradient of the loglikelihood. Each Lagrange multiplier $\mu(m)$ is zero if the corresponding constraint is inactive, and positive otherwise. □

Note that (4.6) expresses that for every admissible variation δa of the parameter (i.e., $a + \delta a \geq 0$), the variation of the likelihood is nonpositive:

$$\sum_{m \in \Lambda_Q} \delta a(m) \operatorname{tr}\left[K_m K_r^{-1}\left(rr^\dagger - K_r\right)K_r^{-1}\right] \leq 0.$$

In the following, we refer to (4.6) as the *trace condition*. For an interior point, all Lagrange multipliers in (4.6) are zero. For the necessary condition (4.6) to be sufficient as well, the Hessian matrix of the loglikelihood needs to be negative definite.

C. Spline Representation

For the representation of functions in the sieve, we select a set of polynomial splines as elementary functions. This permits a good local representation of a function. Besides the standard arguments in favor of such a representation, it should be mentioned that this choice also leads to a considerable reduction in the complexity of the estimation algorithm, as discussed in Section VI.

We consider tensor-product (i.e., separable) polynomial B-splines of degree $L - 1$ in each coordinate. For a comprehensive study of B-splines and tensor-product B-splines, see for instance [24]. Tensor-product B-splines are constructed on $\boldsymbol{U}$ as follows. Let $m,\ M \in \boldsymbol{N}^d$ and define a partition $\boldsymbol{P}(i) := \{x_{i,m(i)},\ -L + 1 \leq m(i) < M(i) + L - 1\}$ on each coordinate axis $1 \leq i \leq d$, such that

$$x_{i,-L+1} \leq \cdots \leq x_{i,0} = u_{\min}(i) < x_{i,1} < \cdots < x_{i,M(i)-1} = u_{\max}(i) \leq x_{i,M(i)} \leq \cdots \leq x_{i,M(i)+L-2}.$$

Denote by $\{B_{i,m(i)}^{(L)}(u(i)),\ -L + 1 \leq m(i) < M(i)\}$ the normalized B-splines associated with each partition $\boldsymbol{P}(i)$, and by $\boldsymbol{P}$ the tensor product of the partitions $\boldsymbol{P}(i)$, $1 \leq i \leq d$. Thus, $\boldsymbol{P}$ is a separable partition and can be written in the form

$$\boldsymbol{P} := \left\{\boldsymbol{P}_m = \left[v_m, v_{m+1}\right],\ -L + 1 \leq m < M + L - 2\right\},$$

where $v_m := (x_{1,m(1)}, \cdots, x_{d,m(d)})$. The tensor-product B-splines defined over $\boldsymbol{P}$,

$$B_m^{(L)}(u) = \prod_{i=1}^{d} B_{i,m(i)}^{(L)}(u(i)), \qquad -L + 1 \leq m < M,\quad u \in \boldsymbol{U},$$

have support set $[v_m,\ v_{m+L}]$. The *overlapping factor*, defined as the number of elementary functions intersecting at any given point, is equal to L^d. The resolution of a spline function at some $u \in \boldsymbol{U}$ is the size of the partition cell to which u belongs. In two dimensions, for $L = 1$, $\{B_m^{(L)}(u)\}$ are nonoverlapping "boxes"; for $L = 2$ and $L = 3$, $\{B_m^{(L)}(u)\}$ are respectively bilinear and biquadratic functions. For a uniform partition, the splines are translates of the basic tensor-product spline $B^{(L)}(u)$ contracted at resolution level $\alpha(i)M(i)$ in each coordinate i, where $\alpha(i) := |u_{\max}(i) - u_{\min}(i)|^{-1}$,

$$B_m^{(L)}(u) = B^{(L)}\left(M.\alpha.(u - u_{\min}) - m\right).$$

The one-dimensional basic B-spline is simply the $(L - 1)$-fold convolution of the characteristic function of [0,1] with itself.

We define a sieve based on splines as follows. The functions in the sieve are tensor product polynomial spline functions. The degree $L - 1$ of the polynomial approximation is chosen according to the desired smoothness of the resulting function. Each subset in the sieve is indexed by the partition $\boldsymbol{P}$ that supports the spline,

$$S_{\boldsymbol{P}}^{(L)} := \left\{S \in \boldsymbol{B} \mid S(u) = \sum_{m=-L+1}^{M-1} a(m) B_m^{(L)}(u),\ a \geq 0\right\}. \tag{4.7}$$

The mesh size of the sieve can be defined as the mesh size of the partition, that is, the size of its largest cell. As the partition is made finer, the size of the sieve subset increases in the space of functions S. The number of elementary functions in the spline representation is $Q := \Pi_{i=1}^{d}(M(i) + L - 1)$. For a uniform partition, (4.7) takes the form

$$S_M^{(L)} := \left\{S \in \boldsymbol{B} \mid S(u) = \sum_{m=-L+1}^{M-1} a(m) B^{(L)} \cdot \left(M.\alpha.(u - u_{\min}) - m\right),\quad a \geq 0\right\}. \tag{4.8}$$

In this instance, the subsets of the sieve are simply indexed by M, and the mesh size is $\|\boldsymbol{P}\| = |\boldsymbol{U}| / |\boldsymbol{M}|$.

D. Conditions for a Sieve

We need to verify that the family of subsets (4.7) is a sieve in $\boldsymbol{B}_+$. In order to do so, conditions S1) and S2) of Section III-B must be satisfied. These conditions are addressed in Lemmas 2 and 3, respectively. The general result is established in Proposition 2.

Lemma 2: Consider the collection of vectors $\{G(n\Delta t, u),\ 0 \leq n < N\}$ indexed by $u \in \boldsymbol{U}$. Assume that any N vectors in this collection are linearly independent, except perhaps for a set of du-measure zero. Then the likelihood function is bounded above almost surely.

Proof: ML estimates of structure-constrained covariance matrices do not always exist. The likelihood function

can be unbounded above, implying that no maximizer exists. Fuhrmann and Miller [11] showed that a necessary and sufficient condition for the likelihood to be unbounded above is that there exist a singular covariance matrix in the set with the observation vector r in its range. Under the assumptions of the lemma, all basis covariance matrices K_m are positive definite. The only singular covariance in the sieve subset has all of its coefficients $\{a(m)\}$ equal to zero. The probability that r lies in the range space of this zero matrix is zero.

Lemma 3: The sieve $\bigcup_P S_P^{(L)}$ is dense in L_+^p, for $L = 1, 2, \cdots$ and $p \geq 1$.

Proof: See Appendix.

Any function in L_+^p can be approximated with arbitrary accuracy by a B-spline with nonnegative coefficients. To obtain the required accuracy it is necessary that the partition size be small enough. This result is well known in approximation theory when there are no constraints on the coefficients. In Lemma 3, it is shown that the conclusion holds under the constraint of nonnegative coefficients.

