DIGITAL SIGNAL PROCESSING 4, 21–39 (1994)

# Total Least Squares, Matrix Enhancement, and Signal Processing*

James A. Cadzow

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

Cadzow, J. A., Total Least Squares, Matrix Enhancement, and Signal Processing, *Digital Signal Processing* 4 (1994), 21–39.

The concept of *least squares* (LS) as applied to a system of generally inconsistent linear equations $A\underline{x} \approx \underline{y}$ plays a central role in algorithms used for solving a variety of important signal processing problems. In the standard LS approach, the smallest unconstrained Euclidean norm perturbation of vector $\underline{y}$ is sought so that the resultant perturbed system of linear equations $A\underline{x} = \underline{y} + \underline{\delta}$ has a solution. Implicit in the LS problem formulation is the assumption that only $\underline{y}$ vector is subject to error. Depending on the nature of the application, however, system matrix $A$ may also be subject to error. When both the system matrix and right side vector are imprecisely known, the use of a least-squares solution can lead to disappointingly poor results. To mitigate the effects of errors in both $A$ and $\underline{y}$ the concept of *total least squares* (TLS) is invoked. In the standard TLS problem, the smallest unconstrained perturbation of the matrix–vector pair $(A, \underline{y})$ is sought so that the perturbed system of equations $[A + \Delta]\underline{x} + = \underline{y} + \underline{\delta}$ has a solution. In this paper, theoretical issues regarding TLS are first addressed in which a *singular value decomposition* approach is used to find the optimal unconstrained choices for the perturbations $\Delta$ and $\delta$. The TLS concept is next extended to the case in which the perturbed matrix–vector pair $(A + \Delta, \underline{y} + \underline{\delta})$ identifying the perturbed linear equations is constrained to satisfy prescribed properties known (or hypothesized) to be possessed in the idealized case. These property constraints take on such disparate forms as requiring that the perturbed auxiliary matrix $[A + \Delta - \underline{y} - \underline{\delta}]$ be positive definite, Hermitian, block Toeplitz, or of a given nonfull rank. Property requirements of this nature characterize a variety of practical signal processing problems. Applications of these concepts to the problems of exponential data modeling and system identification are then made where improvements over more traditional solution procedures are realized. The notion of constrained data perturbations has also been successfully applied to the problems of synthesizing one-dimensional and higher dimensional linear systems, obtaining positive sequence approximations, and missing data interpolation.

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

## 1. LEAST-SQUARES ERROR REVIEW

When formulating a solution procedure for a prescribed signal processing problem, a subsidiary requirement of solving a consistent system of linear equations or of finding a best approximate solution to an inconsistent system of inconsistent linear equations often arises. The individual involved in signal processing research or its application therefore has to have a good grasp of fundamental theoretical issues related to systems of linear equations. With this in mind, let us consider the following generally inconsistent system of linear equations:

$$A\underline{x} \approx \underline{y}. \tag{1}$$

In this representation $A$ is the $m \times n$ *system matrix* while $\underline{x}$ and $\underline{y}$ are $n \times 1$ and $m \times 1$ vectors, respectively. The convention of designating vectors by underlined roman letters (i.e., $\underline{x}$) is here adapted. The elements of vectors $\underline{x}$ and $\underline{y}$ and matrix $A$ may be either real or complex valued. In order that our results be applicable to the broadest class of problems, the more general case of a complex system of linear equations is treated. By taking this approach, we are also treating real systems of linear equations as a special case. Specifically, any theorem related to complex linear equations may be immediately converted to its real case counterpart by the simple process of dropping all complex conjugate operations that appear. This is usually manifested by the replacement of all

1051-2004/94 $5.00

complex conjugate transposition operations by transposition operations so that any conjugate transposed matrix $A^*$ is replaced by the transposed matrix $A^T$. By adopting this approach, we avoid the duplicative effort of treating the real and complex cases separately.

The system of linear equations (1) corresponds to a system of $m$ linear equations in the $n$ unknown elements comprising vector $\underline{x}$. The most commonly occurring problem is concerned with the case of an *overdetermined* system of equations in which there are more equations than unknowns (i.e., $m > n$). Applications involving an *underdetermined* system of equations $m < n$ also arises from time to time but are of less interest. Whatever the case, if there exists a selection of $\underline{x}$ for which relationship (1) is satisfied (i.e., $A\underline{x} = \underline{y}$) then that choice is called a *solution* and the system of equations is said to be *consistent.* If no such choice of $\underline{x}$ exists, the system of equations is said to be *inconsistent.* In a typical application, we are confronted with an inconsistent system of overdetermined linear equations. This inconsistency can be caused by such factors as (i) measurement noise whereby inaccurate values for the elements of $\underline{y}$ and (or) $A$ are used or (ii) the hypothesized linear model only approximately represents the underlying functional relationship between vectors $\underline{x}$ and $\underline{y}$.

### *Least-Squares Error Approximate Solution*

The classical least-squares error (LSE) approach in finding an approximate solution to 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 $\underline{y}$. In particular, the vector $\underline{x}$ is selected so as to make the *equation error vector* defined by $\underline{e} = A\underline{x} - \underline{y}$ as small as possible in the Euclidean norm sense. To pose the least-squares error problem in a format consistent with a total least squares approach, an equivalent interpretation is now described. Specifically, let it be desired to select the smallest Euclidean sized additive perturbation of vector $\underline{y}$ so that the perturbed system of linear equations

$$A\underline{x} = \underline{y} + \underline{\delta} \tag{2}$$

is consistent. The perturbation vector $\underline{\delta}$ is seen to correspond to the *equation error vector* as mentioned above. Thus, in seeking the smallest perturbation of vector $\underline{y}$ so as to render a consistent system of perturbed equations we are simultaneously obtaining the smallest equation error vector and vice versa. Although a number of different means for measuring the size of the perturbation vector are available, the squared Euclidean norm as defined by

$$\|\underline{\delta}\|^2 = \underline{\delta}^*\underline{\delta} \tag{3}$$

is the one invoked in a least-squares approach.[1] The superscript symbol (*) here appearing designates the complex conjugate operation and the subscript 2 on this Euclidean norm (see relationship (4)) has been dropped for reasons of notational simplicity.

The least squares error problem corresponds to selecting perturbation vector $\underline{\delta}$ so as to solve the following constrained optimization problem:

$$\min_{A\underline{x}=\underline{y}+\underline{\delta}} \underline{\delta}^*\underline{\delta}. \tag{5}$$

It is possible to transform this constrained optimization problem into an unconstrained optimization problem which is conceptually simpler to solve. To effect this transformation use is made of the observation that for a consistent system of perturbed equations the perturbation vector must satisfy the identity $\underline{\delta} = A\underline{x} - \underline{y}$. Upon substituting this identity into the criterion $\underline{\delta}^*\underline{\delta}$ being minimized, the following equivalent least-squares functional is obtained:[2]

$$\begin{aligned} f(\underline{x}) &= \underline{\delta}^*\underline{\delta} = (\underline{x}^*A^* - \underline{y}^*)(A\underline{x} - \underline{y}) \\ &= \underline{x}^*A^*A\underline{x} - \underline{x}^*A^*\underline{y} - \underline{y}^*A\underline{x} + \underline{y}^*\underline{y}. \end{aligned} \tag{6}$$

A solution to constrained optimization problem (5) may then be alternatively obtained by finding a vector $\underline{x}^o$ which minimizes this function. This is an unconstrained minimization problem since no restrictions are placed on the choice for an optimum $\underline{x}$. Once an optimum vector $\underline{x}^o$ has been found, the associated minimum Euclidean norm perturbation vector that results in a consistent perturbed system of linear equations is specified by $\underline{\delta}^o = A\underline{x}^o - \underline{y}$.

In order to find a solution for the LSE problem use is made of a fundamental theorem from calculus which states that a necessary condition for a function of $n$ real variables to assume a local minimum at a

[1] The $\ell_p$ norm of the perturbation vector as specified by

$$\|\delta\|_p = [\sum_{k=1}^{m} |\delta(k)|^p]^{1/p} \tag{4}$$

for any choice of $p \geq 1$ serves as a useful choice of perturbation vector size. The selections $p = 1, 2, \infty$ provide the most frequently used selections with $p = 2$ commonly referred to as the Euclidean norm.

[2] No loss of generality is incurred by using the unweighted squared error criterion (6). 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 $f(\underline{x}) = (\underline{x}^*\tilde{A}^* - \tilde{\underline{y}}^*)(\tilde{A}\underline{x} - \underline{y})$.

point is that the $n$ first derivatives of that function evaluated at that point all be zero. With this in mind, the required LS solution is obtained by setting to zero the partial derivatives of functional (6) with respect to the real and imaginary components of $\underline{x}$. This gives rise to an associated consistent system of linear *normal* equations that provide the necessary (and sufficient) condition for minimizing this functional. In a companion paper, the following fundamental theorem captures the essence of the least-squares error solution [4].

THEOREM 1. *Consider the generally inconsistent system of m linear equations in n unknowns as represented by* $A\underline{x} \approx \underline{y}$. *It follows that any choice of the vector* $\underline{x}$ *that satisfies the consistent linear system of* normal *equations*

$$A^*A\underline{x} = A^*\underline{y} \tag{7}$$

*minimizes the squared Euclidean norm of error vector* $\underline{e} = A\underline{x} - \underline{y}$. *Furthermore, all solutions* $\tilde{\underline{x}}$ *to these normal equations have associated equation error vectors that have the same squared Euclidean norm as given by*

$$\|A\tilde{\underline{x}} - \underline{y}\|^2 = \|[I - P_A]\underline{y}\|^2, \tag{8}$$

*where* $P(A)$ *denotes the* $m \times m$ *orthogonal projection matrix whose range space equals the range space of matrix A. The unique minimum Euclidean norm solution to the consistent system of normal equations (7) is specified by*

$$\underline{x}_{LS}^{o} = A^{\dagger}\underline{y}, \tag{9}$$

*where* $A^{\dagger}$ *designates the Moore–Penrose generalized inverse of system matrix A. If A has full column rank n then the associated orthogonal projection matrix and Moore–Penrose generalized inverse matrices are specified by*

$$P_A = A[A^*A]^{-1}A \quad \text{and} \quad A^{\dagger} = [A^*A]^{-1}A^*. \tag{10}$$

## 2. TOTAL LEAST SQUARES

In applications related to the generally inconsistent system of linear equations $A\underline{x} \approx \underline{y}$, the elements of the system matrix $A$ and right side vector $\underline{y}$ are often each subject to error. If the classical least-squares error approach is employed in analyzing these equations, the resultant approximate solution is usually poor in quality because errors in the system matrix have not been taken into account. To mitigate the deleterious effects of these errors, it is possible to straightforwardly modify the perturbation concept invoked in the least-squares approach. This modification gives rise to the so-called *total-least-squares* (TLS) problem. Total least squares has a rich history dating back more than one century. Adcock first studied the univariate version of this problem [2] with other contributions being made by Pearson [24], Koopmans [19], Madansky [22], York [29], and others. A generalization of TLS to the multivariate case (i.e., $n > 1$) was advanced by several authors including Golub and Van Loan [15,16], Sprent [26] (for a more complete listing see [28]). Gleser further generalized TLS to the case of multiple systems of linear equations [14]. A very readable summary of the present state of TLS is to be found in Van Huffel and Vandewalle [28].

