DIGITAL SIGNAL PROCESSING 4, 2–20 (1994)

Least Squares, Modeling, and Signal Processing*

James A. Cadzow

Department of Electrical Engineering, Vanderbilt University, Nashville, Tennessee 37235

Cadzow, J. A., Least Squares, Modeling, and Signal Processing, *Digital Signal Processing* 4 (1994), 2–20.

The concept of *least squares* (LS) as applied to an inconsistent system of linear or nonlinear equations is a fundamental tool in numerical analysis. Furthermore, these techniques have been applied with much success in solving many of the more challenging problems found in signal processing. In the standard LS problem, one seeks a choice for the vector $\underline{x}$ governing the inconsistent system of linear equations $A\underline{x} \approx \underline{y}$ so that these equations are best approximated in the least-squares error sense. In this paper, the concepts underlying an LS solution approach are presented in a tutorial fashion and only a basic knowledge of Euclidean spaces R^n and C^n is presumed. Much of the analysis is related to a linear system of equations where use of the fundamental fact that all Hermitian matrices have a full set of pairwise orthonormal eigenvectors plays a central role. Once the basic LS solution characterization of a linear system of equations has been made, a statistical analysis of this solution is undertaken. Conditions under which the LS solution is particularly sensitive to additive noise are established. This sensitivity can be decreased by using the concept of *reduced rank* approximation where a trade-off between estimation bias and estimation variance is made. The notion of linear least-squares error is then generalized to consider the case whereby the inconsistent system of linear equations $A(\underline{\theta})\,\underline{x} = \underline{y}$ has a system matrix $A(\underline{\theta})$ that depends on a set of real parameters $\underline{\theta}$. It is now desired to select both $\underline{x}$ and $\underline{\theta}$ so as to obtain a best approximate solution. This is shown to lead to a modified LS solution. An important extension of this problem is next made whereby the multiple inconsistent system of linear equations $A(\underline{\theta})\,\underline{x}_k = \underline{y}_k$ for $1 \leq k \leq N$ is best approximated in the least-squares error sense. These concepts are then further generalized to include the task of finding an LS solution to a system of nonlinear equations described by $\underline{F}(\underline{x}, \underline{\theta}) = \underline{y}$.

* This work was sponsored in part by the SDIO/IST and managed by the Office of Naval Research under Grant N00014-92-J-1995.

1051-2004/94 $5.00

1. INTRODUCTION

In a class of problems of interdisciplinary interest, it is desired to estimate values assumed by a set of *primary variables* which provide a salient characterization of a dynamic process (i.e., a system) under investigation. Due to the nature of the process, however, it is often not feasible to directly measure these primary variables. A means for overcoming this dilemma is to measure a set of *auxiliary variables* and then employ a known functional relationship between the primary and auxiliary variables to generate estimates of the primary variables. For example, this approach is frequently invoked by the physician when making a medical diagnosis, by the economist when providing an economic forecast, the meteorologist in making weather forecasts, and the signal processor in performing filtering or deconvolution operations. In the first three of these examples the auxiliary and primary variables correspond to parameters associated with the underlying system, while in the last example the variables correspond to number sequences (i.e., a time series or discrete-time signal). Whatever the case, the effectiveness of this approach is predicated on the correctness of the hypothesized functional relationship between the auxiliary and primary variables.

The above philosophical approach is now mathematically formulated whereby the set of measurably m auxiliary variables are taken to be the components of the $m \times 1$ *observed vector* designated by $\underline{y}$. Similarly, the n primary variables which cannot be directly measured are taken to form the components of the $n \times 1$ *unobserved vector* designated by $\underline{x}$. Although the auxiliary and primary variables are real-valued in most practical applications, in order to study the cases of real- and complex-valued variables in a single setting it is convenient to interpret these variables as being complex-valued so that $\underline{x} \in C^n$ and $\underline{y} \in C^m$. Since real numbers form a subset of the complex num-

bers, the theoretical results to follow are straightforwardly modified to the real data case. This typically entails removing any complex conjugate operations that may appear.

Linear Models

In the most general modeling application, the functional relationship between the observed and unobserved vectors is of a nonlinear nature and takes the form $\underline{y} = f(\underline{x})$ where $f(\ ): C^n \rightarrow C^m$. This expression represents a system of m generally nonlinear equations in n variables. Over the past 2 centuries, a great deal of interest has been directed toward the important special case of a system of linear equations. The eminent mathematician C. F. Gauss laid much of the groundwork in this area, and his works have had a great impact on contemporary numerical analysis and signal processing [9]. This effort has been continued by numerous mathematicians and scientists, reflecting the importance of linear models in quantitative disciplines (e.g., see Refs. [1,2,10,14]). One of the primary purposes of this paper is to put some of the more important results of these studies as they apply to signal processing into a common format. Although the algebraic background needed for this study is minimal, the reader interested in a more in-depth treatment has available many excellent sources (e.g., [11,12,13,18]).

When invoking a linear model, the observed and unobserved vectors are related in the *linear fashion:*

$$\underline{y} = \sum_{k=1}^{n} x(k)\underline{a}_k = A\underline{x} \tag{1}$$

In particular, the observed vector is taken to be a linear combination of the n mode (or signal) vectors $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n$, each contained in C^m that characterize the linear operation with the elements $x(k)$ servicing as the weights (or amplitudes) associated with these modes. The *system matrix* $A \in C^{m \times n}$ appearing in relationship (1) has as its columns the mode vectors, while the $x(k)$ weights associated with these modes form the components of the *unobserved vector* $x \in C^n$.[1]

In a standard application, it is desired to determine whether an unobserved vector $\underline{x}$ exists so that $A\underline{x} = \underline{y}$. The system of linear equations is said to be *consistent* if such an unobserved vector exists, and the vector $\underline{x}$ is called a *solution.* When no such vector exists, the system of linear equations is said to be *inconsistent.* Inconsistency can arise from measurement noise whereby an inaccurate value for $\underline{y}$ is used, or, when the given linear model does not accurately represent the functional relationship between the observed and unobserved vectors. Furthermore, the linear system of equations is said to be *overdetermined* when $m > n$ and to be *underdetermined* when $m < n$. In order to obtain useful estimates of the unobserved variables, it is generally good practice to use significantly more measurements than unknowns (i.e., $m > n$). This leads to a system of overdetermined linear equations which are typically inconsistent. With this in mind, our primary interest is directed toward analyzing systems of inconsistent-overdetermined linear equations.

Least-Squares Error Approximate Solution

The central consideration when investigating the inconsistent system of linear equations $A\underline{x} \approx \underline{y}$ is that of selecting $\underline{x}$ so that $A\underline{x}$ most closely resembles $\underline{y}$. The phrase *most closely resembles* is purposely made vague so as to encompass a variety of legitimate closeness measures. For the purpose of this paper, however, we restrict the measure of closeness to the squared *Euclidean norm* of the error vector $A\underline{x} - \underline{y}$. This measure is commonly referred to as the sum of squared error criterion, as formally defined by

$$\begin{aligned} f(\underline{x}) &= \|A\underline{x} - \underline{y}\|^2 \\ &= (\underline{x}^*A^* - \underline{y}^*)(A\underline{x} - \underline{y}) \\ &= \underline{x}^*A^*A\underline{x} - \underline{x}^*A^*\underline{y} - \underline{y}^*A\underline{x} + \underline{y}^*\underline{y}, \end{aligned} \tag{2}$$

where $\|\underline{x}\| = \sqrt{\underline{x}^*\underline{x}}$ designates the standard *Euclidean norm* of vector $\underline{x}$ and the asterisk symbol (*) complex conjugate transposition. The task at hand is then that of selecting the vector $\underline{x}$ so as to minimize this Euclidean norm criterion.[2]

A basic theorem from calculus indicates that a necessary condition for the vector $\underline{x}$ to minimize squared error criterion $f(\underline{x})$ is that the derivative of this criterion with respect to the real and imaginary components of vector $\underline{x}$ be equal to zero. Setting these deriva-

[1] The more general case of a multiple system of linear equations as specified by $Y = AX$ where X and Y are matrices can be studied in fashion similar to that here presented.

[2] No loss of generality is incurred by using the unweighted squared error criterion (2). If the weighted squared error criterion $f(\underline{x}) = (\underline{x}^*A^* - \underline{y}^*)W(A\underline{x} - \underline{y})$ had instead been employed where W is a Hermitian positive definite weighting matrix, a simple transformation converts this to the standard unweighted criterion. In particular, the weighting matrix is first factored as $W = Q^*Q$ and the substitutions $\tilde{\underline{y}} = Q\underline{y}$ and $\tilde{A} = QA$ yields the equivalent unweighted criterion $f(\underline{x}) = (\underline{x}^*\tilde{A}^* - \tilde{\underline{y}}^*)(\tilde{A}\underline{x} - \tilde{\underline{y}})$.

tives equal to zero is found to yield the following consistent system of linear *normal equations*:

$$A^*A\hat{\underline{x}} = A^*\underline{y}. \tag{3}$$

Any solution to these normal equations results in a least-squares solution. Furthermore, since matrix A^*A comprising the quadratic term in $f(\underline{x})$ is positive-semidefinite, it follows that this necessary condition is also sufficient. The $n \times n$ matrix A^*A characterizing these normal equations is commonly referred to as the *Gram* matrix associated with system matrix A. Upon substitution of any solution $\hat{\underline{x}}$ to these normal equations into squared error criterion (2), the criterion's minimum value is found to be

$$f(\underline{x}) = \underline{y}^*\underline{y} - \hat{\underline{x}}^*A^*A\hat{\underline{x}}. \tag{4}$$

It is to be noted that if the original system of linear equations has a solution then $A\hat{\underline{x}} = \underline{y}$ and this criterion's value is equal to zero. On the other hand, this criterion's minimum value will be positive if the original system of linear equations is inconsistent system.

Range Space, Null Space, and Solution Set

In order to gain an understanding for the more subtle features of a system of linear equations, we now examine the notions of range space, null space, and solution sets. As suggested by relationship (1), we may interpret vector $A\underline{x}$ as being a linear combination of the column vectors of matrix A in which the elements of vector $\underline{x}$ serve as the coefficients of this linear combination. It therefore follows that this system of linear equations has a solution if and only if the vector $\underline{y}$ is expressible as a linear combination of these column vectors. It is useful to formally capture this notion through the *range space* associated with system matrix A as formally specified by

$$\mathcal{R}(A) = \{\underline{y} \in C^m : \text{there exists an } \underline{x} \in C^n \text{ such that } A\underline{x} = \underline{y}\}. \tag{5}$$

The range space is therefore composed of all vectors $\underline{y} \in C^m$ for which a solution to the system of linear equations $A\underline{x} = \underline{y}$ exists. It is a simple matter to establish that the range space is a subspace of C^m with the dimension of this subspace being equal to the largest number of linearly independent column (or row) vectors comprising system matrix A. This dimension is commonly referred to as the *rank* of the system matrix.

The *null space* associated with the system matrix also provides another useful concept in characterizing the basic nature of a linear system of equations. Formally, the null space is defined as

$$\mathcal{N}(A) = \{\underline{x} \in C^n : A\underline{x} = \underline{0}\} \tag{6}$$

and consists of all vectors contained in C^n which map into the zero vector under matrix operation A. It is apparent that the null space is a subspace of C^n. Furthermore, since matrix A has rank r, it follows that each row vector of matrix A is expressible as a linear combination of r linearly independent vectors. Thus, the dimension of the null space must be $n - r$. It is a simple matter to establish the fact that if the system of linear equations $A\underline{x} = \underline{y}$ is consistent then the set of all solutions is contained in the linear variety as specified by

$$S(\underline{y}, A) = \underline{x}_p + \mathcal{N}(A) \quad \text{where} \quad A\underline{x}_p = \underline{y}. \tag{7}$$

In this expression, $\underline{x}_p$ designates any specific solution to the given consistent system of linear equations. Any solution is therefore equal to the sum of a particular solution (i.e., $\underline{x}_p$) and a homogeneous solution (i.e., a vector contained in $\mathcal{N}(A)$). This is a simple restatement of the *superposition theorem* that characterizes linear operators. Upon examination of this solution set, it is clear that the solution to a consistent system of linear equations is unique if and only if the null space is composed of only the zero vector (i.e., $r = n$).

