92-33 J

Structured Total Least Squares and L_2 Approximation Problems*

Bart De Moor†
ESAT—Department of Electrical Engineering
Katholieke Universiteit Leuven
Kardinaal Mercierlaan 94
B-3001 Leuven, Belgium

ABSTRACT

It is shown how structured and weighted total least squares and L_2 approximation problems lead to a "nonlinear" generalized singular value decomposition. An inverse iteration scheme to find a (local) minimum is proposed. The emphasis of the paper is not on the convergence analysis of the algorithm; rather the purpose is to illustrate its use in numerous applications in systems and control, including total least squares with relative errors and/or fixed elements, inverse singular value problems, an errors-in-variables variant of the Kalman filter, impulse response realization from noisy data, H_2 model reduction, H_2 system identification, and calculating the largest stability radius of uncertain linear systems. Several numerical examples are given.

[*] This work was started when the author was a visitor at the Institute of Mathematics and Applications (IMA) of the University of Minnesota in Minneapolis in April 1992 (IMA Preprint Ser. 1036, Sept. 1992). It was supported in part by the IMA with funds provided by the National Science Foundation. Additional support was provided by the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Science Policy Programming (Prime Minister's Office) and by the European Community Research Program ESPRIT, Basic Research Action 3280. The scientific responsibility is assumed by its author, who is a research associate of the Belgian National Fund for Scientific Research (NFWO).

[†] Telephone: 32/16/220931. Fax: 32/16/221855. E-mail: demoor@esat.kuleuven.ac.be. Revised versions will be available via anonymous ftp to gate.esat.kuleuven.ac.be in the file pub/SISTA/demoor/reports/h2-rankdef-final.ps.Z (compressed PostScript format).

© Elsevier Science Publishing Co., Inc., 1993
655 Avenue of the Americas, New York, NY 10010 0024-3795/93/$6.00

1. INTRODUCTION

Let $B(b) = B_0 + b_1 B_1 + \cdots + b_n B_n \in \mathbb{R}^{p\times q}$ be an affine matrix function of the components b_i of the parameter vector $b \in \mathbb{R}^n$, where B_i, $i = 0, 1, \ldots, n$, are fixed given matrices. Let $a \in \mathbb{R}^m$ be a data vector, and w be a given vector of weights. A problem that often occurs in systems and control applications is to find a rank-deficient matrix in the affine set $B(b)$ such that a given quadratic function $[a, b, w]_2^2$ of the parameters b_i is minimized. This will be called the *structured total least squares* (STLS) problem; it can be formalized as

$$\min_{b\in\mathbb{R}^n,\ y\in\mathbb{R}^q} [a, b, w]_2^2 \qquad \text{subject to} \quad B(b)\,y = 0,\ y^t y = 1. \qquad (1)$$

Examples of affine matrix functions are structured matrices such as Toeplitz, Hankel, and Brownian matrices or matrices with certain zero patterns. An example of a matrix function which is nonaffine in its parameters is a Vandermonde matrix.

The main motive of this paper is to show that many signal processing, system identification, and control system design and analysis problems reduce to the solution of an STLS problem.

In Section 2, we present an easy derivation for the unstructured total least squares problem. Our main result is Theorem 1 in Section 3, which states that the solution to the STLS problem follows from that of a nonlinear generalized SVD problem. The derivation parallels the one for the unstructured case. In Section 4, we present several examples of STLS problems, including relative error TLS, TLS with arbitrarily fixed elements, TLS for linearly structured matrices such as Hankel and Toeplitz (including noisy realization and model reduction), and an example from system identification. We also show how one can do an errors-in-variables variation of the Kalman filter and the calculation of the stability margin of an uncertain linear system. In Section 5, we derive a straightforward algorithm that is inspired by inverse iteration to find the smallest singular value and corresponding singular vectors of a matrix. Numerical experiments suggest that it is linearly convergent, although a further theoretical analysis is definitely required. Several numerical examples are given in Section 6.

We will confine ourselves to real data and real STLS problems, although all results can be generalized to the complex field. Our notation is fairly standard.

2. AN EASY DERIVATION OF TOTAL LINEAR LEAST SQUARES

The purpose of this section is to take the simplest total least squares problem, for which the solution is known: it is given in terms of a singular value decomposition (SVD). The idea is that the steps used in this derivation will be exactly the same ones as in the solution of the general problem.

The total least squares problem reduces to finding a rank-deficient matrix approximation B in Frobenius norm to a given matrix A. This can be formulated as

$$\min_{B\in\mathbb{R}^{p\times q},\ y\in\mathbb{R}^q} \|A - B\|_F^2 \qquad \text{subject to} \quad By = 0, \quad y^t y = 1.$$

Obviously, due to the rank deficiency constraint on B, this is not a convex optimization problem. Let $l \in \mathbb{R}^{p\times 1}$ be a vector of Lagrange multipliers, and $\lambda \in \mathbb{R}$ a Lagrange multiplier; then the Lagrangian can be written as

$$\mathscr{L}(B, y, l, \lambda) = \sum_{i=1}^{p}\sum_{j=1}^{q}(a_{ij} - b_{ij})^2 + l^t By + \lambda(1 - y^t y).$$

Let us now show how we can arrive at the well-known SVD solution.

Derivatives

Setting all derivatives equal to zero results in the set of equations[1]

$$A - B = ly^t, \qquad B^t l = y\lambda, \qquad By = 0, \qquad y^t y = 1.$$

Note that the error $A - B$ is a rank one matrix. Also, it is straightforward to show that $\lambda = 0$ from $l^t By = 0$.

Elimination of B

By postmultiplying $A - B$ with y and $(A - B)^t$ with l we find

$$Ay = l, \qquad A^t l = y(l^t l), \qquad y^t y = 1.$$

[1] We absorb all irrelevant constant factors 2 in the Lagrange multipliers.

Normalization

Next normalize l as $x = l/\|l\|$, where we call $\sigma = \|l\|$. Then we find

$$Ay = x\sigma, \qquad x^t x = 1,$$

$$A^t x = y\sigma, \qquad y^t y = 1,$$

which implies that (x, σ, y) must be a singular triplet of A. Observe that $\| A - B\|_F^2 = \sigma^2$, so that we need the singular triplet corresponding to the smallest singular value.

An Orthogonality Property

It is straightforward to derive that

$$B^t(A - B) = 0 \quad \text{and} \quad (A - B)B^t = 0.$$

Construction of B

B can be constructed as $B = A - x\sigma y^t$ and hence is a rank one modification of A.

The approximation in Frobenius norm of a given matrix by one of lower rank has an long history, with its roots going back to Adcock [2, 3] and continuing through Pearson [23], Eckart and Young [12], and Young and Householder [20, 27]. Golub and Van Loan [17, 18] formulate the problem as a generalization of solving $Ay = b$ in a least squares sense for the case where all data A and b (and not only the right hand side b as in least squares) are corrupted by noise. More details (geometric, algebraic, and statistical) on the (unstructured) total least squares problem can be found in [17, 24].

3. STLS AS A "NONLINEAR" GENERALIZED EIGENVALUE PROBLEM

Let us now consider in detail the STLS problem (1), with $m = n$. We take the quadratic criterion

$$[a, b, w]_2^2 = \sum_{i=1}^{m} (a_i - b_i)^2,$$

in which, for the moment, we do not consider weights w (the general case with weights is treated in Section 4 and is a straightforward extension of the results obtained in this section). In order to solve the minimization problem, we follow the same path as outlined in Section 2 for the unstructured case. The Lagrangian is given by

$$\mathscr{L}(b, y, l, \lambda) = \sum_{i=1}^{m} (a_i - b_i)^2 + l^t(B_0 + b_1 B_1 + \cdots + b_m B_m)y$$

$$+ \lambda(1 - y^t y),$$

where $l \in \mathbb{R}^p$ is a vector of Lagrange multipliers and $\lambda \in \mathbb{R}$ is a scalar Lagrange multiplier.

Derivatives

Setting all derivatives equal to zero gives[2]

$$\forall k: \quad a_k - b_k = l^t B_k y, \qquad (B_0 + b_1 B_1 + \cdots + b_m B_m) y = 0,$$

$$(B_0^t + b_1 B_1^t + \cdots + b_m B_m^t) l = y\lambda, \qquad y^t y = 1. \tag{2}$$

From $l^t B(b) y = 0$, it follows directly that $\lambda = 0$.

Elimination of b

Next we eliminate the parameters b by using $b_k = a_l - l^t B_k y$ to find

$$(a_1 B_1 + \cdots + a_m B_m) y = \left[(l^t B_1 y) B_1 + \cdots + (l^t B_m y) B_m \right] y, \tag{3}$$

$$(a_1 B_1^t + \cdots + a_m B_m^t) l = \left[(l^t B_1 y) B_1^t + \cdots + (l^t B_m y) B_m^t \right] l. \tag{4}$$

Observe that terms with B_0 have canceled out. Looking at the right hand sides, we observe that the first right hand side (3) is quadratic in y and linear in l, while the second right hand side (4) is quadratic in l and linear in y.

[2] Again irrelevant constants are absorbed in the Lagrange multipliers.

Let's now concentrate on one term in the right hand side of (3). Without loss of generality, we can take the first one. Define β_{ij} as element (i,j) of B_1, and let $\bar{b}_i^t$ be the ith row of B_1. Then

$$\left[\text{element } k \text{ of } B_1 y(l^t B_1 y)\right] = \sum_{j=1}^{q} \beta_{kj} y_j \sum_{r=1}^{p} \sum_{s=1}^{q} \beta_{rs} l_r y_s$$

$$= \sum_{r=1}^{p} (y^t \bar{b}_k)(\bar{b}_r^t y) l_r.$$

Hence

$$B_1 y(l^t B_1 y) = \begin{pmatrix} y^t \bar{b}_1 \\ y^t \bar{b}_2 \\ \vdots \\ y^t \bar{b}_p \end{pmatrix} \begin{pmatrix} \bar{b}_1^t y & \cdots & \bar{b}_p^t y \end{pmatrix} l.$$

Observe that in the right hand side, the matrix preceeding l is a rank one matrix, and as it is the outer product of a vector with itself, it is nonnegative definite.

Obviously, we can repeat this for each term of (3) to obtain the result that the right hand side of (3) can be written as

$$\sum_{i=1}^{m} \left[B_i(l^t B_i y) \right] y = D_y l. \tag{5}$$

Here, D_y is a symmetric matrix which is a sum of m rank one nonnegative definite matrices; hence D_y itself is nonnegative or positive definite. Its elements are quadratic functions of the components of the vector y. A similar derivation applies for the right hand side of (4):

$$\sum_{i=1}^{m} \left[B_i^t(l^t B_i y) \right] l = D_l y, \tag{6}$$

where D_l is symmetric nonnegative or positive definite, with elements that are quadratic functions of the components of l.

Normalization

Next we define $x = l/\|l\|$ and call $\sigma = \|l\|$. Let D_x be defined in the same way as D_l, by replacing every component of l appearing in D_l by its corresponding component in x. Since the elements of D_l are quadratic functions of the components of l, we find $D_l = D_x \sigma^2$. Next define $A \in \mathbb{R}^{p \times q}$ as $A = \sum_{i=1}^{m} B_i a$. Then we find

$$\begin{aligned} Ay &= D_y x \sigma, & x^t x &= 1, \\ A^t x &= D_x y \sigma, & y^t y &= 1. \end{aligned} \tag{7}$$

We are now ready to prove the following theorem:

THEOREM 1 (STLS as a nonlinear generalized SVD). *Consider the STLS problem*

$$\min_{b \in \mathbb{R}^m,\, y \in \mathbb{R}^q} \sum_{i=1}^{m} (a_i - b_i)^2 \quad \textit{subject to} \quad B(b)y = 0, \quad y^t y = 1,$$

where a_i, $i = 1, \dots, m$, are the components of the data vector $a \in \mathbb{R}^m$, and $B(b) = B_0 + B_1 b_1 + \cdots + B_m b_m$, with B_i, $i = 0, 1, \dots, m \in \mathbb{R}^{p \times q}$, fixed given matrices. The solution is generated as follows:

(a) Find the triplet (u, τ, v) corresponding to the minimal τ that satisfies

$$\begin{aligned} Av &= D_v u \tau, & u^t D_v u &= 1, \\ A^t u &= D_u v \tau, & v^t D_u v &= 1, \end{aligned} \tag{8}$$

where $A = \sum_{i=1}^{m} a_i B_i$. Here D_u is defined via $D_u = \sum_{i=1}^{m} B_i^t(u^t B_i v)u = D_u v$ and is a positive or nonnegative definite matrix, the elements of which are quadratic in the components of u. D_v is defined similarly via $\sum_{i=1}^{m} B_i(u^t B_i v)v = D_v u$ and is positive or nonnegative definite with elements that are quadratic in the components of v.

