REDUCED ORDER IDENTIFICATION: PARAMETER ERROR, CONVERGENCE AND ROBUSTNESS FOR LEAST SQUARES ALGORITHMS

by

Robert J. Minnichelli

Memorandum No. UCB/ERL M87/31

20 May 1987

REDUCED ORDER IDENTIFICATION: PARAMETER ERROR, CONVERGENCE AND ROBUSTNESS FOR LEAST SQUARES ALGORITHMS

by

Robert J. Minnichelli

Memorandum No. UCB/ERL M87/31

20 May 1987

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

REDUCED ORDER IDENTIFICATION: PARAMETER ERROR, CONVERGENCE AND ROBUSTNESS FOR LEAST SQUARES ALGORITHMS

by

Robert J. Minnichelli

Memorandum No. UCB/ERL M87/31

20 May 1987

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

REDUCED ORDER IDENTIFICATION: PARAMETER ERROR, CONVERGENCE AND ROBUSTNESS FOR LEAST SQUARES ALGORITHMS

by

Robert J. Minnichelli

Department of Electrical Engineering
and Computer Science
University of California, Berkeley

ABSTRACT

We consider the implementation of reduced order identifiers to incorporate prior information for DARMA and ARX systems. A bound is given on the ratio between the reduced order parameter error and the full order error. This guarantees that the reduced order algorithm converges at least as fast as the full order algorithm. The algorithm is shown to be robust to improper parameterization (incorrect prior knowlede) with practical error bounds. We extend this result to obtain a new error bound for robustness to unmodelled dynamics for both full order and reduced order algorithms.

Research supported by the Aerospace Corporation, El Segundo, CA 90245

1. INTRODUCTION

We consider the finite dimensional SISO discrete time difference equation of Eq. 2.1. The standard (ARX or DARMA) identification problem requires estimating the coefficients of this equation based on a sample of inputs and outputs. We will be concerned with the implementation of *a priori* constraints on these parameters. In particular, we want to implement a set of affine constraints while taking advantage of the reduction of the dimension of the allowable parameter space. Typically, this situation occurs in systems comprised of known subsystems with unknown interconnection gains and in systems with a few unmodelled (or incompletely modelled) components.

The contributions of this paper are all with respect to least squares identification algorithms. However, the reduced order methodology is fairly independent of the particular algorithm being used, since it simply involves a reparameterization of the original difference equation parameters and the application of the (arbitrary) identification algorithm to the new (and smaller) set of parameters. The notable exceptions to this are fast gain and lattice algorithms, which utilize "all" of the original difference equation structure (see Sec. 2.8). The general reduced parameter structure was first consider by Dasgupta, Anderson and Kaye [8], and more thoroughly by Dasgupta [7]. Further investigations into the general algorithm have been made by Bai and Sastry [2] and Mason and Sastry [13]. All of these authors considered both continuous time and discrete time systems with a variety of identification algorithms. Independent work by Clary [6] concentrated on the least squares algorithm for discrete time systems (as do we) while restricting attention primarily to the Filtered Least Squares algorithm for the special case of *a priori* known poles and zeros (see Sec. 2.6). (None-the-less, when discussing the *interpretation* of the Filtered Least Squares algorithm, as in Sec. 2.7 of this paper, Clary develops the more general structure as considered here and in [2], [7], [8] and [13].) Dasgupta, Anderson and Kaye [8] also consider extensions to bilinear parameter dependence.

Although these papers are fairly recent, the underlying concepts of parameter space reduction via filtering seems to be present throughout the adaptive identification and control literature. For example, in his survey on adaptive control [1], Åström discusses an example which requires the identification of polynomials in the following equation:

$$A_m A_0 = B^-[Ru + Sy]$$

The polynomials $A_m(z^{-1})$ and $A_0(z^{-1})$ are assumed to be known (in fact, they are user selected in this adaptive control design). $B^-(z^{-1})$ contains the non-minimum phase zeros of the plant, and $R(z^{-1})$ and $S(z^{-1})$ are controller parameters that need to be estimated. When B^- is not characterized *a priori*, we are left with a bilinear identification problem. However, when B^- is assumed known, the problem reduces to a linear identification in terms of the 'filtered' signals B^-u and

B^-y. In either case, we also use the filtered output $A_m A_0 y$. This separation of the B^- dynamics from R and S and the use of filtered inputs and outputs essentially incorporates all of the features of the more general parameter reduction procedure.

As the title suggests, we shall establish some results on the convergence and robustness properties of the reduced order algorithms. As the reduced order least squares identification algorithm really does solve a least squares problem, just as the full order algorithm does, all the convergence and robustness properties present for the full order algorithms correspond to analogous properties for the reduced order algorithms. However, these properties will have conditions which, though analogous to the full order conditions, are not necessarily equivalent. For example, under suitable conditions, the full order parameter estimates will converge to the true parameter if the regression vectors are persistently exciting. Under analogous conditions, the reduced order parameter estimates will converge to the true parameter if the *filtered* regression vectors (or pseudo-regression vectors) are persistently exciting. Thus it seems possible to have general situations under which the full order algorithm converges while the corresponding reduced order algorithm does not, unless we can show that persistent excitation of the full order regression vector implies persistent excitation of the filtered regression vector (this is established in Prop. 3.3).

The first objective of this paper is to establish *comparative* convergence results for the reduced order estimation to complement the *analogous* convergence results. That is, we would like to say that the reduced order algorithm converges "faster than" the full order algorithm. In fact, our main result on parameter errors states that the ratio between the reduced order parameter error and the full order error can be bounded above for all time by a persistent excitation ratio (Eq. 3.26). It cannot be bounded below.

The second objective is to establish robustness of the algorithm to both unmodelled dynamics and improper (incorrect) parameterization (or constraint). In fact, the robustness of the algorithm was exhibited by Clary [6]. We derive a new robustness result which is more direct and produces practical error bounds.

In section 4.2, we show that unmodelled dynamics is a special case of improper parameterization. This leads to a new robustness result for full order identifiers and a new bound on the error of the so-called "tuned value."