Corollary: The sieve $\bigcup_P S_P^{(L)}$ is dense in the information sense (3.4) in the set

$$L_\epsilon^1 := \{S \in L^1 \mid S(u) \geq \epsilon\}.$$

The parameter ϵ is an arbitrary positive number.If $N_0 = 0$, ϵ is allowed to be zero.

Proof: See Appendix.

Under the restriction C4) on the lower bound of the functions in the parameter set, S can be approximated with arbitrary accuracy in information by an element of the sieve.

Proposition 2: The collection of subsets (4.7) is a sieve in L_+^p for $L = 1, 2, \cdots$ and $p \geq 1$. This collection is also a sieve in L_ϵ^1 under the information distance.

Proof: From Lemma 2, the likelihood function is bounded above almost surely in each subset of the sieve. Furthermore when $a(m)$ tends to infinity for some m, the loglikelihood (4.3) tends to $-\infty$. These two properties imply the existence of a maximizer in each restricted set, with probability one. Condition S1) is thus satisfied. In Lemma 3 and its corollary, it is shown that condition S2) is satisfied too. □

E. Multiresolution Analysis

Let Ω be a subset of R^d and define $\Lambda_j = \{2^{-j}m \mid 2^{-j}m \in \Omega,\ j \in Z,\ m \in Z^d\}$. This set is a grid in Ω with mesh size 2^{-jd}. The union of all sets $\{\Lambda_j,\ j \in Z\}$, denoted by Λ, is dense in Ω. The spline functions

$$f(u) = \sum_{2^{-j}m \in \Lambda_j} a(m) B^{(L)}(2^j u - m), \qquad u \in U, \quad (4.9)$$

generate a vector space $V_j(\Omega)$, over which the L^2 inner product is introduced. The collection of functions $\{B^{(L)}(2^j u - m),\ 2^{-j}m \in \Lambda_j\}$ form a nonorthogonal basis for $V_j(\Omega)$. The spaces $\{V_j(\Omega),\ j \in Z\}$ form a *multiresolution analysis* of Ω [21].

For a *dyadic* resolution $M = 2^j$, the functions in the sieve set (4.8) have the form (4.9), up to a translation by $u_{\min}$ and a scaling by α. The sequence of sieve subsets forms a multiresolution analysis of U. The particular choice $M = 2^j$ is convenient when the same resolution is required in all coordinates and when the number of knots should be a power of 2. However, it is not required that $M = 2^j$ to obtain a good approximation with our method. It should also be noted that our method is not a multigrid method, in the sense that the estimates at a resolution level are not used in the estimation at another resolution level.

Nonuniform local resolution can be obtained by using a nonuniform grid for the sieve, as in (4.7). However, finding a statistical decision procedure that selects optimal, local resolution levels is still an open problem.

An interesting alternative to the nonuniform spline representation is the wavelet representation. In fact, wavelets can be constructed from spline spaces, as shown by Jaffard and Meyer [17]. Besides the problem of selecting an appropriate wavelet subset for the representation of the intensity function, another major difficulty would arise. The estimation of the wavelet coefficients is to be performed subject to the constraint that the resulting image be nonnegative. In general, this is a formidable computational problem involving a very large number—$|K|$—of constraints. The spline representation chosen here bypasses this difficulty, since the nonnegativity of S is easily enforced by nonnegativity constraints on Q spline coefficients.

V. Rate of Growth of the Sieve

This section is devoted to an analysis of the convergence and consistency properties of the sieve estimator. We determine what rates of growth of the sieve lead to consistent estimates and find optimal rates. Convergence and consistency are measured with respect to the information distance. The mesh size of the sieve is selected by minimizing the estimation error measured in this fashion.

A. Estimation Error Analysis

An asymptotic expression for the estimation error measure (3.3) is given for a general sieve problem. Assume that the subsets of the sieve are parameterized by a Q-vector parameter, as is the case for the spline sieve. Denote these subsets by S_Q, and the ML estimator in S_Q by $\hat{\theta}_Q$. Under some mild regularity conditions [15], for each (fixed) Q, $\hat{\theta}_Q$ converges to an element $\tilde{\theta}_Q$ of S_Q, as $N \to \infty$. The following two lemmas, which are derived from the properties of the directed distance (3.3), are powerful tools for analysis of the error. The proofs are a direct application of known results about the asymptotic χ^2-square distribution of the likelihood ratio of ML estimators [23, Section 6.3.2].

Lemma 4 (Error Decomposition):

$$d^{(N)}(\theta : \hat{\theta}_Q) \sim \bar{d}(\theta : \tilde{\theta}_Q) + \frac{Q}{2N} \qquad \text{as } N/Q \to \infty, \quad (5.1)$$

where the notation $f(N) \sim g(N)$ as $N \to \infty$ means that f *is asymptotic to* g *as* $N \to \infty$, i.e., $\lim_{N\to\infty} f(N)/g(N) = 1$.

Lemma 5: $\tilde{\theta}_Q$ achieves the infimum of $\bar{d}(\theta:\theta^*)$ over all $\theta^* \in S_Q$.

By application of Lemma 5, we simply denote $\bar{d}(\theta:\tilde{\theta}_Q)$ by $\bar{d}(\theta:S_Q)$. This quantity is the sieve approximation error and is independent of N. When Q tends to infinity, $\bar{d}(\theta:S_Q)$ tends to zero if the sieve is dense in the information sense in the parameter set. As seen in the corollary to Lemma 3, this condition holds for the spline sieve.

Lemma 4 shows that $d^{(N)}(\theta:\hat{\theta}_Q)$ can be decomposed in two terms. The first is the sieve approximation error previously discussed. The second is the estimation error within the sieve and is a penalty for high-order models. The simple trade-off (5.1) between these two terms motivated our adoption of (3.3) as a measure of the total estimation error. This decomposition is reminiscent of Akaike's criterion [1]. However, Akaike's selection of the model order is made from the data themselves, whereas in the sieve design problem this selection depends on the data only via the sample size.

B. Criterion for Mesh Size Selection

Definition 2: A sequence $Q(N)$ leads to estimates $\hat{\theta}_{Q,N}$ which are consistent in information if $\lim_{N\to\infty} d^{(N)}(\theta:\hat{\theta}_{Q,N}) = 0$.

Definition 3: A sequence $Q(N)$ is optimal in the information sense if, as $N \to \infty$, $d^{(N)}(\theta:\hat{\theta}_{Q,N}) \le d^{(N)}(\theta:\hat{\theta}_{Q',N})$ for all sequences $Q'(N)$.

The optimal convergence rate of the estimator is obtained by *minimizing the estimation error (5.1) with respect to* Q. We adopt this minimization procedure as a basic principle for selection of the rate of growth $Q(N)$ of the sieve.