(b) The vector y is given as $y = v/\|v\|$.

(c) The components of b are obtained from $b_k = a_k - u^t B_k v \tau$, $k = 1, \dots, m$

Before we present the proof of this theorem, we will first devote a few words to its interpretation. First observe that, if D_u and D_v were constant positive or nonnegative definite matrices (i.e. independent of u and v), the equations of Theorem 1 for a given matrix A would be satisfied by any triplet (u, σ, v) of the *restricted singular value decomposition* (RSVD) of the triplet $(A, D_v^{1/2}, D_u^{1/2})$ (where $D_v^{1/2}$ is a square root of D_v). The RSVD is just an SVD with different positive or nonnegative definite inner products in the column and row space and is extensively studied in [9]. Observe that, still under the assumption that D_u and D_v are constant,

$$\begin{pmatrix} 0 & A \\ A^t & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} D_v & 0 \\ 0 & D_u \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}\tau, \qquad u^tD_vu = 1, \quad v^tD_uv = 1,$$

would be a generalized eigenvalue problem. The "nonlinear" aspect however is the explicit dependence of the weight matrices D_u and D_v on the components of u and v. Still, as we will see, these matrices are always nonnegative definite (and hence correspond to inner products, which are however "position dependent"), and the elements are quadratic functions of the components of u and v.

Proof of Theorem 1. From the equations (7) it follows directly that

$$x^tAy = y^tA^tx = x^tD_y x\sigma = y^tD_x y\sigma. \tag{9}$$

Next observe that from (2)

$$\sum_{i=1}^{m}(a_i - b_i)^2 = \sum_{i=1}^{m}(l^tB_iy)^2 = \sum_{i=1}^{m}(x^tB_iy)^2\sigma^2$$

$$= x^t\sum_{i=1}^{m}\left[B_iy(x^tB_iy)\right]\sigma^2 = x^tD_yx\sigma^2 = x^tAy\sigma. \tag{10}$$

The last equality follows from (5) and (9).

Let the triplet (u, τ, v) solve (8). Then

$$u^tAv = (u^tD_vu)\tau = \tau. \tag{11}$$

Furthermore, by normalizing u and v, and recalling that D_u and D_v are quadratic in the components of u and v, we obtain

$$A\frac{v}{\|v\|} = \frac{D_v}{\|v\|^2}\frac{u}{\|u\|}(\tau\|u\|\,\|v\|),$$

$$A^t\frac{u}{\|u\|} = \frac{D_u}{\|u\|^2}\frac{v}{\|v\|}(\tau\|u\|\,\|v\|).$$

If we now put $x = u/\|u\|$, $y = v/\|v\|$, and $\sigma = \tau\|u\|\,\|v\|$, then from (10), (11) we find

$$x^tAy\sigma = \frac{u^t}{\|u\|}A\frac{v}{\|v\|}(\tau\|u\|\,\|v\|) = \frac{u^t}{\|u\|}\frac{D_v}{\|v^2\|}\frac{u}{\|u\|}(\tau\|u\|\,\|v\|)^2 = \tau^2. \tag{12}$$

This shows that we need to find the minimal τ. It also delivers the result for y. For the components of b we find from (2)

$$b_k = a_k - \frac{u}{\|u\|}B_k\frac{v}{\|v\|}\sigma = a_k - u^tB_kv\tau.$$

This completes the proof. ■

A useful characterization of the optimal solution can be obtained from (2) by observing that

$$l^t\left(\sum_{i=1}^{m}B_ib_i + B_0\right)y = 0 \quad \rightarrow \quad \sum_{i=1}^{m}(a_i - b_i)b_i + l^tB_0y = 0.$$

If $B_0 = 0$, this property says that the vector of residuals $a - b$ is orthogonal to b.

4. EXAMPLES AND APPLICATIONS

In this section, we treat several examples and applications of STLS problems. If all the elements of B are unknown parameters, we will simply use B instead of $B(b)$ as in the previous section.

4.1. *Total Least Squares with Elementwise Relative Weighting*

The elementwise weighted total least squares problem is the following:

$$\min_{B \in \mathbb{R}^{p \times q},\, y \in \mathbb{R}^q} \sum_{i=1}^{p} \sum_{j=1}^{q} (a_{ij} - b_{ij})^2 w_{ij}$$

$$\text{subject to} \quad By = 0, \quad y^t y = 1. \tag{13}$$

There are many applications. In statistics, a table of means based on samples of widely varying sizes should be fitted with weights proportional to the sample sizes as mentioned in [15]. In geology, where one analyses metamorphic mineral assemblages and reactions [14], it is required that the decomposition matrix be weighted elementwise, because some mineral composition data are known more accurately than others. Sometimes, the variance of measurements is not constant (heteroskedasticity). All of these are examples of elementwise weighted TLS problems.

From the Lagrangian, we find $\partial\mathscr{L}/\partial b_{ij} = 0 \to w_{ij}(a_{ij} - b_{ij}) = l_i y_j$, $\partial\mathscr{L}/\partial y_j = 0 \to B^t l = y\lambda$, $\partial\mathscr{L}/\partial l_i = 0 \to By = 0$, and $\partial\mathscr{L}/\partial\lambda = 0 \to y^t y = 1$. Observe that the matrix with elements $w_{ij}(a_{ij} - b_{ij})$ is a rank one matrix. This matrix can be represented as $W \cdot *(A - B)$, where $\cdot *$ is the elementwise product. Obviously, when $w_{ij} = 0$, either l_i or y_j should be 0. We will however consider the case where all weights $w_{ij} \neq 0$ and use $v_{ij} = w_{ij}^{-1}$. Furthermore, define the diagonal matrices $L = \operatorname{diag}(l_i)$ and $Y = \operatorname{diag}(y_i)$.[3] We can then rewrite the set of equations as

$$B = A - LVY, \qquad By = 0, \qquad B^t l = y\lambda, \qquad y^t y = 1,$$

[3] The notation diag(·) should be interpreted in the MATLAB convention: If the entity in parentheses is a matrix, then diag(·) will result in a vector having as its components the diagonal elements of the matrix. If the entity between brackets is a vector, diag(·) delivers a square diagonal matrix with the components of the vector on its main diagonal.

where V is the matrix with v_{ij} as its elements. Observe that

$$\operatorname{rank}(A - B) = \operatorname{rank}(LVY). \tag{14}$$

It follows easily that the Lagrange multiplier λ is 0, since $y^t B^t l = \lambda = 0$. The neat thing is that we can completely eliminate B by postmultiplying the first equation with y and premultiplying it with l^t, which leaves us with the equations $Ay = LVYy$, $A^t l = YV^t Ll$, and $y^t y = 1$. Let $x = l/\|l\|$ and $\|l\| = \sigma$. Also define the diagonal matrices

$$\begin{aligned} D_x &= \operatorname{diag}(V \operatorname{diag}(x)\, x), \\ D_y &= \operatorname{diag}(V^t \operatorname{diag}(y)\, y). \end{aligned} \tag{15}$$

Then we find the equations as in (7) which can be renormalized to get equations as in (8). Note that D_x and D_y are guaranteed to be nonnegative definite if all elements of V are nonnegative (which is the only case that occurs in practice). If all the weights $w_{ij} = 1$, we recover the ordinary TLS problem of Section 2.

Let's now verify that we need the minimal eigenvalue of the nonlinear generalized eigenvalue problem of Theorem 1. We have

$$\sum_{i=1}^{p} \sum_{j=1}^{q} (a_{ij} - b_{ij})^2 w_{ij} = \sum_{i=1}^{p} \sum_{j=1}^{q} x_i^2 y_j^2 v_{ij} \sigma^2 = x^t D_y x \sigma^2 = x^t A y \sigma = \tau^2.$$

The last equality follows from (12).

4.2. *Total Least Squares with Fixed Elements*

When only some of the elements of A can be modified to reduce the rank, we have a total least squares problem with fixed elements. In order to formalize this, we split A as $A = A_1 + A_2$, where A_1 contains the fixed elements and A_2 the elements of A that can be modified. Denote by $\mathscr{I}$ the set of index pairs of fixed elements and by $\mathscr{I}_C$ its complement. The problem now becomes

$$\min_{B \in \mathbb{R}^{p \times q},\, y \in \mathbb{R}^q} \|A - B\|_F^2$$

$$\text{subject to} \quad By = 0, \quad y^t y = 0, \quad B = A_1 + B_2, \quad B_2(i,j) = 0 \ \forall (i,j) \in \mathscr{I}.$$

The problem does not always have a solution, as the following example shows:

EXAMPLE.

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} = A_1 + A_2 = \begin{pmatrix} 1 & 0 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} + \begin{pmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Obviously, by only modifying the element 2, we can never reduce the rank of the matrix A, which is 2.

The Lagrangian function here is $\mathscr{L}(B, y, l, \lambda) = \Sigma\Sigma_{(i,j)\in\mathscr{I}_C}(a_{ij} - b_{ij})^2 + l^tBy + \lambda(1 - y^ty) + \Sigma\Sigma_{(i,j)\in\mathscr{I}} r_{ij}(a_{ij} - b_{ij})$, where $R \in \mathbb{R}^{p\times q}$ is a matrix of Lagrange multipliers with $R(i,j) = 0$ if $(i,j) \in \mathscr{I}_c$. This leads to the set of equations

$$B_2 = A_2 + R - ly^t, \qquad B = A_1 + B_2,$$

$$B_2(i,j) = 0 \quad \forall(i,j) \in \mathscr{I}, \qquad R(i,j) = 0 \quad \forall(i,j) \in \mathscr{I}_c,$$

$$By = 0, \qquad B^tl = y\lambda, \qquad y^ty = 1.$$

Again, we can show that $\lambda = 0$ and eliminate B to obtain $(A+R)^tl = y(l^tl)$, $(A+R)y = l$, and $y^ty = 1$, which after normalization of l as $x = l/\|l\|$, $\|l\| = \sigma$ becomes

$$(A+R)y = x\sigma, \qquad x^tx = 1,$$

$$(A+R)^tx = y\sigma, \qquad y^ty = 1.$$

This is an SVD of the matrix $A + R$. Of course, R is unknown too, but we know it is zero on $\mathscr{I}_C$. The message is that we need to modify A precisely in those positions that we are not supposed to modify!

Another point of view is that we basically have to do with an SVD completion problem: We have to modify the matrix A in such a way that some structural constraints on the singular values and vectors are satisfied. These structural constraints are the following: From $B_2 = A_2 + R - ly^t$ we have that $R(i,j) = l_iy_j = x_i\sigma y_j$ when $(i,j) \in \mathscr{I}$ and that $R(i,j) = 0$ when $(i,j) \in \mathscr{I}_C$. This implies that $B_2(i,j) = A_2(i,j) - l_iy_j$ and hence $A_2 - B_2$ will be of rank one. Using these observations, we can eliminate R and find that

$$Ay = D_y x\sigma, \qquad x^tx = 1,$$

$$A^tx = D_x y\sigma, \qquad y^ty = 1,$$

where D_x and D_y are given by exactly the same expressions as in (15), and where now V is a zero-one matrix which has a 0 at every element $(i,j) \in \mathscr{I}$ and a 1 at every element $(i,j) \in \mathscr{I}_C$. In this sense, the TLS problem with fixed elements can be considered as a limiting case of the weighted TLS problem of Section 4.1.

Let us conclude by pointing out that some special cases of fixed element patterns can be solved explicitly. The case of some columns error-free is considered in [16], while the case of three out of four subblocks error-free is considered in [7, 9].