2. PRELIMINARY RESULTS AND NOTATION

2.1 System Description

We begin with a finite dimensional SISO discrete-time input/output description

$$y_k \approx -a_1 y_{k-1} - \cdots - a_n y_{k-n} + b_1 u_{k-1} + \cdots + b_n u_{k-n} \tag{2.1}$$

The approximate equality implies the possible presence of modelling errors, system disturbances, measurement noise, etc. This will be made more precise in later sections. We can write the right side of Eq. 2.1 in vector form by defining

$$\theta := (a_1 \cdots a_n \; b_1 \cdots b_n)^T \in \mathbb{R}^{2n}$$

$$\phi_k := (-y_{k-1} \cdots -y_{k-n} \; u_{k-1} \cdots u_{k-n})^T \in \mathbb{R}^{2n}$$

Then Eq. 1 becomes

$$y_k \approx \phi_k^T \theta \tag{2.2}$$

2.2 Least Squares Identification

The fixed order identification problem is to determine an estimate for θ in Eq. 2.2, $\hat{\theta}_k$, based on the data available at time k, namely $\{y_s, u_s, s \le k\}$. Here we have restricted our attention to strictly proper systems so that u_k should not affect our estimate of y_k. For purposes of adaptive control, this allows u_k to be determined by the controller *after* the estimate $\hat{\theta}_k$ is formed. The data $\{y_s, u_{s-1}, s \le k\}$ is equivalent to the data $\{y_s, \phi_s, s \le k\}$. We define

$$Y_k := (y_k \; y_{k-1} \cdots y_1)^T \in \mathbb{R}^k$$

$$\Phi_k := (\phi_k \; \phi_{k-1} \cdots \phi_1) \in \mathbb{R}^{2n \times k}$$

Then Eq. 2.2, applied to all the data available at time k, yields[1]

$$Y_k \approx \Phi_k^T \theta \tag{2.3}$$

The least squares estimate for θ at time k based on Eq. 2.3 minimizes

$$J(\hat{\theta}_k) := \| Y_k - \Phi_k^T \hat{\theta}_k \|_2^2 = \sum_{s=1}^{k} (y_s - \phi_s^T \hat{\theta}_k)^2 \tag{2.4}$$

and is given by

$$\hat{\theta}_k = (\Phi_k \Phi_k^T)^{-1} \Phi_k Y_k = (\sum_{s=1}^{k} \phi_s \phi_s^T)^{-1} (\sum_{s=1}^{k} \phi_s y_s) \tag{2.5}$$

[1] Here we assume implicitly that $y_s = u_{s-1} = 0$ for $s \le 0$.

We make Eq. 2.5 recursive by defining

$$M_k := \sum_{s=1}^{k} \phi_s \phi_s^T \in \mathbb{R}^{2n \times 2n}$$

$$\xi_k := \sum_{s=1}^{k} \phi_s y_s \in \mathbb{R}^{2n}$$

Then Eq. 2.5 becomes

$$M_k = M_{k-1} + \phi_k \phi_k^T \tag{2.6}$$

$$\xi_k = \xi_{k-1} + \phi_k y_k \tag{2.7}$$

$$\hat{\theta}_k = M_k^{-1} \xi_k \tag{2.8}$$

In the exact least squares algorithm we start with $M_0=0$ and $\xi_0=0$. We cannot determine an estimate $\hat{\theta}_k$ until M_k becomes nonsingular. In particular, we need $k \geq 2n$.

In the least squares with initial estimate algorithm, we assume an *a priori* estimate $\hat{\theta}_0$ and choose a (small) positive definite matrix M_0. Setting $\xi_0 = M_0 \hat{\theta}_0$, the algorithm (2.6)-(2.8) is defined for all k and minimizes

$$J(\hat{\theta}_k) := (\hat{\theta}_k - \hat{\theta}_0)^T M_0 (\hat{\theta}_k - \hat{\theta}_0) + \sum_{s=1}^{k} (y_s - \phi_s^T \hat{\theta}_k)^2 \tag{2.9}$$

For both algorithms we should determine $\hat{\theta}_k$ through Gaussian elimination rather than inverting M_k. The algorithm (2.6)-(2.8) has $o(n^3)$ computation requirements.

We can reduce the computational requirements to $o(n^2)$ by noticing that M_k is updated in Eq. 2.6 by adding a dyad to M_{k-1}. This allows $P_k := M_k^{-1}$ to be updated directly without any matrix inversions. Eqs. 2.6-2.8 become

$$\hat{\theta}_k = \hat{\theta}_{k-1} + \frac{P_{k-1}\phi_k}{1 + \phi_k^T P_{k-1} \phi_k} [y_k - \phi_k^T \hat{\theta}_{k-1}] \tag{2.10}$$

$$P_k = P_{k-1} - \frac{P_{k-1}\phi_k \phi_k^T P_{k-1}}{1 + \phi_k^T P_{k-1} \phi_k} \tag{2.11}$$

$$P_0 = M_0^{-1} \tag{2.12}$$

Many sources are available for a more detailed derivation of the least squares algorithm, along with a more complete discussion of the many implementation issues and the selection of M_0. Ljung and Söderström [12] is suggested. See also Section 2.8 for further discussion.

2.3 Reduced Order Identification

We now turn our attention toward incorporating *a priori* knowledge of the system into the identification algorithm. To some extent, we were able to do this by choosing $\hat{\theta}_0$. We saw that M_0 could reflect our confidence in $\hat{\theta}_0$. So if M_0 is 'very large' in some direction (equivalently, if P_0 is very small), the orthogonal projection of $\hat{\theta}_k$ onto that direction will remain essentially fixed for a 'long time' at its initial value. We wish to address this problem more formally.

Suppose that θ is not completely undetermined, but instead is known up to a set of m unknown parameters $v_1, \ldots, v_m$. Further suppose θ depends linearly on the parameters. In other words, suppose θ is restricted to an m-dimensional affine sub-space $\Theta \subset \mathbb{R}^{2n}$.

$$\theta = \theta_0 + \theta_1 v_1 + \cdots + \theta_m v_m = \theta_0 + D v \tag{2.13}$$

where $v \in \mathbb{R}^m$ is the (unknown) parameter vector and $D \in \mathbb{R}^{2n \times m}$ is full column rank (injective). (If D is not full column rank, the vectors $\theta_1, \ldots, \theta_m$ are not linearly independent, and some of the parameters v_j are redundant and can be eliminated to obtain a smaller parameterization.) We assume, of course, that θ_0 and D are known *a priori*.