When invoking a total least squares error interpretation to an inconsistent system of linear equations $A\underline{x} \approx \underline{y}$, it is useful to express these equations in their equivalent homogeneous format

$$[A - \underline{y}]\begin{bmatrix}\underline{x}\\ 1\end{bmatrix} \approx \underline{0}, \tag{11}$$

where $\underline{0}$ denotes the zero vector. It is therefore clear that the original system of linear equations is consistent if and only if there exists a $(n + 1) \times 1$ vector whose last component is one that maps into the zero vector under the $m \times (n + 1)$ *augmented matrix* as defined by

$$[A - \underline{y}]. \tag{12}$$

This augmented matrix is obtained by appending the column vector $-\underline{y}$ to the system matrix. If the system matrix $A$ has rank $r$, it follows that the rank of this augmented matrix must be either $r$ or $r + 1$. Thus, a necessary and sufficient condition that the original system of linear equations be consistent is that the rank of the augmented matrix also be $r$. This is an immediate consequence of the observation that under this equivalent rank condition, vector $\underline{y}$ must be expressible as a linear combination of $A$'s column vectors. On the other hand, if $\underline{y}$ cannot be expressed in this linear combination format, then the original system of linear equations must be inconsistent thereby establishing that the augmented matrix has rank $r + 1$. It will be useful to formally emphasis these observations in terms of the singular value decomposition (SVD) of the system matrix $A$ and its associated augmentation $[A - \underline{y}]$. Since the rank of system matrix $A$ has rank $r$, its SVD takes the form

$$A = \sum_{k=1}^{r} \sigma_k \underline{u}_k \underline{v}_k^*. \tag{13}$$

In this SVD representation the positive valued $\sigma_k$ are called *singular values* and without loss of generality are ordered in the standard monotonically nonincreasing fashion $\sigma_k \geq \sigma_{k+1}$. Moreover, the associated $m \times 1$ *left singular vectors* $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_r$ and the $n \times 1$ *right singular vectors* $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_r$ forming the outer-products terms in this SVD representation each form orthonormal vector sets. We are now able to give a basic lemma relating to the consistency of a linear system of equations.

LEMMA 1. *The system of linear equations* $A\underline{x} \approx \underline{y}$ *is inconsistent if and only if every vector in the null space of the augmented matrix* $[A - \underline{y}]$ *has its last component equal to zero. Equivalently, this system of linear equations is inconsistent if and only if* $\text{rank}[A - \underline{y}] = r + 1$ *where the* $r = \text{rank}(A)$. *Moreover, the SVD of the augmented matrix for the case of an inconsistent system of linear equations takes the form*

$$[A - \underline{y}] = \sum_{k=1}^{r+1} \tilde{\sigma}_k \tilde{\underline{u}}_k \tilde{\underline{v}}_k^*, \tag{14}$$

*in which* $\tilde{\sigma}_{r+1}$ *designates the smallest positive singular value.*

### *Perturbed System of Linear Equations*

Since our primary interest is directed towards the case of an inconsistent system of linear equations, the properties expressed Lemma 1 are now used to formulate a solution procedure using a total-least-squares approach. When using the TLS concept, the system matrix and right side vector are each perturbed in the additive fashion $A + \Delta$ and $\underline{y} + \delta$, respectively. These perturbations are chosen so that the resultant perturbed system of linear equations as given by

$$[A + \Delta]\underline{x} = \underline{y} + \underline{\delta} \tag{15}$$

is consistent. It is to be noted that there will exist an uncountable infinite number of such perturbations that result in the required consistency. Since the perturbed system of equations should bear some resemblance to the original system of equations, it is logical to select the smallest possible perturbation that achieves equation consistency. To measure perturbation size, the Frobenius norm of the augmented perturbed matrix $[\Delta - \underline{\delta}]$ is used.[3]

[3] The Frobenius norm of the $M \times N$ matrix $B$ is defined by

$$\|B\|_F = \sqrt{\sum_{m=1}^{M} \sum_{n=1}^{N} |b_{mn}|^2}.$$

From the above remarks, the total least squares problem is concerned with solving the following constrained optimization problem:

$$\min_{[A+\Delta]\underline{x}=\underline{y}+\underline{\delta}} \|[\Delta - \underline{\delta}]\|_F. \tag{16}$$

A solution to this problem is readily obtained by appealing to the observation made in Lemma 1 that a necessary condition for the original system of linear equations to have a solution is that the rank of augmented matrix $[A - \underline{y}]$ by equal to $\text{rank}(A)$. When the augmented matrix has rank $1 + \text{rank}(A)$, however, a solution is not possible. We therefore seek a matrix $[\hat{A} - \underline{\hat{y}}]$ of rank $A$ that lies closest to augmented matrix $[A - \underline{y}]$ in the Frobenius norm sense. These concepts are now formally described.

THEOREM 2. *Consider the system of inconsistent linear equations* $A\underline{x} \approx \underline{y}$ *in which the rank of system matrix A is r and the SVD representation of the rank r + 1 augmented matrix* $[A - \underline{y}]$ *is given by relationship (14). Moreover, let the smallest nonzero singular value* $\tilde{\sigma}_{r+1}$ *be unique with* $\tilde{\underline{v}}_{r+1}$ *being its associated right singular vector. If the last component of* $\tilde{\underline{v}}_{r+1}$ *is nonzero (i.e.,* $\tilde{\underline{v}}_{r+1}(n+1) \neq 0$*), then the unique minimum Frobenius norm perturbation of the given system of linear equations that results in a consistent system of perturbed equations is specified by*

$$[\Delta^\circ - \underline{\delta}^\circ] = -\tilde{\sigma}_{r+1} \tilde{\underline{u}}_{r+1} \tilde{\underline{v}}_{r+1}^*, \tag{17}$$

*and the Frobenius norm of this perturbed auxiliary matrix is*

$$\|[\Delta^\circ - \underline{\delta}^\circ]\|_F = \tilde{\sigma}_{r+1}. \tag{18}$$

*Furthermore, the solution of the resultant consistent system of perturbed linear equations (15) that has the smallest Euclidean norm has its components specified by*

$$\underline{x}_{TLS}(k) = \frac{\tilde{\underline{v}}_{r+1}(k)}{\tilde{\underline{v}}_{r+1}(n+1)} \quad \text{for} \quad 1 \leq k \leq n \tag{19}$$

*and is referred to as the total-least-squares solution.*[4]

[4] The set of all solutions to this perturbed system of consistent linear equations is specified by

$$S_{TLS} = \underline{x}_{TLS} + \mathcal{N}(A + \Delta^\circ),$$

where $\Delta^\circ$ and $\underline{x}_{TLS}$ and specified in relationships (17) and (19), respectively, and $\mathcal{N}(A + \Delta^\circ)$ denotes the null space of matrix $A + \Delta^\circ$.

A proof of this theorem is straightforward since by selecting the TLS perturbations according to relationship (17), the perturbed augmented matrix is identical to expression (14) but with the upper sum limit $r + 1$ replaced by $r$. This reduced rank $r$ perturbed augmented matrix, however, is the closest rank $r$ matrix to the original augmented matrix in the Frobenius norm sense. It is interesting to note that the standard LSE solution (9) can be interpreted as corresponding to a TLS type perturbation in which the constraint $\Delta = 0$ is imposed. This indicates that the Euclidean norm of the perturbation vector $\underline{\delta}$ employed when taking an LS solution approach is *always* at least as large as the Frobenius norm of the perturbation matrix $[\Delta - \underline{\delta}]$ incurred when a TLS solution procedure is invoked.

This theorem is based on the assumption that the smallest singular value of the augmented matrix is unique. Although this assumption is met in most practical applications, in rare situations the minimum positive singular value is multiple (e.g., $\tilde{\sigma}_{r+1} = \tilde{\sigma}_r$). If the smallest nonzero singular value has multiplicity $q$, then at most only one of the $q$ associated right singular vectors need have a nonzero last component for this theorem to apply. A TLS solution is obtained by simply dropping any outerproduct associated with one of the smallest right singular vectors with nonzero last component. Moreover, if more than one of these smallest right singular vectors has a nonzero last component, it is easily shown that there exists an uncountable infinite number of minimum Frobenius norm perturbations that result in a consistent system of perturbed linear equations. On the other hand, if all of the right singular vectors associated with the smallest nonzero singular value have a zero last component, then another solution approach must be pursued. A plausible alternative procedure would be to find all right singular vector in the set $(\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_{r+1})$ that have a nonzero last components. There must exist at least one such vector since the appended right singular vectors form a basis of $C^n$. We then select that right singular value in this restricted set that has the smallest associated singular value. This right singular vector would then constitute a pseudo-TLS solution and the Frobenius norm of the corresponding perturbed auxiliary matrix would be equal to its associated singular value.

### *Special Case:* Rank$(A) = n$

In most practical applications, the rank of the $m \times n$ system matrix $A$ is $n$. If the associated system of linear equations $A\underline{x} \approx \underline{y}$ is inconsistent then the augmented matrix $[A \; \underline{y}]$ has full rank $n + 1$. From Theorem 2 it is seen that the TLS solution has its components specified by

$$\underline{x}_{\mathrm{TLS}}(k) = \frac{\tilde{v}_n(k)}{\tilde{v}_n(n+1)} \quad \text{for} \quad 1 \leq k \leq n \qquad (20)$$

provided that the last component of $\underline{\tilde{v}}_{n+1}$ is nonzero. It is apparent that for this full rank case, the two requirements that (i) $\tilde{\sigma}_n > \tilde{\sigma}_{n+1}$ and (ii) $\tilde{v}_{n+1}(n+1) \neq 0$ will result in the perturbed system of equations (15) having the unique solution and is equivalent to $\tilde{\sigma}_n > \tilde{\sigma}_{n+1}$ [28].

## 3. CONSTRAINED TOTAL LEAST SQUARES

When taking the conventional TLS approach to finding an approximate solution to the system of linear equations $A\underline{x} \approx \underline{y}$, the *only* consideration given in generating the perturbations of system matrix $A$ and vector $\underline{y}$ is that the resultant perturbed system of linear equations be consistent. In various applications of interest, however, the allowable perturbations may be additionally restricted in order that the perturbed augmented matrix $[A + \Delta \; \underline{y} - \underline{\delta}]$ satisfy additional requirements. These requirements can be general in nature and may include both structural and algebraic based constraints. For example, the perturbed augmented matrix may be required to have a Toeplitz structure, to have a prescribed singular value distribution, or (and) to be positive semidefinite. Although the imposition of additional constraints on the perturbations typically results in a much more difficult optimization problem to solve, the advantage thereby accrued can be considerable in that a conceptually superior means for removing errors in the elements of $A$ and $\underline{y}$ is possible. Intuitively, if the allowable perturbations are such that the underlying requirements are satisfied, then errors present in the augmented matrix due to random factors should be mitigated to a greater extent than would be the case if only the consistency constraint were imposed. As is subsequently demonstrated by means of examples, this conjecture is fulfilled in many important applications.