EXAMPLE 1. A commonly occurring problem in signal processing is that of characterizing the basic nature of a time series given a finite set of samples $x(1), x(2), \ldots, x(N)$ of that time series. When invoking the concept of linear prediction to solve this problem, it is postulated that every element of the time series is expressible as a weighted sum of its most p immediate preceding elements, that is,

$$x(k) = a_1x(k-1) + a_2x(k-2) + \cdots + x(k-n) \quad \text{for} \quad n+1 \leq k \leq N. \tag{8}$$

The interval $n + 1 \leq k \leq N$ here appearing is dictated by the fact that the values of the time-series elements appearing in this relationship are known only in this interval (i.e., $x(k)$ is unknown for $k \notin [1, N]$). The $\{a_k\}$ parameters are commonly referred to as prediction coefficients since relationship (8) implies that one can perfectly predict a time-series element from knowledge of its immediate n most recent values. We may express this system of linear prediction equations in the matrix form

$$A\underline{x} = \underline{y}, \tag{9}$$

where

$$A = \begin{bmatrix} x(n) & x(n-1) & \cdots & x(1) \\ x(n+1) & x(n) & \cdots & x(2) \\ \vdots & \vdots & \cdots & \vdots \\ x(N-1) & x(N-2) & \cdots & x(N-n) \end{bmatrix},$$

$$\underline{y} = \begin{bmatrix} x(n+1) \\ x(n+2) \\ \vdots \\ x(N) \end{bmatrix}, \quad \underline{x} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}. \quad (10)$$

It is to be noted that this system of linear prediction equations will have a solution in the overdetermined case $N - n > n$ only if the given data is perfectly expressible as a linear combination of exponential signals of order n or less. In most applications, however, this is not the case, and a prediction coefficient vector $\underline{x}$ is sought for which this system of equations is best approximated. In the next section the issue of obtaining a best approximate solution is addressed.

2. FUNDAMENTAL THEOREM

As indicated in the last section, one of the first issues to be addressed when studying a system of linear equations is to determine whether or not a solution exists. It is seen that this task is equivalent to determining if $\underline{y} \in \mathcal{R}(A)$. Depending on the answer to this question, we then need to either find a solution(s) or find a useful approximate solution(s). To determine whether $\underline{y} \in \mathcal{R}(A)$, it is useful to generate a set of vectors (i.e., a basis) that spans the same subspace as do the column vectors of system matrix A. Under the assumption that A has rank r where $r \leq \min(m, n)$, it follows that any such basis must be composed of exactly r linearly independent vectors. Let a specific basis selection be given by

$$\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_r. \quad (11)$$

Although these basis vectors could be comprised of any r linearly independent column vectors of system matrix A, we do not so restrict the basis choice. It therefore follows that any vector contained in the range space can be expressed either as a linear combination of the n column vectors of the system matrix A or as a linear combination of the r vectors comprising this basis. We express this observation as

$$\mathcal{R}(A) = [\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n] = [\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_r], \quad (12)$$

where the bracket notation $[\cdot]$ represents the subspace composed of all linear combinations of the vectors enclosed within the brackets. It is important to note that there exist many nontrivial different bases that span the range space and the choice made is generally unimportant when a standard least squares solution is sought. When employing noise reduction adaptations of the least squares solution procedure, however, the choice of the particular basis used can be critical. We shortly address this important selection process.

The basis vectors (11) to be employed are next used to form the column vectors of the $m \times r$ *basis matrix* as specified by

$$B = [\underline{b}_1 \ \vdots \ \underline{b}_2 \ \vdots \ \cdots \ \vdots \ \underline{b}_r]. \quad (13)$$

Since the subspace spanned by the column vectors of matrices A and B are identical, any vector expressible as a linear combination of the column vectors of A is also expressible as a linear combination of the column vectors of matrix B. Since each column vector of A is expressible as a linear combination of the $\{\underline{b}_k\}$ basis vectors, it follows that there exists an $r \times n$ full rank matrix R such that

$$A = BR. \quad (14)$$

The system matrix has therefore been decomposed into the product of the basis matrix B and the *coefficient matrix* R. The required coefficient matrix is readily obtained by first left multiplying each side of expression (14) by B^*. Each side of this matrix identity is then left multiplied by the inverse of the full rank matrix B^*B to obtain the required expression

$$R = [B^*B]^{-1}B^*A. \quad (15)$$

It is noted that the commonly used QR decomposition of a matrix is a special case of decomposition (14). In a QR decomposition, the B matrix is composed of orthonormal vectors, while R is a nonsingular upper triangular matrix. A QR decomposition is realized by performing a Gram–Schmidt orthogonalization on the column vectors of matrix A.

Orthogonal Projection Matrix

It is important to note that there exist an uncountable number of distinct basis matrices B that span the same subspace as does system matrix A. Associated with each of these basis matrices is a unique companion coefficient matrix R as specified by relationship (15). For each such matrix pair, the product BR is equal to the underlying system matrix A. Depending on the nature of the problem at hand, however, some choices of the basis matrix are superior to others. Fur-

thermore, upon substituting matrix relationship (15) into decomposition (14), the following fundamental matrix identity is established:

$$A = P_A A. \tag{16}$$

Here the $m \times m$ matrix P_A here appearing is specified by

$$P_A = B[B^*B]^{-1}B^*. \tag{17}$$

P_A is an *orthogonal projection matrix* since it possesses the two prerequisite properties of being Hermitian (i.e., $P_A = P_A^*$) and idempotent (i.e., $P_A^2 = P_A$). Furthermore, since B is full rank, it follows that the range space of this orthogonal projection matrix is identical with the range space of system matrix A. Thus, P_A is the unique orthogonal projection matrix with range space $\mathcal{R}(A)$. The subscript A has been appended to P_A to explicitly recognize this property. It is important to reemphasize that although there exists an uncountable infinite number of different B basis matrices associated with system matrix A, each will generate the same orthogonal projection matrix P_A as specified by relationship (17).

Orthogonal Decomposition

Perhaps the most useful feature of an orthogonal projection matrix is the ability that it provides in decomposing a vector into the sum of two vectors that are orthogonal. It is recalled that the standard *inner product* between any two vectors in C^m is defined by

$$\langle \underline{y}_1, \underline{y}_2 \rangle = \underline{y}_1^* \underline{y}_2 = \sum_{k=1}^{m} \bar{y}_1(k) y_2(k) \tag{18}$$

Two vectors contained in C^m are said to be *orthogonal* if their inner product is zero. Furthermore, if the vectors $\underline{y}_1, \underline{y}_2 \in C^m$ are orthogonal, the *Pythagorean Theorem* states that

$$\|\underline{y}_1 + \underline{y}_2\|^2 = \|\underline{y}_1\|^2 + \|\underline{y}_2\|^2, \tag{19}$$

where $\|\underline{y}\|^2 = \langle \underline{y}, \underline{y} \rangle$ designates the squared Euclidean norm. This theorem is proven by direct substitution in which the orthogonality assumption is incorporated. A visual depiction of this decomposition principal is shown in Fig. 1.

We now use the Pythagorean Theorem in conjunction with orthogonal projection matrix P_A to provide an insight into obtaining a best approximate solution to an inconsistent system of linear equations. In particular, let the observed vector appearing in relationship (1) be decomposed as

$$\begin{aligned} \underline{y} &= \underline{y} + P_A \underline{y} - P_A \underline{y} \\ &= P_A \underline{y} + [I - P_A]\underline{y}. \end{aligned} \tag{20}$$

The two vectors $P_A \underline{y}$ and $[I - P_A]\underline{y}$ comprising this decomposition are orthogonal since their inner product equals zero, that is,

$$\begin{aligned} [P_A \underline{y}]^* [I - P_A]\underline{y} &= \underline{y}^* P_A^* [I - P_A]\underline{y} \\ &= \underline{y}^* P_A [I - P_A]\underline{y} \\ &= \underline{y}^* [P_A - P_A P_A]\underline{y}] = \underline{\mathbf{0}}. \end{aligned} \tag{21}$$

The Hermitian and idempotency properties of the orthogonal projection matrix have been used in arriving at this orthogonality condition. The vector $P_A \underline{y}$ in decomposition (20) lies in the range space of A, while vector $[I - P_A]\underline{y}$ lies in the orthogonal complement of this range space.

Moore-Penrose Generalized Inverse

Using the fact that the vectors $A\underline{x}$ and $P_A \underline{y}$ are contained in $\mathcal{R}(A)$ and that vector $[I - P_A]\underline{y}$ is contained in $\mathcal{R}(A)^{\perp}$, the Pythagorean Theorem (19) indicates that

$$\|A\underline{x} - \underline{y}\|^2 = \|A\underline{x} - P_A \underline{y}\|^2 + \|[I - P_A]\underline{y}\|^2 \tag{22}$$

for any $\underline{x} \in C^n$ in which orthogonal decomposition (20) for vector $\underline{y}$ has been employed. Moreover, using the fact that the linear system of equations $A\underline{x} = P_A \underline{y}$ is consistent, it is seen that the first term on the right side can always be made zero. Thus, the smallest value assumed by squared error criterion (22) is equal to $\|[I - P_A]\underline{y}\|^2$. Moreover, any $\underline{x}$ that satisfies

$$A\underline{x} = P_A \underline{y} \tag{23}$$

will achieve this minimum. In the set of solutions to

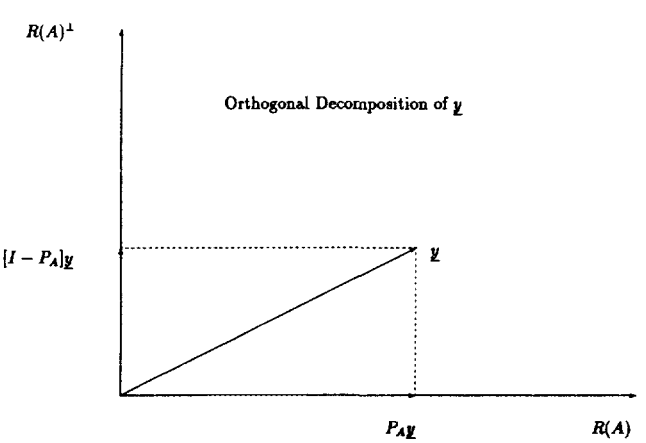

FIG. 1. Orthogonal decomposition of vector $\underline{y}$.

this consistent system of linear equations, a logical selection would be one which has the smallest Euclidean norm. Specifically, it is desired to find a vector that minimizes $\underline{x}^*\underline{x}$ subject to constraint (23). Using Lagrange multiplier techniques, the unique minimum Euclidean norm best approximate solution is found to be

$$\underline{x}^o = A^\dagger \underline{y}, \tag{24}$$

where $A^\dagger$ is the $n \times m$ *Moore–Penrose* generalized inverse matrix as specificed by [6]

$$A^\dagger = R^*[RR^*]^{-1}[B^*B]^{-1}B^*. \tag{25}$$

This Moore–Penrose generalized inverse may be expressed in the product form $A^\dagger = R^\dagger B^\dagger$ where $R^\dagger = R^*[RR^*]^{-1}$ and $B^\dagger = [B^*B]^{-1}B^*$ correspond to the Moore–Penrose generalized inverses of matrices R and B, respectively. Although the number of decompositions of A of form BR is uncountable, each gives rise to the same Moore–Penrose generalized inverse matrix when employing relationship (25). Upon left-multiplying the left side of this relationship by A and the right side by its equivalent BR, we obtain the additional important identity

$$P_A = AA^\dagger. \tag{26}$$

Thus, the orthogonal projection associated with a matrix is readily generated once knowledge of its Moore–Penrose generalized inverse is available. The following theorem summarizes these observations.