We re-write Eq. 2.2

$$y_k \simeq \phi_k^T \theta = \phi_k^T \theta_0 + \phi_k^T D v \tag{2.14}$$

We define a new 'pseudo-output' and 'pseudo-regression' vector by

$$z_k := y_k - \phi_k^T \theta_0 \in \mathbb{R}$$

$$\psi_k := D^T \phi_k \in \mathbb{R}^m$$

Then Eq. 2.14 becomes

$$z_k \simeq \psi_k^T v \tag{2.15}$$

Comparing Eq. 2.15 with Eq. 2.2, we derive the least squares estimate of v based on Eq. 2.15 immediately.

$$\Psi_k := (\psi_k \cdots \psi_1) \in \mathbb{R}^{m \times k}$$

$$Z_k := (z_k \cdots z_1)^T \in \mathbb{R}^k$$

$$N_k := \Psi_k \Psi_k^T = \sum_{s=1}^{k} \psi_s \psi_s^T \in \mathbb{R}^{m \times m}$$

$$\zeta_k := \Psi_k Z_k = \sum_{s=1}^{k} \psi_s z_s \in \mathbb{R}^m$$

$$N_k = N_{k-1} + \psi_k \psi_k^T \tag{2.16}$$

$$\zeta_k = \zeta_{k-1} + \psi_k z_k \tag{2.17}$$

$$\hat{v}_k = N_k^{-1}\zeta_k \tag{2.18}$$

where the least squares estimate $\hat{v}_k$ minimizes the cost

$$J(\hat{v}_k) = \|Z_k - \Psi_k^T \hat{v}_k\|_2^2 = \sum_{s=1}^{k} (z_s - \psi_s^T \hat{v}_k)^2 \tag{2.19}$$

Note that

$$Z_k - \Psi_k^T v = Y_k - \Phi_k^T \theta \tag{2.20}$$

so that the minimization in Eq. 2.19 is the same as the minimization in Eq. 2.4 with $\hat{\theta}_k$ restricted to $\theta_0 + \mathrm{Im}(D)$. As before, Eq. 2.16-2.18 can be modified to eliminate the matrix inversion.

$$\hat{v}_k = \hat{v}_{k-1} + \frac{Q_{k-1}\psi_k}{1 + \psi_k^T Q_{k-1}\psi_k}[z_k - \psi_k^T \hat{v}_{k-1}] \tag{2.21}$$

$$Q_k = Q_{k-1} - \frac{Q_{k-1}\psi_k \psi_k^T Q_{k-1}}{1 + \psi_k^T Q_{k-1}\psi_k} \tag{2.22}$$

$$Q_0 = N_0^{-1} \tag{2.23}$$

If a desirable M_0 and $\theta_0 \in \Theta$ have already been determined, we choose N_0 and $\hat{v}_0$ so that

$$N_0 = D^T M_0 D \tag{2.23a}$$

$$\hat{\theta}_0 = \theta_0 + D\,\hat{v}_0 \tag{2.23b}$$

2.4 Error Model

We make the approximate equality in Eqs. 2.1, 2.2 and 2.15 precise by defining the 'error'

$$e_k := y_k - \phi_k^T \theta = z_k - \psi_k^T v$$

Eq. 2.2 and Eq. 2.15 become

$$y_k = \phi_k^T \theta + e_k \tag{2.24}$$

$$z_k = \psi_k^T v + e_k \tag{2.25}$$

We consider two types of models in this paper:

(1) DARMA (Deterministic Auto-Regressive, Moving Average exogenous input): $e_k \equiv 0$.

(2) ARX ((stochastic) Auto-Regressive with eXogenous Input): e_k is a 'white noise' process (this will be made precise in later sections).

2.5 Transfer Function Formulation of Reduced Order Identification

The transfer function description of the n^{th} order ARX model is

$$Y(z) = \frac{n(z^{-1})}{d(z^{-1})}U(z) + \frac{1}{d(z^{-1})}E(z) \tag{2.26}$$

where $n(\cdot)$ and $d(\cdot)$ are polynomials of degree n and the constant terms are zero and one, respectively.[2] In terms of the coefficients defined in Eq. 2.1, we have

$$n(z^{-1}) = b_1 z^{-1} + \cdots + b_n z^{-n} \tag{2.27}$$

$$d(z^{-1}) = 1 + a_1 z^{-1} + \cdots + a_n z^{-n} \tag{2.28}$$

Now suppose that $n(\cdot)$ and $d(\cdot)$ are parameterized by $\nu_1, \ldots, \nu_m$ as follows:

$$\frac{n(z^{-1})}{d(z^{-1})} = \frac{n_0(z^{-1}) + \sum_{j=1}^{m} \nu_j n_j(z^{-1})}{d_0(z^{-1}) + \sum_{j=1}^{m} \nu_j d_j(z^{-1})} \tag{2.29}$$

where d_j, n_j are known polynomials in z^{-1}:

$$d_0(z^{-1}) = 1 + a_1^0 z^{-1} + \cdots + a_n^0 z^{-n}$$

$$d_j(z^{-1}) = a_1^j z^{-1} + \cdots + a_n^j z^{-n} \qquad j = 1, \ldots, m$$

$$n_j(z^{-1}) = b_1^j z^{-1} + \cdots + b_n^j z^{-n} \qquad j = 0, \ldots, m$$

Now we form the vectors:

$$\theta_j := [\, a_1^j \cdots a_n^j \; b_1^j \cdots b_n^j \,]^T \in \mathbb{R}^{2n} \qquad j = 0, \ldots, m$$

$$\theta := [\, a_1 \cdots a_n \; b_1 \cdots b_n \,]^T \in \mathbb{R}^{2n}$$

and we get

$$\theta = \theta_0 + \nu_1 \theta_1 + \cdots + \nu_m \theta_m \tag{2.30}$$

as in Eq. 2.13.

Conclusion. We see that the affine polynomial parameterization of the transfer function in Eq. 2.29 corresponds precisely (i.e. with identical parameters) to the affine parameterization of the ARX coefficients in Eq. 2.30.

---

[2] Alternately, we could consider $\bar{n}(z) = z^n n(z^{-1})$ and $\bar{d}(z) = z^n d(z^{-1})$. Then $\bar{n}(\cdot)$ is of degree n-1 and $\bar{d}(\cdot)$ is monic of degree n.