With the above comments in mind, we now examine the *constrained total-least-squares* (CTLS) problem as defined by

$$\min_{[A+\Delta \; \underline{y}-\underline{\delta}] \in \mathcal{M}} \| [\Delta - \underline{\delta}] \|_{\mathrm{F}}. \qquad (21)$$

In this problem formulation, $\mathcal{M}$ designates the *composite feasible matrix* set and is composed of all matrices contained in $C^{m \times (n+1)}$ that satisfy the prescribed conditions (or properties). In the standard TLS version of this more general problem, the only condition imposed is that the perturbed augmented matrix $[A + \Delta \; \underline{y} - \underline{\delta}]$ must correspond to a consistent system of

linear equations. In the more general CTLS problem this consistency requirement as well as other requirements are typically imposed. The CTLS problem here being considered is very general in nature and is not to be confused with a similarly named problem which considered only a specific type of linear constraint [1]. In retrospect, this latter problem might have been more properly called the linear constrained total-least-squares problem. With this in mind, we hereafter refer to optimization problem (21) as the CTLS problem.

In many practical applications, a closed form solution to CTLS problem (21) is generally precluded due to the highly complex nature of the composite feasible matrix set $\mathcal{M}$. In such instances one must appeal to nonlinear programming techniques to obtain a solution in an iterative fashion. We now formulate the CTLS problem so as to make it amenable to a nonlinear programming solution. Implicit in this approach is the hypothesis that the composite feasible matrix set can be expressed as the intersection of simpler feasible matrix sets. Specifically, let $M_1, M_2, \ldots, M_q$ designate $q$ sets of matrices contained in $C^{m\times(n+1)}$ where set $M_k$ containing all matrices that satisfy a given condition (or property). For example, $M_1$ might denote all $m \times (n+1)$ Toeplitz matrices. In order that optimization problem (21) have a solution, it is tacitly assumed that the intersection of these $q$ matrix sets as designated by

$$M = M_1 \cap M_2 \cap \cdots \cap M_q \tag{22}$$

is nonempty. Clearly, the composite feasible matrix set $M$ is composed of all matrices contained in $C^{m\times(n+1)}$ that possess each of the $q$ properties.

Corresponding to each individual property set $M_k$ there exists an associated operator $P_k(\cdot)$ as formally defined by

$$P_k(X) = \{Y \in C^{m\times(n+1)}: \|Y - X\|_F = \min_{\hat{X}\in M_k} \|\hat{X} - X\|_F\}. \tag{23}$$

The set $P_k(X)$ is seen to be composed of all $m \times (n+1)$ matrices contained in $M_k$ (or having property $k$) that lie closest to $X$ in the Frobenius norm sense. The mapping $P_k(\cdot)$ is a *projection operator* since it possesses the prerequisite *idempotent* property $P_k^2 = P_k$. Depending on the nature of matrix set $M_k$, the set $P_k(X)$ may be composed of a single matrix or of many matrices. Thus, the projection mapping $P_k(\cdot)$ may not be of the traditional one-to-one type. Furthermore, for many relevant matrix properties sets the associated projection mapping is nonlinear.

For many commonly employed matrix property sets $M_k$, it is subsequently shown that a convenient formulation of the associated $P_k(\cdot)$ projection operator exists. As suggested previously, however, when two or more of these relatively simple property sets are employed to form the composite feasible matrix set (22), the associated composite projection operator $P_M(\cdot)$ generally does not have a closed form representation. This is manifested in optimization problem not having a convenient closed form solution. An effective means for achieving a solution in such cases is described in the next section.

## 4. METHOD OF SUCCESSIVE PROJECTIONS

The *method of successive projections* provides an useful means for solving a class of optimization problems which includes the CTLS problem as a special case. In the method of successive projections, the underlying optimization problem is formulated in a vector space setting. Specifically, the entities of interest are taken to be vectors contained in a normed vector space $X$ with the norm of any vector $\underline{x} \in X$ being denoted by $\|\underline{x}\|$. The approximation problem defined on this space is formally given by

$$\min_{\underline{y}\in M_1 M_2 \cdots M_q} \|\underline{y} - \underline{x}\|, \tag{24}$$

where $\underline{x} \in X$ is a fixed vector. In effect, a solution to this problem corresponds a vector $\underline{y}^o$ contained in each of the sets $M_1, M_2, \ldots, M_q$ that best approximates of $\underline{x}$ in the minimum norm sense. The individual set $M_k$ is composed of all vectors contained in normed vector space $X$ that satisfy a prescribed property imposed on the approximating vector. Thus, the given vector $\underline{x}$ is to be approximated by a vector that possesses the $q$ properties associated with the sets $M_1, M_2, \ldots, M_q$. For the purposes of this paper, the underlying vector space $X$ is taken to be the set of $m \times n$ matrices and the norm employed is the Frobenius norm. In this section's discussion, we do not so restrict $X$ in order to formulate the method of successive projections in a more general setting.

Associated with each of the sets $M_k$ that in part comprise the optimization problem there is a projection operator as defined by

$$P_k(\underline{x}) = \{\underline{y} \in M_k: \|\underline{y} - \underline{x}\| = \min_{\underline{z}\in M_k} \|\underline{z} - \underline{x}\|\}. \tag{25}$$

The set $P_k(\underline{x})$ is seen to be composed of all vectors contained in $M_k$ that lie closest to $\underline{x}$ in the normed sense. Projection operator $P_k(\cdot)$ therefore projects $\underline{x}$

onto all vectors contained in $M_k$ that lie closest to $\underline{x}$. A little thought will convince oneself that any solution to optimization problem (24) must be a fixed point of the composite projection operator $P = P_q P_{q-1} \cdots P_1$ (i.e., $P(\underline{y}) = \underline{y}$).

For many of the more interesting approximation problems of form (24), there will not exist a convenient closed form solution due to the extreme nonlinearity of the composite projection operator $P$. This being the case, a solution to the approximation problem must be obtained in an iterative fashion using nonlinear programming techniques. The method of *successive projections* constitutes a particularly effective means for solving the approximation problem. A typical iteration of this algorithm takes the form

$$x_k \in P_q P_{q-1} \cdots P_1(x_{k-1}) \quad \text{for} \quad k \geq 1, \quad (26)$$

where $\underline{x}_k$ designates the approximate solution at the $k$th iteration. The initial vector used in this algorithm is set equal to the vector being approximated (i.e., $\underline{x}^0 = \underline{x}$). The process of generating the vector $\underline{x}_k$ from $\underline{x}_{k-1}$ employing this algorithm is to be implemented in the following manner. First, the set $P_1(\underline{x}_{k-1})$ is found. This set consists of all vectors contained in $M_1$ that lie closest to $\underline{x}_{k-1}$ in the minimum norm sense. Next, the set $P_2(P_1(\underline{x}_{k-1}))$ is formed and is composed of all vectors contained in $M_2$ that lie closest to one or more of the vectors in the set $P_1(\underline{x}_{k-1})$. It is important to note that although each of the vectors in $P_2(P_1(\underline{x}_{k-1}))$ is contained in $M_2$, the projection operation $P_2$ typically projects the vector $P_1(\underline{x}_{k-1})$ outside of $M_1$. This process is continued until each of the $q$ projection operations have been invoked according to relationship (25). One iteration of the algorithm is completed by arbitrarily selecting one vector from vector set $P_q P_{q-1} \cdots P_1(\underline{x}_{k-1})$ to be $\underline{x}_k$. It is to be noted that if the individual projection mappings $P_k$ are each point-to-point mappings, the vector $\underline{x}_k$ generated in this fashion is unique.

It has been shown that the under different conditions imposed on the $M_k$ sets, the successive projection algorithm produces a vector sequence that contains a convergent subsequence which converges to a solution of optimization problem (24). As original formulated, the $q$ sets were taken to be closed subspaces. For this case, Halperin established that the successive projection algorithm converged to a solution of approximation problem (24) [18]. This result is readily generalized to include the case in which the $M_k$ sets are closed linear varieties. Youla and Webb then generalized these results to the case in which the sets $M_k$ are each closed and convex [30]. Although a solution to the optimum solution could not be ensured in this case, they showed that the method of successive projections produces a vector sequence which always contains a subsequence that converges to a vector possessing the $q$ prerequisite properties.

The author further generalized Youla and Webb's result to the case in which each of the sets $M_k$ is only required to be closed [5]. The importance of this generalization follows from the fact that many of the sets used in practical applications are closed but not convex. For example, the set of $m \times n$ matrices that have a given rank is closed but not convex. The author has proven that the method of successive projections algorithm (26) produces a vector sequence that always contains a subsequence that converges to a vector that possessing the $q$ prerequisite properties provided that

1. the mapping $P_q P_{q-1} \cdots P_1$ is a *closed mapping.*
2. the mapping $P_q P_{q-1} \cdots P_1$ is distant reducing relative to a reference signal $\underline{x}_r$.
3. the set of vectors $\{\underline{y}\}$ satisfying $\|\underline{y} - \underline{x}_r\| \leq \|\underline{x} - \underline{x}_r\|$ comprises a closed and bounded set where $\underline{x}$ designates the signal being approximated and $\underline{x}_r$ denotes a fixed reference vector.

The reference vector referred to in the third condition is often set to the zero vector. Extensive experience with a variety of approximation problems indicates that the successive projection algorithm as applied to closed property sets typically generates a useful approximate solution to optimization problem (24). A closed mapping is a generalization of the notion of continuity as applied to standard point-to-point mappings [32], that is

DEFINITION 1. The point-to-set mapping $P$ is said to be closed at $\underline{x}$ if the assumptions (i) $\underline{x}_k \to \underline{x}$, and, (ii) $\underline{y}_k \to \underline{y}$ with $\underline{y}_k \in P_q P_{q-1} \cdots P_1(\underline{x}_k)$ implies that $\underline{y} \in P_q P_{q-1} \cdots P_1(x)$.

### *Matrix Enhancement Algorithm*

As is shown in the next several sections, the ability to use nonconvex sets can be *critical* in many practical applications involving the cleansing of empirically gathered data. The method of successive projections when applied to data that is expressed in the form of data matrices shall hereafter be referred to as the *matrix enhancement algorithm.* This matrix enhancement algorithm therefore takes the form

$$X_k \in P_q P_{q-1} \cdots P_1(X_{k-1})$$

$$\text{for} \quad k = 1, 2, 3, \cdots, 1, \quad (27)$$

where $X_k$ is a matrix that designates the algorithm's result after $k$ iterations. If the conditions previously specified in this section are satisfied, then the matrix sequence generated according to this algorithm is

guaranteed to have a subsequence that converges to matrix which lies within each of the matrix sets $M_k$ associated with the projection operators $P_k$ for $k = 1, 2, \ldots, q$. Principal among these requirements is that these projection operators be each closed.