C. Application to the Spline Sieve

Let the intensity function S play the role of the parameter θ in the previous analysis. We study the case of a uniform resolution and derive an expression for the directed distance from the true parameter S, to the ML estimator $\hat{S}$ in the sieve (4.8). By application of (5.1),

$$d^{(N)}(S:\hat{S}) \sim \bar{d}(S:S_M^{(L)}) + \frac{Q}{2N} \qquad \text{as } N/Q \to \infty, \tag{5.2}$$

where $Q = \Pi_{i=1}^{d}(M(i) + L - 1) \sim |M|$ as $M \to \infty$. The convergence rate (5.2) is to be determined for given resolutions $M(N)$. The optimal convergence rate depends on the smoothness properties of the true intensity function. Fast convergence can be expected if the intensity function is smooth and splines of appropriate smoothness are used to represent it. Saturation results apply, that is, increasing the smoothness of the spline beyond the degree of smoothness of the function does not improve convergence.

The smoothness of B-splines is determined by their order L. In the following analysis, the smoothness of the intensity function is characterized in terms of the d-dimensional Sobolev spaces L_q^2. Functions in L_q^2 as well as their partial derivatives (in the sense of distributions) up to order q belong to L^2.

Proposition 3 gives an upper bound on the convergence rate for $\hat{S}$, as M tends to infinity componentwise. This upper bound is obtained by performing an analysis of the spline approximation error (3.4), in the case of uniformly spaced knots. The bound is tight in the sense that the constants involved in the final expression can be improved upon, but not the rate of convergence. The quantities

$$R^{(L)}(i) := C_L |U|^{-1} \int_U \left| \frac{|U_i|^L \partial^L S/\partial u(i)^L}{S(u) + N_0/E_G} \right|^2 du, \qquad i = 1, \cdots, d \tag{5.3}$$

that appear in (5.4) are independent of N and of M. We call them the Lth *roughness indexes* of S in their respective coordinates.

Proposition 3: If Condition C) is satisfied and if $S \in L_q^2$, then, for splines of order $L \le q$, the convergence rate of $\hat{S}$ is upper-bounded by

$$\hat{d}^{(N)}(S:\hat{S}) \sim \frac{1}{2}\left[\sum_{i=1}^{d} \frac{R^{(L)}(i)}{M(i)^{2L}} + |M|/N\right] \qquad \text{as } M \to \infty \text{ and } N/|M| \to \infty, \tag{5.4}$$

and the estimator is consistent in information, if and only if $M \to \infty$ and $|M| = o(N)$. For splines of order $L > q$, there is saturation of the performance, i.e., $\hat{d}^{(N)}(S:\hat{S})$ is given by (5.4) in which $L = q$.

Proof: See Appendix.

We have established an upper bound for the convergence rate for $\hat{S}$. This expression can now be minimized to determine an approximation to the optimal rate of growth of the sieve.

Proposition 4: Under the assumptions of Proposition 3, the optimal convergence rate of $\hat{S}$, for a spline of order $L \le q$, is upper-bounded by

$$\hat{d}^{(N)}(S:\hat{S}) = AN^{-2L/(2L+d)}, \tag{5.5}$$

where $A = \frac{1}{2}\bar{R}^{d/(2L+d)}[d(2L)^{-2L/(2L+d)} + (2L)^{d/(2L+d)}]$ is an upper bound on the optimal coefficient. The optimal rate of growth of the sieve is approximated by minimizing (5.4) over M:

$$M_{\text{opt}} = BN^{1/(2L+d)}, \tag{5.6}$$

with $B(i) = (2L)^{1/(2L+d)}(R^{(L)}(i)\bar{R}^{-d/(2L+d)})^{1/2L}$, $i = 1, \cdots, d$, and $\bar{R} = (\Pi_{i=1}^{d} R^{(L)}(i))^{1/d}$. Increasing spline orders L lead to better convergence rates, with saturation for $L \ge q$.

Proof: The optimal rate of growth of the sieve is obtained by minimizing $d^{(N)}(S:\hat{S})$ over M, for fixed N. Letting M be equal to its optimal value yields (5.5). The convergence rate, $O(N^{-2L/(2L+d)})$, $L \le q$, improves with increasing L. It follows that the optimal spline order is q. □

The results presented in Propositions 3 and 4 indicate that the estimates obtained with our sieve are consistent in information. The optimal convergence rate is obtained when the order of the spline is equal to the degree of smoothness of the

intensity function. Smoother intensity functions allow better convergence rates.

The approach described in this section makes use of a priori information on the smoothness properties of the intensity function. Even if such information is not available, consistent estimates can be obtained by letting $M = o(N^{1/d})$. Naturally, the convergence rates achieved in this fashion are not as good.

VI. Discrete Processing

A. Discrete Model

For the purpose of a digital implementation of the estimation algorithm, we consider a discrete approximation to the model (2.2). Assume that the process is discrete in u. Define a uniform sampling $\{u_k,\ 0 \le k < K\}$, $K \in N^d$, of U, let $U_k := [u_k u_{k+1}]$, $0 \le k < K$, and define the *discrete process*

$$c(k) := |U_k|^{-1/2} \int_{U_k} z(u)\, du, \qquad 0 \le k < K. \quad (6.1)$$

Next, define N-vectors r and w with nth entries equal to $r(n\Delta t)$ and $w(n\Delta t)$, respectively. Also define a vector c with entries indexed by k: $c := \{c(k),\ 0 \le k < K\}$. This vector has length $|K|$. Similarly, let $|U_k|^{1/2} G(n\Delta t, u_k)$ be the (n, k) entry of a $N \times |K|$ matrix $\Gamma^\dagger$. Approximating the integral (2.2) with a Riemann sum, we obtain

$$r = \Gamma^\dagger c + w. \quad (6.2)$$

The statistical model for the discrete problem is as follows. $\{c(k),\ 0 \le k < K\}$ defined in (6.1) is a zero-mean, orthogonal Gaussian random process with diagonal covariance

$$\Sigma(k, k') = E\big[c(k)c^*(k')\big] = \sigma^2(k)\delta_{kk'}, \qquad 0 \le k, k' < K, \quad (6.3)$$

where $\{\sigma^2(k),\ 0 \le k < K\}$ is the *discrete intensity function*. From (2.3), (6.1), and (6.3), each sample of the discrete intensity function is given by $\sigma^2(k) = |U_k|^{-1} \int_{U_k} S(u)\, du$ and is an approximation to $S(u_k)$. The discrete approximation to $r(\cdot)$ converges to the stochastic integral (2.2) in the mean-square sense as K tends to infinity. The covariance matrix for the vector r is derived from (6.2) and (6.3):

$$K_r = \Gamma^\dagger \Sigma \Gamma + N_0 I_N. \quad (6.4)$$

Equations (6.2) and (6.4) define the statistical model in the discrete case.