2.6 Example: Known Poles and Zeroes

We consider an ARX model with some known dynamics characterized by r poles $\{p_j\}_{j=1}^{r}$ and s zeroes $\{\zeta_j\}_{j=1}^{s}$. We define the 'known' polynomials:

$$d^k(z^{-1}) := (1-p_1z^{-1})(1-p_2z^{-1})\cdots(1-p_rz^{-1}) = 1 + a_1^kz^{-1} + \cdots + a_r^kz^{-r}$$

$$n^k(z^{-1}) := (1-\zeta_1z^{-1})(1-\zeta_2z^{-1})\cdots(1-\zeta_sz^{-1}) = 1 + b_1^kz^{-1} + \cdots + b_s^kz^{-s}$$

Next, we define 'unknown' polynomials which contain the remaining dynamics.

$$d^u(z^{-1}) := 1 + a_1^uz^{-1} + \cdots + a_{n-r}^uz^{-n+r}$$

$$n^u(z^{-1}) := b_1^uz^{-1} + \cdots + b_{n-s}^uz^{-n+s}$$

Our unknown parameter vector consists of the $2n-r-s$ coefficients of $d^u(\cdot)$ and $n^u(\cdot)$, and the combined dynamics are given by Eqs. 2.31 and 2.32.

$$d(z^{-1}) = d^u(z^{-1})d^k(z^{-1}) = d^k(z^{-1}) + \sum_{j=1}^{n-r} a_j^u(z^{-j}d^k(z^{-1})) \tag{2.31}$$

$$n(z^{-1}) = n^u(z^{-1})n^k(z^{-1}) = \sum_{j=1}^{n-s} b_j^u(z^{-j}n^k(z^{-1})) \tag{2.32}$$

Conclusion. A collection of known poles and known zeroes of a transfer function does indeed induce an affine parameterization of the transfer function polynomials. This parameterization is in terms of the *coefficients* of the unknown polynomials.

We now show that the implementation of the identification algorithm for the parameterization shown above can be reduced to a the particularly simple and attractive structure shown in Fig. 1. Using previous notation, we would implement the identification algorithms for $a_1^u, \ldots, a_{n-r}^u, b_1^u, \ldots, b_{n-s}^u$ by forming $m = 2n-r-s$ linear filters to generate ψ_k.

$$\Psi_j(z) = \begin{cases} -z^{-j}d^k(z^{-1})Y(z) & j=1,\ldots,n-r \\ z^{j-n+q}n^k(z^{-1})U(z) & j=n-s+1,\ldots,m \end{cases} \tag{2.33}$$

We need one additional filter to generate z_k:

$$Z(z) = d^k(z^{-1})Y(z) \tag{2.34}$$

The example of known poles and zeroes has a special property that the filters have many common elements. $Y(z)$ is always filtered by $d^k(z^{-1})$ and $U(z)$ is always filtered by $n^k(z^{-1})$. To reduce computation, we can filter these signals first, then apply the simple delay operators that constitute the remaining filter dynamics. In fact, when we do this, we see that the remaining filtering and identification is just the usual (full order) identification algorithm for $2n-r-s$ parameters. Combining the filtering and the usual least-squares is called the filtered least squares algorithm (introduced by Clary in [6]) and is illustrated in Fig. 1. Notice that the noise input e_k is transformed

properly to the filtered identification.

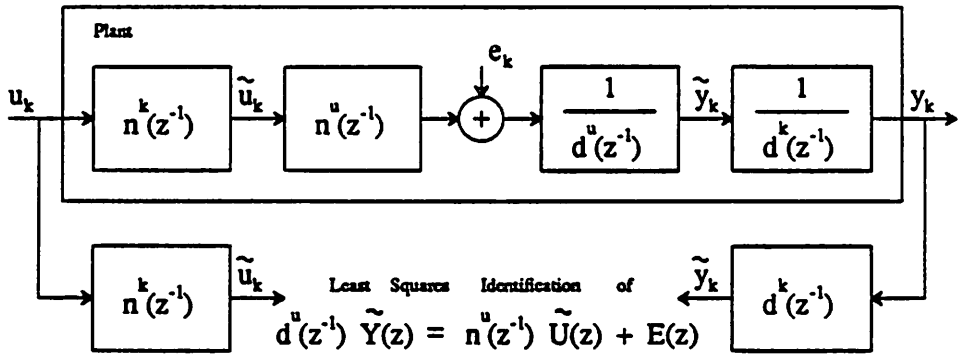

Figure 1.
Schematic Diagram of the Filtered Least Squares Algorithm

2.7 Interpretation of Reduced Order Identification

We intend to show that the reduced order identification yields precisely the same estimate as the full order identification algorithm without matrix inverse (Eqs. 2.10-2.12) if P_0 and $\hat{\theta}_0$ satisfy:

$$\mathrm{Im}(P_0) \subset \mathrm{Im}(D) \tag{2.37}$$

$$\hat{\theta}_0 \in \Theta \tag{2.38}$$

Intuitively, making P_0 singular in certain directions is like letting M_0 become unbounded in certain directions. Thus 'infinite' weight is placed on certain linear combinations of $\hat{\theta}_k - \hat{\theta}_0$ in the cost function of Eq. 2.9, resulting in an estimate which satisfies

$$\hat{\theta}_k - \hat{\theta}_0 \in \mathrm{Im}(D) \tag{2.39}$$

The feasibility of the argument lies in two facts which follow by inspection of Eqs. 2.10 and 2.11.

(1) Eq. 2.11 implies $\mathrm{Im}(P_{k-1}) \subset \mathrm{Im}(D) \Rightarrow \mathrm{Im}(P_k) \subset \mathrm{Im}(D)$.

(2) Eq. 2.10 implies $\hat{\theta}_k - \hat{\theta}_{k-1} \in \mathrm{Im}(P_{k-1})$.