To illustrate a typical application of the matrix enhancement algorithm, let us consider the constrained TLS problem associated with the inconsistent system of linear equations $A\underline{x} \approx \underline{y}$. It is desired to perturb the $m \times n$ system matrix $A$ and $m \times 1$ vector $\underline{y}$ so as to achieve consistency. In particular, it is desired to find the smallest Frobenius norm perturbation of the $m \times (n + 1)$ auxiliary matrix

$$X = [A - \underline{y}] \tag{28}$$

so that the perturbed auxiliary matrix is contained in each of the sets $M_1, M_2, \ldots, M_q$ of $m \times (n + 1)$ complex valued matrices. If $P_k$ designates the projection operator associated with set $M_k$, then the matrix enhancement algorithm (27) produces a convergent subsequence that converges to a matrix which lies within each of the sets $M_1, M_2, \ldots, M_q$. This convergence behavior is contingent on the projection operators satisfying the previously stated requirements. The matrix used to initiate the algorithm is set equal to the auxiliary matrix (28) being perturbed, that is $X_0 = X$. The utility of this signal enhancement algorithm is dependent on our ability to implement the $P_k$ projection mappings in a computational viable fashion for matrix properties that identify practical applications. In the next three sections we describe how the projection mappings are implemented for important matrix properties.

## 5. RANK "$p$" MATRIX APPROXIMATION

The singular valued decomposition (SVD) of data generated matrices plays an increasingly important role in contemporary signal processing applications. In particular, we now examine some fundamental SVD properties of a general complex valued $m \times n$ data matrix $X$ of rank $r$. In accordance with our previous discussion, the SVD representation for this matrix takes the form

$$X = \sum_{k=1}^{r} \sigma_k \underline{u}_k \underline{v}_k^*. \tag{29}$$

The augmented data matrix (28) provides a specific example of how such a matrix may arise. In signal processing applications, it is frequently desired to find a matrix of rank $p$ that best approximates $X$ is the Frobenius norm sense. Eckart and Young provided a convenient solution to this problem as now formally stated [11].

THEOREM 3. *Let $X$ be a generally complex valued $m \times n$ matrix of rank $r$ whose SVD representation is given by expression (29). If $p \leq r$, it then follows that a matrix of rank $p$ that best approximates $X$ in the Frobenius and Euclidean norm sense is given by truncating this SVD representation to its largest $p$ outer products, that is*[5]

$$X^{(p)} = P^p(X) = \sum_{k=1}^{p} \sigma_k \underline{u}_k \underline{v}_k^*. \tag{30}$$

*This best rank $p$ approximation is unique if and only if $\sigma_p > \sigma_{p+1}$. Furthermore, the projection mapping $P^p$ is closed and continuous if $\sigma_p > \sigma_{p+1}$, and, is closed but not continuous if $\sigma_p = \sigma_{p+1}$. The Frobenius and Euclidean norms of the resultant approximation error are given by*

$$\|X - X^{(p)}\|_F = \sqrt{\sum_{k=p+1}^{r} \sigma_k^2},$$

and

$$\|X - X^{(p)}\|_2 = \sigma_{p+1} \tag{31}$$

A proof of the conditions needed for projection operator $P^{(p)}$ to be closed and continuous is found in Ref. [23]. This theorem states that the best rank $p$ matrix approximation (30) is unique if and only if $\sigma_p > \sigma_{p+1}$. When the smallest positive singular value has multiplicity greater than one, however, there will exist an infinite number of distinct rank $p$ matrices that optimally approximate $X$. It therefore follows that the mapping from $X$ into $X^{(p)}$ is a point-to-point mapping if $\sigma_p > \sigma_{p+1}$ and is a point-to-set mapping when $\sigma_p = \sigma_{p+1}$. The point-to-set case raises a number of theoretical as well as practical issues that must be addressed in using an SVD matrix representation. The following related theorem is also useful in signal processing applications related to correlation matrices.

THEOREM 4. *Let the generally complex valued $m \times n$ matrix $X$ of rank $r$ have SVD representation (29). The $m \times n$ matrix which best approximates $X$ in the Frobenius norm sense and has its smallest $r - p$ nonzero eigenvalues equal is given by*

$$X_{(p)} = \sum_{k=1}^{p} \sigma_k \underline{u}_k \underline{v}_k^* + \sigma \sum_{k=p+1}^{r} \underline{u}_k \underline{v}_k^*, \tag{32}$$

[5] The Euclidean norm of matrix $A$ is defined to be $\|A\| = \sup \|A\underline{x}\| / \|\underline{x}\|$.

*where*

$$\sigma = \frac{1}{r-p} \sum_{p+1}^{r} \sigma_k. \tag{33}$$

*Furthermore, the projection mapping $P_{(p)}: X \to X_{(q)}$ is closed and continuous if $\sigma_p > \delta_{p+1}$, and is closed but not continuous if $\sigma_p = \sigma_{p+1}$.*

In typical signal processing applications of the SVD, the distribution of the singular values of a data matrix is often used to determine model order information when analyzing empirical data. Ideally, the gap between the so-called signal level and noise level singular values (i.e., $\sigma_p - \sigma_{p+1}$) is large enough so that questions of uniqueness and continuity of mapping do not arise. Unfortunately, in challenging applications (e.g., the detection of multiple sinusoids whose frequencies are closely spaced) the gap can be very small, thereby leading to possible undesirable algorithmic sensitivities.

## 6. LINEAR STRUCTURED MATRICES

In various applications, the matrix $X$ under consideration is known to have its elements functionally dependent on a set of parameters. To illustrate this point, a listing of some typically matrix classes which fall into this category are given in Table 1. In each case, there exists a functional interdependence between the matrix elements since they are each dependent on a set of real valued parameters. For example, a $m \times n$ Toeplitz matrix is completely specified by the $m + n - 1$ parameters identifying its first row and first column elements. In this paper we shall be interested in classes of matrices whose elements are linearly dependent on a set of parameters. We now formalize this concept.

DEFINITION 2. Let $x_{ij}(\underline{\theta})$ for $1 \leq i \leq m$ and $1 \leq j \leq n$ designate a set of $mn$ functions that are dependent on the real valued components of parameter vector $\underline{\theta} = [\theta_1 \cdots \theta_p]^T$ in which $p < mn$. Furthermore, consider the class of all $m \times n$ matrices whose components are governed by the functional relationships

$$X(i,j) = x_{ij}(\underline{\theta}) \quad \text{for} \quad 1 \leq i \leq m \text{ and } 1 \leq j \leq n \tag{34}$$

for specific choices of the parameter vector $\underline{\theta}$. This matrix class is said to have a structure induced by the functions $x_{ij}(\underline{\theta})$ and to have $p$ degrees of freedom. If these functions are linear in the $p$ parameters then the matrix class $\mathcal{M}$ is said to have a *linear structure.*

**TABLE 1**

Structured Matrices

| Matrix class | Matrix elements |
|---|---|
| Hermitian | $x(i,j) = \bar{x}(j,i)$ |
| Toeplitz | $x(i+1, j+1) = x(i,j)$ |
| Hankel | $x(i+1, j) = x(i, j+1)$ |
| Circulant | $x(i+1, j) = x(i, j-1)$ with $X(i+1, 1) = X(i, n)$ |
| Vandermonde | $x(i,j) = x(1,j)^i$ |

### *Linear Structured Matrix Approximation*

In what is to follow, we are concerned with the task of finding a matrix of a specified linear structure that lies closest to a given matrix $X$ of the same size. The importance of linear structured matrices is made evident by noting that each of the matrix classes given in Table 1 is so characterized with the exception of Vandermonde matrices. A little thought indicates that any matrix possessing a linear structure can be uniquely expressed in the decomposed format

$$\hat{X}(\underline{\theta}) = \sum_{k=1}^{p} \theta_k X_k \tag{35}$$

with the fixed matrices $X_1, X_2, \ldots, X_p$ constituting a basis for the subspace of linear structured matrices under consideration. The nature of these basis matrices depends on the particular linear structure being characterized.[6] The matrix approximation problem under consideration can be formally expressed as

$$\min_{\underline{\theta} \in R^p} \| X - \sum_{k=1}^{p} \theta_k X_k \|_F. \tag{36}$$

It is therefore desired to select the parameter vector $\underline{\theta}$ so as to minimize the squared Frobenius normed functional specified by

$$f(\underline{\theta}) = \| X - \sum_{k=1}^{p} \theta_k X_k \|_F^2$$

$$= \text{trace}\{ [X - \sum_{k=1}^{p} \theta_k X_k]^* [X - \sum_{k=1}^{p} \theta_k X_k] \}. \tag{37}$$

In arriving at this result, use has been made of the fact that the squared Frobenius norm of matrix $A$ is equal to the trace of $A^*A$. Upon carrying out the matrix

[6] For instance, it is demonstrated in Example 1 that the basis matrices corresponding to the class of $m \times n$ Toeplitz matrices are given by those $m + n - 1$ matrices of size $m \times n$ that have all zero elements except for ones that appear along a specific diagonal.

multiplications comprising this functional, it is found that

$$f(\underline{\theta}) = \text{trace}\{X^*X\} - \sum_{k=1}^{p} \theta_k \,\text{trace}\{X_k^*X\} - \sum_{k=1}^{p} \theta_k \,\text{trace}\{XX_k^*\} + \sum_{k=1}^{p}\sum_{m=1}^{p} \theta_k\theta_m \,\text{trace}\{X_k^*X_m\} = \text{trace}\{X^*X\} - 2\underline{\theta}^{\mathrm{T}}\underline{b} + \underline{\theta}^{\mathrm{T}}C\underline{\theta}, \tag{38}$$

where the elements of the real valued vector $\underline{b}$ and real valued matrix $C$ are given by

$$b(k) = 2\,\text{Real}[\text{trace}\{X_k^*X\}]$$

and

$$C(m, n) = \text{trace}\{X_m^*X_n\}, \quad \text{for} \quad 1 \leq k, m, n \leq p. \tag{39}$$

The matrix of the specified linear structure that best approximates the given matrix $X$ is therefore obtained by selecting $\underline{\theta}$ to minimize functional (38). Upon setting the gradient of this functional with respect to $\underline{\theta}$ equal to the zero vector, the optimum parameter vector is found to satisfy the consistent system of linear equations

$$\text{Real}\{C\}\underline{\theta}^{o} = \underline{b}, \tag{40}$$

where real$\{C\}$ designates the *real part* of matrix $C$. Substituting this optimal parameter choice back into the relationship (35), the matrix of the prerequisite linear structure that most closely approximates $X$ in the Frobenius norm sense is given by

$$\hat{X}(\underline{\theta})^{o} = \sum_{k=1}^{p} \theta_k^{o} X_k. \tag{41}$$

EXAMPLE 1. To illustrate the above procedure, let us consider the specific case of the class of real $3 \times 2$ Toeplitz matrices. It is observed that any such matrix can always be uniquely represented as

$$\begin{bmatrix} \theta_1 & \theta_2 \\ \theta_3 & \theta_1 \\ \theta_4 & \theta_3 \end{bmatrix} = \theta_1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} + \theta_2 \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} + \theta_3 \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} + \theta_4 \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}. \tag{42}$$

The four matrices appearing in this representation constitute a basis for the space of real $3 \times 2$ Toeplitz matrices. In accordance with expression (41), the closest Toeplitz matrix in the Frobenius norm sense is obtained by first computing for the vector $\underline{b}$ and matrix $C$ specified by relationships (39). It is a simple matter to show that

$$\underline{b} = \begin{bmatrix} X(1,1) + X(2,2) \\ X(1,2) \\ X(2,1) + X(3,2) \\ X(3,1) \end{bmatrix}$$

$$C = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{43}$$

Since matrix $C$ is nonsingular, the unique Toeplitz matrix that approximates a general $3 \times 2$ matrix $X$ is associated with the parameter vector

$$\underline{\theta}^{o} = \begin{bmatrix} \frac{1}{2}[X(1,1) + X(2,2)] \\ X(1,2) \\ \frac{1}{2}[X(2,1) + X(3,2)] \\ X(3,1) \end{bmatrix}. \tag{44}$$

It is seen that the components of this vector correspond to the means of the diagonals of the matrix being approximated.