THEOREM 1. *Consider the system of m linear equations in n unknowns as represented by* $A\underline{x} = \underline{y}$ *where the* $m \times n$ *system matrix A has rank r. Let B be any* $m \times r$ *matrix whose column vectors comprise a basis for the range space of A and* $P_A = B[B^*B]^{-1}B^*$ *be the associated projection matrix. It then follows that any choice of the vector* $\underline{x}$ *that satisfies the consistent system of linear equations*

$$A\underline{x} = P_{A\underline{y}} \tag{27}$$

minimizes the Euclidean norm of the error vector $\underline{e} = A\underline{x} - \underline{y}$. *Furthermore, all solutions to this consistent system of equations (i.e.,* $A\underline{\hat{x}} = P_A\underline{y}$*) result in the same squared Euclidean norm as given by*

$$\|A\underline{\hat{x}} - \underline{y}\|^2 = \|[I - P_A]\underline{y}\|^2. \tag{28}$$

Furthermore, the unique minimum Euclidean norm solution to the consistent system of linear equations (27) is specified by

$$\underline{x}^o = A^\dagger \underline{y}, \tag{29}$$

where $A^\dagger$ *designates the Moore–Penrose generalized inverse of the system matrix A as specified in relationship (25).*

Upon perusal of this theorem, it is clear that the system of linear equations $A\underline{x} = \underline{y}$ is consistent if and only if $\underline{y}$ satisfies the fixed point relationship $P_A\underline{y} = \underline{y}$. Another interpretation is that consistency follows if and only if $\underline{y}$ is an eigenvector of the orthogonal projection matrix P_A with an associated eigenvalue of one.

3. SINGULAR VALUE DECOMPOSITION

In the last section, a characterization of the solutions or LSE approximate solutions to a linear system of equations was presented that depended on the selection of any basis that spans the column space of the underlying system matrix. Intuitively, a more salient characterization might result if the particular basis employed captured more of the essential features of the system matrix other than merely spanning its range space. The *singular value decomposition* (SVD) provides the most widely used such choice. As is now shown, the SVD enables us to decompose any matrix into a sum of rank one outerproduct matrices which are pairwise orthogonal. This orthogonal decomposition plays a central role in characterizing the basic nature of the system matrix.

To motivate the concept of the SVD representation of a rectangular matrix, it is recalled that a least-squares error solution to the original system of linear equations satisfies the consistent system of normal equations $A^*A\underline{\hat{x}} = A^*\underline{y}$. The nature of a least squares error solution is therefore dictated by the characteristics of the associated $n \times n$ Gram matrix A^*A. This Gram matrix is seen to be both positive-semidefinite and Hermitian. The Hermitian property indicates that the eigenvalues of this Gram matrix are strictly real and that it possesses a full set of n eigenvectors which can always be selected pairwise orthonormal. The positive-definite property implies that the eigenvalues of this Gram matrix are nonnegative. Furthermore, if the system matrix A has rank r, it must follow that A^*A also has rank r. This further implies that this Gram matrix has r positive eigenvalues and a zero eigenvalue of multiplicity $n - r$. This eigenanalysis therefore takes the form

$$A^*A\underline{v}_k = \sigma_k^2\underline{v}_k \quad \text{for} \quad 1 \leq k \leq n, \tag{30}$$

where without loss of generality the eigenvalues are arranged in the monotonically nonincreasing fashion

$$\sigma_1^2 \geq \sigma_2^2 \geq \cdots \sigma_r^2 > 0$$

and

$$\sigma_{r+1}^2 = \cdots = \sigma_n^2 = 0, \tag{31}$$

while the $m \times 1$ eigenvectors are selected to be pairwise orthonormal so that

$$\underline{v}_i^* \underline{v}_k = \delta(k - i). \tag{32}$$

Upon left multiplying each side of eigenrelationship (30) by A there results the identity $AA^*A\underline{v}_k = \sigma_k^2 A\underline{v}_k$. From this identity it follows that the vectors $A\underline{v}_k$ are eigenvectors of the companion $m \times m$ positive-semidefinite Hermitian matrix AA^* with the same positive σ_k^2 eigenvalues of matrix A^*A. It is a simple exercise to show that the $A\underline{v}_k$ eigenvectors associated with the r positive eigenvalues are orthogonal. Moreover, the scalar multiples of the positive eigenvalue associated eigenvectors generated according to $\underline{u}_k = A\underline{v}_k/\sigma_k$ for $1 \leq k \leq r$ are readily shown to have a Euclidean norm of one. It therefore follows that the vectors $\underline{u}_k$ satisfying the relationship

$$A\underline{v}_k = \sigma_k \underline{u}_k \quad \text{for} \quad 1 \leq k \leq r \tag{33}$$

are orthonormal eigenvectors of AA^* with associated positive eigenvalues σ_k^2. Furthermore, since the rank of AA^* is r, it follows that this Hermitian matrix has a zero eigenvalue of multiplicity $m - r$ with an associated full set of eigenvectors that span the null space of A^*.

We are now in a position to generate the SVD of matrix A. This entails right-multiplying each side of relationship (33) by $\underline{v}_k^*$ and then summing the resultant products according to

$$\sum_{k=1}^{r} A\underline{v}_k \underline{v}_k^* = \sum_{k=1}^{r} \sigma_k \underline{u}_k \underline{v}_k^*. \tag{34}$$

The upper sum limit on the left side summation is now increased from r to n since it is recalled that $A\underline{v}_k$ equals the zero vector for $r + 1 \leq k \leq n$ as is evident from eigenrelationships (30) and (31). Furthermore, since matrix A does not depend on the sum index k, it may be factored outside the left side summation to give rise to

$$A \sum_{k=1}^{n} \underline{v}_k \underline{v}_k^* = \sum_{k=1}^{r} \sigma_k \underline{u}_k \underline{v}_k^*. \tag{35}$$

Finally, using the fact that the eigenvectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ comprise an orthonormal basis for C^n, it follows that the matrix right-multiplying A on the left side is equal to the identity matrix. It has therefore been established that the $m \times n$ matrix A can be decomposed into the following weighted sum of outer-products:

$$A = \sum_{k=1}^{r} \sigma_k \underline{u}_k \underline{v}_k^*. \tag{36}$$

This decomposition constitutes the SVD representation of matrix A in which the r positive scalars σ_k are referred to as *singular values* while the $m \times 1$ vectors $\underline{u}_k$ and $n \times 1$ vectors $\underline{v}_k$ for $1 \leq k \leq r$ are called the *left singular* and *right singular* vectors, respectively. The vectors comprising the set of left singular vectors and the set of right singular vectors are each pairwise orthonormal.

When conducting as analysis, it is often convenient to express SVD representation (36) in its equivalent matrix form

$$A = U\Sigma V^*, \tag{37}$$

where the $m \times r$ matrix U has as its columns the left singular vectors, the $n \times r$ matrix V has as its columns the right singular vectors, and Σ is an $r \times r$ diagonal matrix with the positive singular values being its diagonal elements. Due to the pairwise orthogonality of the left and right singular vectors, it follows that $U^*U = V^*V = I_r$. This SVD representation provides a convenient means for characterizing the range space of matrix A. In particular, upon setting $B = U$ and $R = \Sigma V^*$, we have a specific decomposition of the system matrix as represented by general expression (14). Substitution of these specific choices into relationships (17) and (25), it follows that the associated orthogonal projection matrix is expressed as

$$P_A = UU^*, \tag{38}$$

while the Moore–Penrose generalized inverse matrix is given by

$$A^\dagger = V\Sigma^{-1}U^*. \tag{39}$$

These SVD-based expressions provide a specific means for generating the projection matrix and Moore–Penrose generalized inverse employed in Fundamental Theorem 1. It must be emphasized, however, that there are uncountably many choices for the basis matrix B other than $B = U$ that accomplish the same objective. For a variety of numerically based reasons, the SVD approach is normally preferred.

We may also use the SVD representation for the system matrix to obtain an explicit characterization of the null space of A. In particular, this null space is seen to be equivalent to the set of vectors $\{\underline{x}\}$ which satisfy the homogeneous relationship $A\underline{x} = U\Sigma V^*\underline{x}$

$= \underline{0}$. Since matrices U and Σ each have full rank r, it follows that $\mathcal{N}(A)$ is equivalent to the null space of V^*. Since matrix V has rank r, it follows that there exists exactly $n - r$ linearly independent vectors that map into the zero vector under V^*. A useful procedure for identifying these vectors is to augment the right singular (orthonormal vectors) $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_r$ by $\underline{v}_{r+1}$, $\underline{v}_{r+2}, \ldots, \underline{v}_n$ such that the combined set of vectors constitutes an orthonormal basis for C^n. It follows that the augmented vector set constitutes a basis for the null space of V^* and therefore A, that is,

$$\mathcal{N}(A) = \{\underline{x} \in C^n : \underline{x} = \sum_{k=r+1}^{n} \alpha_k \underline{v}_k\}. \tag{40}$$

There are other additional dividends to be accrued when using an SVD representation for the system matrix. For example, the concept of rank reduced approximations of system matrices plays a vital role in contemporary signal processing. The *Frobenius norm* of a general $m \times n$ matrix is defined by

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} \tag{41}$$

and is seen to equal the square root of the sum of squared component magnitudes of the matrix. A simple computation shows that this measure is equal to the square root of the trace of matrix product A^*A, that is,

$$\begin{aligned} \|A\|_F &= \sqrt{\text{trace}(A^*A)} \\ &= \sqrt{\sum_{k=1}^{r} \sigma_k^2}, \end{aligned} \tag{42}$$

where $\{\sigma_k\}$ denote the positive singular values associated with the rank r matrix A. In arriving at this result use has been made of the SVD representation $A = U\Sigma V^*$ which gives $A^*A = V\Sigma U^* U \Sigma V^*$. The identities $U^*U = I$ and $\text{trace}(V\Sigma^2 V^*) = \text{trace}(\Sigma^2)$ are then employed to arrive at relationship (42).

Reduced Risk Matrix Approximation

In seeking to cleanse noise corrupted data, a widely used procedure is to generate rank reduced data matrices. Specifically, let the rank r matrix A have the SVD representation (36). It is now desired to find a matrix of reduced rank $p < r$ that lies closest to A in the Frobenius norm sense. This entails finding a $m \times n$ matrix C that solves the following optimization problem:

$$\min_{\text{rank}C=p} \|A - C\|_F. \tag{43}$$

Eckart and Young have shown that the solution to this problem is given by [7]

$$A^{(p)} = \sum_{k=1}^{p} \sigma_k \underline{u}_k \underline{v}_k^*. \tag{44}$$

Thus, the closest rank p approximation of matrix A entails truncating the SVD of this matrix to its largest p outerproducts. It is apparent that this rank p Frobenius norm approximation is unique if and only if $\sigma_p > \sigma_{p+1}$.

The relative size of the error in approximating a matrix by a reduced rank p matrix is a key consideration in selecting the integer p. It directly follows from expression (42) that the *normalized* Frobenius norm of this approximation is given by

$$\begin{aligned} \rho(p) &= \frac{\|A - A^{(p)}\|_F}{\|A\|_F} \\ &= \sqrt{\frac{\sigma_{p+1}^2 + \sigma_{p+2}^2 + \cdots + \sigma_r^2}{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2}}. \end{aligned} \tag{45}$$

This normalized measure always takes on values in the interval [0, 1] with values closer to zero (one) indicating that $A^{(p)}$ provides a good (poor) rank p approximation of A.