4.3. *Weighted total least squares*

The componentwise weighting of Section 4.1 can be generalized to a pairwise weighted total least squares problem as follows:

$$\min_{B\in\mathbb{R}^{p\times q},\, y\in\mathbb{R}^q} \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{p}\sum_{l=1}^{q}(a_{ij} - b_{ij})w_{ijkl}(a_{kl} - b_{kl})$$

$$\text{subject to} \quad By = 0, \quad y^ty = 1.$$

From the Lagrangian we find $\partial\mathscr{L}/\partial b_{rs} = 0 \to \Sigma_{i=1}^{p}\Sigma_{j=1}^{q}(w_{rsij} + w_{ijrs})(a_{ij} - b_{ij}) = l_ry_s$, which can be written as $\Sigma_{i=1}^{p}\Sigma_{j=1}^{q}w_{ij}^{rs}(a_{ij} - b_{ij}) = l_ry_s$, where w^{rs} is a 1 × pq row vector. Next define the matrix $W \in \mathbb{R}^{pq\times pq}$ which has the row vectors ω^{rs} as its rows. Then, assuming invertibility of W,

$$\operatorname{vec}(A - B) = W^{-1}\begin{pmatrix} ly_1 \\ ly_2 \\ \vdots \\ ly_q \end{pmatrix},$$

where vec(·) stacks the columns of the matrix in a long pq × 1 vector. Call $W^{-1} = V$, and partition it in $p \times p$ blocks as

$$V = \begin{pmatrix} V^{11} & V^{12} & \cdots & V^{1q} \\ V^{21} & V^{22} & \cdots & V^{2q} \\ \vdots & \vdots & & \vdots \\ V^{q1} & V^{q2} & \cdots & V^{qq} \end{pmatrix}.$$

Then we can eliminate B to find $Ay = (V^{11}l\ V^{12}l\ \cdots\ V^{1q}l)yy_1 + \cdots + (V^{q1}l\ \cdots\ V^{qq}l)yy_q$. We can write out element i of Ay as

$$(Ay)_i = l_1 \left[y^t \begin{pmatrix} V_{i1}^{11} & V_{i1}^{12} & \cdots & V_{i1}^{1q} \\ \vdots & \vdots & & \vdots \\ V_{i1}^{q1} & V_{i1}^{q2} & \cdots & V_{i1}^{qq} \end{pmatrix} y \right]$$

$$+ l_2 \left[y^t \begin{pmatrix} V_{i2}^{11} & V_{i2}^{12} & \cdots & V_{i2}^{1q} \\ \vdots & \vdots & & \vdots \\ V_{i2}^{q1} & V_{i2}^{q2} & \cdots & V_{i2}^{qq} \end{pmatrix} y \right] + \cdots,$$

which is a linear function in the elements of l and quadratic in the elements of y. Similarly, element j of l^tA can be written as $(l^tA)_j = (l^tV^{j1}l)y_1 + \cdots + (l^tV^{jq}l)y_q$, so that

$$Ay = \begin{pmatrix} y^tU^{11}y & y^tU^{12}y & \cdots & y^tU^{1q}y \\ y^tU^{21}y & y^tU^{22}y & \cdots & y^tU^{2q}y \\ \vdots & \vdots & & \vdots \\ y^tU^{q1}y & y^tU^{q2}y & \cdots & y^tU^{qq}y \end{pmatrix} l,$$

$$A^tl = \begin{pmatrix} l^tV^{11}l & l^tV^{12}l & \cdots & l^tV^{1q}l \\ l^tV^{21}l & l^tV^{22}l & \cdots & l^tV^{2q}l \\ \vdots & \vdots & & \vdots \\ l^tV^{q1}l & l^tV^{q2}l & \cdots & l^tV^{qq}l \end{pmatrix} y,$$

where we define the matrix U^{kl} as $U_{ij}^{kl} = V_{kl}^{ij}$. Normalizing l as $x = l/\|l\|$ with $\sigma = \|l\|$ leads to the problem (7) or, via a renormalization into vectors u and v, to the nonlinear generalized eigenvalue problem (8).

In many statistical problems of practical interest, the noise covariance matrix is known. When all data in a $p \times q$ matrix A are noisy, the noise covariance matrix is a $pq \times pq$ positive definite matrix. If the noise on the data is zero mean normally distributed with this known covariance matrix, the pairwise weighted TLS problem corresponds to a maximum likelihood estimation formulation.

4.4. *An Errors-in-Variables Variant of the Kalman Filter*

A particularly interesting application of STLS is what may be called an errors-in-variables variant of the Kalman filter. For simplicity, we consider the first order linear time-invariant system

$$x_{k+1} = ax_k + v_k,$$
$$z_k = cx_k + w_k,$$

where $a \in \mathbb{R}$, $c \in \mathbb{R}$, x_k is the state and z_k the output at time k, and v_k and w_k are the process and measurement noises respectively. For subsequent time instants, one can write down a set of linear equations as

$$\begin{pmatrix} z_0 \\ 0 \\ z_1 \\ 0 \\ z_2 \\ 0 \\ \vdots \end{pmatrix} = \begin{pmatrix} c & 0 & 0 & 0 & \cdots \\ a & -1 & 0 & 0 & \cdots \\ 0 & c & 0 & 0 & \cdots \\ 0 & a & -1 & 0 & \cdots \\ 0 & 0 & c & 0 & \cdots \\ 0 & 0 & a & -1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ \vdots \end{pmatrix} + \begin{pmatrix} w_0 \\ v_0 \\ w_1 \\ v_1 \\ w_2 \\ v_2 \\ \vdots \end{pmatrix}. \tag{16}$$

The Kalman filter assumes that the model is known exactly (i.e., a and c are exact) and also requires knowledge of the covariance structure of the noise. The Kalman filter is nothing else than a clever way of updating the least squares solution to the above set of equations, by exploitation of the special and sparse structure of the data matrix every time a new measurement or state is added. The associated Riccati difference equation updates the Schur complement of the lower part of the matrix with a and c.

If however also a and c are not known precisely, it makes sense to consider the following minimization problem:

$$\min_{z_k, t_k, \alpha, \gamma, \hat{x}_k} (\gamma - c)^2 + (\alpha - a)^2 + \sum_{k=0}^{p} (s_k - z_k)^2 + \sum_{k=0}^{p} t_k^2 \tag{17}$$

subject to the linear equations

$$\hat{x}_{k+1} = \alpha \hat{x}_k + t_k,$$

$$s_k = \gamma \hat{x}_k, \qquad k = 0, 1, 2, \dots, p. \quad (18)$$

These equations can be written as

$$\begin{pmatrix} s_0 & \gamma & 0 & 0 & \cdots \\ t_0 & -\alpha & 1 & 0 & \cdots \\ s_1 & 0 & \gamma & 0 & \cdots \\ t_1 & 0 & -\alpha & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} -1 \\ \hat{x}_0 \\ \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \vdots \end{pmatrix} = 0. \qquad (19)$$

The interpretation is the following: Not only do we assume the presence of process noise and measurement noise [which motivates the modification of the left hand side in (16)], but also we approximate the system (a, c) with (α, γ). The result is a system (α, γ) given by (18), which is driven by a process noise sequence t_k but without measurement noise. Of course, one could include additional weights in (17) to emphasize the relative importance of each quadratic term in the criterion or to exploit *a priori* information concerning the noise covariance structure or model errors if e.g.

$$\mathbf{E}\left[\begin{pmatrix} a - \alpha \\ c - \gamma \end{pmatrix} (a - \alpha \quad c - \gamma)\right]$$

were known, for instance from an identification algorithm.

With the constraints (19), we associate $2p + 2$ Lagrange multipliers $l_1, \dots, l_{2p+2}$. We find easily the following equations by taking derivatives of the Lagrangian. Note that all the results are implicitly dependent on the time horizon p, a dependence that is denoted between square brackets. For instance, the estimate of γ after $p = 2$ measurements is denoted by $\gamma[2]$. We illustrate the results here for $p = 2$:

$$\begin{aligned} s_0[2] - y_0 &= l_1[2], \qquad t_0[2] = l_2[2], \\ s_1[2] - y_1 &= l_3[2], \qquad t_1[2] = l_4[2], \\ s_2[2] - y_2 &= l_5[2], \qquad t_2[2] = l_6[2], \\ \gamma[2] - c &= -l_1[2]\hat{x}_0[2] - l_3[2]\hat{x}_1[2] - l_5[2]\hat{x}_2[2], \\ \alpha[2] - a &= l_2[2]\hat{x}_0[2] + l_4[2]\hat{x}_1[2] + l_6[2]\hat{x}_2[2]. \end{aligned}$$

If we denote

$$A[2] = \begin{pmatrix} y_0 & c & 0 & 0 & 0 \\ 0 & -a & 1 & 0 & 0 \\ y_1 & 0 & c & 0 & 0 \\ 0 & 0 & -a & 1 & 0 \\ y_2 & 0 & 0 & c & 0 \\ 0 & 0 & 0 & -a & 1 \end{pmatrix}, \qquad l[2] = \begin{pmatrix} l_1[2] \\ l_2[2] \\ l_3[2] \\ l_4[2] \\ l_5[2] \\ l_6[2] \end{pmatrix},$$

$$y[2] = \begin{pmatrix} -1 \\ x_0[2] \\ x_1[2] \\ x_2[2] \\ x_3[2] \end{pmatrix},$$

it is now tedious though straightforward to show that

$$A[2]y[2] = D_{y[2]}l[2] \qquad (20)$$

where

$$D_{y[2]} = \begin{pmatrix} 1 + \hat{x}_0^2 & 0 & \hat{x}_1\hat{x}_0 & 0 & \hat{x}_2\hat{x}_0 & 0 \\ 0 & 1 + \hat{x}_0^2 & 0 & \hat{x}_1\hat{x}_0 & 0 & \hat{x}_2\hat{x}_0 \\ \hat{x}_0\hat{x}_1 & 0 & 1 + \hat{x}_1^2 & 0 & \hat{x}_1\hat{x}_2 & 0 \\ 0 & \hat{x}_0\hat{x}_1 & 0 & 1 + \hat{x}_1^2 & 0 & \hat{x}_1\hat{x}_2 \\ \hat{x}_0\hat{x}_2 & 0 & \hat{x}_1\hat{x}_2 & 0 & 1 + \hat{x}_2^2 & 0 \\ 0 & \hat{x}_0\hat{x}_2 & 0 & \hat{x}_1\hat{x}_2 & 0 & 1 + \hat{x}_2^2 \end{pmatrix}_{[2]}.$$

Observe that $D_{y[2]}$ is a rank two modification of a diagonal matrix. Also we find that

$$(A[2])^t l[2] = D_{l[2]} y[2], \tag{21}$$

where

$$D_{l[2]} = \begin{pmatrix} \|l\|^2 & 0 & 0 & 0 & 0 \\ 0 & l_1^2 + l_2^2 & l_1 l_3 + l_2 l_4 & l_1 l_5 + l_2 l_6 & 0 \\ 0 & l_1 l_3 + l_2 l_4 & l_3^2 + l_4^2 & l_3 l_5 + l_4 l_6 & 0 \\ 0 & l_1 l_5 + l_2 l_6 & l_3 l_5 + l_4 l_6 & l_5^2 + l_6^2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}_{[2]}.$$

Observe that $D_{l[2]}$ is a rank three matrix. The lower 4 × 4 part can be factored as

$$\begin{pmatrix} l_1 & l_2 \\ l_3 & l_4 \\ l_5 & l_6 \\ 0 & 0 \end{pmatrix}_{[2]} \begin{pmatrix} l_1 & l_3 & l_5 & 0 \\ l_2 & l_4 & l_6 & 0 \end{pmatrix}_{[2]}.$$

Together, the equations (20) and (21), via the equations (7), show that the errors-in-variables Kalman filtering problem reduces to the "nonlinear" generalized SVD problem described in Theorem 1.

If we add one more measurement equation, the set of equations gets updated. For each measurement update, we increase the time horizon index between square brackets by 0.5. We then find the equation

$$A[2.5] y[2.5] = D_{y[2.5]} l[2.5], \tag{22}$$

where

$$A[2.5] = \begin{pmatrix} A[2] \\ (y_3 \quad 0 \quad 0 \quad 0 \quad c) \end{pmatrix},$$

and $D_{y[2.5]}$ now becomes a 7 × 7 matrix of the form

$$D_{y[2.5]} = \begin{pmatrix} 1+\hat{x}_0^2 & 0 & \hat{x}_1\hat{x}_0 & 0 & \hat{x}_2\hat{x}_0 & 0 & \hat{x}_3\hat{x}_0 \\ 0 & 1+\hat{x}_0^2 & 0 & \hat{x}_1\hat{x}_0 & 0 & \hat{x}_2\hat{x}_0 & 0 \\ \hat{x}_0\hat{x}_1 & 0 & 1+\hat{x}_1^2 & 0 & \hat{x}_1\hat{x}_2 & 0 & \hat{x}_3\hat{x}_1 \\ 0 & \hat{x}_0\hat{x}_1 & 0 & 1+\hat{x}_1^2 & 0 & \hat{x}_1\hat{x}_2 & 0 \\ \hat{x}_0\hat{x}_2 & 0 & \hat{x}_1\hat{x}_2 & 0 & 1+\hat{x}_2^2 & 0 & \hat{x}_3\hat{x}_2 \\ 0 & \hat{x}_0\hat{x}_2 & 0 & \hat{x}_1\hat{x}_2 & 0 & 1+\hat{x}_2^2 & 0 \\ \hat{x}_0\hat{x}_3 & 0 & \hat{x}_1\hat{x}_3 & 0 & \hat{x}_2\hat{x}_3 & 0 & 1+\hat{x}_3^2 \end{pmatrix}_{[2.5]}.$$

Also

$$(A[2.5])^t l[2.5] = D_{l[2.5]} y[2.5], \tag{23}$$

where $D_{l[2.5]}$ becomes

$$D_{l[2.5]} = \begin{pmatrix} \|l\|^2 & 0 & 0 & 0 & 0 \\ 0 & l_1^2 + l_2^2 & l_1 l_3 + l_2 l_4 & l_1 l_5 + l_2 l_6 & l_7 l_1 \\ 0 & l_1 l_3 + l_2 l_4 & l_3^2 + l_4^2 & l_3 l_5 + l_4 l_6 & l_3 l_7 \\ 0 & l_1 l_5 + l_2 l_6 & l_3 l_5 + l_4 l_6 & l_5^2 + l_6^2 & l_5 l_7 \\ 0 & l_1 l_7 & l_3 l_7 & l_5 l_7 & l_7^2 \end{pmatrix}_{[2.5]}.$$

The lower 4 × 4 part can be factored as

$$\begin{pmatrix} l_1 & l_2 \\ l_3 & l_4 \\ l_5 & l_6 \\ l_7 & 0 \end{pmatrix}_{[2.5]} \begin{pmatrix} l_1 & l_3 & l_5 & l_7 \\ l_2 & l_4 & l_6 & 0 \end{pmatrix}_{[2.5]}.$$

Taking together Equations (20), (21), (22), and (23), we can now get a picture of a recursive updating of the nonlinear generalized SVD problem that solves the errors-in-variables Kalman filter. The details however will not be worked out here.