Appealing to the results of the above example, it is readily established the best Toeplitz approximation to a general $m \times n$ matrix is obtained by first determining the sampled means of each of its $m + n - 1$ diagonals and then using these sampled means as entries for the corresponding diagonals of the best approximating Toeplitz matrix. A similar statement holds for obtaining the approximating Hankel matrix in which the term diagonal is replaced by antidiagonal. Other properties related to linear structured matrices are found in Ref. [5].

## 7. POSITIVE SEMIDEFINITE MATRICES

In this section an examination of some salient characteristics of the class of generally complex valued $n \times n$ *positive semidefinite* matrices is made. The $n \times n$ matrix $X$ is said to be positive semidefinite if the associated quadratic form inequality as specified by

$$\underline{x}^*X\underline{x} \geq 0 \tag{45}$$

is satisfied for all vectors $\underline{x} \in C^n$. Furthermore, if the only vector that causes this quadratic form to be zero is the zero vector, then $A$ is said to be *positive definite*. Since a positive quadratic form is real valued, this

further implies that any positive semidefinite matrix must also be Hermitian. Our interest in positive semidefinite matrices is motivated by their frequent appearance in studies related to quantitative oriented disciplines. As examples of there importance in signal processing applications, the following matrices are each positive semidefinite: (i) autocorrelation matrices associated with wide-sense stationary time series, (ii) orthogonal projection matrices appearing in optimization problems, and (iii) weighting matrices used in quadratic criterion related to optimization problems.

With the importance of positive semidefinite matrices in mind, the basic problem of finding a positive semidefinite matrix that lies closest to a given Hermitian matrix $X$ in the Frobenius norm sense is now considered.[7] The ingredients of a solution to this problem are made evident upon examining the eigenanalysis of $X$ as given by

$$X\underline{x}_k = \lambda_k \underline{x}_k \quad \text{for} \quad 1 \leq k \leq n. \tag{46}$$

Since $X$ is Hermitian, it follows that the $\lambda_k$ eigenvalues are exclusively real. Furthermore, there always exist a full set of $n$ eigenvectors which can be always chosen orthonormal (i.e., $\underline{x}_k^* \underline{x}_m = \delta(k - m)$). With this characterization, the following lemma is readily proven.

LEMMA 2. *Let $X$ be a Hermitian matrix contained in $C^{n\times n}$ whose eigenanalysis is specified by relationship (46). Furthermore, let the eigenvalues be ordered in the monotonically nonincreasing fashion $\lambda_k \geq \lambda_{k+1}$ in which the first p eigenvalues are positive and the last $n - p$ are nonpositive. It then follows that Hermitian matrix $X$ can be uniquely decomposed as*

$$X = \sum_{k=1}^{p} \lambda_k \underline{x}_k \underline{x}_k^* + \sum_{k=p+1}^{n} \lambda_k \underline{x}_k \underline{x}_k^* = X^+ + X^-, \tag{47}$$

*in which Hermitian matrix $X^+ = \sum_{k=1}^{p} \lambda_k \underline{x}_k \underline{x}_k^*$ is positive semidefinite with rank p while Hermitian matrix $X^- = \sum_{k=p+1}^{n} \lambda_k \underline{x}_k \underline{x}_k^*$ is negative semidefinite.*

Examination of this theorem indicates that any Hermitian matrix may be uniquely be decomposed into the sum of a positive and negative semidefinite Hermitian matrix as specified by (47). From this decomposition, the solution to the best approximating positive semidefinite matrix immediately follows and is characterized as follows.

[7] No loss of generality is incurred by assuming that $X$ is Hermitian, since if it is not then it is replaced by its closest Frobenius Hermitian image $(X + X^*)/2$.

THEOREM 5. *Let $X$ be a Hermitian matrix contained in $C^{n\times n}$ whose SVD representation is given by expression (47). The unique positive semidefinite matrix that lies closest to $X$ in the Frobenius and Euclidean norm sense is given by the following truncated SVD mapping:*

$$X^+ = P^+(X) = \sum_{k=1}^{p} \lambda_k \underline{x}_k \underline{x}_k^*. \tag{48}$$

*The projection operator $P^+$ defined by relationship (48) is both closed and continuous. The Frobenius norm of the error associated with this optimum positive semidefinite matrix approximation is given by*

$$\| X - X^+ \|_F = \sqrt{\sum_{k=p+1}^{n} \sigma_k^2}. \tag{49}$$

*Similarly, an orthogonal projection matrix which lies closest to $X$ in the Frobenius and Euclidean norm sense is specified by*

$$X^{op} = P^{op}(X) = \sum_{k:\lambda_k \geq 0.5} \underline{x}_k^* \underline{x}_k. \tag{50}$$

*This closest orthogonal projection matrix is unique provided that none of the eigenvalues of $X$ are equal to 0.5. Moreover, projection operator $P^{op}$ closed for any distribution of eigenvalues. The Frobenius norm of the error associated with this optimum projection matrix approximation is given by*

$$\| X - X^+ \|_F = \sqrt{\sum_{k:\lambda_k \geq 0.5}^{n} [\sigma_k - 0.5]^2}. \tag{51}$$

This theorem's proof is a direct consequence of the fact that the Frobenius and Euclidean norm of the matrices $X - Y$ and $Q^*[X - Y]XQ$ are equal for any choice of the unitary matrix $Q$ and Hermitian matrix $Y$. For our purposes, $Q$ is set equal to the $n \times n$ unitary matrix whose columns are equal to the $n$ orthonormal eigenvectors of matrix $X$. For this selection $D = Q^*XQ$ is a diagonal matrix whose diagonal elements are equal to the eigenvalues of $X$ so that $Q^*[X - Y]Q = D - Q^*YQ$. It then follows that the closest positive semidefinite choice for matrix $Y$ that minimizes the Frobenius norm of $Q^*[X - Y]Q$ is made according to relationship (48). This corresponds to truncating the SVD representation of $X$ to its positive singular value outer products. In a similar fashion, the closest orthogonal projection matrix is obtained by replacing each singular value by one if the singular value is greater than or equal to 0.5 and by zero otherwise. This closest idempotent Hermitian matrix is unique provided that none of the eigenvalues of $X$ are equal

to 0.5. Moreover, projection operator $P^{op}$ is closed for any distribution of eigenvalues.

EXAMPLE 2. Let us consider the problem of finding a positive semidefinite Toeplitz matrix that lies closest to the matrix

$$X = \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}.$$

Although this matrix is positive semidefinite, it does not possess the specified Toeplitz structure. To find a positive semidefinite Toeplitz matrix that provides an acceptably accurate approximation it is necessary to apply nonlinear programming techniques. The matrix enhancement algorithm governed by

$$X_{k+1} = P^{+}P_{T}(X_k)$$

has guaranteed convergence to a positive semidefinite Toeplitz matrix approximation. In this algorithm $P^{+}$ corresponds to the positive semidefinite projection operator (48) while $P_T$ denotes the Toeplitz projection operator heretofore described. Using matrix $X$ as initial condition, it is found that the matrix sequence thereby generated converges in two iterations to

$$\hat{X} = \begin{bmatrix} 3.0811 & 2.9230 \\ 2.9230 & 3.0811 \end{bmatrix}.$$

Convergence was deemed to have occurred when the normed matrix error $\|X_n - X\|/\|X\|$ became less than $10^{-9}$. It is to be noted that the positive semidefinite Toeplitz matrix which lies closest to $X$ in Frobenius norm sense is specified by

$$\hat{X}^{o} = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix}.$$

This example nicely illustrates the point that although the matrix enhancement algorithm does not have guaranteed convergence to the closest matrix approximation (for the properties of this example), it typically results in sufficiently good substitution. A modification of the matrix enhancement algorithm that has guaranteed convergence to the optimal solution was proposed by Dykstra [10].

### *Spectral Estimation Application*

In the field of spectral estimation, the basic goal is that of estimating the spectrum associated with a wide-sense stationary time series. This estimation is to be based on a finite set of contiguous samples of the time series as exemplified by

$$x(1), x(2), \ldots, x(N) \tag{52}$$

This estimation process is normally explicitly or implicitly begun by forming estimates of the time series' autocorrelation lags $r_{xx}(k) = E\{x(n+k)x(n)^*$ for $k = 0, \pm 1, \pm 2, \ldots$. The spectrum of the wide-sense stationary time series is defined to be the Fourier transform of the autocorrelation lag sequence. With these remarks in mind, we shall now examine a typical spectral estimation approach which employs the standard unbiased estimate of the associated auto correlation lags generated from the finite samples (52) as specified by

$$\hat{r}_{xx}(k) = \frac{1}{N-k}\sum_{m=1}^{N-k} x(k+m)\bar{x}(m) \quad \text{for} \quad 0 \leq k \leq N-1. \tag{53}$$

Using these correlation lag estimates, the following $n \times n$ Hermitian-Toeplitz structured *correlation matrix* estimate is formed

$$\hat{R}_{xx} = \begin{bmatrix} \hat{r}_{xx}(0) & \hat{r}_{xx}^{*}(1) & \cdots & \hat{r}_{xx}^{*}(n-1) \\ \hat{r}_{xx}(1) & \hat{r}_{xx}(0) & \cdots & \hat{r}_{xx}^{*}(n-2) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{r}_{xx}(n-1) & \hat{r}_{xx}(n-2) & \cdots & \hat{r}_{xx}(0) \end{bmatrix}. \tag{54}$$

This correlation matrix estimate serves as an approximation of the underlying correlation matrix $R_{xx}$ which is typically unknown to the signal processor. Under the ergodic assumption, this correlation matrix estimate converges to $R_{xx}$ as the number of observations $N$ approaches infinity. In the practical case where $N$ is finite, however, this estimate serves as only a relatively crude approximation of $R_{xx}$. To obtain a superior estimate, we now employ the matrix enhancement algorithm to hopefully remove a significant amount of the estimation error. For this purposes, the matrix properties (or sets $M_k$) employed that characterize the unknown correlation matrix are that (i) $R_{xx}$ is positive semidefinite and (ii) $R_{xx}$ possesses a Toeplitz–Hermitian structure. Although correlation matrix estimate (54) possesses the prerequisite Toeplitz–Hermitian structure, it often fails to be positive semidefinite. Our task is to then find an approximation of this correlation matrix estimate that possesses these two properties. To achieve this objective, we employ the matrix enhancement algorithm

$$\hat{R}_{xx}(k) = P_{TH}P^{+}\hat{R}_{xx}(k-1) \quad \text{for} \quad k \geq 1, \tag{55}$$

in which the initial matrix estimate $\hat{R}_{xx}(0)$ is given by

expression (54). Since the closest Toeplitz–Hermitian projection mapping $P_{TH}$ and the closest positive semidefinite definite projection mapping $P^+$ are each closed, it follows that the matrix sequence generated by this algorithm produces a subsequence that converges to a matrix which is positive semidefinite and has a Toeplitz–Hermitian structure. The matrix to which this sequence converges may then be used in any of a variety of spectral estimation methods to obtain hopefully improved spectral estimates relative to that achieved with the initial estimate (54).