So we see that Eqs. 2.37 and 2.38 do guarantee that $\hat{\theta}_k$ remains in Θ for all k. The proposition below shows that the processes themselves are identical. In the sequel, we use $\ddot{\theta}_k$ to denote the parameter estimate that results from the reduced order estimate $\hat{v}_k$:[3]

[3] We shall never use the dot notation for time derivatives in this paper, so $\ddot{\theta}_k$ should not create any confusion.

$$\ddot{\theta}_k := \theta_0 + D\hat{v}_k$$

Proposition: Consider the algorithm (2.21)-(2.22), with $\ddot{\theta}_k := \theta_0 + D\hat{v}_k$. Let $P_0 = DQ_0D^T$, $\hat{\theta}_0 = \theta_0 + D\hat{v}_0$, and consider the algorithm (2.10)-(2.11).[4] Then $\ddot{\theta}_k = \hat{\theta}_k$ for all k.

Proof: First, we show by induction that $P_k = DQ_kD^T$ for all $k \geq 0$:

$$P_{k-1} = DQ_{k-1}D^T \Rightarrow P_k = P_{k-1} - \frac{P_{k-1}\phi_k\phi_k^T P_{k-1}}{1+\phi_k^T P_{k-1}\phi_k}$$

$$= DQ_{k-1}D^T - \frac{DQ_{k-1}(D^T\phi_k)(\phi_k^T D)Q_{k-1}D^T}{1+(\phi_k^T D)Q_{k-1}(D^T\phi_k^T)}$$

$$= D\left[Q_{k-1} - \frac{Q_{k-1}\psi_k\psi_k^T Q_{k-1}}{1+\psi_k^T Q_{k-1}\psi_k}\right]D^T$$

$$= DQ_kD^T$$

Another inductive argument establishes the result:

$$\hat{\theta}_{k-1} = \theta_0 + D\hat{v}_{k-1} \Rightarrow \hat{\theta}_k = \hat{\theta}_{k-1} + P_k\phi_k[y_k - \phi_k^T\hat{\theta}_{k-1}]$$

$$= \theta_0 + D\hat{v}_{k-1} + DQ_k(D^T\phi_k)[y_k - \phi_k^T\theta_0 - \phi_k^T D\hat{v}_{k-1}]$$

$$= \theta_0 + D[\hat{v}_{k-1} + Q_k\psi_k(z_k - \psi_k^T\hat{v}_{k-1})]$$

$$= \theta_0 + D\hat{v}_k \qquad \square$$

The equivalence between reduced order identification and the 'singular' full order identification (with $P_0 = DQ_0D^T$) is presented for conceptual purposes only. The singular identification should not be implemented, as:

(1) it does not retain the benefit of reduced computation; and

(2) it is numerically ill-conditioned, and as soon as the eigenvalues of P_k wander away from zero, the constraint $\hat{\theta}_k \in \Theta$ may be violated.

[4] Here 'algorithm' refers only to the equations 2.10 and 2.11 applied 'blindly.' The algorithm doesn't really apply, since P_0 is not positive definite.

2.8 Fast Gain Algorithms and Lattice Filters

We have emphasized that the reduced order algorithm solves a least squares problem analogous (in fact, equivalent) to the full order algorithm, and so inherits all of the analogous properties. However, the full order algorithm has more structure than is required to apply a least squares criterion, and more structure than we used in Section 2.2. This additional structure is *not* inherited by the reduced order algorithms, and the structure allows the full order algorithm (2.10)-(2.12) to be reduced to an $o(n)$ process (linear computational growth). We conclude that for large systems with only a few constraints (i.e. when the reduced order system is still large), it may cost less to implement the full order algorithm.

To demonstrate the additional structure we consider the simplified auto-regressive case (no exogenous input). Then

$$\theta := (a_1 \cdots a_n)^T$$

$$\phi_k := (-y_{k-1} \cdots -y_{k-n})^T$$

$$M_k := \sum_{s=1}^{k} \phi_s \phi_s^T$$

$$\xi_k := \sum_{s=1}^{k} \phi_s y_s$$

and M_k and ϕ_k have the following "shift" structure:

$$M_k = \begin{bmatrix} \sum_{s=1}^{k} y_{s-1}y_{s-1} & \sum_{s=1}^{k} y_{s-1}y_{s-2} & \cdots & \sum_{s=1}^{k} y_{s-1}y_{s-n} \\ \sum_{s=1}^{k} y_{s-2}y_{s-1} & \sum_{s=1}^{k-1} y_{s-1}y_{s-1} & \cdots & \sum_{s=1}^{k-1} y_{s-1}y_{s-n} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{s=1}^{k} y_{s-n}y_{s-1} & \cdots & \sum_{s=1}^{k} y_{s-n}y_{s-n} & \\ & \sum_{s=1}^{k-1} y_{s-n}y_{s-1} & \cdots & \sum_{s=1}^{k-1} y_{s-n}y_{s-n} \end{bmatrix} = M_{k-1}$$

$$\phi_k = \begin{bmatrix} y_{k-1} \\ y_{k-2} \\ \vdots \\ y_{k-n} \\ y_{k-n-1} \end{bmatrix} = \phi_{k-1}$$

For the full ARX or DARMA cases M_k and ϕ_k will have a "block shift" structure. In either case, such systems can be recursively solved with linear computational growth in n using methods related to the solutions of Toeplitz systems. It is easily seen that upon reparameterization, the reduced order identifier may lose this valuable property. We note that the Filtered Least Squares example of Section 2.6 *does* inherit the shift structure.

We will not discuss the fast gain and lattice algorithms any further. The derivations are quite involved. An adequate discussion of these algorithms is given by Ljung and Söderström [12], with additional references given there.

3. PARAMETER ERROR AND CONVERGENCE

In this section, we will review a few properties of identifiability and parameter convergence for least squares algorithms. We will then compare convergence for reduced order identification against convergence of full order schemes. In Section 4 we will give a robustness result for the case when the *a priori* information is inaccurate.

As an illustration of the type of result we'll be studying, we first consider the case of the *exact* least squares algorithm (Eqs. 2.6-2.8) applied to a *deterministic* system. This rather simple analysis also illustrates the intuition behind the claim that reduced order identifiers might in general have faster convergence. Suppose that $\{\phi_1, \ldots, \phi_{2n}\}$ are linearly independent vectors. Then the exact least squares estimate $\hat{\theta}_k$ is defined and is exactly equal to θ for all $k \geq 2n$; that is, the algorithm converges exactly in $2n$ steps. With the reduced order identification under analogous conditions, $\ddot{\theta}_k$ converges exactly to θ in m steps (where $m < n$). The resulting improvement and the underlying mechanism are very clear in this case.