4. STATISTICAL ANALYSIS

In most applications, modeling expression (1) provides only an approximation to the actual relationship between the observed and the unobserved variables. This inaccuracy arises primarily from the three factors: (i) inaccurate measurements of the observed variables $\underline{y}$, (ii) incorrect entries of the system matrix A, and (iii) the inappropriateness of a linear model. When invoking a standard least-squares error estimation of the unobserved variables, it is assumed that the system matrix is known with complete accuracy but the observed vector is subject to error. We now examine the standard least-squares method from a probability viewpoint [15].

In the standard least-squares solution approach, it is assumed that the underlying observed and unobserved vectors are perfectly related according in a linear fashion. Due to an imperfect measurement or modeling process, however, the observed vector is contaminated in an additive fashion giving rise to the corrupted system of linear equations

$$\hat{\underline{y}} = A\underline{x} + \underline{w}, \tag{46}$$

where vector $\underline{w}$ represents measurement error. In

most practical applications of interest, it is generally found that the perturbed observed vector $\underline{\hat{y}}$ does not lie in the range space of A. Thus, the perturbed system of linear equations ($A\underline{x} = \underline{\hat{y}}$) is typically inconsistent even though the underlying noise free system $A\underline{x} = \underline{y}$ is consistent.

Our primary objective is to seek a useful estimate of the unobserved vector $\underline{x}$ and noise-free vector $\underline{y}$ given the noise corrupted observed vector $\underline{\hat{y}} = \underline{y} + \underline{w}$. If it is suspected that the size of the measurement error vector $\underline{w}$ is small in comparison to the underlying vector $\underline{y}$, an intuitively appealing procedure for finding a useful estimate of the unobserved vector is to apply least squares estimation methods. From the results given in Theorem 1, it follows that the least-squares estimate of the unobserved vectors is given by

$$\begin{aligned}\underline{\hat{x}}_{\mathrm{LS}} &= A^{\dagger}\underline{\hat{y}} \\ &= A^{\dagger}A\underline{x} + A^{\dagger}\underline{w}, \end{aligned} \tag{47}$$

where $A^{\dagger}$ designates the Moore–Penrose generalized inverse of system matrix A.

Random Error Model

It is possible to provide a useful second-order statistical analysis for this estimation problem. In this analysis, the underlying vector $\underline{x}$ is taken to be unknown thereby implying no a priori information concerning it is presumed. Furthermore, the error vector $\underline{w}$ is taken to be random with an expected value of zero and covariance matrix R_{ww}, that is,

$$E\{\underline{w}\} = \underline{0}$$

and

$$R_{ww} = E\{[\underline{w} - E\{\underline{w}\}][\underline{w} - E\{\underline{w}\}]^{*}\}, \tag{48}$$

where E designates the expected value operator. It then follows that the corrupted observed vector $\underline{\hat{y}}$ has expected value $A\underline{x}$ and covariance matrix $R_{\hat{y}\hat{y}} = R_{ww}$. In order to mitigate the deleterious effects caused by the additive noise vector, the LS estimate described by relationship (47) is commonly employed. In fact, it follows that if the additive noise is Gaussian with covariance matrix $\sigma^2 I$ (i.e., white noise), then this LS estimate correspond to the *maximum likelihood estimate* of $\underline{x}$.

A measure of the quality of the LS estimate (47) is obtained by evaluating its expected value. Using the fact that the additive noise vector has zero mean and that $E\{\underline{y}\} = A\underline{x}$, we have

$$\begin{aligned} E\{\underline{\hat{x}}_{\mathrm{LS}}\} &= A^{\dagger}E\{\underline{\hat{y}}\} \\ &= A^{\dagger}A\underline{x} \\ &= VV^{*}\underline{x} \\ &= \sum_{k=1}^{r} (\underline{v}_k^{*}\underline{x})\underline{v}_k, \end{aligned} \tag{49}$$

where the SVD representation $A = U\Sigma V^{*}$ for the system matrix has been employed. Ideally, the expected value of an estimate should be equal to the quantity being estimated which in this case is $\underline{x}$. With this in mind, the *bias* vector associated with LS estimator (47) is formally specified by

$$\begin{aligned}\underline{b} &= \underline{x} - E\{\underline{\hat{x}}_{\mathrm{LS}}\} \\ &= \sum_{k=r+1}^{n} (\underline{v}_k^{*}\underline{x})\underline{v}_k. \end{aligned} \tag{50}$$

The LS estimator is said to be *unbiased* if this bias vector is equal to the zero vector. Clearly, we are assured of an unbiased estimate only if the rank of the $m \times n$ system matrix A is equal to n. Fortunately, in most practical applications the system matrix has a full column rank, thereby resulting in an unbiased LS estimate.

The expected value of an estimator provides insight into the average value assumed by the estimate when a large number of realizations of the underlying experiment are conducted. It is conceivable that two different estimators may have the same expected value yet one may produce a broader spread of experimental realizations than the other. Other things being equal, an estimator that produces numerical realizations on a large number of trials that are closely clustered about its expected values are to be preferred. To measure the degree of estimator variability, knowledge of the estimator's covariance matrix is essential. The covariance matrix of the LS estimator (47) is formally given by

$$\begin{aligned} R_{\hat{x}_{\mathrm{LS}}\hat{x}_{\mathrm{LS}}} &= E\{[\underline{\hat{x}}_{\mathrm{LS}} - E\{\underline{\hat{x}}_{\mathrm{LS}}\}][\underline{\hat{x}}_{\mathrm{LS}} - E\{\underline{\hat{x}}_{\mathrm{LS}}\}]^{*}\} \\ &= E\{[A^{\dagger}\underline{w}][A^{\dagger}\underline{w}]^{*}\} \\ &= A^{\dagger}R_{ww}A^{\dagger *}. \end{aligned} \tag{51}$$

It is to be noted that the LS estimate has the effect of decreasing the noise corruption since $R_{ww} \geq A^{\dagger}R_{ww}A^{\dagger}$. This is further made evident by noting that the *mean square value* of the LS estimate is given by

$$\begin{aligned} E\{\|\underline{\hat{x}}_{\mathrm{LS}} - E\{\underline{\hat{x}}_{\mathrm{LS}}\}\|^2\} &= E\{\|A^{\dagger}\underline{w}\}\|^2 \\ &= E\{\|V\Sigma^{-1}U^{*}\underline{w}\|^2 \\ &= \mathrm{trace}(V\Sigma^{-1}U^{*}R_{ww}U\Sigma^{-1}V^{*}). \end{aligned} \tag{52}$$

In arriving at this result, use has been made of the

readily established identity $\underline{x}^*\underline{z} = \text{trace}(\underline{z}\underline{x}^*)$ which holds for all same-dimensioned vectors $\underline{x}$ and $\underline{z}$.

Reduced Rank Estimator

Useful insight into the effectiveness of the LS estimate is obtained when the additive noise vector $\underline{w}$ is a zero mean white process. In this case, the noise covariance matrix is specified by

$$R_{ww} = \sigma^2 I, \tag{53}$$

where σ^2 designates the variance of each component of $\underline{w}$. If this noise covariance is substituted into the mean squared measure, it follows that

$$E\{\|\hat{\underline{x}}_{\text{LS}} - E\{\hat{\underline{x}}_{\text{LS}}\}\|^2\} = \sigma^2 \sum_{k=1}^{r} \frac{1}{\sigma_k}. \tag{54}$$

It is seen that the mean square associated with the LS estimate is directly proportional to the additive noise variance and inversely proportional to the reciprocal of the positive singular values of system matrix A. This expression indicates that the mean squared error can become quite large if one or more of the nonzero singular values of A are close to zero. Thus, the direct employment of a least squares solution can lead to poor estimation performance.

To mitigate the deleterious effects resulting from small positive singular values, it would be desirable to remove the offending terms associated with these singular values. In particular, suppose that a preliminary SVD analysis of the system matrix indicate that $r - p$ of the positive singular values are "small" enough to cause an unacceptably large mean square value. To excise the impact of these small singular values, let us consider the following reduced rank p approximation of the system matrix:

$$A^{(p)} = \sum_{k=1}^{p} \sigma_k \underline{u}_k \underline{v}_k^*. \tag{55}$$

As indicated in the last section, this matrix corresponds to the closest rank p Frobenius norm approximation of system matrix A. The Moore–Penrose generalized inverse and projection matrices associated with this rank p approximation are therefore given by

$$A^{\dagger(p)} = \sum_{k=1}^{p} \frac{1}{\sigma_k} \underline{v}_k \underline{u}_k^* \quad \text{and} \quad P_A^{(p)} = \sum_{k=1}^{p} \underline{u}_k \underline{u}_k^*. \tag{56}$$

As is now shown, a reduction of the mean square value is achievable if the reduced rank system matrix (55) is used instead of A. The LS estimator associated with the reduced order system matrix is formally defined by

$$\begin{aligned} \hat{\underline{x}}_{\text{LS}}^{(p)} &= A^{\dagger(p)} \underline{y} \\ &= \sum_{k=1}^{p} (\underline{v}_k^* \underline{x}) \underline{v}_k + \sum_{k=1}^{p} \left(\frac{\underline{u}_k^* \underline{w}}{\sigma_k} \right) \underline{v}_k \\ &= \underline{x} - \sum_{k=p+1}^{r} (\underline{v}_k^* \underline{x}) \underline{v}_k + \sum_{k=1}^{p} \left(\frac{\underline{u}_k^* \underline{w}}{\sigma_k} \right) \underline{v}_k. \end{aligned} \tag{57}$$

Examination of this expression indicates that the LS estimate is composed of an additive term due exclusively to the signal vector $\underline{x}$ and an additive term due to the noise vector $\underline{w}$. It is useful to study the effects of employing a reduced rank approximation on these signal and noise components. The expected value of this reduced rank LS estimate is seen to be

$$E\{\hat{\underline{x}}_{\text{LS}}^{(p)}\} = \underline{x} - \sum_{k=p+1}^{r} (\underline{v}_k^* \underline{x}) \underline{v}_k. \tag{58}$$

It is therefore concluded that the bias of the reduced order LS estimate is specified by

$$\underline{b}^{(p)} = \sum_{k=p+1}^{n} (\underline{v}_k^* \underline{x}) \underline{v}_k. \tag{59}$$

Unfortunately, the use of a reduced order estimator is seen to result in a generally larger bias vector than is the case of the full order estimator, that is,

$$\|\underline{b}^{(p)}\|^2 - \|\underline{b}\|^2 = \sum_{k=p+1}^{r} |\underline{v}_k^* \underline{x}|^2. \tag{60}$$

This difference is clearly positive if at least one of the inner products $\underline{v}_k^* \underline{x}$ is nonzero for $r > p$.

Although the reduced order LS estimate leads to an inferior behavior in bias behavior, its employment results in a mean square error that can be considerably smaller than its full order counterpart. In particular, the mean square error of the reduced rank estimator (57) is found to be

$$E\{\|\underline{x} - \underline{b}^{(p)} - \hat{\underline{x}}_{\text{LS}}^{(p)}\|^2\} = \sigma^2 \sum_{k=1}^{p} \frac{1}{\sigma_k}. \tag{61}$$

Upon comparison of this reduced order mean square error expression with the full-order expression (54), it is seen that the reciprocal terms associated with the smaller valued singular values $\sigma_{p+1}, \ldots, \sigma_r$ are removed. Thus, the possibility of trading off a modest increase in bias for a significant decrease in means square error is indicated. The reduced rank principle

has been successfully applied by numerous investigators to an important and diverse set of applications.