## 8. EXPONENTIAL MODELING

One of the more important applications of matrix enhancement is that of approximating empirical data by a linear combination of exponentials. In particular, let there be given the finite set of time-series observations

$$x(1), x(2), \ldots, x(N). \tag{56}$$

It is well known that this data set can be modeled exactly as a linear combination of $p$ or fewer exponential signals if and only if there exists $a_k$ coefficients such that the following homogeneous relationships of order $p$ are satisfied:

$$x(n) + a_1 x(n-1) + \cdots + a_p x(n-p) = 0 \quad \text{for} \quad p+1 \leq n \leq N. \tag{57}$$

The $a_k$ terms are often referred to a prediction coefficients. Most relevant data modeling applications are concerned with the overdetermined case in which the number of homogeneous equations $N - p$ exceed by a wide margin the number of prediction coefficients $p$ to be determined (i.e., $N - p \gg p$). For the practical case in which the data is not perfectly modeled as a linear combination of $p$ or fewer exponential signals, these homogeneous equations have no solution.

For the development that follows it is useful to formulate the ideal homogeneous relationships (57) in their equivalent matrix–vector format

$$X\underline{a} = \underline{0}. \tag{58}$$

In this representation $\underline{a} = [1\ a_1\ a_2 \cdots a_p]^T$ is the $(p+1) \times 1$ *prediction coefficient vector* and $X$ is the corresponding $(N-p) \times (p+1)$ *data matrix* as specified by

$$X = \begin{bmatrix} x(p+1) & x(p) & \cdots & x(1) \\ x(p+2) & x(p+1) & \cdots & x(2) \\ \vdots & \vdots & \vdots & \vdots \\ x(N) & x(N-1) & \cdots & x(N-p) \end{bmatrix} \tag{59}$$

This data matrix is seen to have a *Toeplitz* structure since the elements along any of its diagonals are identical. Furthermore, if homogeneous relationship (58) is to have a nontrivial solution, it follows that the rank of data matrix $X$ must be equal to or less than $p$. These salient properties play a critical role in various exponential modeling algorithms and are formally recognized in the following lemma.

LEMMA 3. *The finite set of time series samples $x(1), x(2), \ldots, x(N)$ is exactly represented by a qth order exponential time series if and only if the SVD of the associated $(N-p) \times (p+1)$ Toeplitz structured data matrix (59) has q nonzero singular values where $q \leq p$.*

The exponential modeling characterization spelled out in this lemma provides the conditions under which the given data is exactly represented by an exponential model. In most practical applications, however, the data can only be approximately represented by an exponential model of reasonably small order. For such situations, we can employ the matrix enhancement algorithm to slightly perturb the given data matrix so that the associated perturbed data set is perfectly modeled by an exponential model of order $q$. To achieve this objective, we need to introduce matrix properties consistent with this goal. Two obvious properties which the idealized data matrix must possess are

*Property 1.* Data matrix $X$ has rank $q$.

*Property 2.* Data matrix $X$ has a Toeplitz structure.

The rank $q$ projection operator $P^{(q)}$ corresponding to Property 1 is governed by relationship (30) and is implemented by first computing the SVD of data matrix $X$ and then truncating this SVD representation to its $p$ largest singular value weighted outerproducts. Theorem 3 states that projection mapping $P^{(q)}$ is closed so it can be employed in a matrix enhancement algorithm. The Toeplitz projection mapping $P_T$ corresponding to Property 2 is also closed. It is recalled that when Toeplitz projection operator $P_T$ is applied to a general matrix $X$ it yields a Toeplitz matrix whose diagonal elements are equal to the average value of the corresponding diagonal elements of matrix $X$.

Since projection operators $P_T$ and $P^{(q)}$ are each closed, it follows that the matrix enhancement algorithm as specified by

$$X_k = P_T P^{(q)}(X_{k-1}) \quad \text{for} \quad k \geq 1 \qquad (60)$$

is ensured to have a subsequence that converges to a rank $q$ Toeplitz matrix. The initial data matrix used in this iterative scheme is set equal to the given data matrix (59) so that $X_0 = X$. To begin this algorithm, the rank $q$ approximation of the data matrix $X$ is first computed. The corresponding rank $q$ matrix $P^{(q)}(X)$ is generally non-Toeplitz in structure. To recover the required Toeplitz structure, we next apply projection mapping $P_T$ to matrix $P^{(q)}(X)$ to complete the first iteration of the matrix enhancement algorithm. It is generally found that this new Toeplitz structured data matrix $X_1 = P_T P^{(q)}(X)$ has full rank. It is closer to a rank $q$ matrix, however, than was the original data matrix $X$. The first iteration has therefore led to a data matrix whose elements comprise a data sequence that is more compatible with a $q$th order exponential model than was the original data. Often, this first iteration is sufficient for modeling applications.

To obtain a data sequence that is exactly representable by a $q$th order exponential model, we may continue this iterative process in an obvious manner. In particular, one sequentially computes the data matrices $X_{k+1} = P_T P^{(q)}(X_k)$ for $k = 0, 1, 2, \cdots$ until the data matrix $X_{k+1}$ is deemed to have a rank that is sufficiently close to $q$. This stopping condition could be implemented by computing the ratios of the $(q+1)$st to $q$th singular values of $X_{k+1}$ (i.e., $\sigma_k(q+1)/\sigma_k(q)$. If this ratio is found to be sufficiently close to zero then matrix $X_{k+1}$ is said to have an approximate rank of $q$. It has been empirically determined that this algorithm converges in a rapid fashion and typically takes from three to ten iterations to converge for moderately sized data matrices. More importantly, the resulting enhanced data matrix has component data elements that generally provide a better representation of the underlying exponential signal components than did the original data. This enhancement process has therefore effectively stripped away noise that contaminated the original data. A particularly important special case of the exponential modeling problem is now addressed.

### *Sinusoidal Signal Identification*

In a surprisingly large number of fundamental signal processing applications, the primary objective is that of identifying sinusoidal components present in noise contaminated data. For example, multiple plane waves incident on an equispaced linear array give rise to sinusoidal steering vectors. To identify sinusoidal signals, a widely employed procedure is to first form the data matrix whose upper and lower halves correspond to the forward and backward prediction equations associated with the finite length data (56). The *forward prediction equations* take the homogeneous form

$$x(n) + a_1 x(n-1) + \cdots + a_p x(n-p) \quad \text{for} \quad p+1 \leq n \leq N, \qquad (61)$$

while the *backward prediction equations* are specified by the homogeneous relationships

$$\bar{x}(n) + a_1 \bar{x}(n+1) + \cdots + a_p \bar{x}(n+p) \quad \text{for} \quad 1 \leq n \leq N-p. \qquad (62)$$

If the data under analysis is composed of a linear combination of pure complex sinusoids of order $q$ or less where $q \leq p$ it is well known that there exist a choice for the $a_k$ predictor coefficients so that homogeneous relationships (61) and (62) are each satisfied. This observation forms the basis for many contemporary sinusoidal identification algorithms.

The algebraic properties associated with sinusoidal modeling are best described by expressing forward and backward prediction relationships (61) and (62) in their equivalent matrix format, that is

$$\hat{X}\underline{a} = \underline{0}. \qquad (63)$$

In this relationship $\hat{X}$ is the $2(N-p) \times (p+1)$ combined forward–backward data matrix as specified by

$$\hat{X} = \begin{bmatrix} X \\ \cdots \\ \bar{X}J_{p+1} \end{bmatrix}, \qquad (64)$$

where $X$ designates the $(N-p) \times (p+1)$ Toeplitz structured forward prediction data matrix (59) and is associated with forward prediction equations (61). Similarly, $\bar{X}J_{p+1}$ denotes the Hankel structured backward prediction data matrix in which $J_{p+1}$ is the $(p+1) \times (p+1)$ *order reversal matrix* whose elements are all zero except for $p+1$ ones that appear along its main antidiagonal while $\bar{X}$ denotes the conjugate of $X$. This backward prediction matrix arises from the backward prediction equations (62).

If the data under analysis is noise free and composed of $q$ complex sinusoids, then the block Toeplitz–Hankel data matrix $\hat{X}$ has rank $q$ provided that $p \geq q$. To identify the component sinusoids in this ideal pure sinusoidal case, we simply find a prediction coefficient vector $\underline{a}$ with first component one that sat-

isfies homogeneous relationship (63). The components of this vector are then used to form the auxiliary polynomial

$$A(z) = 1 + \sum_{n=1}^{p} a_n^o z^n = (1 - z_1 z^{-1})(1 - z_2 z^{-1}) \cdots (1 - z_p z^{-1}). \quad (65)$$

A component sinusoid is associated with each of the $q$ roots of $A(z)$ that lie on the unit circle (i.e., $z_k = e^{j\theta_k}$) with the sinusoid frequencies corresponding to the $\theta_k$ angles of these roots.

When the data is noise contaminated or the data is not perfectly represented as a linear combination of $q$ complex sinusoids, data matrix $\check{X}$ typically has full rank $p + 1$. In this more realistic case, homogeneous relationship (63) has no nontrivial solution and we must appeal to alternate solution procedures. Various methods for modeling the given data as a linear combination of $q$ sinusoids (with $q < p$) have been proposed. Two related methods based on the idealized case of data matrix $\check{X}$ being of rank $q$ have been effective for this purpose. In the method developed by this author, the rank $q$ approximation of data matrix (64) is first computed. This rank $q$ approximation is then decomposed as

$$\check{X}^{(q)} = [\underline{\check{x}}_1, \check{X}_r], \quad (66)$$

in which $\underline{\check{x}}_1$ designates the first column of $\check{X}^{(q)}$ and $\check{X}_r$ its remaining $p$ columns. Finally, the proposed model coefficient vector estimate is then specified by [6]

$$\underline{a}_c^o = -\frac{1}{\underline{e}_1^T [\check{X}_r]^\dagger \underline{\check{x}}_1} [\check{X}_r^\dagger] \underline{\check{x}}_1 \quad (67)$$

where $^\dagger$ designates the Moore–Penrose generalized inverse operator. In a similar fashion, the Kumaresan and Tufts method is based on the decomposition of the data matrix given by

$$\check{X} = [\underline{x}_1, X_r], \quad (68)$$

in which $\underline{x}_1$ designates the first column of $\check{X}$ and $X_r$ its remaining $p$ columns. The corresponding Kumaresan–Tufts coefficient vector estimate is then given by [21,27]

$$\underline{a}_{KT}^o = -[X_r^{(q)}]^\dagger \underline{x}_1. \quad (69)$$

Kumaresan and Tufts have demonstrated that their method has a maximum-likelihood performance. It is interesting to note that the first selection (67) for identifying the $a_k$ parameters are TLS solutions to a linear system of equations. This TLS association was observed by Rahman and Yu [25] and later shown to be equivalent to expression (67) [20].