In more general situations, it will be difficult to guarantee *improvement* in performance, but at the very least we would like to show that no performance is lost by implementing a reduced order algorithm.

3.1 Identifiability

Consider the deterministic, zero-initial-state system with input/output maps:

$$h(\theta) : \{u_k\}_{k=0}^{\infty} \mapsto \{y_k\}_{k=0}^{\infty} \tag{3.1}$$

We think of h as a map from the parameter θ to the space of input/output maps. Then we say that $\theta \in \mathbb{R}^{2n}$ is identifiable if for all $\theta' \in \mathbb{R}^{2n}$

$$h(\theta) = h(\theta') \Rightarrow \theta = \theta' \tag{3.2}$$

Likewise, we may consider the input/output maps as a function of the reduced parameter:

$$g(v) : \{u_k\}_{k=0}^{\infty} \mapsto \{y_k\}_{k=0}^{\infty} \tag{3.3}$$

with, of course,

$$g(v) = h(\theta_0 + D v). \tag{3.4}$$

Then $v \in \mathbb{R}^m$ is identifiable if for all $v' \in \mathbb{R}^m$

$$g(v) = g(v') \Rightarrow v = v' \tag{3.5}$$

First we note that θ is identifiable if and only if the corresponding n^{th} order transfer function has no pole-zero cancellations. The following result was first shown by Bai and Sastry [2].

Theorem 3.1 Suppose $\theta = \theta_0 + D\nu$, and θ is identifiable. Then ν is identifiable if and only if D is injective.

Proof: $\Rightarrow$: Suppose D is not injective. Let $\bar{\nu} \in \mathrm{Ker}(D)$. Then $\nu + \bar{\nu}$ corresponds to the same parameter vector θ as ν (i.e. $\theta_0 + D\nu = \theta_0 + D(\nu + \bar{\nu}) = \theta$), and thus the same input/output map.

$\Leftarrow$: Suppose D is injective, and $\nu \neq \nu'$ have the same input/output map. Then $\theta_0 + D\nu \neq \theta_0 + D\nu'$ while $\theta = \theta_0 + D\nu$ and $\theta_0 + D\nu'$ have the same input/output map. This contradicts the identifiability of θ. ⊠

We conclude that identifiability cannot be hindered by using *a priori* information in the manner proposed.

Remark. We have shown that ν is identifiable whenever θ is (since D was required to be injective). However, ν may be identifiable even when θ is not, as in Example 3.2. The identifiability situation is thus improved by incorporating prior information using the reduced order algorithm.

Example 3.2 Consider the transfer function

$$h(z^{-1}) = \frac{1+.5z^{-1}}{(1+.3z^{-1})(1+\nu z^{-1})}$$

Then ν is always identifiable, but θ is identifiable if and only if $\nu \neq .5$.

Remark. Identifiability simply insures that there is a unique parameter to which, under 'some ideal conditions,' (usually persistent excitation) we might expect $\hat{\theta}_k$ to converge. In the absence of identifiability, we may still be able to obtain convergence to a submanifold of parameter values with identical input/output properties. We may speculate that when input/output properties are the primary object of interest (as is the case with many controller designs, for example) that this may suffice, depending on 'how' $\hat{\theta}_k$ converges to the submanifold. Mason has studied this type of 'convergence' in [13].

Remark. For the ARX case we have two input/output maps to consider: $u \rightarrow y$ and $e \rightarrow y$. The ARX structure requires the noise/output map to be all-pass (no zeros). The denominator of the noise/output transfer function is equal to the denominator of the input/output transfer function (see Eq. 2.26). The denominator is always identifiable from the $e \rightarrow y$ map. Knowing the denominator, the numerator is always identifiable from the $u \rightarrow y$ map (i.e. pole/zero cancellations can be accurately detected). We conclude that full order ARX systems are always identifiable; reduced order systems are identifiable if and only if D is injective.

3.2 Convergence of Least Squares Algorithms

We now return to the full order least squares identification problem, with notation as defined in Section 2.2. A standard result of least squares identification gives sufficient conditions for parameter convergence (see, for example, Kumar and Varaiya [11], p. 203).

Theorem 3.2 Suppose $\{e_k\}_{k=1}^{\infty}$ is a sequence of uncorrelated zero-mean random variables with bounded second mean. Then $\hat{\theta}_k \xrightarrow{a.s.} \theta$ on the set of ω for which:

$$\text{(a)} \quad \sigma_{\min} M_k \to +\infty$$

$$\text{(b)} \quad \limsup_{k\to\infty} \frac{\sigma_{\max} M_k}{\sigma_{\min} M_k} < \infty$$

Convergence becomes uniform in time if we strengthen conditions (a) and (b) to a persistent excitation condition:

Definition: The deterministic sequence $\{\phi_k\}_{k=0}^{\infty}$ is persistently exciting (p.e.) if there are positive real constants α_1, α_2 and a positive integer l such that for any $k \in \mathbb{N}$

$$\alpha_1 I < \sum_{j=k}^{k+l} \phi_j \phi_j^T < \alpha_2 I \tag{3.6}$$

Clearly a persistently exciting signal satisfies conditions (a) and (b) in Theorem 3.1, since

$$\sum_{j=k}^{k+l} \phi_j \phi_j^T = M_{k+l} - M_k \tag{3.7}$$

The set of ω for which $\phi_k(\omega)$ is persistently exciting is denoted Ω_{pe}^{ϕ}.

We now determine the parameter error by analyzing $\hat{\theta}_k$ using the fact that $y_k = \phi_k^T \theta + e_k$.

$$\begin{aligned}
\hat{\theta}_k &= M_k^{-1}(\zeta_0 + \sum_{s=1}^{k} \phi_s y_s) \\
&= M_k^{-1}(\zeta_0 + \sum_{s=1}^{k} \phi_s \phi_s^T \theta + \sum_{s=1}^{k} \phi_s e_s) \\
&= M_k^{-1}(M_0 \hat{\theta}_0 + (M_k - M_0)\theta + M_k^{-1} \sum_{s=1}^{k} \phi_s e_s) \\
&= \theta + M_k^{-1} M_0 (\hat{\theta}_0 - \theta) + M_k^{-1} \sum_{s=1}^{k} \phi_s e_s
\end{aligned} \tag{3.8}$$

We define the parameter errors:

$$\underset{\sim}{\theta}_k := \hat{\theta}_k - \theta = \underset{\sim}{\theta}_k^{ic} + \underset{\sim}{\theta}_k^{e}$$

$$\underset{\sim}{\theta}_k^{ic} := M_k^{-1} M_0 (\hat{\theta}_0 - \theta)$$

$$\underset{\wedge}{\theta}_k^e := M_k^{-1} \sum_{s=1}^{k} \phi_s e_s$$

with $\underset{\wedge}{y}_k$, $\underset{\wedge}{y}_k^{ic}$, $\underset{\wedge}{y}_k^e$, $\underset{\cdot\cdot}{\theta}_k$, $\underset{\cdot\cdot}{\theta}_k^{ic}$ and $\underset{\cdot\cdot}{\theta}_k^e$ defined analogously.