In the discrete version of the sieve estimation problem, ML estimates of the unknown parameters a in (4.4) are sought with the basis covariance matrices given by a discrete version of (4.5):

$$K_m = \Gamma^\dagger \Psi_m \Gamma, \qquad m \in \Lambda_Q, \quad (6.5)$$

where $\Psi_m = \operatorname{diag}\{\psi_m(u_k)\}_{0 \le k < K}$.

For fixed N, K_m in (4.5) is the limit of (6.5) as $K \to \infty$. It is thus equivalent to maximize the loglikelihood for the continuous or the discrete model, in the limit as $K \to \infty$. These results are consistent with the asymptotic equivalence of the two models previously mentioned. The existence of a ML estimator in each sieve subset can be proven by extending the proof of Lemma 2 to the discrete case [23, Section 7.6.2].

B. EM Algorithm

The trace condition (4.6) is a nonlinear equation in the Q unknown parameters a. In general, it cannot be solved in closed form, and a numerical method must be used. The EM algorithm of Dempster, Laird and Rubin [10] is an alternating maximization algorithm, based on the concept of complete and incomplete data spaces, that yields a sequence of estimates having nondecreasing likelihood. We propose the following EM algorithm, based on the discrete model (6.2). In the formulation of this algorithm, it is not required that the elementary functions be splines. Write the discrete process c as the sum of Q independent processes: $c = \sum_{m \in \Lambda_Q} c_m$, where c_m is a $|K|$-vector, sample of a zero-mean Gaussian process with diagonal covariance $a(m)\Psi_m$. The support set of the elementary function $\psi_m(u_k)$ in the k-domain is denoted by D_m. We define the complete data as $(\{c_m,\ m \in \Lambda_Q\};\ w)$. Because the Q processes $\{c_m\}$ are independent, the loglikelihood for the complete data is simply

$$\begin{aligned} l_{cd}(a) &:= \ln p(c : a) \\ &= \sum_{m \in \Lambda_Q} \ln p(c_m : a) = -\sum_{m \in \Lambda_Q} \ln \det\big(a(m)\Psi_m\big) \\ &\quad - \sum_{m \in \Lambda_Q} c_m^\dagger \big(a(m)\Psi_m\big)^{-1} c_m. \end{aligned}$$

Maximizing the conditional expectation of $l_{cd}(a)$ with respect to $a(m)$ at step p of the algorithm, we obtain the updated estimate for the following step:

$$\hat{a}(m)^{(p+1)} = \frac{1}{|D_m|} \sum_{k \in D_m} \frac{E\big[|c_m(k)|^2 \,|\, r, \hat{a}^{(p)}\big]}{\psi_m(u_k)}. \quad (6.6)$$

As shown in the Appendix, after evaluating the conditional expectation of $|c_m(k)|^2$ in (6.6), we obtain the update equations at stage p of the algorithm:

$$K_r^{(p)} = \sum_{j \in \Lambda_Q} \hat{a}(j)^{(p)} K_j + N_0 I_N \quad (6.7)$$

$$Z^{(p)} = K_r^{(p)-1}\big(rr^\dagger - K_r^{(p)}\big)K_r^{(p)-1}, \quad (6.8)$$

$$\hat{a}(m)^{(p+1)} = \hat{a}(m)^{(p)} + \frac{1}{|D_m|}\big(\hat{a}(m)^{(p)}\big)^2 \operatorname{Tr}\big[K_m Z^{(p)}\big], \qquad m \in \Lambda_Q. \quad (6.9)$$

As appears from (6.6), the estimates are guaranteed to be nonnegative. The trace appearing in the update equation (6.9) is the mth component of the gradient of the loglikelihood; therefore, the stable points of the algorithm satisfy the necessary Kuhn–Tucker conditions of Proposition 1. The complexity of each step of the algorithm is $N^3 + QN^2$.

A fast version of this algorithm can be obtained for the special discretization $|K| = N$, for which an orthogonalization of the observations exists under Condition C). It can be

shown that the resulting estimator is consistent in information and even offers the same convergence rate as the estimator for the true model [23, Section 8]. In fact, this special discrete approximation introduces an error $O(N^{-1})$ under the information distance—see Lemma 1. This error is dominated by the estimation error, which is $O(N^{-2q/(2q+d)})$.

C. Estimation Algorithm, $|K| = N$

When condition C) holds, K_r, like K_m, can be written as the product of three square matrices, the middle one being diagonal:

$$K_r = \Gamma^\dagger(\Sigma + N_o/E_G I_N)\Gamma. \tag{6.10}$$

Thus, the change of variables

$$p := (\Gamma^\dagger)^{-1} r, \tag{6.11}$$

defines an orthogonalization of the observations[1]. Using (6.10), the first update equation (6.7) becomes

$$\Sigma^{(p)} = \sum_{j\in\Lambda_Q} \hat{a}(j)^{(p)} \Psi_j. \tag{6.12}$$

Next, introducing the expressions (6.5), (6.8), and (6.10) in the update equation (6.9), we obtain for the trace in the right-hand side

$$\mathrm{Tr}\left[K_m K_r^{(p)-1}\left(rr^\dagger - K_r^{(p)}\right)K_r^{(p)-1}\right] = \sum_{k\in D_m} \frac{\psi_m(u_k)\left(|p(k)|^2 - \sigma^2(k)^{(p)} - N_0/E_G\right)}{\left(\sigma^2(k)^{(p)} + N_0/E_G\right)^2}. \tag{6.13}$$

Another update equation is obtained from (6.9) and (6.13):

$$\hat{a}(m)^{(p+1)} = \hat{a}(m)^{(p)} + \frac{1}{|D_m|}\left(\hat{a}(m)^{(p)}\right)^2 \cdot \sum_{k\in D_m} \frac{\psi_m(u_k)}{\left(\sigma^2(k)^{(p)} + N_0/E_G\right)^2} \times \left(|p(k)|^2 - \sigma^2(k)^{(p)} - N_0/E_G\right), \quad m\in\Lambda_Q. \tag{6.14}$$

There is a special instance of the discrete model for which closed-form expressions can be derived. When the elementary functions are indicator functions over their support, we obtain

$$\hat{a}(m) = \max\left\{\frac{1}{|D_m|}\sum_{k\in D_m} |p(k)|^2 - N_0/E_G,\ 0\right\}, \quad m\in\Lambda_Q.$$

Complexity: In general, the complexity of the preprocessing (6.11) is N^2. If Γ defines a d-dimensional Fourier transform from c to r, as occurs in spectrum estimation or in the radar problem when a stepped-frequency waveform is used as a transmitted signal [27], the complexity of the preprocessing is only $N \log_2 N$.

The number of additions and multiplications to be performed in (6.12) and (6.14) is equal to $\Sigma_{m\in\Lambda_Q} |D_m|$. For the spline representation, the overlapping factor of the elementary functions is equal to L^d, so that $\Sigma_{m\in\Lambda_Q} |D_m| = L^d |K| = L^d N$. In this instance, the complexity of the update equations is simply $O(N)$ and the proportionality constant is the overlapping factor of the splines.