Non-SVD Reduced Rank Approximation

The reduced rank SVD representation for the system matrix provides the standard tool for achieving a reduction in estimation variance at the cost of increased bias. It should be noted, however, that other reduced rank procedures can be employed for attaining this trade-off. In particular, it is recalled that the system matrix A can always be represented in the product form BR with the column vectors of basis matrix B chosen to span the same subspace as the column vectors of A. A rank k approximation of the basis matrix B is designated by $B^{(k)}$ and its columns are composed of any set of k columns from the basis matrix. It then follows that a rank k approximation for the system matrix is specified by

$$A^{(k)} = B^{(k)}R^{(k)}, \quad \text{where}$$
$$R^{(k)} = [B^{(k)*}B^{(k)}]^{-1}B^{(k)*}A. \quad (62)$$

It must be emphasized that this rank k approximation of A corresponds to the SVD-based approximation if and only if B is equal to U or a column rearrangement of U. Since there exists an uncountable number of distinct BR representations for the system matrix, it is logical to inquire whether there exists any advantages in using a non-SVD-based choice. For example, in a given application it may be possible to obtain a decomposition that leads to a superior trade-off between estimation bias and estimation variance than does an SVD decomposition. This possibility is currently under investigation.

5. NONLINEAR LEAST-SQUARES MODELING

In many applications of interest, the phenomenon under investigation can be represented by a system of linear equations in which the elements of the system matrix are known functions of a set of real-valued parameters. This is illustrated by problems found in array processing where the parameters correspond to directions-of-arrival angles and in linear data modeling where feedforward and feedback coefficients serve as the parameters. Whatever the case, we are interested in analyzing a system of equations that is expressible as

$$A(\underline{\theta})\underline{x} \approx \underline{y}, \quad (63)$$

where $A(\underline{\theta})$ is an $m \times n$ matrix whose components are known functions of the real parameters $\theta_1, \theta_2, \ldots, \theta_p$. These parameters comprise the elements of the *parameter vector* $\underline{\theta} \in R^p$. Due to a variety of reasons already alluded to, this system of equations will be inconsistent. Our task is to then find a selection of the parameter vector $\underline{\theta}$ and unobserved vector $\underline{x}$ so that $A(\underline{\theta})\underline{x}$ best approximates $\underline{y}$ in the Euclidean norm sense. This entails solving the following squared Euclidean norm optimization problem:

$$\min_{\underline{x}\in C^n} \min_{\underline{\theta}\in R^p} \|\underline{y} - A(\underline{\theta})\underline{x}\|^2. \quad (64)$$

A closed form solution to this optimization problem is generally not feasible due to the highly nonlinear fashion in which the entities $\underline{x}$ and $\underline{\theta}$ appear. It is then necessary to employ nonlinear programming techniques to numerically find a solution. Since the computational load of any nonlinear programming algorithm is directly dependent on the number of parameters being optimized, it behooves us to decrease this number whenever possible. With this in mind, we now appeal to the theory presented in Sections 2 and 3 to effectively remove the unobserved vector in the optimization process. In particular, for any selection of $\underline{\theta}$ (optimum or nonoptimum) it follows from Theorem 1 that an associated optimum selection of the unobserved vector is specified by

$$\hat{\underline{x}}_{LS}(\underline{\theta}) = A(\underline{\theta})^{\dagger}\underline{y}. \quad (65)$$

Upon substitution of this selection into expression (64), the optimization problem simplifies to

$$\begin{aligned} &\min_{\underline{x}\in C^n} \min_{\underline{\theta}\in R^p} \|\underline{y} - A(\underline{\theta})\underline{x}\|^2 \\ &\quad = \min_{\underline{\theta}\in R^p} \|\underline{y} - A(\underline{\theta})A(\underline{\theta})^{\dagger}\underline{y}\|^2 \\ &\quad = \min_{\underline{\theta}\in R^p} \|[I - P(\underline{\theta})]\underline{y}\|^2, \end{aligned} \quad (66)$$

where the identity $P(\underline{\theta}) = A(\underline{\theta})A(\underline{\theta})^{\dagger}$ as specified by relationship (26) has been used. Thus the original problem entailing minimization with respect to the vectors $\underline{x}$ and $\underline{\theta}$ has been reduced to a minimization with respect to vector $\underline{\theta}$. Once the optimal choice for $\underline{\theta}$ that solves problem (66) has been found, relationship (65) is then used to find the associated optimal selection for $\underline{x}$.

Although the number of variables being optimized has been reduced by the above procedure, the resultant decreased dimensioned optimization problem is generally very nonlinear in $\underline{\theta}$. We must therefore resort to a nonlinear programming algorithm to approximate an optimal solution. Relationship (66) indicates that the functional to be minimized is specified by

$$f(\underline{\theta}) = \| [I - P(\underline{\theta})]\underline{y} \|^2. \quad (67)$$

The term $[I - P(\underline{\theta})]\underline{y}$ appearing in this functional is seen to correspond to the equation error $A\underline{x}_{LS} - \underline{y}$ associated with the least-squares error choice of $\underline{x}$. Many of the more widely used nonlinear programming algorithms used for minimizing functionals such as (67) employ the method of *descent*. In a descent based algorithm, the present estimate of the solution $\underline{\theta}$ is additively perturbed to $\underline{\theta} + \underline{\delta}$ where $\underline{\delta}$ is referred to as the *perturbation vector*. The basic task is to select the perturbation vector so that the *improving condition*

$$f(\underline{\theta} + \underline{\delta}) < f(\underline{\theta}) \quad (68)$$

is satisfied. There are a variety of different procedures for selecting the perturbation vector to satisfy this improvement condition. A particularly effective method is now developed.

First-Order Analysis

If the perturbation vector is sufficiently small in size, a Taylor series expansion of the perturbed criterion can be made in which only the first two terms of the expansion are retained. This expansion is specified by

$$\begin{aligned} f(\underline{\theta} + \underline{\delta}) &= \| [I - P(\underline{\theta} + \underline{\delta})]\underline{y} \|^2 \\ &\approx \left\| \left[I - P(\underline{\theta}) - \sum_{k=1}^{p} \frac{\partial P(\underline{\theta})}{\partial \theta_k} \delta_k \right] \underline{y} \right\|^2 \\ &= \left\| [I - P(\underline{\theta})]\underline{y} - \sum_{k=1}^{p} \left[\frac{\partial P(\underline{\theta})}{\partial \theta_k} \underline{y} \right] \delta_k \right\|^2 \\ &= \| [I - P(\underline{\theta})]\underline{y} - L(\underline{\theta})\underline{\delta} \|^2, \end{aligned} \quad (69)$$

where $\partial P(\underline{\theta})/\partial \theta_k$ are $m \times m$ matrices, while the δ_k entities are the components of the perturbation vector $\underline{\delta}$. The $m \times p$ *Jacobian matrix* $L(\underline{\theta})$ here appearing is specified by

$$L(\underline{\theta}) = \left[\frac{\partial P(\underline{\theta})}{\partial \theta_1} \underline{y} \vdots \frac{\partial P(\underline{\theta})}{\partial \theta_2} \underline{y} \vdots \cdots \vdots \frac{\partial P(\underline{\theta})}{\partial \theta_p} \underline{y} \right]. \quad (70)$$

This Jacobian matrix provides a means for obtaining a first-order approximating of the effect which incremental changes made in the prevailing parameter vector have on the Euclidean norm criterion being minimized.

A logical choice for the perturbation vector would be one that minimizes the approximation of the Euclidean norm criterion as specified by relationship (69). An expression for this optimum selection is achieved by expanding this squared error criterion to obtain

$$\begin{aligned} f(\underline{\theta} + \underline{\delta}) = f(\underline{\theta}) &- \underline{\delta}^{\mathrm{T}} L(\underline{\theta})^* [I - P(\underline{\theta})]\underline{y} \\ &- \underline{y}^* [I - P(\underline{\theta})] L(\underline{\theta})\underline{\delta} + \underline{\delta}^{\mathrm{T}} L(\underline{\theta})^* L(\underline{\theta})\underline{\delta}. \end{aligned} \quad (71)$$

Upon setting the gradient of this expression with respect to the real vector $\underline{\delta}$ equal to the zero vector, the optimal selection for the perturbation vector is found to be

$$\underline{\delta}^{\circ} = [\mathrm{Real}\{ L(\underline{\theta})^* L(\underline{\theta}) \}]^{\dagger} \mathrm{Real}\{ L(\underline{\theta})^* [I - P(\underline{\theta})]\underline{y} \}. \quad (72)$$

It can happen that the perturbation vector arising from this computation will be relatively large in size thereby putting into question the validity of approximation (69) which generally holds only for small sized perturbations. To ensure a sufficiently small perturbation the scaled perturbation vector $\alpha\underline{\delta}^{\circ}$ is instead used. It can be shown that improvement condition (68) can always be met by selecting the *step size* scalar α to be sufficiently small positive. With these thoughts in mind, the basic steps of the *linearization* descent algorithm are given in Table 1.

TABLE 1

Nonlinear Programming Algorithm

Step	Description
1.	Generate an initial estimate of $\underline{\theta}$.
2.	Evaluate Euclidean norm criterion $\|[I - P(\underline{\theta})]\underline{y}\|^2$.
3.	Determine the Jacobian matrix $L(\underline{\theta})$.
4.	Compute the optimum perturbation $\underline{\delta}^{\circ}$.
5.	Evaluate the perturbed Euclidean norm criterion $\|[I - P(\underline{\theta} + \alpha\underline{\delta}^{\circ})]\underline{y}\|^2$ for $\alpha = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \cdots$ until an improving value is found.
6.	Evaluate stopping condition(s) for algorithm. If these conditions are not satisfied, set $\underline{\theta} = \underline{\theta} + \alpha^{\circ}\underline{\delta}^{\circ}$ and go to Step 2.

Computation of Jacobian Matrix

In order to implement the linearization matrix, it is necessary to compute the Jacobian matrix as specified in relationship (70). This in turn entails the need to determine the partial derivatives of the projection matrix $P(\underline{\theta})$ with respect to the parameter vector components. As is now shown, there exists a convenient closed form solution for these partial derivatives. To begin this development, the partial derivative of the projection matrix identity $P(\underline{\theta}) = P(\underline{\theta})^2$ is first taken with respect to θ_k to give

$$\frac{\partial P(\underline{\theta})}{\partial \theta_k} = \frac{\partial P(\underline{\theta})}{\partial \theta_k} P(\underline{\theta}) + P(\underline{\theta}) \frac{\partial P(\underline{\theta})}{\partial \theta_k}$$

$$= \frac{\partial P(\underline{\theta})}{\partial \theta_k} P(\underline{\theta}) + \left[\frac{\partial P(\underline{\theta})}{\partial \theta_k} P(\underline{\theta}) \right]^*. \quad (73)$$

The second line of this identity is a direct consequence of the fact that the projection matrix is Hermitian, which in turn implies that its derivative with respect to θ_k must also be Hermitian.

To compute the terms on the right-hand side of relationship (73), each side of the matrix identity $P(\underline{\theta})A(\underline{\theta}) = A(\underline{\theta})$ as specified in Eq. (16) is differentiated with respect to θ_k. Upon carrying out this differentiation and then rearranging terms, we have

$$\frac{\partial P(\underline{\theta})}{\partial \theta_k} A(\underline{\theta}) = [I - P(\underline{\theta})] \frac{\partial A(\underline{\theta})}{\partial \theta_k}. \quad (74)$$

Right-multiplying each side of this relationship by the Moore–Penrose generalized inverse $A(\underline{\theta})^\dagger$ and using the identity $P(\underline{\theta}) = A(\underline{\theta})A(\underline{\theta})^\dagger$ as specified by expression (26) results in

$$\frac{\partial P(\underline{\theta})}{\partial \theta_k} P(\underline{\theta}) = [I - P(\underline{\theta})] \frac{\partial A(\underline{\theta})}{\partial \theta_k} A(\underline{\theta})^\dagger. \quad (75)$$

This expression is then substituted into relationship (73) to obtain the partial derivative expressions

$$\frac{\partial P(\underline{\theta})}{\partial \theta_k} = \left[[I - P(\underline{\theta})] \frac{\partial A(\underline{\theta})}{\partial \theta_k} A(\underline{\theta})^\dagger \right] + \left[[I - P(\underline{\theta})] \frac{\partial A(\underline{\theta})}{\partial \theta_k} A(\underline{\theta})^\dagger \right]^* \quad \text{for} \quad 1 \leq k \leq q. \quad (76)$$

Under the assumption that one is able to determine the partial derivatives $\partial A(\underline{\theta})/\theta_k$, each entity in this expression is computed and then substituted into relationship (70) to obtain the required Jacobian matrix.