It goes without saying that the simple first order example we have given here can be extended to general matrices $A \in \mathbb{R}^{n \times n}$ and $C \in \mathbb{R}^{l \times n}$, with arbitrary given covariance matrices for the model and the noise. Even structure of A and C (such as the requirement that certain elements be zero or equal to each other) can be taken into account, as well as time-varying models.

4.5. *Approximation by a Rank-Deficient Hankel Matrix*

Consider the problem of approximating a given data sequence $a \in \mathbb{R}^{p+q-1}$ by $b \in \mathbb{R}^{p+q-1}$ so that $\sum_{i=1}^{p+q-1}(a_i - b_i)^2$ is minimized subject to $By = 0$, $y^t y = 1$, where B is a $p \times q$ Hankel matrix with the elements of b. The rank deficiency of the Hankel matrix B ensures that b is the impulse response of a finite dimensional linear system of order $q - 1$ at most.

If the sequence a is itself the impulse response of a higher dimensional system, our problem corresponds to *model reduction*. For $p \to \infty$ we get model reduction in the H_2-norm. If the sequence a is a given data sequence (which is not an impulse response, but for instance a noise-corrupted one), one might consider this problem as a *noisy realization problem*. Rank-deficient (block-)Hankel matrices are the key issue in realization problems, which consist in modeling a given set of data by impulse responses from finite dimensional time-invariant linear systems. Applications occur in system identification, modal analysis, biomedical signal processing (such as NMR), etc.

The "classical" realization algorithms such as [21] use the SVD to find a rank-deficient approximation to a full rank Hankel matrix. But the approximation itself does not have the required structure.

Attempts have been made to restore the structure by finding the closest Hankel matrix (in Frobenius norm) to the rank-deficient approximation (which is simply obtained by replacing the antidiagonals by the average of their elements). However, this new Hankel matrix is no longer rank-deficient. One then iterates by again computing the truncated SVD and again obtaining the closest Hankel matrix, etc. It can be shown that this process converges [4, 5], but we will show with an example in Section 6.3 that the solution does not satisfy any H_2 optimality condition.

Another approach is developed by Abatzoglou et al. [1], where the matrix structure is taken into account by parametrizing it with (0, 1) matrices as Fh, where F contains only 0 and 1 and Fh is the Hankel matrix formed with the elements of the vector h. Then a gradient-based minimization approach is derived to solve an equation which is similar to our Equation (30).

The Lagrangian function is $\mathscr{L}(b, y, l) = \sum_{i=1}^{p+q-1}(a_i - b_i)^2 + l^t By + \lambda(y^t y - 1)$ with B Hankel. Setting all derivatives to zero results in the set of equations (a convolution is denoted by $\star$)

$$a - b = l \star y,$$

which means that

$$\begin{aligned} a_1 - b_1 &= l_1 y_1, \\ a_2 - b_2 &= l_1 y_2 + l_2 y_1, \\ a_3 - b_3 &= l_1 y_3 + l_2 y_2 + l_3 y_1, \\ &\vdots \\ a_{p+q-1} - b_{p+q-1} &= l_p y_q, \end{aligned}$$

and

$$B^t l = y\lambda, \qquad y^t y = 1, \qquad By = 0.$$

Note that we have $2p + 2q + 1$ unknowns (the elements of b, l, y, and λ) and exactly $2p + 2q + 1$ equations. The first equation is a convolution which represents $p + q$ equations. It is straightforward to find that $\lambda = 0$ because $l^t By = \lambda = 0$. Let B be the $p \times (q + 1)$ Hankel matrix formed with the elements of b. Then

$$A - B = \begin{pmatrix} l_1 & l_2 & \cdots & l_{p-2} & l_{p-1} & l_p & 0 & \cdots & 0 & 0 & 0 \\ l_2 & l_3 & \cdots & l_{p-1} & l_p & 0 & 0 & \cdots & 0 & 0 & l_1 \\ l_3 & l_4 & \cdots & l_p & 0 & 0 & 0 & \cdots & 0 & l_1 & l_2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ l_p & 0 & \cdots & 0 & 0 & l_1 & l_2 & \cdots & l_{p-3} & l_{p-2} & l_{p-1} \end{pmatrix}$$
$$\times \begin{pmatrix} y_1 & y_2 & \cdots & y_q \\ 0 & y_1 & \cdots & y_{q-1} \\ 0 & 0 & \cdots & y_{q-2} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & y_1 \\ \vdots & \vdots & & \vdots \\ y_q & 0 & \cdots & 0 \\ y_{q-1} & y_q & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ y_2 & y_3 & \cdots & 0 \end{pmatrix}, \tag{24}$$

which means that the difference $A - B$ is the product of a Hankel and a Toeplitz matrix (note the "circulant" structure in both matrices).

A useful property of this factorization, when it is postmultiplied with a vector z, is illustrated here for the case $p = 4$, $q = 3$:

$$\begin{pmatrix} l_1 & l_2 & l_3 & l_4 & 0 & 0 \\ l_2 & l_3 & l_4 & 0 & 0 & l_1 \\ l_3 & l_4 & 0 & 0 & l_1 & l_2 \\ l_4 & 0 & 0 & l_1 & l_2 & l_3 \end{pmatrix} \begin{pmatrix} y_1 & y_2 & y_3 \\ 0 & y_1 & y_2 \\ 0 & 0 & y_1 \\ 0 & 0 & 0 \\ y_3 & 0 & 0 \\ y_2 & y_3 & 0 \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \\ z_2 \end{pmatrix}$$

$$= \begin{pmatrix} z_0 & z_1 & z_2 & 0 & 0 & 0 \\ 0 & z_0 & z_1 & z_2 & 0 & 0 \\ 0 & 0 & z_0 & z_1 & z_2 & 0 \\ 0 & 0 & 0 & z_0 & z_1 & z_2 \end{pmatrix} \begin{pmatrix} y_1 & 0 & 0 & 0 \\ y_2 & y_1 & 0 & 0 \\ y_3 & y_2 & y_1 & 0 \\ 0 & y_3 & y_2 & y_1 \\ 0 & 0 & y_3 & y_2 \\ 0 & 0 & 0 & y_3 \end{pmatrix} \begin{pmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \end{pmatrix}$$

$$= T_z T_y^t l, \tag{25}$$

where T_z and T_y are banded Toeplitz matrices with the elements of z and y. It shows that the Hankel-Toeplitz vector product is converted into a Toeplitz-Toeplitz vector product. We now use this property to eliminate the matrix B. Postmultiplying $A - B$ with y results in $Ay = D_y l$, where D_y is $p \times p$ banded symmetric positive definite Toeplitz of the form $D_y = T_y T_y^t$. Hence, its elements are quadratic functions of the components of y. Similarly, we find by postmultiplying $A^t - B^t$ with l that $A^t l = D_l y$, where D_l is a $q \times q$ symmetric positive definite Toeplitz matrix of a form generated in the same way from the elements of l as D_y from the elements of y: $D_l = T_l T_l^t$. If we normalize l so that $l/\|l\| = x$ and $\|l\| = \sigma$, we have exactly the equations as in (7).

If we are given a linear system of order n and we want to approximate it by one of order $q < n$ (model reduction), the algorithm proposed in this paper can be converted via the z-transform to a z-domain iteration. Details and additional references can be found in [11].

One might also consider to minimize the Frobenius norm $\|A - B\|_F^2$ where both A and B are Hankel, subject to $By = 0$ and $y^t y = 1$. For $p \to \infty$ and $q \to \infty$ this is called the Hilbert-Schmidt-Hankel norm, and it has some very interesting features (see [19]). For this object function, one can show that the equations as in (7) are obtained with

$$\begin{aligned} D_x &= T_x W T_x^t, \\ D_y &= T_y W T_y^t. \end{aligned} \tag{26}$$

Here, T_x and T_y are band Toeplitz matrices as in (25), while W is a diagonal weighting matrix which takes into account the relative frequency of the elements in the Hankel structure:

$$W = \operatorname{diag}\Big[1 \quad \tfrac{1}{2} \quad \tfrac{1}{3} \quad \cdots \quad \tfrac{1}{4} \quad \tfrac{1}{3} \quad \tfrac{1}{2} \quad 1\Big].$$

It is straightforward to show that the orthogonality property now becomes

$$(a - b)^t W^{-1} b = 0.$$

4.6. *System Identification*

If a dynamic time-invariant linear system with one input and one output has q poles and r zeros, its inputs u_k and outputs y_k will be related by

$$\begin{pmatrix} y_1 & y_2 & \cdots & y_q & u_1 & u_2 & \cdots & u_r \\ y_2 & y_3 & \cdots & y_{q+1} & u_2 & u_3 & \cdots & u_{r+1} \\ y_3 & y_4 & \cdots & y_{q+2} & u_3 & u_4 & \cdots & u_{r+2} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ y_p & y_{p+1} & \cdots & y_{p+q-1} & u_p & u_{p+1} & \cdots & u_{p+r-1} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_q \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_r \end{pmatrix} = 0.$$

Here, the coefficients α_i and β_j are the coefficients of the transfer function of the system. The data matrix here is a concatenation of two Hankel matrices. We assume that $p \geqslant q + r$.

When the input-output data are obtained from measurements, they will be corrupted by noise. The noise variance can vary greatly, as the magnitude of the signals can range over several orders of magnitude. Also, some of the elements of the input sequence might be known *a priori* to be equal to zero. If, for instance, $q = 1$ and $r \geqslant 2$, the sequence β_j is a finite impulse response, and part of the input sequence (when starting up from initial conditions zero for instance) might be zero. Another possibility is that certain elements of the input-output sequence might be missing due to unreliable sensor readings.

In all these cases the double Hankel data matrix will not be rank-deficient, and one could try to approximate the given double Hankel data matrix by a rank-deficient one, replacing the measured y_k by a sequence z_k and μ_k by a sequence t_k. This can be formulated as an approximation problem:

$$\min_{z_k, t_k, \alpha_i, \beta_i} \sum_{k=1}^{p+q-1} (y_k - z_k)^2 w_k + \sum_{k=1}^{p+r-1} (u_k - t_k)^2 r_k$$

subject to

$$Za + Tb = 0 \quad \text{and} \quad a^t a + b^t b = 1.$$

Here w_k and r_k are user-defined weights, and Z and T are Hankel matrices generated from the sequences z_k and t_k. The vectors a and b contain the elements α_i and β_j. We will also use $v_k = 1/w_k$, $s_k = 1/r_k$.

Instead of working out the general result in full detail, it suffices to consider a particular example, where $p = 4$, $q = 3$, $r = 2$. Let l be a vector of Lagrange multipliers associated with $Za + Tb = 0$. Then we find, by setting the derivatives of the Lagrangian to zero,

$$\begin{aligned}
&y_1 - z_1 = v_1(l_1 \alpha_1), && u_1 - t_1 = s_1(l_1 \beta_1),\\
&y_2 - z_2 = v_2(l_1 \alpha_2 + l_2 \alpha_1), && u_2 - t_2 = s_2(l_1 \beta_2 + l_2 \beta_1),\\
&y_3 - z_3 = v_3(l_1 \alpha_3 + l_2 \alpha_2 + l_3 \alpha_1), && u_3 - t_3 = s_3(l_2 \beta_2 + l_3 \beta_1),\\
&y_4 - z_4 = v_4(l_2 \alpha_3 + l_3 \alpha_2 + l_4 \alpha_1), && u_4 - t_4 = s_4(l_3 \beta_2 + l_4 \beta_1),\\
&y_5 - z_5 = v_5(l_3 \alpha_3 + l_4 \alpha_2), && u_5 - t_5 = s_5(l_4 \beta_2).\\
&y_6 - t_6 = v_6(l_4 \alpha_3). &&
\end{aligned}$$

Observe that the difference sequences $y_k - z_k$ and $u_k - t_k$ are both obtained from "weighted" convolutions of the sequence l_k with the sequences α_i and β_j. It can also be shown that the Lagrange multiplier associated with the constraint $a^t a + b^t b = 1$ must be zero.