It is to be noted that although the two coefficients vectors solutions (67) and (69) are similar, the latter approach excludes the first column of $X$ in the rank $q$ approximation. As such, it does not take full advantage of the rank $q$ reduction data cleansing SVD operation and is therefore marginally inferior to the first method (e.g., see Refs. [9,20]). In both methods, the sinusoid frequencies estimates are generated by substituting the components of the prediction coefficient vectors (67) or (69) into polynomial (65). A component sinusoid is then associated with each root of $A(z)$ that lies within a user prescribed distance of the unit circle (i.e., $\|z_k| - 1| < \epsilon$).

Although the above two algorithms are effective in identifying sinusoidal components, application of a data cleansing matrix enhancement algorithm can significantly improve their performance. An appropriate matrix enhancement algorithm for this purpose is specified by

$$X_k = P_{TH} P^{(q)}(X_{k-1}) \quad \text{for} \quad k \geq 1. \quad (70)$$

In this algorithm $P_{TH}$ designates the projection operator that determines the unique block Toeplitz–Hankel matrix lying closest to a given matrix in the Frobenius norm sense. It is implemented in a fashion identical with $P_T$. Mapping $P^{(q)}$ corresponds to the aforementioned closest rank $q$ projection operator. These two projection operators correspond to the properties known to be possessed by the data matrix when the data being analyzed is perfectly represented as a linear combination of complex sinusoids of order $q$.

Upon application of matrix enhancement algorithm (70) to the given data matrix, a matrix sequence is thereby generated which has a subsequence that converges to a cleansed data matrix which has possesses the two prerequisite properties of being block Toeplitz–Hankel and having rank $q$. This cleansed data matrix may therefore be used in either algorithms (67) or (69) to obtain improved sinusoidal frequency estimates. The example to follow illustrates the benefits of this precleansing data operation.

EXAMPLE 3. To illustrate the effectiveness of the matrix enhancement algorithm, let us consider the following data set

$$x(n) = e^{j(2\pi f_1 n + \theta_1)} + e^{j(2\pi f_2 n + \theta_2)} + w(n) \quad \text{for} \quad 1 \leq n \leq 25 \quad (71)$$

with $f_1 = 0.50$, $f_2 = 0.52$, $\theta_1 = \pi/4$, $\theta_2 = 0$ and $w(n)$ is

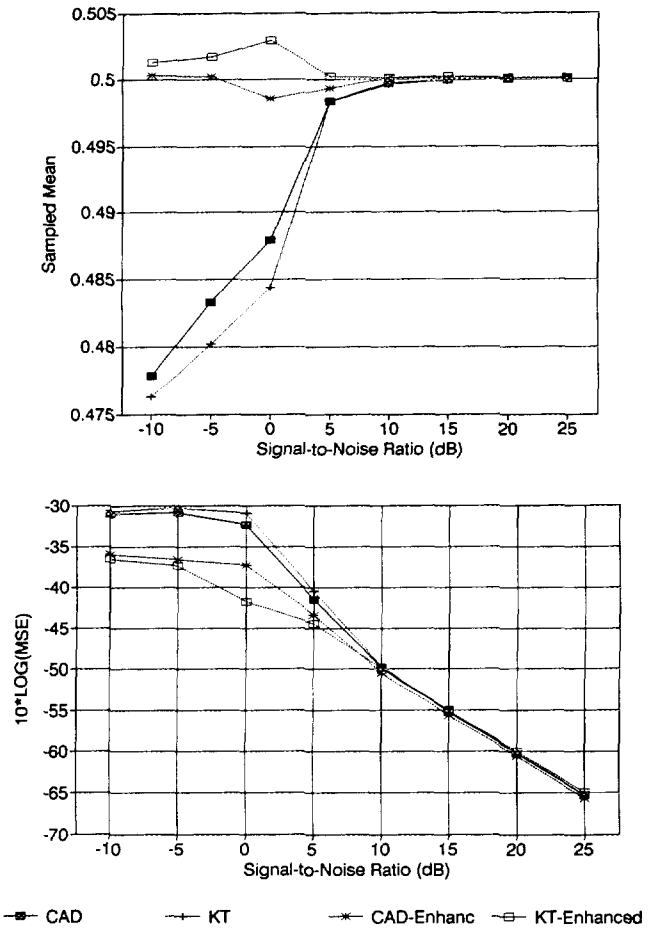


FIG. 1. Sampled mean and MSE statistic for noise-corrupted sinusoidal data.

Gaussian white noise whose uncorrelated real and imaginary components have standard deviation $\sigma$. The individual sinusoidal signal-to-noise ratio (SNR) for this time series is therefore $10 \log(1/2\sigma^2)$. One hundred statistically independent runs (different noise samples) of this times series are next made at several SNRs. The two prediction coefficient vector estimates (67) and (69) are then computed for each of the trial runs with a choice of $p = 17$ (the choice advocated in [27] and $q = 2$ to yield unenhanced frequency estimates.

Statistics relating to sampled means and mean squared error for the unenhanced frequency estimates of the frequency parameter $f_1 = 0.50$ for the one hundred trial runs made at several SNRs are summarized Fig. 1. Similar statistics for the parameter $f_2 = 0.52$ were obtained and are therefore not shown. From these statistics it is seen that for high SNRs the two methods yield virtually identical performance with method (67) being marginally better at SNRs less than 10 dB. Furthermore, each estimate is basically unbiased for SNRs exceeding 0 dB.

The effectiveness of the proposed precleansing matrix enhancement procedure is next determined. In particular, algorithm (70) is applied to the noise contaminated data. After 15 iterations, the enhanced data matrix with the prescribed block Toeplitz-Hankel structure and approximate rank $q$ is substituted into expressions (67) and (69). The statistics associated with the enhanced estimates are also shown in Fig. 1. From these results it is apparent that the enhancement process has provided a significant improvement at low signal-to-noise ratios in reducing bias and MSE. It has been shown that sinusoidal detection performance is also improved [9].

## 9. SYSTEM IDENTIFICATION

An application of interdisciplinary interest is concerned with the linear recursive modeling of *excitation–response* data. This problem is more commonly referred to as *system identification*. For purposes of presentation simplicity, we here only deal with the case in which the data is dependent on a single time variable. The procedure to be now described, however, is readily extended to the multidimensional time variable case (see Ref. ([7]). In the one-dimensional time case as developed in Ref. [8], there is given the data pair sequence

$$(x_n, y_n) \quad \text{for} \quad 0 \leq n \leq N, \tag{72}$$

where $x(n)$ and $y(n)$ are identified as the excitation and response data, respectively. This data set is said to be recursively related if there exist choices for the $a_k$ and $b_k$ coefficients such that the following linear recursive relationship of order $(p, q)$ is satisfied:

$$y(n) + \sum_{k=1}^{p} a_k y(n-k) = \sum_{k=0}^{q} b_k x(n-k) \quad \text{for} \quad 0 \leq n \leq N. \tag{73}$$

In specifying the time interval $0 \leq n \leq N$ over which this recursive relationship holds, it has been tacitly assumed that the data pairs are identically zero prior to $n = 0$. If this is not the case, then the time interval over which relationship (73) holds must be changed to $\max(p, q) \leq n \leq N$. Modification of the analysis to follow for this case is straightforward and therefore not given.

It will be convenient to represent recursive relationships (73) in matrix format so as to take advantage of algebraic properties that characterize the data. This matrix representation takes the form

$$\begin{bmatrix} y(0) & 0 & \cdots & 0 \\ y(1) & y(0) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ y(N- & y(N-1) & \cdots & y(N-p) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix}$$

$$= \begin{bmatrix} x(0) & 0 & \cdots & 0 \\ x(1) & x(0) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ x(N) & x(N-1) & \cdots & x(N-q) \end{bmatrix} \begin{bmatrix} b_o \\ b_1 \\ \vdots \\ b_q \end{bmatrix}, \tag{74}$$

or equivalently

$$Y_p \underline{a}_p = X_q \underline{b}_q. \tag{75}$$

In this latter representation, $Y_p$ and $X_q$ are referred to as the $(N+1) \times (p+1)$ *response matrix* and the $(N+1) \times (q+1)$ *excitation matrix*, respectively. Similarly, $\underline{a}_p$ and $\underline{b}_q$ are the recursive coefficient vectors identifying the recursive operator with $\underline{a}_p(1) = 1$. With this preliminary development, the basic properties characterizing recursively related data are now formally spelled out (see Ref. [8]).

THEOREM 6. *Let the excitation-response data $(x(n), y(n))$ for $0 \leq n \leq N$ be related through a reduced-order recursive relationship of order $(\bar{p}, \bar{q})$ in which $\bar{p} \leq p$, $\bar{q} \leq q$, and $p + q < N$. It then follows that the extended order recursive relationship (75) always has a solution in which the constraint $a_p(1) = 1$ is satisfied. Moreover, if the excitation and response matrices are full rank so that* $\text{rank}[X_q] = q + 1$ *and* $\text{rank}[Y_p] = p + 1$, *it then follows that the null space of the $(N+1) \times (p+q+2)$ composite excitation-response data matrix*

$$D_{p,q} = [Y_p \vdots -X_q] \tag{76}$$

*has dimension $s = 1 + \min(p - \bar{p}, q - \bar{q})$. Furthermore, the set of all solutions to relationship (75) in which $a_p(1) = 1$ is specified by*

$$S = \left\{ \begin{bmatrix} \underline{a}_p \\ \underline{b}_q \end{bmatrix} = \frac{1}{\underline{e}_1^T V \underline{x}} V \underline{x} \text{ for all } \underline{x} \in C^s \right.$$
$$\left. \text{in which } \underline{e}_1^T V \underline{x} \text{ is nonzero} \right\}, \tag{77}$$

*where $\underline{e}_1$ designates the $(p+q+2) \times 1$ vector whose components are all zero except for its first which is one while $V$ is a $(p+q+2) \times s$ matrix whose columns are composed of any set of linearly independent vectors that span the null space of the composite excitation-response matrix (76).*[8] *Furthermore, the transfer function associated with any solution contained in the linear variety set (77) reduces to (after common pole-zero cancellation) the underlying reduced-order transfer function of order $(\bar{p}, \bar{q})$. The minimum Euclidian norm solution contained in solution set (77) is given by*

$$\begin{bmatrix} \underline{a}_p^o \\ \cdots \\ \underline{b}_q^o \end{bmatrix} = \frac{1}{\underline{e}_1^T V V^* \underline{e}_1} V V^* \underline{e}_1. \tag{78}$$

In using the algebraic characteristics of the composite excitation–response matrix to form a rational model of empirical data, there is much to be gained by using an over-ordered model (i.e., $\tilde{p} < p$ and $\tilde{q} < q$)). By taking this overordered approach, the resultant recursive model parameters estimates are made less sensitive to quirks in the empirical data. A more detailed explanation of this concept is found in Ref. [9].