QR and SVD Representations

From a computational perspective, an effective means for determining the partial derivatives is to first make a QR decomposition of the prevailing rank r system matrix $A(\underline{\theta})$, that is,

$$A(\underline{\theta}) = Q(\underline{\theta})R(\underline{\theta}). \quad (77)$$

In this decomposition $Q(\underline{\theta})$ is an $m \times r$ matrix whose columns are orthonormal so that $Q(\underline{\theta})^*Q(\underline{\theta}) = I$, while $R(\underline{\theta})$ is an $r \times r$ invertible upper triangular matrix. Moreover, using this QR decomposition in conjunction with relationship (25) for the Moore–Penrose generalized inverse of the system matrix and expression (26) for the associated orthogonal projection matrix, we have

$$A(\underline{\theta})^\dagger = R(\underline{\theta})^{-1}Q(\underline{\theta})^* \quad \text{and}$$

$$P(\underline{\theta}) = Q(\underline{\theta})Q(\underline{\theta})^*. \quad (78)$$

Substitution of these two relationships into Eq. (76) then yields the QR decomposition representation for the prerequisite partial derivatives. Moreover, if the SVD representation for system matrix $A(\underline{\theta}) = U(\underline{\theta})\Sigma(\underline{\theta})V(\underline{\theta})^*$ is employed, the expressions for the Moore–Penrose generalized inverse matrix and projection matrix are given by

$$A(\underline{\theta})^\dagger = V(\underline{\theta})\Sigma(\underline{\theta})^{-1}U(\underline{\theta})^*$$

and

$$P(\underline{\theta}) = U(\underline{\theta})U(\underline{\theta})^* \quad (79)$$

6. NONLINEAR LEAST-SQUARES MODELING: MULTIPLE SAMPLES

In many practical applications involving nonlinear least squares modeling, there is given multiple samples of the data under analysis. For example, in array signal processing one generally has available multiple time samples of the *snapshot vector* characterizing the array's sensor signals. We may directly extend the concepts developed in the last section to treat this important class of modeling problems. In particular, let us consider the case in which there is given multiple samples of the nonlinear phenomenon governed by

$$\underline{y}_k = A(\underline{\theta})\underline{x}_k + \underline{w}_k \quad \text{for} \quad 1 \leq k \leq N, \quad (80)$$

where $A(\underline{\theta})$ is an $m \times n$ matrix which is a function of the model parameter vector $\underline{\theta} \in R^p$ and $\underline{w}_k \in C^m$ represents additive noise. The task at hand is to select the model parameter vector and the $\{\underline{x}_k\}$ that are most compatible with the given data (80). To measure data compatibility, we employ the standard squared error criterion as specified by

$$f(\underline{\theta}, \{\underline{x}_k\}) = \sum_{k=1}^{N} \|A(\underline{\theta})\underline{x}_k - \underline{y}_k\|^2. \quad (81)$$

The objective is to select model parameter vector and the vectors $\{\underline{x}_k\}$ so as to minimize this squared error criterion.

Upon examination of criterion (81) and using the results from the last section, it is noted that for any

choice of the model parameter vector the optimal selection of the $\{\underline{x}_k\}$ vectors are given by

$$\underline{x}_k^o = A(\underline{\theta})^\dagger \underline{y}_k \quad \text{for} \quad 1 \leq k \leq N, \tag{82}$$

where $A(\underline{\theta})^\dagger$ designates the Moore–Penrose generalized inverse of $A(\underline{\theta})$. Substitution of this optimum selection into the above squared error criterion yields

$$f(\underline{\theta}, \{\underline{x}_k^o\}) = \sum_{k=1}^{N} \|[I - P(\underline{\theta})]\underline{y}_k\|^2 \quad \text{for } 1 \leq k \leq N, \tag{83}$$

where $P(\underline{\theta})$ denotes the orthogonal projection matrix whose range space is identical to the range space of $A(\underline{\theta})$. This substitution process has effectively reduced the minimization problem to that of selecting the model parameter vector $\underline{\theta}$.

Since squared error criterion (83) is a highly nonlinear function of $\underline{\theta}$, it is again necessary to use nonlinear programming techniques to iteratively determine an optimal choice for the model parameter vector. If a descent algorithm is to be used, a procedure for obtaining an effective perturbation vector for the prevailing model parameter vector is required. To achieve such a selection, a first-order analysis of the perturbed squared error criterion is made. This analysis takes the form

$$\begin{aligned} f(\underline{\theta} + \underline{\delta}, \{\underline{x}_k^o\}) &= \sum_{k=1}^{N} \|[I - P(\underline{\theta} + \underline{\delta})]\underline{y}_k\|^2 \\ &\approx \sum_{k=1}^{N} \|[I - P(\underline{\theta})]\underline{y}_k - L_k(\underline{\theta})\underline{\delta}\|^2 \\ &= \sum_{k=1}^{N} \underline{y}_k^*[I - P(\underline{\theta})]\underline{y}_k - \underline{\delta}^T L_k^*(\underline{\theta})[I - P(\underline{\theta})]\underline{y}_k \\ &\quad - \underline{y}_k^*[I - P(\underline{\theta})]L_k(\underline{\theta})\underline{\delta} + \underline{\delta}^T L_k^*(\underline{\theta})\underline{\delta} L_k(\underline{\theta})\underline{\delta}. \end{aligned} \tag{84}$$

In this expression, $L_k(\underline{\theta})$ designates the Jacobian matrix associated with the orthogonal projection matrix $P(\underline{\theta})$ as specified in expression (70) with $\underline{y}$ replaced by $\underline{y}_k$. Upon setting to zero the gradient of this first-order approximation of the perturbed squared error criterion, it is found that the optimum choice for the minimizing model parameter vector is given by

$$\underline{\delta}^o = \frac{1}{N}\Big[\sum_{k=1}^{N} \text{Real}\{L_k^*(\underline{\theta})L_k(\underline{\theta})\}\Big]^{-1} \text{Real}\Big\{\sum_{k=1}^{N} L_k^*(\underline{\theta})[I - P(\underline{\theta})]\underline{y}_k\Big\}. \tag{85}$$

The nonlinear programming algorithm described in Table 1 is then employed in which selection (85) for the perturbation vector is substituted in Step 4 while the multiple sample criterion (83) is used in Steps 2 and 5.

It is possible to further generalize the above results whereby a nonlinear function of the squared error criterion is incoporated. The results of this generalization are captured in the following theorem.

THEOREM 2. *Let the set of vector measurements be governed by the relationships*

$$\underline{y}_k = A(\underline{\theta})\underline{x}_k + \underline{w}_k \quad \text{for} \quad 1 \leq k \leq N, \tag{86}$$

where $A(\underline{\theta})$ is a $m \times n$ matrix which is a function of the model parameter vector $\underline{\theta} \in R^p$ and $\underline{w}_k \in C^m$ represents additive noise. It is now desired to select the model parameter vector and the vectors $\{\underline{x}_k\}$ so as to minimize the nonlinear squared error criterion

$$f(\underline{\theta}, \{\underline{x}_k\}) = \sum_{k=1}^{N} \psi(\|A(\underline{\theta})\underline{x}_k - \underline{y}_k\|^2), \tag{87}$$

where $\psi(x)$ is a monotonically increasing function of its nonnegative argument so that $\psi(x_1) > \psi(x_2)$ for all $0 \leq x_1 < x_2$ and $\psi(0) = 0$. The optimum choice for the $\{\underline{x}_k\}$ vectors are given by the standard least-squares error selection

$$\underline{x}_k^o = A(\underline{\theta})^\dagger \underline{y}_k \quad \text{for} \quad 1 \leq k \leq N. \tag{88}$$

Furthermore, the selection of the perturbation vector that minimizes a first-order approximation of the nonlinear squared error criterion (87) is specified by[3]

$$\underline{\delta}^o = \Big[\sum_{k=1}^{N} \dot{\psi}(\|[I - P(\underline{\theta})]\underline{y}_k\|^2)\, Real\{L_k^*(\underline{\theta})L_k(\underline{\theta})\}\Big]^{-1} \times Real\Big\{\sum_{k=1}^{N} \dot{\psi}(\|[I - P(\underline{\theta})]\underline{y}_k\|^2) L_k^*(\underline{\theta}) \times [I - P(\underline{\theta})]\underline{y}_k\Big\}, \tag{89}$$

where $\dot{\psi}(x)$ designates the derivative of the nonlinear function $\psi(x)$.

The validity of this theorem is easily proven using elementary reasoning. For instance, the optimum selection (88) is a direct consequence of the monotoni-

[3] The perturbed nonlinear squared error criterion with the optimum selection (88) is given by $f(\underline{\theta} + \underline{\delta}, \{\underline{x}_k^o\}) = \sum_{k=1}^{N} \psi(\|[I - P(\underline{\theta})]\underline{y}_k\|^2)$.

cally increasing assumption of the nonlinear function $\psi(x)$. Moreover, the optimum selection of the perturbation vector is readily obtained by making a Taylor series expansion of perturbed criterion $f(\underline{\theta} + \underline{\delta}, \{\underline{x}_k^o\})$ about the point $(\underline{\theta}, \{\underline{x}_k^o\})$. It is interesting to note that the optimum selection for the perturbation vector (89) is in agreement with the standard least-squares error choice (85) since in that case $\dot{\psi}(x) = 1$ for $x > 0$.

Utilization of a nonlinear squared error criterion of form (87) can be employed to achieve significant improvement in parameter estimation performance over that obtained with the standard least-squares approach. For example, if it is known that a subset of the data vector samples are not as reliable as others (i.e., data outliers), then a choice for $\psi(x)$ which saturates for increasing values of x will tend to mitigate the effects of the more unreliable data. As an example, the function $\psi(x) = b[1 - e^{-ax}]u(x)$ possesses this property where a and b are positive parameters. The parameter b controls the saturation level, while a determines the rate of saturation. The derivative of this function as given by $\dot{\psi}(x) = ae^{-ax}u(x)$ is seen to approach zero for significant large positive values of x. This derivative behavior is noteworthy since those model error terms $[I - P(\underline{\theta})]\underline{y}_n$ which are large in size (i.e., potential data outliers) do not greatly influence the perturbation vector (89). A judicious choice for a and b can therefore result in significantly improved modeling performance relative to that achieved with the classical least squares error approach when data outliers are present. The benefits of this approach have been demonstrated for the direction-of-arrival problem in which the saturating function $\psi(x)$ is taken to be the *sigmoid* function [19].

7. LINEAR RECURSIVE MODELING AND IDENTIFICATION

To illustrate the concepts developed in the last section, we now examine the important application problem of establishing whether there exists a functional relationship between the elements of two time series. In particular, let there be given the following finite-length time-series sample pair:

$$(x_n, y_n) \quad \text{for} \quad 1 \leq n \leq N. \tag{90}$$

Based on the physical laws governing the process from which these time-series samples arose or on simple intuition, it is hypothesized that the elements of the individual time series $\{x_n\}$ and $\{y_n\}$ are interdependent. This interdependency is based on the hypothesis that the x_n variables give rise to the variables y_n in a cause–effect manner. To test this conjecture, one may employ any number of linear or nonlinear models. A commonly invoked model is governed by a linear recursive system of order (p, q) as described by (e.g., see Ref. [4,5,8,16,17])

$$\hat{y}_n(\underline{a}, \underline{b}) = -\sum_{k=1}^{p} a_k \hat{y}_{n-k}(\underline{a}, \underline{b}) + \sum_{k=0}^{q} b_k x_{n-k}. \tag{91}$$

This model's response has been explicitly expressed as a function of the model parameter vectors $\underline{a}$ and $\underline{b}$ whose a_k and b_k components are taken to be real valued for presentation simplicity. A straightforward modification of the procedure to be now given can be made for the case of complex valued parameters.