From the previous discussion, the global complexity of the EM algorithm requiring N_{EM} iterations for convergence is $N^2 + N_{\mathrm{EM}} L^d N$ or $N \log_2 N + N_{\mathrm{EM}} L^d N$ if Γ is a discrete Fourier transform.

VII. Simulation Results

In order to visualize the performance of the multiresolution estimation algorithm, we have performed a set of simulations for a phantom target in a radar imaging problem. This target is a sphere rotating about a point along the line of sight. The scattering function for this simple object can be calculated [26]. A discrete version ($|K| = 128 \times 128$ pixels) of this image is displayed in Fig. 1(a). Independent samples of the discrete reflectivity process are produced using a Gaussian random number generator. Each of these samples is zero-mean and has variance equal to the corresponding sample of the scattering function image. The transmitted signal is a stepped-frequency waveform [27]. The return echo is generated from the radar equation (6.2), with $N = |K| = 128^2$. The data are corrupted by an additive noise with variance equal to 60. The maximal intensity of the scattering function is equal to 300 and its mean value to 7.

In Fig. 1(b), we show the conventional estimate $|p|^2$ of the scattering function. The background noise is relatively important, making it difficult to identify the contour of the target. In Fig. 1(c)–(g), we show a multiresolution estimation obtained with the algorithm described in Section VI-C. The images are bilinear splines defined on a uniform grid and estimated at increasing resolution levels, $M(1) = M(2) = 8$, 16, 32, 64, and 128. These figures show the trade-off between resolution and estimation accuracy. For the case $M(1) = M(2) = 8$, the resolution is unacceptably coarse. At the other extreme, $M(1) = M(2) = 128$, the discrete model is no longer regularized.

VIII. Conclusion

We have investigated the class (1.1) of estimation problems, with application to radar imaging and spectrum estimation. The problem is to find ML estimates of an unknown, nonnegative intensity function. This problem is ill-posed; we have constructed stable estimates using a method of sieves. Our approach is based on a spline representation for the unknown function. We have proposed a tractable algorithm for estimating the function subject to nonnegativity constraints. Estimates can be produced at different resolution levels, so the method offers a capability for multiresolution ML estimation.

[1] The squared magnitude of p is the periodogram in spectrum estimation [19], and is analogous to the output of the conventional processing in radar imaging [27].

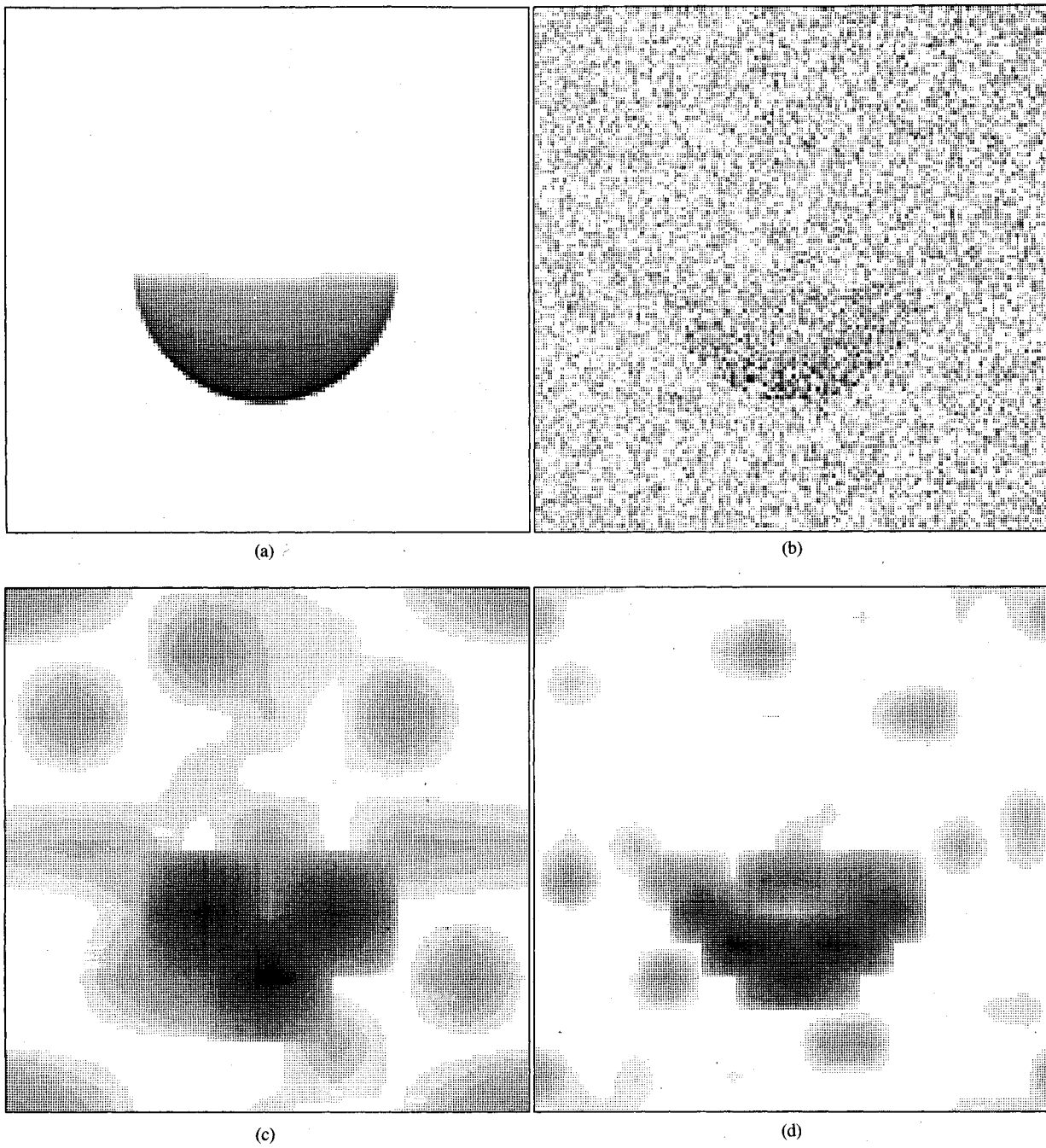

Fig. 1. Multiresolution estimation. (a) Scattering function for the rotating sphere. (b) Conventional estimate. (c) $M(1) = M(2) = 8$. (d) $M(1) = M(2) = 16$.