Next we eliminate z_k and t_k using $Za + Tb = 0$. Let Y and U be Hankel matrices with the outputs and inputs. Then we find

$$\begin{aligned}
(Y \quad U)\begin{pmatrix} a \\ b \end{pmatrix} &= (D_a + D_b) l,\\
\begin{pmatrix} Y^t \\ U^t \end{pmatrix} l &= \begin{pmatrix} D_l^y & 0 \\ 0 & D_l^u \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix},\\
a^t a + b^t b &= 1,
\end{aligned}$$

where D_a is a positive definite matrix obtained as $D_a = T_a V T_a^t$ and $D_b = T_b S T_b^t$. Here T_a and T_b are band Toeplitz matrices as in (25), and $V = \operatorname{diag}(v_i)$, $S = \operatorname{diag}(s_i)$. Note that we can take (some or all of) the s_i zero if (some or all of) the inputs are noise-free. The matrices D_l^u and D_l^y are defined similarly.

Let us conclude by pointing out that when the data are generated by an exact linear time-invariant system, but corrupted by additive white noise which is zero mean normally distributed, the STLS method here will also provide the maximum likelihood estimates.

4.7. *Approximation by a Rank-Deficient Toeplitz Matrix*

Let $A \in \mathbb{R}^{p \times p}$ be a symmetric Toeplitz matrix with elements a_i, $i = 1, \ldots, p$ of full rank p. Minimizing $\sum_{i=1}^{p}(a_i - b_i)^2$ subject to $By = 0$ and $y^t y = 1$, where also $B \in \mathbb{R}^{p \times p}$ is required to be symmetric Toeplitz, leads to a similar generalized matrix decomposition to that in Theorem 1. Because of the symmetry in the problem, the normalized vector of Lagrange multipliers x will be equal to y, and hence the generalized SVD problem reduces to a generalized symmetric eigenvalue problem of the form

$$Ax = D_x x \sigma, \qquad x^t x = 1,$$

where the matrix D_x is positive definite. It has a spectral structure which is illustrated here for the case $p = 6$:

$$D_x = \begin{pmatrix}
x_1 & x_2 & x_3 & x_4 & x_5 & x_6\\
x_2 & x_1 + x_3 & x_4 & x_5 & x_6 & 0\\
x_3 & x_2 + x_4 & x_1 + x_5 & x_6 & 0 & 0\\
x_4 & x_3 + x_5 & x_2 + x_6 & x_1 & 0 & 0\\
x_5 & x_4 + x_6 & x_3 & x_2 & x_1 & 0\\
x_6 & x_5 & x_4 & x_3 & x_2 & x_1
\end{pmatrix} \times (\cdot)^t.$$

It can be seen that the left factor is Hankel + Toeplitz. A similar structure is obtained for every p.

Interestingly enough, it can be shown, using the results in [6], that the matrix D_x will be rank-deficient. Since A and D_x are symmetric Toeplitz matrices, it follows that the minimal eigenvector of $Ax = D_x x\sigma$ will be highly structured. Indeed, the fact that each eigenvector of a symmetric Toeplitz matrix is either reciprocal (symmetric around its midpoint) or antireciprocal (skew-symmetric around its midpoint) can trivially be extended to our generalized symmetric Toeplitz eigenvalue problem. This structure implies the rank deficiency of D_x, which is of rank $p/2$ for p even and rank $(p \pm 1)/2$ for p odd.

4.8. *Giving a Matrix a Specified Singular Value*

Suppose we want to find a matrix B which is close to A such that B has a specified singular value β. This can be formulated as

$$\min_{B\in\mathbb{R}^{p\times q},\, u\in\mathbb{R}^{p},\, \nu\in\mathbb{R}^{q}} \sum_{i=1}^{p}\sum_{j=1}^{q}(a_{ij}-b_{ij})^2 w_{ij}$$

$$\text{subject to}\quad Bv = u\beta,\quad B^t u = v\beta,\quad u^t u = 1,$$

where w_{ij} are user-defined weights. Here u and v are singular vectors corresponding to the singular value β (it is easy to show that $u^t u = v^t v$ always). Introducing Lagrange multiplier vectors $l \in \mathbb{R}^p$ and $k \in \mathbb{R}^q$ results in the set of equations

$$(a_{ij}-b_{ij})w_{ij} = l_i v_j + k_j u_i,\qquad B^t l = k\beta,\qquad -l\beta + Bk = u\lambda,$$

$$Bv = u\beta,\qquad B^t u = v\beta,\qquad u^t u = 1.$$

After some manipulation, one can show that $\lambda = 0$, $l = u$, and $v = k$, so that the nonlinear generalized SVD problem becomes

$$\begin{pmatrix} -\beta I_p & A \\ A^t & -\beta I_q \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} D_u & 0 \\ 0 & D_v \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}\sigma,$$

where $D_u = \mathrm{diag}(V^t\,\mathrm{diag}(u)\,u)$ and $D_v = \mathrm{diag}(V\,\mathrm{diag}(v)\,v)$. The matrix B can be reconstructed as $B = A - \mathrm{diag}(u)\,V\,\mathrm{diag}(v)\,\sigma$. Note that if V is a matrix consisting of all ones, the answer is basically given in terms of the SVD of A. As with the rank-deficient case (where $\beta = 0$), there is not always a solution.

4.9. *The Largest Stability Radius of an Uncertain Linear System*

The following type of problem occurs in the determination of the minimum distance to instability, or parameter margin, for linear time-invariant systems that are subject to rational variations of certain parameters: Consider a linear time-invariant closed-loop system of order q with $\dot{x} = F(r)x$, where $x \in \mathbb{R}^q$ is the state and $F(r) \in \mathbb{R}^{q\times q}$ is a rational matrix function of a parameter vector r of size n. Then the two parameter margin $R_2(M, r)$ of $F(r)$ is defined as

$$R_2(M, r) = \min_{r\in\mathbb{R}^n}\left\{\|r\|_2 \middle| \det[I_N + M\Delta(r)] = 0\right\} \tag{27}$$

with

$$\Delta(r) = \mathrm{blockdiag}\begin{bmatrix} r_1 I_{N_1} & \cdots & r_n I_{N_n}\end{bmatrix}.$$

Here M is a constant real $N \times N$ matrix which is constructed from the entries of the matrix $F(r)$, and the dimensions N_i of the blocks of the block-diagonal matrix $\Delta(r)$ add up to N. Full details can be found in [13] and the references there.

In general, the s-parameter margin for the system $\dot{x} = F(r)x$ is defined as

$$R_s(F) = \min_{r\in\mathbb{R}^q}\left\{\|r\|_s \middle| F(r) \text{ is unstable}\right\}.$$

In Doyle's μ-analysis, the singularity constraint is expressed as $\det[j\omega I - F(r)] = 0$, where ω is the frequency. In order to solve this problem, a frequency sweep is required. The pitfall with Doyle's μ is that the problem is not continuous in the input data r when these are real. However, for many robust stability problems, the singularity constraint can be written as a singularity constraint on a real, frequency-independent matrix, which reduces to the formulation as in (27). In [13] this problem is solved for two parameters ($n = 2$) in the l_2-norm, but our technique will work for any number n of parameters.

Obviously, (27) is a special case of the following problem:

$$\min_{b\in\mathbb{R}^n} b_1^2 + \cdots + b_n^2$$

$$\text{subject to}\quad \left(B_0 + \sum_{i=1}^{n} B_i b_i\right) y = 0,\quad y^t y = 1,$$

where the two constraints express the rank deficiency of $B = B_0 + \sum_{i=1}^{n} B_i b_i$ (the matrices B_i are not necessarily square, but in the stability radius problem they are). It is straightforward to convert this problem to (7) in which

$$D_x = \sum_{k=1}^{n} (B_k^t x)(B_k^t x)^t,$$

$$D_y = \sum_{k=1}^{n} (B_k y)(B_k y)^t,$$

and $A = -B_0$. Obviously, D_x and D_y are nonnegative (or positive) definite matrices.

5. AN ALGORITHM

We now present an algorithm to solve the set of nonlinear equations (7) [with some slight modifications, this algorithm can solve (8) instead]. The iteration number is indexed between square brackets.

We will use the QR decomposition of A:

$$A = \begin{pmatrix} \underbrace{Q_1}_{p \times q} & \underbrace{Q_2}_{p \times (p-q)} \end{pmatrix} \begin{pmatrix} \underbrace{R}_{q \times q} \\ 0 \end{pmatrix}. \tag{28}$$

We can decompose $l = x\sigma$ as $l = Q_1 z + Q_2 w$ for certain vectors z and w. From (7) we find

$$\begin{pmatrix} R^t & 0 & 0 \\ Q_2^t D_y Q_1 & Q_2^t D_y Q_2 & 0 \\ Q_1^t D_y Q_1 & Q_1^t D_y Q_2 & -R \end{pmatrix} \begin{pmatrix} z \\ w \\ y \end{pmatrix} = \begin{pmatrix} D_l y \\ 0 \\ 0 \end{pmatrix}. \tag{29}$$

A possible algorithm to calculate the smallest eigenvalue and corresponding eigenvector of a symmetric matrix is by inverse iteration. Instead however of calculating the minimal eigenvalue in each step, we could also perform only one step of an inverse iteration scheme and then update the weighting matrices D_x and D_y. (A variation of this algorithm could be to perform $t > 1$ steps of inverse iteration with fixed D_x and D_y. However, here we will only use $t = 1$.) This is achieved in the following iteration, which is just an iterative way of solving Equation (29).[4]

[4] As a matter of fact, there are several other possibilities for solving this set of equations iteratively (e.g. Gauss-Seidel-like or SOR-like variants), but we only analyse one particular version here.

INVERSE ITERATION ALGORITHM.

```
**Initialization**: Choose $x^{[0]}$, $y^{[0]}$, $\sigma^{[0]}$, and normalize $\|x^{[0]}\| = 1$, $\|y^{[0]}\| = 1$.
**For $k = 1$ till convergence**:
    1. $z^{[k]} = R^{-t} D_{x^{[k-1]}} y^{[k-1]}$.
    2. $w^{[k]} = -(Q_2^t D_{y^{[k-1]}} Q_2)^{-1} (Q_2^t D_{y^{[k-1]}} Q_1) z^{[k]}$.
    3. $x^{[k]} = Q_1 z^{[k]} + Q_2 w^{[k]}$.
    4. $x^{[k]} = x^{[k]} / \|x^{[k]}\|$.
    5. $y^{[k]} = R^{-1} Q_1^t D_{y^{[k-1]}} x^{[k]}$.
    6. $\sigma^{[k]} = \|y^{[k]}\|$.
    7. $y^{[k]} = y^{[k]} / \sigma^{[k]}$.
    8. Convergence test using $x^{[k]}$, $y^{[k]}$, $\sigma^{[k]}$.
```

As the numerical experiments below demonstrate, the convergence rate of this algorithm seems to be linear. Intuitively, this can be understood when we assume that D_x and D_y are invertible. In that case, we can eliminate the vector $x = D_y^{-1} A y / \sigma$ and write

$$\left(A^t D_y^{-1} A \right) y = D_x y \sigma^2, \tag{30}$$

which is a generalized eigenvalue problem in which the weights D_x and D_y are quadratic functions of the elements of x and y. If D_x is invertible, we might even convert this generalized eigenvalue problem into a symmetric eigenvalue problem as

$$T_{xy}\left(D_x^{1/2} y\right) = \left(D_x^{-1/2} A^t D_y^{-1} A D_x^{-1/2}\right)\left(D_x^{1/2} y\right) = \left(D_x^{1/2} y\right) \sigma^2,$$

where $D_x^{1/2}$ is a symmetric square root of D_x. Note that T_{xy} is a symmetric positive definite matrix, which implies that all its eigenvalues are real positive. As we will see in the numerical examples below, $D_{x^{[k]}}$ and $D_{y^{[k]}}$ converge rapidly to matrices that are approximately constant, which implies that also T_{xy} is approximately constant, so that we are basically iterating with T_{xy}^{-1}.