The persistent excitation condition immediately tells us that $\underset{\wedge}{\theta}_k^{ic}$ decreases deterministically as $1/n$. To analyze $\underset{\wedge}{\theta}_k^e$ we use stochastic convergence theory.

First we make some assumptions about the error sequence $\{e_k\}_{k=0}^{\infty}$. Let F_k be the σ-algebra generated by the past $\{y_s, u_s, e_s, s \le k\}$ and assume

(A1) $E\{e_k | F_{k-1}\} = 0, \; k=0,1,...$

(A2) $E\{e_k^2 | F_{k-1}\} \le 1, \; k=0,1,...$

In our case, F_{k-1} is also the σ-algebra generated by $\{\phi_s, s \le k\}$. Note that (A1) and (A2) will be satisfied if:

(B1) $E\{e_k | e_j\} = 0 \quad \forall j < k$

(B2) $E\{e_k^2 | e_j\} \le 1 \quad \forall j < k$

(B3) u_k is a deterministic function of past inputs and outputs.

Now we note that (for a given, fixed sample path) if $\{\phi_k\}_{k=0}^{\infty}$ is persistently exciting with bounds α_1 and α_2, then $\|\phi_k\|^2 \le \alpha_2$ for all k. Thus we have

$$\|\underset{\wedge}{\theta}_k^e\| \le \|M_k^{-1}\| \; \|\sum_{j=1}^{k} \phi_j e_j\| \le \frac{l}{\alpha_1 (k-l)} \; \|\sum_{j=1}^{k} \phi_j e_j\| \tag{3.9}$$

and assumptions (A1) and (A2) guarantee that $\{\phi_k e_k\}_{k=1}^{\infty}$ is a sequence of uncorrelated zero-mean random variables with variance bounded by α_2. We conclude that:

(1) $\underset{\wedge}{\theta}_k^e \underset{q.m.}{\overset{a.s.}{\rightarrow}} 0$ (Strong Law of Large Numbers)

(2) $E[\, \|\underset{\wedge}{\theta}_k^e\|^2 \,] \le \dfrac{\alpha_2 l^2 k}{\alpha_1^2 (k-l)^2}$

Having established the convergence results for full order estimation, analogous results follow immediately for the reduced order systems. For example, assuming appropriate noise statistics,

$$\hat{v}_k \overset{a.s.}{\rightarrow} v \;\; \text{on} \;\; \Omega_{pe}^{w} \tag{3.10}$$

We want to establish more than this. Indeed, we intend to show that convergence properties cannot be destroyed by a reduced parameterization. So in relation to Eq. 3.10, since we know that the full order estimate converges on Ω_{pe}^{ϕ}, we want to show that $\Omega_{pe}^{w} \supset \Omega_{pe}^{\phi}$. In fact, we would really like to show that if the full order algorithm converges at a certain rate, then the

reduced order estimates *must* converge at the same rate or faster. In the next 2 subsections, we shall establish the following results:

(1) $\{\phi_k\}_{k=0}^{\infty}$ p.e. $\Rightarrow$ $\{\psi_k\}_{k=0}^{\infty}$ p.e.; in fact, the persistent excitation requirement is generically weaker for reduced order identification.

(2) $\underset{\cdot\cdot}{\theta}_k^{\epsilon}$ is a projection of $\underset{\sim}{\theta}_k^{\epsilon}$ onto $\mathrm{Im}(D)$ (see Eq. 3.24).

(3) $\{\phi_k\}_{k=0}^{\infty}$ p.e. implies

$$\frac{\|\underset{\cdot\cdot}{\theta}_k^{\epsilon}\|}{\|\underset{\sim}{\theta}_k^{\epsilon}\|} \le \sqrt{\frac{\alpha_2}{\alpha_1}}\sqrt{\frac{k+l+\sigma_{\min}(M_0)/\alpha_2}{k-l}} \underset{k\to\infty}{\to} \sqrt{\frac{\alpha_2}{\alpha_1}} \tag{3.11}$$

if $\underset{\sim}{\theta}_k^{\epsilon} \neq 0$ and $\underset{\cdot\cdot}{\theta}_k^{\epsilon} = 0$ if $\underset{\sim}{\theta}_k^{\epsilon} = 0$. Of course, this implies that the asymptotic rate of convergence of $\ddot{\theta}_k$ is at least as fast as that of $\hat{\theta}_k$.

3.3 Persistent Excitation Comparisons

The main result is quite straight-forward. Recall that in our framework, D is defined to be injective.

Proposition 3.3 If $\{\phi_k\}_{k=0}^{\infty}$ is persistently exciting then $\{\psi_k\}_{k=0}^{\infty}$ is persistently exciting.