We have addressed the problem of selection of the mesh size of the sieve for the important class of splines defined on uniform grids. The mesh size is a measure of the resolution of the estimates. Our starting point is a measure of discrepancy between the parameter-function and its estimate. This measure is derived from the Kullback–Leibler information between the probability measures associated with each function. For the class of sieves considered here, the estimation error for the sieve-constrained maximum-likelihood estimator can be decomposed into two terms, the first being a measure of the distance between the parameter and the sieve subset, and the second a penalty term, as shown by (5.2). Minimizing the total error yields a criterion which determines the optimal rate of growth of the sieve. This criterion is expected to find applications in a broad class of estimation problems where the unknown parameter is a whole function.

Our estimation method is also applicable to splines defined on nonuniform grids. In this instance, the resolution is allowed to vary spatially. How the local resolution should be selected is still an open problem. The development of a

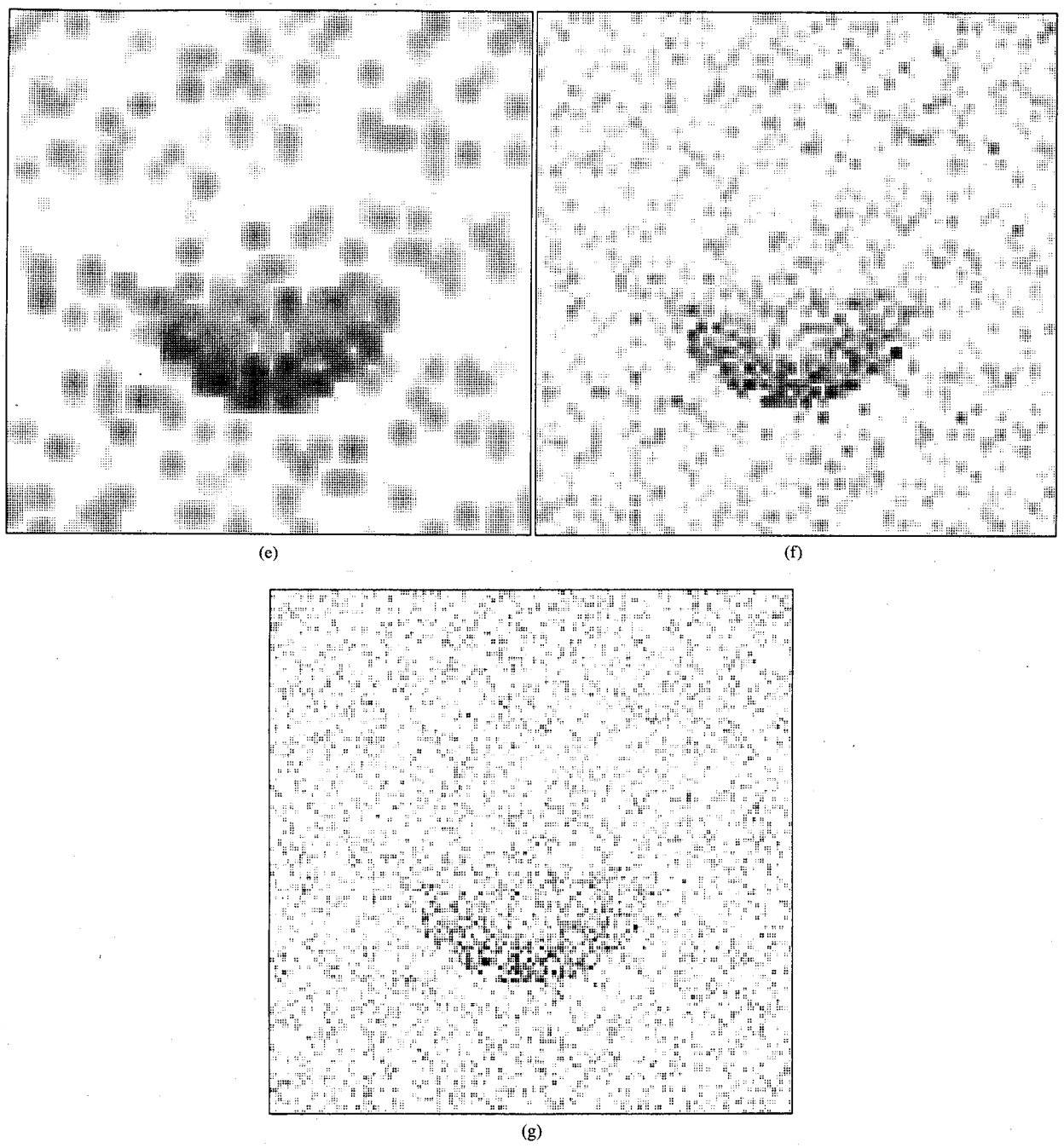

Fig. 1. Multiresolution estimation. (e) $M(1) = M(2) = 32$. (f) $M(1) = M(2) = 64$. (g) $M(1) = M(2) = 128$.

systematic decision mechanism for the local resolution represents in our view one of the most exciting challenges for expanding our work.

APPENDIX

A. Proof of Lemma 3

Since the linear combinations of tensor products of functions in $\{L_+^p(U_i),\ 1 \le i < d\}$ with nonnegative coefficients are dense in L_+^p, the proof for the case $d = 1$ carries immediately to higher dimensions. Let $L^p := L^p(U)$ and $L_+^p := L_+^p(U)$.

It is sufficient to show that the set of spline functions defined on uniform partitions is dense in L_+^p. For a uniform partition of the u-axis with mesh size $\Delta u = |U|/M$, the spline subsets are given by

$$S_M^{(L)} = \left\{ S \mid S(u) = \sum_{m=-L+1}^{M-1} a(m) B^{(L)} \cdot \left(M(u - u_{\min})/\Delta u - m\right),\quad a \ge 0 \right\},$$

where $B^{(L)}(u)$ are the basic B-splines. First we prove that $\bigcup_M S_M^{(L)}$ is dense in $L_+^p \cap C^1$, where C^1 denotes the space of continuously differentiable functions. For every S in $L_+^p \cap C^1$ and every $\epsilon > 0$, we construct an approximation $\bar{S}_L$ in $S_M^{(L)}$ such that $\|S - \bar{S}_L\|_{L^p}$

$< \epsilon$. This approximation has only positive coefficients. Define $u_m := u_{\min} + m\Delta u$ and

$$\bar{S}_L(u) = \sum_{m=-L+1}^{M-1} a(m) B^{(L)}(M(u-u_{\min})/\Delta u - m),$$
$$\text{with } a(m) := \begin{cases} S(u_m) & : m \geq 0 \\ S(u_0) & : m < 0. \end{cases} \quad \text{(A.1)}$$

Then,

$$\|S - \bar{S}_L\|_{L^p} = \left(\int_U |S - \bar{S}_L|^p \, du \right)^{1/p}$$
$$= \left(\sum_{m=0}^{M-1} \int_{u_m}^{u_{m+1}} |S - \bar{S}_L|^p \, du \right)^{1/p}. \quad \text{(A.2)}$$