This observation implies that, asymptotically, the convergence rate is linear and will be governed by the two leading eigenvalues λ_1 and λ_2 of T_{xy}^{-1}. In particular, one can then show that with $v^{[k]} = D_{x^{[k]}}^{1/2} y^{[k]} / \| D_{x^{[k]}}^{1/2} y^{[k]} \|$,

$$\frac{\log_{10}\left(v_{k+2}^t v_{k+1}\right)}{\log_{10}\left(v_{k+1}^t v_k\right)} = \left(\frac{\lambda_2^{[k]}}{\lambda_1^{[k]}} \right)^2.$$

Hence, the cosine of the angle between two iteration vectors decreases linearly. A natural convergence test is to monitor the difference between two consecutive iterates, as e.g. $\| y^{[k]} - y^{[k-1]} \|$.

Another implication is that $\log_{10} [\sigma_{\min}(B^{[k]})$ will decrease linearly as a function of the iteration number k, where $B^{[k]} = B_0 + \sum_{i=1}^{m} B_i b_i^{[k]}$ with $b_i^{[k]} = a_k - x^{[k]} B_k y^{[k]} \sigma^{[k]}$. This provides an even better, though much more expensive, convergence test. A good initial guess might be provided by the singular triplet corresponding to the smallest singular value of A.

6. SOME NUMERICAL EXAMPLES

All examples below were generated in MATLAB.

6.1. *Relative Error Total Least Squares*

Let $A \in \mathbb{R}^{p \times q}$, and B be a rank-deficient approximation of it. The relative error r_{ij} for each element is defined as $r_{ij} = |a_{ij} - b_{ij}| / |b_{ij}|$. This implies that $|a_{ij} - b_{ij}| / |a_{ij}| = r_{ij} / (1 + r_{ij})$. If we now choose as weights in the problem (13), $w_{ij} = 1/a_{ij}^2$, we are minimizing

$$\sum_{i=1}^{p} \sum_{j=1}^{q} \frac{r_{ij}^2}{(1 + r_{ij})^2},$$

which for small relative errors is approximately equal to $\sum_{i=1}^{p} \sum_{j=1}^{q} r_{ij}^2$. As an example, consider the linear fit in two dimensions of a set of measurements, ranging over several order of magnitudes. Typically, the measurement errors are relative and not absolute, so that it is more meaningful to minimize the sum of squares of the relative errors than the sum of squares of the absolute errors. An example is shown in Figure 1.

6.2. *Total Least Squares with Fixed Elements*

Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 5 & 6 \\ 5 & 6 & 7 & 1 \\ 2 & 3 & 5 & 8 \\ 5 & 3 & 2 & 1 \end{pmatrix}$$

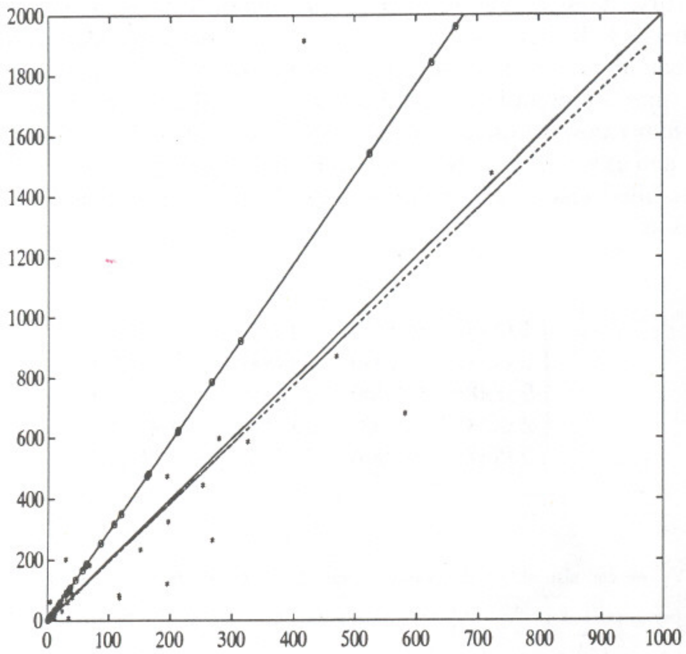

FIG. 1. The unadorned full line represents exact data. The asterisks are 50 noisy observations. Obviously, for large data values, the errors get large too. The dashed line is the weighted total least squares solution with $w_{ij} = 1/a_{ij}^2$ as weights, which approximately minimizes the sum of squared relative errors. The full line with circles is the unweighted TLS solution (which minimizes the sum of squares of absolute errors). Although it would be more meaningful to compare the null spaces instead of the ranges of the matrices A and B (see [10] for an explanation), still, we see here clearly that the relative error fit is much better than the absolute error fit.

and the four inverse weight matrices

$$V_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad V_2 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix},$$

$$V_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \qquad V_4 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}.$$

For the first three weighting matrices, a solution can be found with other algorithms as well. For instance, for V_1, since we are only allowed to modify the last column, a simple least squares regression will do. For V_2, one could use the approach described in [16], while for V_3, the results in [7, 9] apply.[5] For the four cases, we used as initial vectors $x_0 = (1\ 1\ 1\ 1\ 1)^t$ and $y_0 = (1\ 1\ 1\ 1)^t$, and as a convergence criterion the test $\sigma_4(B^{[k]}) \leqslant 10^{-13}$, where $B^{[k]}$ is the modified matrix at iteration k. For V_1, the matrix B is found in two iterations as

$$B = \begin{pmatrix} 1.0000 & 2.0000 & 3.0000 & 2.4330 \\ 2.0000 & 1.0000 & 5.0000 & 7.0258 \\ 5.0000 & 6.0000 & 7.0000 & 3.9158 \\ 2.0000 & 3.0000 & 5.0000 & 4.4731 \\ 5.0000 & 3.0000 & 2.0000 & -0.6019 \end{pmatrix}.$$

[5] For V_3, we can also obtain the solution from the SVD of a Schur complement as follows: Let A be partitioned as $\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ according to the structure of V_3. Then we need to find D such that $\| A_{22} - D \|_F^2$ is minimized subject to

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & D \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 0 \quad \text{and} \quad y_1^t y_1 + y_2^t y_2 = 1.$$

Using vectors of Lagrange multipliers l_1 and l_2, one easily finds

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} - l_2 y_2^t \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 0,$$

$$\begin{pmatrix} A^t & A_{21}^t \\ A_{12}^t & A_{22}^t - y_2 l_2^t \end{pmatrix} \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} = 0.$$

Observe that D is a rank one modification of A_{22}. If A_{11} is square-invertible, we can eliminate y_1 and l_1. Calling $\|l_2\| = \beta$, $x = l_2/\beta$, $\|y_2\| = \alpha$, $y = y_2/\alpha$, and $\alpha\beta = \sigma$, we find

$$\left(A_{22} - A_{21} A_{11}^{-1} A_{12} \right) y = x\sigma, \qquad x^t x = 1,$$

$$\left(A_{22} - A_{21} A_{11}^{-1} A_{12} \right)^t x = y\sigma, \qquad y^t y = 1,$$

which implies that (x, σ, y) is a singular triplet of the Schur complement. The general solution when A_{11} is not square or not invertible is explored in [7, 9].

It is easy to verify that the last column is equal to $A_1 A_1^\dagger A_2$ where A_1 contains the first three columns of A and A_2 is the fourth column of A. It follows from (14) that $A - B$ is a rank one matrix.

For V_2, the approximating matrix B of rank 3 and the matrix R of Section 4.2 are found in 13 iterations as

$$B = \begin{pmatrix} 1.0000 & 2.0000 & 3.4722 & 3.7987 \\ 2.0000 & 1.0000 & 3.6830 & 6.5615 \\ 5.0000 & 6.0000 & 6.0947 & 1.3860 \\ 2.0000 & 3.0000 & 5.9952 & 7.5757 \\ 5.0000 & 3.0000 & 2.9396 & 0.5994 \end{pmatrix},$$

$$R = \begin{pmatrix} -0.0128 & 0.4438 & 0 & 0 \\ 0.0358 & -1.2379 & 0 & 0 \\ 0.0246 & -0.8509 & 0 & 0 \\ -0.0270 & 0.9354 & 0 & 0 \\ -0.0255 & 0.8831 & 0 & 0 \end{pmatrix},$$

with corresponding triplets $x^t = (-0.2190,\ 0.6107,\ 0.4198,\ -0.4615,\ -0.4357)$, $\sigma = 3.0996$, and $y^t = (0.0189,\ -0.6539,\ 0.6957,\ -0.2966)$. It can be verified that (x, σ, y) is the *third* singular triplet of $A + R$ (and not the fourth). For V_3, we find in 10 iterations

$$B = \begin{pmatrix} 1.0000 & 2.0000 & 3.0000 & 4.0000 \\ 2.0000 & 1.0000 & 5.0000 & 6.0000 \\ 5.0000 & 6.0000 & 5.0494 & 2.1037 \\ 2.0000 & 3.0000 & 5.7907 & 7.5526 \\ 5.0000 & 3.0000 & 3.9366 & -0.0958 \end{pmatrix}$$

The same solution can be obtained via the SVD of the Schur complement. It follows from (14) that rank$(A - B) = 2$.

For the checkerboard weighting matrix V_4, 17 iterations were required to find the matrices B and R:

$$B = \begin{pmatrix} 1.4482 & 2.0000 & 3.6558 & 4.0000 \\ 2.0000 & 2.5895 & 5.0000 & 6.2960 \\ 5.0246 & 6.0000 & 7.0360 & 1.0000 \\ 2.0000 & 2.2966 & 5.0000 & 7.8690 \\ 4.9885 & 3.0000 & 1.9832 & 1.0000 \end{pmatrix},$$

$$R = \begin{pmatrix} 0 & 1.1099 & 0 & 0.2067 \\ 0.6419 & 0 & 0.9392 & 0 \\ 0 & 0.0610 & 0 & 0.0114 \\ -0.2840 & 0 & -0.4156 & 0 \\ 0 & -0.0284 & 0 & -0.0053 \end{pmatrix},$$

with corresponding triplet $x^t = (-0.5379,\ 0.7703,\ -0.0296,\ -0.3409,\ 0.0138)$, $\sigma = 2.5663$, and $y^t = (0.3247,\ -0.8040,\ 0.4751,\ -0.1497)$, which is the *fourth* singular triplet of $A + R$. Convergence patterns for the checkerboard matrix V_4 are shown in Figure 2; the evolution of the components of $y^{[k]}$ is shown in Figure 3.

6.3. *Noisy Realization*

The noisy realization problem is to approximate a given sequence a by a sequence b such that the $p \times q$ Hankel matrix B with the elements of b is rank-deficient. This ensures that b is the impulse response of a system of order at most $q - 1$.

We only present a small example here, mainly to show that the iteration proposed in [4] does *not* satisfy any of the optimality properties of Theorem

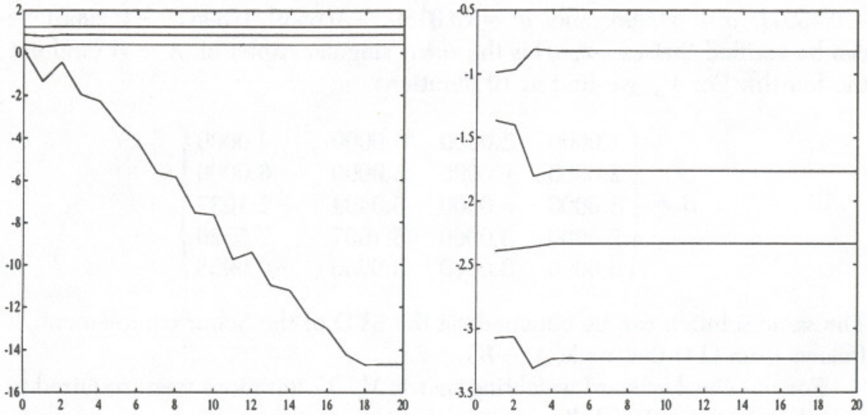

FIG. 2. Convergence patterns (here for V_4 of Section 6.2, but they are typical for all examples in this paper). The x-axis contains the iteration k. The left plot is the logarithm of the singular values of the approximating matrix $B^{[k]}$ as a function of k, while the right plot contains the eigenvalues of T_{xy}^{-1} as a function of k. The interpretation as an inverse power method allows us to estimate the asymptotic convergence rate, which will be governed by the largest two eigenvalues of the matrix T_{xy}^{-1}. As a consequence the smallest singular value of $B^{[k]}$ decreases linearly (at least asymptotically) with a slope determined by $\lambda_2^{[k]}/\lambda_1^{[k]}$. Observe that after five iterations we have already reached the asymptotic regime. Also note that the other singular values of $B^{[k]}$ are approximately constant, as would be the case in the unstructured inverse power method.