The linear recursive model parameters are to be selected so that the sequence $\hat{y}_n(\underline{a}, \underline{b})$ generated by relationship (91) best approximates the given measurements $y_1, y_2, \ldots, y_N$. In keeping with the squared error criterion being employed in this paper, we seek a selection of the parameter vectors $\underline{a}$ and $\underline{b}$ so as to minimize the sum of squared errors criterion

$$f(\underline{a}, \underline{b}) = \|\underline{y} - \underline{\hat{y}}(\underline{a}, \underline{b})\|^2, \tag{92}$$

where $\underline{y}$ and $\underline{\hat{y}}(\underline{a}, \underline{b})$ are each $N \times 1$ vectors whose components correspond to the given measurement (90) and the model response (91), respectively. We now formulate this modeling problem to put it into the form described in the last section.

The model response elements (91) are dependent on the a_k and b_k parameters in a highly nonlinear manner due to the recursive (feedback) nature of this recursive model. In order to minimize squared error criterion (92), it is therefore necessary to employ nonlinear programming methods for obtaining an optimum selection for these $p + q + 1$ parameters. In order to make such an approach computational efficient, a scheme for decreasing the number of variables to be optimized from $p + q + 1$ to p is now described. This entails evaluating the components of an auxiliary sequence as governed by

$$w_n(\underline{a}) = x_n - \sum_{k=1}^{p} a_k w_{n-k}(\underline{a}) \quad \text{for} \quad 1 \leq n \leq N. \tag{93}$$

This auxiliary sequence elements $\{w_n(\underline{a})\}$ have been explicitly expressed as a function of the $\underline{a}$ parameter vector to emphasize this dependency. It has here been implicitly assumed that the elements of the time se-

ries $\{x_n\}$ and $\{y_n\}$ are identically zero for all indices less than or equal to zero.[4] It is readily shown that the linear recursive model (91) response is given by

$$\hat{y}_n(\underline{a}, \underline{b}) = b_0 w_n(\underline{a}) + b_1 w_{n-1}(\underline{a}) + \cdots + b_q w_{n-q}(\underline{a}). \tag{94}$$

An evaluation of this expression for $n = 1, 2, \ldots, N$ is next made and put into the vector format

$$\underline{\hat{y}}(\underline{a}, \underline{b}) = A(\underline{a})\underline{b}. \tag{95}$$

The full rank $N \times (q + 1)$ lower triangular Toeplitz matrix $A(\underline{a})$ here appearing is specified by

$$A(\underline{a}) = \begin{bmatrix} w_1(\underline{a}) & 0 & \cdots & 0 \\ w_2(\underline{a}) & w_1(\underline{a}) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ w_{q+1}(\underline{a}) & w_q(\underline{a}) & \cdots & w_1(\underline{a}) \\ \vdots & \vdots & \vdots & \vdots \\ w_N(\underline{a}) & w_{N-1}(\underline{a}) & \cdots & w_{N-q}(\underline{a}) \end{bmatrix}. \tag{96}$$

The implications of relationship (95) are noteworthy in that the model response is seen to depend on the b_k parameters in a linear fashion. If this vector response expression is substituted into squared error criterion (92), it directly follows that

$$f(\underline{a}, \underline{b}) = \|\underline{y} - A(\underline{a})\underline{b}\|^2. \tag{97}$$

This functional is seen to correspond to the squared error criterion associated with the inconsistent system of linear equations $A(\underline{a})\underline{b} \approx \underline{y}$. We may therefore employ the result of the last few sections to obtain an optimum recursive model. In particular, fundamental Theorem 1 indicates that for any selection of the parameter vector $\underline{a}$, the unique minimizing selection for $\underline{b}$ is given by

$$\underline{\hat{b}}_{LS}(\underline{a}) = A(\underline{a})^{\dagger}\underline{y}. \tag{98}$$

Furthermore, upon substitution of this optimum selection into criterion (97) and then dividing this criterion by $\|\underline{y}\|$, we obtain the normalized sum of errors criterion

$$f_{norm}(\underline{a}) = \frac{f(\underline{a}, \underline{\hat{b}}_{LS}(\underline{a}))}{\|\underline{y}\|^2} = \frac{\|[I - P(\underline{a})]\underline{y}\|^2}{\|\underline{y}\|^2}, \tag{99}$$

where $P(\underline{a}) = A(\underline{a})A(\underline{a})^{\dagger}$ corresponds to the orthogonal projection matrix associated with matrix $A(\underline{a})$. The purpose of dividing the original criterion by $\|\underline{y}\|$ is to provide a measure for judging the goodness of the recursive model which is independent of the magnitude of the data being modeled. In particular, this normalized criterion is seen to take on values exclusively in the interval [0,1] with values close to zero (one) indication a good (poor) recursive model approximation.

The linear recursive modeling problem has therefore been formulated into the problem considered in the last section. In order to determine the vector $\underline{a}$ that minimizes squared error criterion (99), it will be necessary to employ a nonlinear programming method. If the linearization algorithm as described in Table 1 is to be used, it is necessary to compute the partial derivatives of $P(\underline{a})$ with respect to the components a_k. From relationship (76), this is in turn seen to entail the determination of the derivatives of $A(\underline{a})$ with respect to the components a_k. For example, the partial derivative of $A(\underline{\theta})$ with respect to a_1 is from expression (96) found to be

$$\frac{\partial A(\underline{a})}{\partial a_1} = \begin{bmatrix} \frac{\partial w_1(\underline{a})}{\partial a_1} & 0 & \cdots & 0 \\ \frac{\partial w_2(\underline{a})}{\partial a_1} & \frac{\partial w_1(\underline{a})}{\partial a_1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial w_{q+1}(\underline{a})}{\partial a_1} & \frac{\partial w_q(\underline{a})}{\partial a_1} & \cdots & \frac{\partial w_1(\underline{a})}{\partial a_1} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial w_N(\underline{a})}{\partial a_1} & \frac{\partial w_{N-1}(\underline{a})}{\partial a_1} & \cdots & \frac{\partial w_{N-q}(\underline{a})}{\partial q_1} \end{bmatrix}. \tag{100}$$

The partial derivatives $\partial w_n(\underline{a})/\partial a_1$ appearing in this expression are obtained by taking the derivative of auxiliary sequence relationship (93) with respect to a_1. This differentiation gives rise to the following recursive relationship:

$$\frac{\partial w_n(\underline{a})}{\partial a_1} = -w_{n-1}(\underline{a}) - \sum_{k=1}^{p} a_k \frac{\partial w_{n-k}(\underline{a})}{\partial a_1}. \tag{101}$$

[4] If this assumption cannot be justified, then the auxiliary time series is alternatively evaluated on the smaller integer set $p + 1 \leq n \leq N$ where all samples of the time series are known.

The required entries to partial derivative matrix (100) are then obtained by evaluating this recursive expression for $1 \leq n \leq N$ in which zero initial conditions are assumed.

The partial derivatives of $A(\underline{\theta})$ with respect to the general coefficient a_k are readily shown to be given by

$$\frac{\partial A(\underline{a})}{\partial a_k} = S^{k-1} \frac{\partial A(\underline{a})}{\partial a_1} \quad \text{for} \quad 2 \leq k \leq p, \quad (102)$$

where S designates the $N \times (N-1)$ *down-shifting* matrix whose elements are all zero except for ones that appear along the diagonal immediately below the main diagonal (i.e., $S(i,j) = \delta(i-j-1)$). Thus, the partial derivative $\partial A(\underline{a})/\partial a_k$ is obtained by simply appending $(k-1)$ rows of zeros to the top of $\partial A(\underline{a})/\partial a_1$ and simultaneously dropping the last $(k-1)$ rows.

Initial Parameter Selection

An important consideration for the successful employment of any nonlinear programming algorithm for the minimization of a functional dependent on a set of parameters is the selection of a good set of parameters values to initiate the algorithm (e.g., see Step 1 in Table 1). If the initial parameters are chosen too far from their unknown optimum values, the distinct possibility exists that the algorithm will converge to a poor relative minimum of the functional. In many minimization problems there exists no systematic method for making a good initial selection. Fortunately, this is not the case for the linear recursive modeling problem here being considered. A method for making an initial parameter selection that typically leads to effective recursive modeling is now presented [16].

If the recursive model being employed is of high quality, it follows that the time series $\{y_n\}$ and $\{\hat{y}_n(\underline{a}, \underline{b})\}$ will be almost identical. Under this assumption, let us replace the element $\hat{y}(\underline{a}, \underline{b})$ appearing in recursive model (91) by y_n. Since these two time series are not identical, it follows that this substitution results in the *residual error sequence* as defined by

$$\epsilon_n = y_n + \sum_{k=1}^{p} a_k y_{n-k} - \sum_{k=0}^{q} b_k x_{n-k} \quad \text{for} \quad 1 \leq n \leq N. \quad (103)$$

The a_k and b_k parameters are to be chosen so as to minimize the sum of residual squared errors. To effect this minimization, it is useful to express these residual errors in the vector format

$$\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ y_1 & 0 & \cdots & 0 \\ y_2 & y_1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ y_p & y_{p-1} & \cdots & y_1 \\ \vdots & \vdots & \cdots & \vdots \\ y_{N-1} & y_{N-2} & \cdots & y_{N-p} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} - \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ x_2 & x_1 & \cdots & 0 \\ x_3 & x_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ x_{q+1} & x_q & \cdots & x_1 \\ \vdots & \vdots & \cdots & \vdots \\ x_{N-1} & x_{N-2} & \cdots & x_{N-q-1} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_q \end{bmatrix}, \quad (104)$$

or, more compactly,

$$\underline{\epsilon} = \underline{y} + [Y - X] \begin{bmatrix} \underline{a} \\ \underline{b} \end{bmatrix}. \quad (105)$$

From Fundamental Theorem 1 it follows that the selection of the parameter vectors $\underline{a}$ and $\underline{b}$ that minimize the sum of squared residual errors is given by

$$\begin{bmatrix} \underline{a}^o \\ \underline{b}^o \end{bmatrix} = -[Y - X]^{\dagger} \underline{y} = -\begin{bmatrix} Y^T Y & -Y^T X \\ -X^T Y & X^T X \end{bmatrix}^{-1} \begin{bmatrix} Y^T \underline{y} \\ -X^T \underline{y} \end{bmatrix}. \quad (106)$$

It has been empirically found that when this initial parameter vector selection is used in Step 1 of the algorithm described in Table 1, the recursive model that iteratively arises is generally acceptable in most applications.