We use the following property of B-splines: The value of a B-spline function on a partition interval is bounded by the size of the L coefficients "nearby" [9, Section 11]:

$$\min\{a(m-L+1), \cdots, a(m)\} \leq \bar{S}_L(u)$$
$$\leq \max\{a(m-L+1), \cdots, a(m)\}, \; u_m \leq u < u_{m+1}.$$

From this property, there exists an integer m^* in $[m-L+1, m]$ such that for u in $[u_m, u_{m+1}]$,

$$|S(u) - \bar{S}_L(u)|$$
$$\leq |S(u) - a(m^*)|$$
$$\leq \int_{u_{m-L+1}}^{u_{m+1}} |S'(u)| \, du$$
$$\leq (L\Delta u)^{1-1/p} \left(\int_{u_{m-L+1}}^{u_{m+1}} |S'(u)|^p \, du \right)^{1/p},$$

by application of Holder's inequality. From (A.2), it follows that

$$\|S - \bar{S}_L\|_{L^p} \leq (L\Delta u)^{1-1/p} \cdot \left(\sum_{m=0}^{M-1} \Delta u \int_{u_{m-L+1}}^{u_{m+1}} |S'(u)|^p \, du \right)^{1/p}$$
$$\leq L\Delta u \|S'\|_{L^p} = L\frac{|U|}{M} \|S'\|_{L^p}.$$

This equation shows that for every $\epsilon > 0$, there exist M large enough and an approximation $\bar{S}_L \in S_M^{(L)}$ such that

$$\|S - \bar{S}_L\|_{L^p} < \epsilon. \quad \text{(A.3)}$$

Hence the sieve is dense in $L_+^p \cap C^1$. Since every function in L_+^p can be approximated arbitrarily well in the L^p-norm by a function in $L_+^p \cap C^1$, the sieve is also dense in L_+^p. □

B. Proof of Corollary to Lemma 3

First, we derive an inequality between the information distance and the L^1 norm for any two functions S_1 and S_2 in L_ϵ^1. By definition,

$$\bar{d}(S_1 : S_2) = |U|^{-1} \int_U \left[-\ln \frac{S_1(u) + N_0/E_G}{S_2(u) + N_0/E_G} - 1 + \frac{S_1(u) + N_0/E_G}{S_2(u) + N_0/E_G} \right] du$$
$$\leq |U|^{-1} \int_U \left[\left| -\ln \frac{S_1(u) + N_0/E_G}{S_2(u) + N_0/E_G} \right| + \left| \frac{S_1(u) - S_2(u)}{S_2(u) + N_0/E_G} \right| \right] du.$$

From Taylor's theorem, there exists some $S^*(u)$ such that

$$\ln[S_1(u) + N_0/E_G] = \ln[S_2(u) + N_0/E_G] + \frac{S_1(u) - S^*(u)}{S_2(u) + N_0/E_G},$$

with $|S_1(u) - S^*(u)| \leq |S_1(u) - S_2(u)|$. Hence,

$$\bar{d}(S_1 : S_2) \leq 2|U|^{-1} \int_U \frac{|S_1(u) - S_2(u)|}{S_2(u) + N_0/E_G} \, du$$
$$\leq \frac{2|U|^{-1}}{\epsilon + N_0/E_G} \|S_1 - S_2\|_{L^1}. \quad \text{(A.4)}$$

Next, for every S in L_ϵ^1 and every $\eta > 0$, we construct an approximation $\bar{S}$ in $S_M^{(L)}$ such that $\bar{d}(S:\bar{S}) < \eta$. Our approximation is given by (A.1) with $L = 1$ and belongs to L_ϵ^1. By application of (A.3), M can be chosen large enough so that $\|S - \bar{S}\|_{L^1} < \eta(\epsilon + N_0/E_G)|U|/2$, for any given η. From (A.4), this implies that $\bar{d}(S:\bar{S}) < \eta$. □

C. Proof of Proposition 3

The function $\phi(x) := -\ln x - 1 + x$ that appears in (3.4) attains its minimum at $x = 1$. Its Taylor series expansion as $x \to 1$ is given by $\phi(x) \sim \frac{1}{2}(x-1)^2 + O((x-1)^3)$.

By application of Lemma 5 and the corollary to Lemma 3, there exists M large enough so that the infinite-sample ML estimate $\tilde{S}_M$ in $S_M^{(L)}$ approximates S arbitrarily well in information. Then, applying the Taylor series expansion for $\phi(x)$ to (3.4), we obtain

$$\bar{d}(S : \tilde{S}_M) \sim \frac{1}{2}|U|^{-1} \int_U \left[\frac{S(u) - \tilde{S}_M(u)}{\tilde{S}_M(u) + N_0/E_G} \right]^2 du$$
$$\sim \frac{1}{2}|U|^{-1} \int_U \left[\frac{S(u) - \tilde{S}_M(u)}{S(u) + N_0/E_G} \right]^2 du$$
$$\sim \frac{1}{2}|U|^{-1} \sum_{m \in \Lambda_Q} |S(v_m) + N_0/E_G|^{-2} \cdot \int_{P_m} |S(u) - \tilde{S}_M(u)|^2 \, du, \quad \text{as } M \to \infty. \quad \text{(A.5)}$$

Since expressions for the best approximant $\tilde{S}_M$ are not available, our approach is to compute an upper bound on (A.5) instead. Such a bound is obtained by evaluating the error $d(S:\bar{S})$ that occurs for some "reasonable" spline approximant $\bar{S}$. For all $S \in L_q^2$ and all $L \leq q$, there exists a Lth-order spline approximation $\bar{S}$ of S that satisfies the following property. For all $m \in \Lambda_Q$, the squared approximation error

$$\int_{P_m} |S(u) - \bar{S}(u)|^2 \, du \quad \text{(A.6)}$$

is bounded by [24, Section 12.3]

$$\frac{C_L}{d} \left\{ \sum_{i=1}^{d} (\Delta u_i)^L \left(\int_{P_m} |\partial_i^L S|^2 \, du \right)^{1/2} \right\}^2, \quad \text{(A.7)}$$

where C_L is a constant depending only on L, $\partial_i^L S := \partial^L S / \partial u(i)^L$, $i = 1, \cdots, d$, and $\Delta u_i := |U_i|/M(i)$, $i = 1, \cdots, d$, are the mesh sizes along the axes. If $L > q$, (A.6) is still upper-bounded by (A.7), but with q in place of L. Therefore, we consider only the

case $L \le q$. Using the inequality $(\sum_{i=1}^{d} x_i)^2 \le d \sum_{i=1}^{d} x_i^2$, we now bound (A.7) by

$$C_L \sum_{i=1}^{d} (\Delta u_i)^{2L} \int_{P_m} |\partial_i^L S|^2 \, du.$$

Thus the directed distance (A.5) is upper-bounded by

$$\begin{aligned}\hat{d}(S:\tilde{S}_m) &\sim \frac{1}{2} C_L |U|^{-1} \sum_{m \in \Lambda_Q} |S(v_m) + N_0/E_G|^{-2} \\ &\quad \sum_{i=1}^{d} (\Delta u_i)^{2L} \int_{P_m} |\partial_i^L S|^2 \, du \\ &\sim \frac{1}{2} C_L |U|^{-1} \sum_{m \in \Lambda_Q} \sum_{i=1}^{d} (\Delta u_i)^{2L} \\ &\quad \cdot \int_{P_m} \left| \frac{\partial_i^L S}{S(u) + N_0/E_G} \right|^2 du \\ &\sim \frac{1}{2} \sum_{i=1}^{d} \frac{R_i^{(L)}}{M(i)^{2L}} \qquad \text{as } M \to \infty, \end{aligned} \tag{A.8}$$

where we have introduced the roughness indices (5.3). Then (5.4) follows directly from (A.8) and (5.2). □