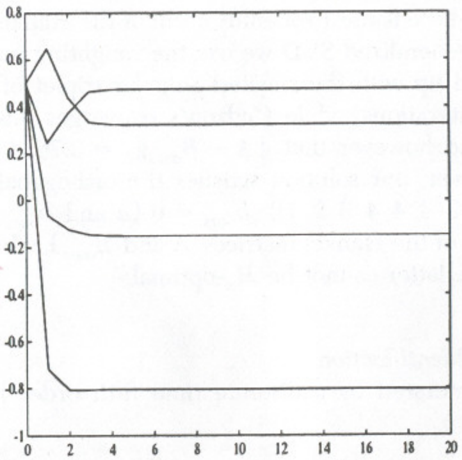

FIG. 3. Evolution of the components of the vector $y^{[k]}$ as a function of the iteration number k for the checkerboard pattern V_4.

1. Consider the Hankel matrix:

$$A = \begin{pmatrix} 3 & 4 & 2 & 1 \\ 4 & 2 & 1 & 5 \\ 2 & 1 & 5 & 6 \\ 1 & 5 & 6 & 7 \\ 5 & 6 & 7 & 1 \\ 6 & 7 & 1 & 2 \end{pmatrix}$$

and its two rank-deficient approximants B_{ours} and B_{Cadzow}:

$$B_{\text{ours}} = \begin{pmatrix} 3.4535 & 3.5356 & 2.0027 & 1.4871 \\ 3.5356 & 2.0027 & 1.4871 & 4.0396 \\ 2.0027 & 1.4871 & 4.0396 & 7.0785 \\ 1.4871 & 4.0396 & 7.0785 & 5.9951 \\ 4.0396 & 7.0785 & 5.9951 & 1.7211 \\ 7.0785 & 5.9951 & 1.7211 & 1.6138 \end{pmatrix},$$

$$B_{\text{Cadzow}} = \begin{pmatrix} 2.9449 & 3.2181 & 2.1356 & 1.6930 \\ 3.2181 & 2.1356 & 1.6930 & 4.0012 \\ 2.1356 & 1.6930 & 4.0012 & 7.0541 \\ 1.6930 & 4.0012 & 7.0541 & 6.1020 \\ 4.0012 & 7.0541 & 6.1020 & 1.5591 \\ 7.0541 & 6.1020 & 1.5591 & 1.6209 \end{pmatrix}.$$

The norm we use here is the Frobenius norm of the Hankel matrices. Hence, in the nonlinear generalized SVD we use the weighting matrices D_x and D_y as in (26). Started up with the smallest singular triplet of A, our algorithm converges in 14 iterations, while Cadzow's converges (also linearly) in 125 iterations. We find however that $\| A - B_{\text{ours}} \|_F = 3.7614 < \| A - B_{\text{Cadzow}} \|_F = 3.8503$. Moreover, our solution satisfies the orthogonality property $(a - b_{\text{ours}})^t \cdot \operatorname{diag}([1\ 2\ 3\ 4\ 4\ 4\ 3\ 2\ 1]) \cdot b_{\text{ours}} = 0$ (a and b_{ours} are 9×1 vectors with the numbers of the Hankel matrices A and B_{ours}), while Cadzow's result doesn't; hence the latter cannot be H_2-optimal.

6.4. *System Identification*

Data were generated by a discrete time fifth order linear system with transfer function

$$h(z) = \frac{0.0202 z^5 + 0.6844 z^4 - 0.6182 z^3 - 0.3996 z^2 + 0.3631 z + 0.0580}{z^5 - 1.1000 z^4 - 0.5200 z^3 + 0.8360 z^2 + 0.3931 z - 0.5206}.$$

The input was generated as rounded Gaussian white noise of mean zero and variance 100 (`round(10*rand(100,1))`). The input and output were then corrupted by white noise with variance 25. We call the signals obtained in this way u_k and y_k, respectively, and their fits t_k and z_k; they are plotted in Figure 4. Models with $q = r$ were fitted, where q ranged from 2 to 9.

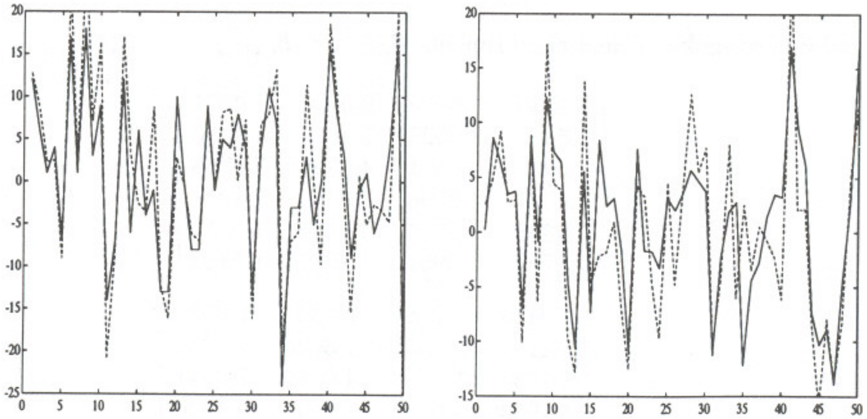

FIG. 4. The original noiseless input is the full line in the left plot; the noiseless output is the full line in the right plot. The noise-corrupted versions are dashed lines.

This means that we have models from order 1 up to order 8. The number p of rows of the double block Hankel matrix was adapted from 57 to 50, so as to keep the number of data used in the identification constant (i.e. $p + q - 1 = 58$). The starting vectors were always the left and right singular vectors corresponding to the smallest singular value of the $p \times (q + r)$ double Hankel matrix $[Y\ U]$. The misfit-versus-complexity tradeoff is illustrated in Table 1 and Figure 5.

6.5. *Symmetric Toeplitz Matrix*

Consider the L_2 approximation of the sequence $a = (4, -9, 7, 4, 2, 4, -8, 1, 3, 4, 4, 4, 5, -7, 2, 0, 8, 4, 8, 5, 5, -5, 3)$ by a sequence b such that the corresponding symmetric Toeplitz matrix B is rank-deficient. For this example, the set of equations becomes a "nonlinear" generalized eigenvalue problem of the form $Ax = D_x x \sigma$ (where we have to allow σ to be negative). When the initial vector $x_0 = (1, 1, 1, \ldots, 1)$, the final vector x is (read from left to right):

$$\begin{array}{ccccc}
(5.6215\text{E} - 02 & 1.0602\text{E} - 01 & 1.3074\text{E} - 01 & 1.7144\text{E} - 01 & 1.7480\text{E} - 01 \\
1.2414\text{E} - 01 & 5.8171\text{E} - 02 & -4.7079\text{E} - 02 & -1.9742\text{E} - 01 & -3.1512\text{E} - 01 \\
-3.9844\text{E} - 01 & -4.2549\text{E} - 01 & -3.9844\text{E} - 01 & -3.1512\text{E} - 01 & -1.9742\text{E} - 01 \\
-3.9844\text{E} - 01 & -4.2549\text{E} - 01 & -3.9844\text{E} - 01 & -3.1512\text{E} - 01 & -1.9742\text{E} - 01 \\
1.3074\text{E} - 01 & 1.0602\text{E} - 01 & 5.6215\text{E} - 02), & &
\end{array}$$

which is a symmetric vector (it is symmetric around its midpoint). The corresponding matrix D_x is of rank 11. The value of σ is -0.31411, and the

TABLE 1

Order $q - 1$	$\Sigma_{i=1}^{58}[(y_i - z_i)^2 + (u_i - t_i)^2]$
1	39.8519
2	37.1859
3	36.2391
4	26.8839
5	25.5382
6	23.3237
7	21.6729
8	17.8117

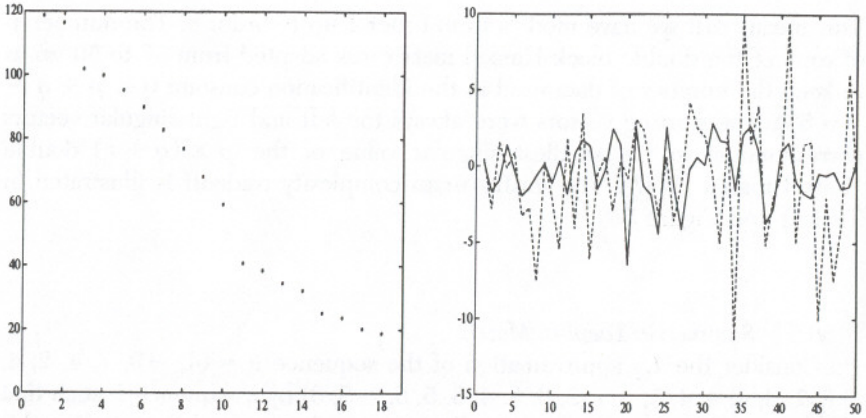

FIG. 5. The left plot shows the singular values of the double Hankel matrix with $p = 50$ and $q = r = 9$. Theoretically, there should be a gap in the singular spectrum. Without noise, the rank of the double Hankel matrix would be the rank of U (which is 9) plus the order of the system (which is 5); see e.g. [22]. Hence the rank of the double Hankel matrix would be 14. Obviously, with noise there is no clear gap which would allow us to determine the order (which is needed in the SVD subspace algorithms for system identification in e.g. [25]). Therefore, it makes sense to consider several models and evaluate them on the tradeoff between misfit and complexity. The right plot shows the output error, which is the difference between y_k and z_k for order 8 (full line) and order 1 (dashed line). Clearly, the misfit is smaller for the more complex model.

approximating sequence b is

$$\begin{array}{rrrrr}
(4.3141\mathrm{E}+00 & -8.4158\mathrm{E}+00 & 7.4642\mathrm{E}+00 & 4.2910\mathrm{E}+00 & 2.0930\mathrm{E}+00 \\
3.9055\mathrm{E}+00 & -8.2423\mathrm{E}+00 & 6.6583\mathrm{E}-01 & 2.6373\mathrm{E}+00 & 3.6655\mathrm{E}+00 \\
3.7333\mathrm{E}+00 & 3.8213\mathrm{E}+00 & 4.9098\mathrm{E}+00 & -7.0165\mathrm{E}+00 & 2.0353\mathrm{E}+00 \\
6.2470\mathrm{E}-02 & 8.0678\mathrm{E}+00 & 4.0602\mathrm{E}+00 & 8.0459\mathrm{E}+00 & 5.0295\mathrm{E}+00 \\
5.1063\mathrm{E}+00 & -4.9925\mathrm{E}+00 & 3.0020\mathrm{E}+00). & &
\end{array}$$

If however we take as an initial guess the right singular vector of A corresponding to the smallest singular value, we find for x

$$\begin{array}{rrrrr}
(2.5312\mathrm{E}-01 & 6.7382\mathrm{E}-02 & 2.0092\mathrm{E}-01 & -7.3734\mathrm{E}-02 & -5.6324\mathrm{E}-02 \\
-2.4955\mathrm{E}-01 & -1.0956\mathrm{E}-01 & -3.8270\mathrm{E}-01 & -3.0173\mathrm{E}-01 & -2.6237\mathrm{E}-01 \\
-4.2434\mathrm{E}-02 & -2.1264\mathrm{E}-16 & 4.2434\mathrm{E}-02 & 2.6237\mathrm{E}-01 & 3.0173\mathrm{E}-01 \\
3.8720\mathrm{E}-01 & 1.0956\mathrm{E}-01 & 2.4955\mathrm{E}-01 & 5.6324\mathrm{E}-02 & 7.3734\mathrm{E}-02 \\
-2.0092\mathrm{E}-01 & -6.7382\mathrm{E}-02 & -2.5312\mathrm{E}-01), & &
\end{array}$$

which is an antireciprocal vector (skew-symmetric around its midpoint). Here $\sigma = -0.37403$, and the corresponding approximating sequence c (say) is

$$\begin{array}{rrrrr}
(4.3740\mathrm{E}+00 & -8.5377\mathrm{E}+00 & 7.4489\mathrm{E}+00 & 4.0990\mathrm{E}+00 & 2.0225\mathrm{E}+00 \\
3.7079\mathrm{E}+00 & -8.3635\mathrm{E}+00 & 5.1886\mathrm{E}-01 & 2.5860\mathrm{E}+00 & 3.6942\mathrm{E}+00 \\
3.7905\mathrm{E}+00 & 3.9767\mathrm{E}+00 & 5.0352\mathrm{E}+00 & -6.7883\mathrm{E}+00 & 2.1559\mathrm{E}+00 \\
2.2477\mathrm{E}-01 & 8.0795\mathrm{E}+00 & 4.1223\mathrm{E}+00 & 7.9986\mathrm{E}+00 & 5.0077\mathrm{E}+00 \\
4.9205\mathrm{E}+00 & -5.0255\mathrm{E}+00 & 2.9521\mathrm{E}+00). & &
\end{array}$$

We find that

$$\|a-b\| = 1.1401 < \|a-c\| = 1.2146.$$

This example illustrates an important point: *There is no guarantee that the algorithm will find a global optimum upon convergence*. In other words, the results may be dependent on the initial estimate.