EXAMPLE 2. *When synthesizing frequency discrimination filters, the zero-phase ideal lowpass, bandpass, and highpass filters often serve as objectives to be approximated. For each of these filters, the symmetric infinite length sequence*

$$h(n) = \frac{\sin(n\omega_c)}{n\pi} \quad (107)$$

TABLE 2

Recursive Lowpass Filter Approximation

Model order	$f_{\text{norm}}(\underline{a}^{0})$	$f_{\text{norm}}(\underline{a}^{\text{opt}})$	Number of iterations
(3, 3)	1.3278×10^{-2}	7.7083×10^{-3}	17
(6, 6)	4.3181×10^{-4}	3.2079×10^{-5}	8
(9, 9)	3.2942×10^{-6}	4.2189×10^{-8}	8
(12, 12)	4.1673×10^{-8}	1.1953×10^{-10}	8

plays a central role.[5] *The recursive model as governed by relationship* (91) *is now employed to approximate the truncated causal component of this sequence as defined by* $y(n) = h(n-1)$ *for* $1 \leq n \leq N$ *in which* $\omega_c = 0.25\pi$ *and* $N = 128$. *In this approximation, the* $N \times 1$ *vector* $\underline{x}$ *has all its components equal to zero except for its first, which equals one.*

In employing the linearization algorithm outlined in Table 1, the initial parameters were selected according to expression (106). *Furthermore, the stopping condition was invoked whenever successive values of the normalized criterions'* (99) *had a relative change of less than 0.0001 from one iteration to the next (i.e.,* $f_{norm}(\underline{a}_k) - f_{\text{norm}}(\underline{a}_{k-1}) < 10^{-4} f_{\text{norm}}(\underline{a}_{k-1})$). *Under these conditions, the linearization algorithm was employed with a choice of* $N = 128$ *and for various values of the order parameters* (p, q). *The results of the algorithm are described in Table 2. It is apparent from these tables that the parameter initialization selection (106) proved satisfactory and the algorithm improved on these initial choices by a considerable factor. A plot of the unit-impulse response that arises from the linear recursive model of order (6,6) is shown in Fig. 2, as is the ideal unit-impulse response (107). These responses are virtually indistinguishable over the interval* [0, 127], *indicative of the quality of approximation.*

8. GENERALIZED NONLINEAR LEAST-SQUARES MODELING

It is possible to straightforwardly modify the concepts presented in Section 5 to treat a more general modeling problem. In particular, let there be given a vector $\underline{y} \in C^m$ that is to be approximated by the vector entity $\tilde{F}(\underline{x}, \tilde{\underline{\theta}})$ where $\tilde{F}(\cdot)$ corresponds to a nonlinear mapping from $C^n \times R^{\tilde{p}}$ into C^m. The accuracy of this model is measured by the squared error criterion

$$\tilde{f}(\underline{x}, \tilde{\underline{\theta}}) = \|\underline{y} - \tilde{F}(\underline{x}, \tilde{\underline{\theta}})\|^2. \tag{108}$$

The vectors $\underline{x}$ and $\tilde{\underline{\theta}}$ are to be selected so as to minimize this squared error criterion. For notational simplification, it is useful to form the real $(2n + \tilde{p}) \times 1$ parameter vector $\underline{\theta}$ whose components are comprised of the real and imaginary components of $\underline{x}$ and the components of $\tilde{\underline{\theta}}$. Using this representation, the squared error criterion (108) can be equivalently expressed as

$$f(\underline{\theta}) = \|\underline{y} - F(\underline{\theta})\|^2, \tag{109}$$

where $F(\underline{\theta}) = \tilde{F}(\underline{x}, \tilde{\underline{\theta}})$, $f(\underline{\theta}) = \tilde{f}(\underline{x}, \tilde{\underline{\theta}})$, and, $\underline{\theta} = [\tilde{\underline{\theta}}^T, \text{Real}\{\underline{x}\}^T, \text{Imag}\{\underline{x}\}^T]$. Upon comparison of this criterion with criterion (67), they are seen to be of a similar form with the projection $P(\underline{\theta})\underline{y}$ being replaced by nonlinear mapping $F(\underline{\theta})$. In order to minimize criterion (108), we shall therefore employ the descent approach taken in Section 5.

If the prevailing value of the parameter vector $\underline{\theta}$ is additively perturbed by $\underline{\delta}$, the perturbed squared error criterion (109) is approximated by truncating the Taylor series expansion of $F(\underline{\theta} + \underline{\delta})$, to its first two terms, that is,

$$\begin{aligned} f(\underline{\theta} + \underline{\delta}) &= \|\underline{y} - F(\underline{\theta} + \underline{\delta})\|^2 \\ &\approx \left\| \underline{y} - F(\underline{\theta}) - \sum_{k=1}^{p} \frac{\partial F(\underline{\theta})}{\partial \theta_k} \delta_k \right\|^2 \\ &= \|\underline{y} - F(\underline{\theta}) - L(\underline{\theta})\underline{\delta}\|^2, \end{aligned} \tag{110}$$

where $\partial F(\underline{\theta})/\partial \theta_k$ are $m \times 1$ vectors while the δ_k entities are the components of the perturbation vector $\underline{\delta}$.

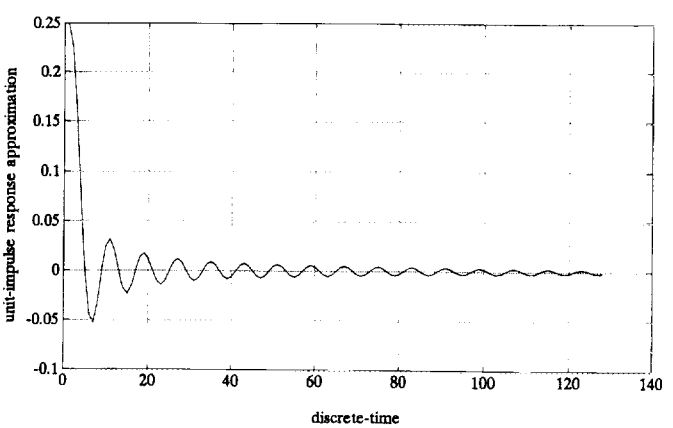

FIG. 2. Plot of the ideal and recursive unit-impulse responses.

[5] Sequence (107) corresponds to the inverse Fourier transform of the ideal zero-phase lowpass filter with cutoff frequency ω_c whose transfer function is equal to one in the frequency interval $[-\omega_c, \omega_c]$ and is zero for all other frequencies in the interval $[-\pi, \pi]$.

In this truncated Taylor series expansion, the $m \times p$ *Jacobian matrix* $L(\underline{\theta})$ is specified by

$$L(\underline{\theta}) = \left[\frac{\partial F(\underline{\theta})}{\partial \theta_1} \vdots \frac{\partial F(\underline{\theta})}{\partial \theta_2} \vdots \cdots \vdots \frac{\partial F(\underline{\theta})}{\partial \theta_p} \right]. \quad (111)$$

This Jacobian matrix provides a first-order approximation of the effect that incremental changes made in the prevailing parameter vector have on the Euclidean norm criterion being minimized.

A logical choice for the perturbation vector would be one that minimizes the approximation of the Euclidean norm criterion as specified by relationship (110). To obtain an expression for this selection, let us expand this squared error criterion is obtain

$$f(\underline{\theta} + \underline{\delta}) = f(\underline{\theta}) - \underline{\delta}^{\mathrm{T}} L(\underline{\theta})^* [\underline{y} - F(\underline{\theta})] - [\underline{y} - F(\underline{\theta})]^* L(\underline{\theta}) \underline{\delta} + \underline{\delta}^{\mathrm{T}} L(\underline{\theta})^* L(\underline{\theta}) \underline{\delta}. \quad (112)$$

Upon setting the gradient of this expression with respect to the real vector $\underline{\delta}$ equal to the zero vector, the optimal selection for the perturbation vector is found to be

$$\underline{\delta}^{\circ} = [\mathrm{Real}\{L(\underline{\theta}^* L(\underline{\theta})\}]^{\dagger} \mathrm{Real}\{L(\underline{\theta})^* [\underline{y} - F(\underline{\theta})]. \quad (113)$$

9. CONCLUSION

A development of basic principles related to the characterization and use of systems of linear and nonlinear equations for modeling data has been presented. Emphasis has been given to the notions of projection matrices and Moore–Penrose generalized inverses.

REFERENCES

1. Adcock, R. J. A problem in least square. *The Analyst*, **5** (1878), 53–54.
2. Anderson, T. W. The 1982 Wald memorial lectures: Estimating linear statistical relationships. *Ann. Statist.* **12** (1984), 1–45.
3. Cadzow, J. A. Signal processing via least squares error modeling. *IEEE ASSP Mag.* (Oct. 1990), 12–31.
4. Cadzow, J. A. Recursive digital filter synthesis via gradient based algorithms. *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34,** No. 5 (Oct. 1976), 349–355.
5. Cadzow, J. A., and Solomon, O. M. Algebraic approach to system identification. *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34,** No. 3 (June 1988), 462–469.
6. Campbell, S. L., and Meyer, C. D. *Generalized Inverses of Linear Transformations.* Pitman, London, 1979.
7. Eckart, G., and Young, G. The approximation of one matrix by another of lower rank. *Psychometrica* **1** (1936), 211–218.
8. Evans, E. G., and Fischl, R., Optimal least squares time-domain synthesis of recursive digital filters. *IEEE Trans. Audio Electroacoust.* **21** (Feb. 1973), 61–65.
9. Gauss, C. F. Theoria combinationis observationum erroribus minimis obnoxiae. *Comment. Soc. Reg. Sci. Gotten. Recent.* **5** (1823), 33–90.
10. Gleser, L. J. Estimation in a multivariate "errors in variables" regression model: Large sample results. *Ann. Statist.* **9** (1981), 24–44.
11. Golub, G. H., and Van Loan, C. F. *Matrix Computations.* John Hopkins Univ. Press, Baltimore, 1993.
12. Lancaster, P., and Tismenetsky, M. *The Theory of Matrices,* 2nd Ed., Academic Press, Orlando, 1985.
13. Ortega, J. M. *Matrix Theory.* Plenum Press, New York, 1987.
14. Pearson, K., On lines and planes of closest fit to points in space. *Philos. Mag.* **2** (1901), 559–572.
15. Scharf, L. L. *Statistical Signal Processing.* Addison-Wesley, Reading, Maryland, 1990.
16. Shanks, J. L. Recursion filters for digital processing. *Geophysics* **XXXII,** No. 1 (Feb. 1967), 33–51.
17. Steiglitz, K., and McBride, L. E. Technique for the identification of linear systems. *IEEE Trans. Automat. Control* **10** (Feb. 1973), 61–65.
18. Stewart, G. W. *Introduction to Matrix Computations.* Academic Press, New York, 1973.
19. Yardimci, Y., and Cadzow, J. A. Robust data modeling through nonlinear least squares. Submitted for publication.

JAMES A. CADZOW was born in Niagara Falls, NY, on January 3, 1936. He received the B.S. and M.S. degrees in electrical engineering from the State University of New York at Buffalo in 1958 and 1963, respectively, and the Ph.D. degree from Cornell University, Ithaca, NY in 1964. From 1958 to 1963 he was associated with the USARDL, Fort Monmouth, NJ; Bell Aerosystems, Wheatfield, NY; and Cornell Aeronautical Laboratories, Buffalo, NY. He was a professor electrical engineering at SUNY at Buffalo from 1964–1977 and at Virginia Polytechnic Institute from 1977–1981. In 1981, Professor Cadzow was appointed Research Professor of Electrical Engineering at Arizona State University and served in that role until 1988 when he accepted a Centennial Professorship at Vanderbilt University. In addition, he was a visiting professor of electrical engineering at Stanford University, Stanford, CA, from 1968–1969, a visiting professor and National Institute of Health fellow at the Department of Biomedical Engineering, Duke University, Durham, NC, from 1972–1973 and a visiting professor of electrical engineering at the University of California, San Diego, La Jolla, CA, from 1987–1988. Professor Cadzow has authored the textbooks *Foundations of Digital Signal Processing and Data Analysis* (Macmillan, New York, 1987), *Signals, Systems and Transforms* (Prentice-Hall, Englewood Cliffs, NJ, 1985), *Discrete-Time Systems* (Prentice-Hall, Englewood Cliffs, NJ, 1973), and *Discrete-Time and Computer Control Systems* (Prentice-Hall, Englewood Cliffs, NJ, 1970). His research interests include signal processing, communication and control theory, system identification and modeling, and neural networks. Dr. Cadzow is an IEEE fellow and served as Chairman and Vice Chairman of the *Spectral Estimation and Modeling* Technical Committee of ASSP and he is now Associate Editor for the *IEEE ASSP Magazine* and the *Journal of Time Series Analysis.* He is also a member of Sigma Xi and Phi Kappa Phi.