D. Derivation of the EM Equations

Define $\hat{c}_m^{(p)} := E[c_m \mid r, \hat{a}^{(p)}]$. The conditional expectation in (6.6) is the kth diagonal element of the matrix

$$\begin{aligned} &E\left[c_m c_m^\dagger \mid r, \hat{a}^{(p)}\right] \\ &\quad = E\left[\left(c_m - \hat{c}_m^{(p)}\right)\left(c_m - \hat{c}_m^{(p)}\right)^\dagger \mid r, \hat{a}^{(p)}\right] \\ &\qquad + \hat{c}_m^{(p)} \hat{c}_m^{(p)\dagger} = K_{c_m c_m}^{(p)} - K_{c_m r}^{(p)} K_r^{(p)-1} K_{r c_m}^{(p)} \\ &\qquad + \left(K_{c_m r}^{(p)} K_r^{(p)-1} r\right)\left(K_{c_m r}^{(p)} K_r^{(p)-1} r\right)^\dagger, \end{aligned} \tag{A.9}$$

where we denote by K_{xy} the conditional correlation $E[xy^\dagger \mid \hat{a}^{(p)}]$ of two random vectors x and y. Now, the expectations are evaluated from the model (6.2), taking into account the independence of the c_m's. We obtain

$$\begin{aligned} K_{c_m r}^{(p)} &= E\left[c_m r^\dagger \mid \hat{a}^{(p)}\right] = \hat{a}(m)^{(p)} \Psi_m \Gamma, \\ K_r^{(p)} &= E\left[r r^\dagger \mid \hat{a}^{(p)}\right] = \sum_{j \in \Lambda_Q} \hat{a}(j)^{(p)} \Gamma^\dagger \Psi_j \Gamma + N_0 I_N, \\ K_{c_m c_m}^{(p)} &= E\left[c_m c_m^\dagger \mid \hat{a}^{(p)}\right] = \hat{a}(m)^{(p)} \Psi_m. \end{aligned}$$

Taking the kth diagonal element of (A.9) and substituting in (6.6) yields (6.8) and (6.9) after some algebraic manipulations.

References

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. 19, pp. 716–723, June 1974.

[2] T. W. Anderson, "Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices," in *Essays in Probability and Statistics*, R. C. Bose *et al.*, Eds. Chapel Hill, NC: Univ. of North Carolina Press, 1970, vol. 1, pp. 1–24.

[3] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: Wiley, 1984.

[4] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Lecture Notes in Statistics. New York: Springer Verlag, 1988.

[5] J. P. Burg, "Maximum-entropy spectral analysis," presented at *Proc. 37th Meeting Soc. Exploration Geophysicists*, Oklahoma City, OK, Oct. 1967.

[6] J.P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices," *Proc. IEEE*, vol. 70, pp. 963–974, Sept. 1982.

[7] Y. Chow and U. Grenander, "A sieve method for the spectral density," *Ann. Statist.*, vol. 13, no. 3, pp. 998–1010, 1985.

[8] I. Csiszar, "Why least-squares and maximum-entropy?," *Math. Inst. Hungarian Academy of Sci.*, preprint no. 19/1989, to appear in *Ann. Statist.*

[9] C. deBoor, *A Practical Guide to Splines*. New York: Springer-Verlag, 1978.

[10] A. D. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. B39, no. 1, pp. 1–37, 1977.

[11] D. R. Fuhrmann and M. I. Miller, "On the existence of positive-definite maximum-likelihood estimates of structured covariance matrices," *IEEE Trans. Inform. Theory*, vol. 34, pp. 722–729, July 1988.

[12] N. R. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution (An introduction)," *Ann. Math. Statist.*, pp. 152–177, 1963.

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

[14] U. Grenander and G. Szego, *Toeplitz Forms and Their Applications*. New York: Chelsea, 1984.

[15] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions," *Proc. 5th Berkeley Symp. Math. Statist. and Probab.* Los Angeles, CA: Univ. of California Press, vol. 1, 1967, pp. 221–233.

[16] F. Itakura and S. Saito, "Analysis-synthesis telephony based on the maximum likelihood method," presented at *Proc. 6th Int. Conf. Acoust. C*, Tokyo, Japan, 1968, pp. 17–20.

[17] S. Jaffard and Y. Meyer, "Bases d'ondelettes dans des ouverts de R^n," *J. Mathematiques Pures et Appliquees*, vol. 68, pp. 95–108, 1989.

[18] L. K. Jones, "Approximation-theoretic derivation of logarithmic entropy principles for inverse problems and unique extension of the maximum entropy method to incorporate prior knowledge," *SIAM J. Appl. Math.*, vol. 49, no. 2, pp. 650–661, 1989.

[19] S. M. Kay and S. L. Marple, "Spectrum estimation—A modern perspective," *Proc. IEEE*, vol. 69, pp. 1380–1419, Nov. 1981.

[20] D. L. Luenberger, *Introduction to Linear and Nonlinear Programming*. Reading, MA: Addison Wesley, 1973.

[21] S. G. Mallat, "Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$," *Trans. Amer. Math. Soc.*, vol. 315, no. 1, pp. 69–87, 1989.

[22] M. I. Miller and D. L. Snyder, "The role of likelihood and entropy in incomplete-data problems: Applications to estimating point-process intensities and Toeplitz constrained covariances," *Proc. IEEE*, vol. 75, pp. 892–907, July 1987.

[23] P. Moulin, "A method of sieves for radar imaging and spectrum estimation," D.Sc. dissert., Washington Univ., St. Louis, MO, 1990.

[24] L. L. Schumaker, *Splines Functions: Basic Theory*. New York: Wiley, 1981.

[25] 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. Inform. Theory*, vol. 35, pp. 536–548, May 1989.

[26] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part III*. New York: Wiley, 1971.

[27] D. R. Wehner, *High Resolution Radar*. Norwood, MA: Artech House, 1987.