6.6. *Giving a Matrix a Specified Singular Value*

Consider the matrix A with the inverse weight matrix V:

$$A = \begin{pmatrix} 16 & -15 & 1 & -1 \\ 8 & 3 & 12 & 8 \\ 1 & -8 & 13 & 11 \\ 13 & 3 & -1 & 19 \end{pmatrix}, \qquad V = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}.$$

Suppose we want to assign a singular value β to A, by keeping its elements in positions $(1,1)$, $(2,4)$, $(3,1)$, and $(4,3)$ fixed, which is reflected by the zero-one structure of V. In Figure 6 we show some results on finding the closest matrix B which coincides with A on the zero support of V and which has a specified singular value $\beta \in [0, 40]$. The convergence behavior is shown in Figure 7.

7. CONCLUSIONS

While many of the problems we have discussed here have been considered separately in the literature, it was not recognized that all of these problems reduce to our main result of Theorem 1, which is the major contribution of this paper.

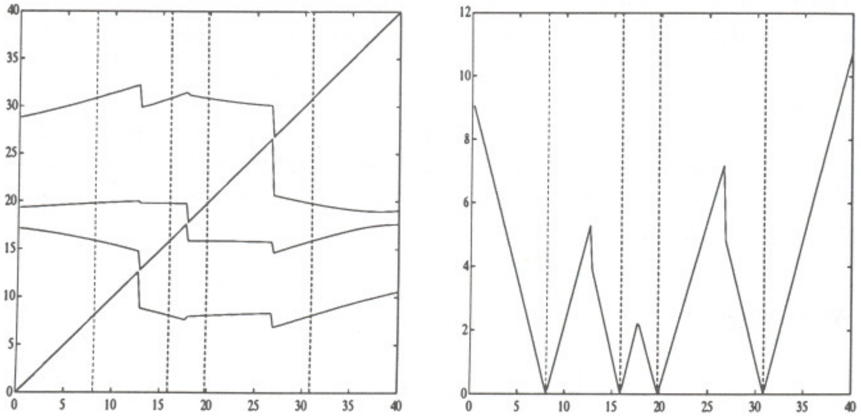

FIG. 6. In both plots, the x-axis contains values of $\beta \in [0, 40]$ with steps of 0.2. The left plot shows the singular values of the approximating matrix B. The right plot contains $\|A - B\|_F$ after convergence. In both plots, the vertical dashed lines are at the positions of the singular values of A. For values of β smaller than $\sigma_4(A)$, β is the smallest singular value of B. Somewhere in between the fourth and third singular values of A, β becomes the third singular value of B. In between $\sigma_3(A)$ and $\sigma_2(A)$, it becomes the second, and for large values it is the dominant singular value of the approximation B. Observe that $\|A - B\|_F$ becomes zero when β hits a singular value of A.

We have only presented one rude outline of an algorithm in this paper, and much more work needs to be done on possible refinements, accelerations, and proofs of (local) convergence. There are several other algorithms that have been considered to solve related problems such as NIPALS (nonlinear iterative partial least squares; see references in [15]), the Newton-method-based algorithm of [1], homotopy methods, etc.

An expensive part of the inverse power algorithm is the explicit calculation of the matrix Q_2 (especially if p gets large). One can however avoid it by applying an inverse iteration scheme to the nonlinear generalized eigenvalue problem (30), where a $q \times q$ matrix needs to be inverted in which the inverse of D_y, which is $p \times p$, appears. D_y however is in many cases a highly structured matrix. Therefore, clever accelerations are possible, such as e.g. fast QR via displacement rank concepts. In the system identification example, one could exploit the special banded Toeplitz structure of the matrices D_x and D_y using the FFT.

In the H_2 model reduction problem, for $p \to \infty$, we have discovered that all the signals involved in the iteration can be modeled as infinite impulse responses from rational transfer functions. Therefore, one can map the iterations to the z-domain using z-transforms and iterate in the z-domain. In this way, we get rid of the large number p (for details, see [11]). There are some good acceleration techniques that could be used in connection with the power method (such as the use of Chebyshev polynomials (see e.g. [8]) or shifts (see e.g. [17] or other techniques described in [24]).

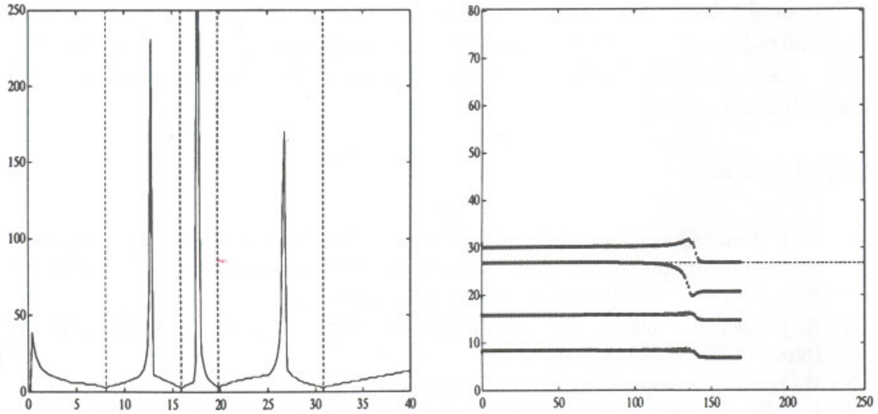

FIG. 7. For each iteration, the initial vector $x^{[0]}$ is the vector x obtained from the previous value of β. The left plot shows, as a function of β, the number of iterations required to reach convergence (defined by $\|By\| - \beta\| < 1\text{E-}010$, where y is the right singular vector approximation). It can be seen that between each pair of singular values of A, there is a certain value of β for which the number of iterations gets large. For instance, between $\sigma_2(A)$ and $\sigma_1(A)$, for $\beta = 26.8$, 170 iterations are required to achieve convergence, even when a good initial estimate is available from the previous iteration. For a value of β which is a little bit smaller, say 26.6, β would be the second singular value of B. However, for this value of $\beta = 26.8$, it becomes the first singular value of B. This magic switching moment is depicted in the right plot, where one finds the singular values of the approximating matrix B during the iteration for this value of $\beta = 26.8$, which is the dashed line. It can be seen that β starts off being the second singular value, but then after approximately 100 iteration steps decides to become the first singular value.

It is an open problem to assess the importance of the fact that D_u and D_v (or D_x and D_y) are positive or nonnegative definite and what role this property plays in the convergence behavior of the algorithm. [If for instance the inverse weight matrix V contains negative elements, so that $\text{diag}(V \,\text{diag}(y)\, y)$ is indefinite, there is no convergence at all.] Other issues to be investigated are order and data recursive updating, for instance in the identification application and in the errors-in-variables variant of the Kalman filter.

A final observation to be explored concerns the resemblance of our generalized nonlinear SVD problem to the nonlinear eigenvalue problems in [26], where the "total least" problem is analyzed for other norms (e.g. total least absolute value).

REFERENCES

1 T. J. Abatzoglou, J. M. Mendel, and G. A. Harada, The constrained total least squares technique and its application to harmonic superresolution, *IEEE Trans. Signal Process.* SP-39(5):1070–1087 (May 1991).

2 R. J. Adcock, Note on the method of least squares, *The Analyst* IV(6):183–184 (Nov. 1877).

3 R. J. Adcock, A problem in least squares, *The Analyst* V(2):53–54 (Mar. 1878).

4 J. Cadzow, Signal enhancement: A composite property mapping algorithm, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-36:49–62 (1988).

5 J. A. Cadzow and D. M. Wilkes, Enhanced sinusoidal and exponential data modeling, in *SVD and Signal Processing, II: Algorithms, Analysis and Applications* (R. J. Vaccaro, Ed.), Elsevier Science, 1991, pp. 335–352.

6 G. Cybenko, On the eigenstructure of Toeplitz matrices, *IEEE Trans. Acoust. Speech and Signal Process.* ASSP-32(4):918–921 (Aug. 1984).

7 J. Demmel, The smallest perturbation of a submatrix which lowers the rank and constrained total least squares problems, *SIAM J. Numer. Anal.* 24(1):199–206 (1987).

8 B. De Moor and J. Vandewalle, An adaptive singular value decomposition algorithm based on generalized Chebyshev recursion, in *Proceedings of the Conference on Mathematics in Signal Processing*, Univ. of Bath, 17–19 Sept. 1985 (T. S. Durrani, J. B. Abbiss, J. E. Hudson, R. N. Madan, J. G. McWhirter, and T. A. Moore, Eds.), Clarendon, Oxford, 1987, pp. 607–635.

9 B. De Moor and G. Golub, The restricted singular value decomposition: Properties and Applications, *SIAM J. Matrix Anal. Appl.* 12:401–425 (1991).

10 B. De Moor, The Singular Value Decomposition and Long and Short Spaces of Noisy Matrices, ESAT-SISTA Report 1990-38, Dept. of Electrical Engineering, Katholieke Univ. Leuven, Belgium, Dec. 1990, 32 pp.; *IEEE Trans. Signal Process.*, Sept. 1993.

11 B. De Moor, P. Van Overschee, and G. Schelfhout, H_2 Model Reduction for SISO Systems, ESAT-SISTA Internal Report 1992-30, Dept. of Electrical Engineering, Katholieke Univ. Leuven, Belgium; to be presented at 12th IFAC World Congress, Sydney, Australia, 1993; also IMA Preprint Ser. 1035, Inst. for Mathematics and its Applications, Univ. of Minnesota, Sept. 1992.

12 C. Eckart and G. Young, The approximation of one matrix by another of lower rank, *Psychometrika* 1:211–218 (1936).

13 L. El Ghaoui, Fast computation of the largest stability radius for a two-parameter linear system, *IEEE Trans. Automat. Control* 37(7):1033–1037 (July 1992).

14 G. W. Fisher, Matrix analysis of metamorphic mineral assemblages and reactions, *Contrib. Mineral. and Petrol.* 102:69–77 (1989).

15 K. R. Gabriel and S. Zamir, Lower rank approximation of matrices by least squares with any choice of weights, *Technometrics* 21, No. 4 (Nov. 1979).

16 G. Golub, A. Hoffman, and G. Stewart, A generalization of the Eckart-Young-Mirsky approximation theorem, *Linear Algebra Appl.* 88/89:317–327 (1987).

17 G. H. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins U.P., Baltimore, 1989.

18 G. H. Golub and C. F. Van Loan, An analysis of the total least squares problem, *SIAM J. Numer. Anal.* 17, No. 6 (Dec. 1980).

19 B. Hanzon, The area enclosed by the (oriented) Nyquist diagram and the Hilbert-Schmidt-Hankel norm of a linear system, *IEEE Trans. Automat. Control* 37(6):835–839 (June 1992).

20 A. S. Householder and G. Young, Matrix approximation and latent roots, *Amer. Math. Monthly* 45:165–171 (1938).

21 S. Y. Kung, A new identification and model reduction algorithm via singular value decomposition, in *Proceedings of the 12th Asilomar Conference on Circuits, Systems and Computers*, Pacific Grove, Calif., 1978, pp. 705–714.

22 M. Moonen, B. De Moor, L. Vandenberghe and J. Vandewalle, On- and off-line identification of linear state space models, *Internat. J. Control* 49(1):219–239 (1990).

23 K. Pearson, On lines and planes of closest fit to systems of points in space, *Philos. Mag 6th ser.* 2:559–572.

24 S. VanHuffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, Frontiers Appl. Math. 9, SIAM, Philadelphia, 1991, 300 pp.

25 P. Van Overschee and B. De Moor, N4SID: Subspace Algorithms for the Identification of Combined Deterministic-Stochastic Systems, ESAT-SISTA Report 1992-34, Dept. of Electrical Engineering, Katholieke Univ. Leuven, Belgium, Feb. 1992; *Automatica*, special issue on Statistical Signal Processing and Control, to appear.

26 G. A. Watson, On a class of algorithms for total approximation, *J. Approx. Theory* 45(3):219–231 (Nov. 1985).

27 G. Young, Matrix approximation and subspace fitting, *Psychometrika* 2(1):21–25 (Mar. 1937).

Received 25 August 1992; final manuscript accepted 20 February 1993