### *Matrix Enhanced System Identification*

From the above development, it follows that when the given observations $\{(x(n), y(n))\}$ are perfectly represented by a recursive relationship of order $(\bar{p}, \bar{q})$, the excitation-response data matrix satisfies the two properties

*Property 1.* $D_{p,q}$ has a lower triangular block Toeplitz structure.

*Property 2.* $D_{p,q}$ has $s = 1 + \min(p - \bar{p}, q - \bar{q})$ of its singular values equal to zero.

In most practical applications, the given data observations are not perfectly represented by a low-order recursive relationship. This is typically manifested in the composite data matrix being full rank. To use the concept of matrix enhancement to achieve a suitably good approximate recursive model, we could suitably modify the given excitation–response data so that the modified data has an associated composite data matrix that satisfies the above two properties. A logical choice for a matrix enhancement algorithm associated with this objective is given by

$$D_k = P_{LT} P^{(p+q+2-s)}(D_{k-1}) \quad \text{for} \quad k \geq 1. \tag{79}$$

This algorithm is initiated by the selection $D_o$ being set equal to the original excitation-response data matrix (76). We have dropped the subscript $p, q$ in the composite data matrix in order to simplify notation. Projection operator $P_{LT}$ has the dual task of (i) find-

[8] As an example the required column vectors can be set equal to the $s$ right singular vectors associated with the zero singular value of the composite excitation–response matrix $D_{p,q}$.

ing the closest block Toeplitz matrix and (ii) setting to zero the upper triangular elements associated with the of submatrices $Y$ and $-X$ in keeping with the prerequisite structure (74).

The theory related to the matrix enhancement algorithm ensures that the composite data matrix sequence (79) contains a subsequence that converges to a composite data matrix that satisfies the prerequisite lower triangular block Toeplitz structure and null space dimension $s$ properties. The recursive coefficient vectors as specified by relationship (78) when applied to the convergent composite data matrix typically gives a satisfactory recursive model of the given data. It should be noted that in some applications, it is known that either the excitation or the response data is accurate and should not be perturbed when applying the projection operator $P^{(p+q+2-s)}$. This is readily accomplished by inserting the original block after projection mapping $P^{(p+q+2-s)}$ has been applied to $D_{k-1}$ as is now demonstrated.

*Recursive System Design*

In various applications, it is desired to approximate the dynamics of a given linear operator (e.g., an ideal low-pass filter) by a linear recursive system. If such an approximation can be achieved, then the linear operator can be effectively replaced by means of a computational efficient linear recursive operation. In the approach to be taken, use is made of the observation that the dynamics of the linear operator being approximated and a linear recursive system are similar if and only if their associated unit-impulse responses are themselves similar. With this in mind, let $h_d(0), h_d(1), h_d(2), \cdots$ designate the generally infinite length unit-impulse response of the causal linear operator. Since numerical methods are to be used, it is first necessary to appropriately truncate this impulse response to a finite length, that is,

$$h_d(0), h_d(1), \cdots, h_d(N). \tag{80}$$

The integer $N$ must be selected sufficiently large so that this truncated unit-impulse response has essentially the same dynamics as its untruncated counterpart being approximated.

To obtain a linear recursive approximation of the truncated ideal behavior (80), we simply set $y(n) = h_d(n)$ and $x(n) = \delta(n)$ in relationship (74) to give

$$D_{p,q}\begin{bmatrix}\underline{a}_p\\ \underline{b}_q\end{bmatrix} = \underline{0} \quad \text{where} \quad D_{p,q} = [H_p \vdots -\Delta_q]. \tag{81}$$

In this expression first component of $\underline{a}_p$ is constrained to be one and the $(N+1)\times(q+1)$ excitation matrix $\Delta_q$ has all zero elements except for ones that appear along its main diagonal. The restricted structure of the excitation matrix is due to the nature of the unit-impulse excitation. Since the truncated unit-impulse response is almost never perfectly represented by a linear recursive operation, it follows that this system of homogeneous linear equations is inconsistent. In this case, it is desired to select the filter coefficient vectors $\underline{a}_p$ and $\underline{b}_q$ so as to best satisfy this homogeneous relationship in the LSE sense. Under the constraint that the first component of $\underline{a}_p$ is one, it is readily shown that the LSE solution is specified by

$$\begin{bmatrix}\underline{a}_p^o\\ \underline{b}_q^o\end{bmatrix} = \frac{1}{\underline{e}_1^T[D_{p,q}^* D_{p,q}]^{-1}\underline{e}_1}[D_{p,q}^* D_{p,q}]^{-1}\underline{e}_1, \tag{82}$$

where $\underline{e}_1$ designates the standard basis vector whose first component is one and its remaining components are zero.

It is generally found that the filter coefficient vectors obtained from relationship (82) provide a close approximation to the ideal behavior. By employing the matrix enhancement, it is possible to improve upon this approximation. For this application, the signal enhancement algorithm makes use of the fact that the given unit-impulse response data is perfectly represented by a linear recursive system if and only if the composite unit-impulse response-excitation matrix $D_{p,q}$ possesses the two properties of having: (i) a lower triangular block Toeplitz structure and (ii) a null space of dimension at least equal to of one. We may therefore use the signal enhancement algorithm (79) with $s \geq 1$ to iteratively obtain a nearby composite unit-impulse response-excitation matrix that possesses these two properties. When implementing the projection operator $P_{LT}$, the excitation matrix thereby resulting is replaced by $\Delta_q$ in order to enforce the unit-impulse excitation constraint. The coefficients of the corresponding linear recursive system are then obtained by employing expression (78) with the prescribed selection of $s$ on this nearby composite unit-impulse response-excitation matrix. Examples illustrating the modelling improvement accrued using this approach are found in Ref. [9].

## 10. CONCLUSION

The concepts of constrained total least squares and matrix enhancement have been developed and applied to a number of important signal processing problems. In addition to the problems described in this paper, the matrix enhancement algorithm has been successfully used for the task of data interpolation, deconvolution, and high-dimensional filter syn-

thesis. In order to employ the concept of matrix enhancement to its fullest extent, it is incumbent on the user to innovatively introduce matrix properties that characterize the underlying data matrix.

## REFERENCES


1. Abatzoglou, T. J., Mendel, J. M., and Harada, G. A. The constrained total least squares technique and its application to harmonic superresolution. *IEEE Trans. Acoust. Speech Signal Process.* **AASP-39** (May 1991), 1070–1087.
2. Adcock, R. J. A problem in least square. *The Analyst* **5** (1878), 53–54.
3. Anderson, T. W. The 1982 Wald memorial lectures: Estimating linear statistical relationships. *Ann. Statist.* **12** (1984), 1–45.
4. Cadzow, J. A. Least squares, modeling, and signal processing. *Digital Signal Process.*, in press.
5. Cadzow, J. A. Signal enhancement: A composite property mapping algorithm. *IEEE Trans. Acoust. Speech Signal Process.* **AASP-36** (Jan. 1988), 49–62.
6. Cadzow, J. A. Spectral estimation: An overdetermined rational model equation approach. *Proc. IEEE, Special Issue on Spectral Analysis* (Sept. 1982), 907–939.
7. Cadzow, J. A., and Chen, T. C. Algebraic approach to two-dimensional recursive digital filter design. *Asilomar Conf., Monterey, CA, Nov. 1987.*
8. 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.
9. Cadzow, J. A., and Wilkes, D. M. Enhanced sinusoidal and exponential data modeling. *Elsevier Signal Processing Journal, Special Issue on SVD.*
10. Dykstra, R. L. An algorithm for restricted least squares regression. *J. Am. Statist. Assoc.* **78** (1983), 837–842.
11. Eckart, G., and Young, G. The approximation of one matrix by another of lower rank. *Psychometrica* **1** (1936), 211–218.
12. Ekstrom, M. P., Twogood, R. E., and Woods, J. W. Two-dimensional recursive filter design—A spectral factorization approach. *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-28,** No. 1 (Feb. 1980), 16–26.
13. Gauss, C. F. Theoria combinationis observationum erroribus minimis obnoxiae. *Comment. Soc. Reg. Sci. Gotten. Recent.* **5** (1823), 33–90.
14. Gleser, L. J. Estimation in a multivariate "errors in variables" regression model: Large sample results. *Ann. Statist.* **9** (1981), 24–44.
15. Golub, G. H. Some modified matrix eigenvalue problems. *SIAM Rev.* **15** (1987), 318–344.
16. Golub, G. H., and Van Loan, C. F. An analysis of the total least squares problem. *SIAM J. Numer. Anal.* **17** (1980), 883–893.
17. Golub, G. H., and Van Loan, C. F. *Matrix Computations,* 2nd ed. Johns Hopkins Univ. Press, Baltimore, 1989.
18. Halperin, I. The product of projection operators. *Acta Sci. Math.* **23** (1962), 96–99.
19. Koopmans, T. C. *Linear Regression Analysis of Economic Time Series.* De Erven F. Bohn, N. V. Haarlem, The Netherlands, 1937.
20. Hua, Y., and Sarkar, T. K. On the total least squares linear prediction method for frequency estimation. *IEEE Trans. Acoust. Speech Signal Process.* **AASP-38** (Dec. 1990), 2187–2189.
21. Kumaresan, R. Estimating the parameters of exponentially damped/undamped sinusoidal signals in noise. Ph.D. dissertation, University of Rhode Island, Kingston, RI, August 1982.
22. Madansky, A. The fitting of straight lines when both variables are subject to error. *J. Am. Statist. Assoc.* **54** (1959), 173–205.
23. Mittelmann, H. D., and Cadzow, J. A. Continuity of closest rank-$p$ approximations to matrices. *IEEE Trans. Acoust. Speech Signal Process.* **AASP-35** (Aug. 1987), 1211–1212.
24. Pearson, K. On lines and planes of closest fit to points in space. *Philos. Mag.* **2** (1901), 559–572.
25. Rahman, M. A., and Yu, K. B. Total least squares approach for frequency estimation using linear prediction. *IEEE Trans. Acoust. Speech Signal Process.* **AASP-35** (Oct. 1990), 1440–1454.
26. Sprent, P. *Models in Regression and Related Topics.* Methuen, London, 1969.
27. Tufts, D. W., and Kumaresan, R. Estimation of frequencies of multiple sinusoids: making linear prediction perform like maximum likelihood. *Proc. IEEE, Special Issue on Spectral Analysis* (Sept. 1982), 975–989.
28. Van Huffel, S., and Vandewalle, J. *The Total Least Squares Problem—Computational Aspects and Analysis.* SIAM, Philadelphia, 1991.
29. York, D. Least squares fitting of a straight line. *Can. J. Phys.* **44** (1966), 1079–1086.
30. Youla, D. C., and Webb, H. Image restoration by the method of convex projections: Part 1—Theory. *IEEE Trans. Med. Imaging* **M-1** (Oct. 1982), 81–94.
31. Wilkes, D. M. "On eigenspace recursions." Submitted for publication.
32. Zangwill, W. I. *Nonlinear Programming: A Unified Approach.* Prentice-Hall, Englewood Cliffs, NJ, 1969.



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 of 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.