Proof:

$$\sum_{j=k}^{k+l}\psi_j\psi_j^T = \sum_{j=k}^{k+l}D^T\phi_j\phi_j^TD = D^T\left[\sum_{j=k}^{k+l}\phi_j\phi_j^T\right]D \tag{3.12}$$

So $$\sigma_{\max}\left(\sum_{j=k}^{k+l}\psi_j\psi_j^T\right) \le \sigma_{\max}^2(D)\sigma_{\max}\left(\sum_{j=k}^{k+l}\phi_j\phi_j^T\right)$$ and

$\sigma_{\min}\left(\sum_{j=k}^{k+l}\psi_j\psi_j^T\right) \ge \sigma_{\min}^2(D)\sigma_{\min}\left(\sum_{j=k}^{k+l}\phi_j\phi_j^T\right)$. If $\{\phi_k\}_{k=0}^{\infty}$ is persistently exciting with l, α_1 and α_2 as defined in Eq. 3.6, we have

$$\alpha_1\sigma_{\min}^2(D) \le \sum_{j=k}^{k+l}\psi_j\psi_j^T \le \alpha_2\sigma_{\max}^2(D) \qquad \square$$

The converse is certainly not true. Even if $\{\phi_k\}_{k=0}^{\infty}$ is not persistently exciting, $\{\psi_k\}_{k=0}^{\infty}$ may be. The persistent excitation condition requirement is weaker for the reduced order parameterization than for full order systems. To make this more precise, we define the persistent excitation nullity to be the set of directions in which $\{\phi_k\}_{k=0}^{\infty}$ is <u>not</u> persistently exciting:

$$N_{pe}^{\phi} := \left\{x \in \mathbb{R}^{2n} : (\forall \beta_1,\beta_2>0)\ (\forall l \in \mathbb{N})\ (\exists k \in \mathbb{N})\ \sum_{j=k}^{k+l}|\phi_j^Tx|^2 \notin [\beta_1,\beta_2]\right\} \tag{3.13}$$

It follows from the definition that $\{\phi_k\}_{k=0}^{\infty}$ is persistently exciting if and only if $N_{pe}^{\phi} = \{0\}$.

Proposition 3.4 $\{\psi_k\}_{k=0}^{\infty}$ is persistently exciting if and only if $N_{pe}^{\phi} \cap \mathrm{Im}(D) = \{0\}$.

Proof: $\Rightarrow$: Suppose $0 \neq x \in N_{pe}^{\phi} \cap \mathrm{Im}(D)$. Choose $y \in \mathbb{R}^m$ with $x = Dy$. Then $\phi_j^T x = \psi_j^T y$ and $(\forall \beta_1, \beta_2 > 0)\ (\forall l \in \mathbb{N})\ (\exists k \in \mathbb{N})\ \sum_{j=k}^{k+l} |\psi_j^T y|^2 \notin [\beta_1, \beta_2]$. So $\{\psi_k\}_{k=0}^{\infty}$ is not persistently exciting.

$\Leftarrow$: Suppose $\{\psi_k\}_{k=0}^{\infty}$ is not persistently exciting. Then $(\forall \beta_1, \beta_2 > 0)\ (\forall l \in \mathbb{N})\ (\exists y \neq 0)$ $(\exists k \in \mathbb{N})\ \sum_{j=k}^{k+l} |\phi_j^T D y|^2 = \sum_{j=k}^{k+l} |\psi_j^T y|^2 \in [\beta_1, \beta_2]$. So $0 \neq Dy \in N_{pe}^{\phi} \cap \mathrm{Im}(D)$. ⊠

N_{pe}^{ϕ} is certainly closed under scalar multiplication. However, N_{pe}^{ϕ} need not be a linear subspace of $\mathbb{R}^{2n}$.

Example 3.5 $\phi_k \in \mathbb{R}^2$, $e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\{\phi_k\} = \{e_1, e_2, e_2, e_1, e_1, e_1, e_2, e_2, e_2, e_2, e_1, \ldots\}$. Here $\alpha_1 e_1 + \alpha_2 e_2 \in N_{pe}^{\phi} \Leftrightarrow \alpha_1 = 0$ or $\alpha_2 = 0$. So N_{pe}^{ϕ} is just the coordinate axes of $\mathbb{R}^2$. ⊠

N_{pe}^{ϕ} becomes a linear space if we restrict our attention to stationary sequences.

Definition: A sequence of vectors $\{v_k\}_{k=0}^{\infty}$ is stationary if the following limit exists *uniformly* in k:

$$R_v(m) := \lim_{p \to \infty} \frac{1}{p} \sum_{j=k}^{k+p-1} v_j v_{j+m}^T \tag{3.14}$$

(i.e. $(\forall m \in \mathbb{N})\ (\forall \varepsilon > 0)\ (\exists p_0 \in \mathbb{N})\ (\forall p > p_0)\ (\forall k \in \mathbb{N})\ \left\| R_v(m) - \frac{1}{p} \sum_{j=k}^{k+p} v_j v_{j+m}^T \right\| < \varepsilon$.)

Lemma 3.6 $\{v_k\}_{k=0}^{\infty}$ stationary $\Rightarrow \exists p, \alpha_2$ such that $\forall k\ \ \sigma_{\max}\left(\sum_{j=k}^{k+p} v_j v_j^T\right) < \alpha_2$.

Proof: Choose any $\varepsilon > 0$. Then $\exists p$ such that $\forall k\ \ \left\| R_v(0) - \frac{1}{p+1} \sum_{j=k}^{k+p} v_j v_j^T \right\| < \varepsilon$. So $\alpha_2 = (p+1)\|R_v(0)\| + \varepsilon$ will do the job. ⊠

We conclude that when discussing persistent excitation of stationary signals, we only need to be concerned with the lower bound in Eq. 3.6 (the upper bound is satisfied automatically).

Proposition 3.7 $\{\phi_j\}_{k=0}^{\infty}$ stationary $\Rightarrow N_{pe}^{\phi} = \mathrm{Ker}(R_{\phi}(0))$.

Proof: $N_{pe}^{\phi} \subset \mathrm{Ker}(R_{\phi}(0))$: Fix $x \in N_{pe}^{\phi}$ and let $y = \frac{x}{\|x\|}$. Clearly $y \in N_{pe}^{\phi}$. Choose $\varepsilon > 0$. Then $\exists p < \infty$ such that $\forall k\ \ \left\| R_{\phi}(0) - \frac{1}{p} \sum_{j=k}^{k+p-1} \phi_j \phi_j^T \right\| < \frac{\varepsilon}{2}$. Since $\|y\| = 1$, $\forall k$ $\left| y^T R_{\phi}(0) y - \frac{1}{p} \sum_{j=k}^{k+p-1} |\phi_j^T y|^2 \right| < \frac{\varepsilon}{2}$. Choosing $l = p$ and $\alpha_1 = \frac{p\varepsilon}{2}$ in the definition of N_{pe}^{ϕ}, $\exists k_0$ such that $\frac{1}{p} \sum_{j=k_0}^{k_0+p-1} |\phi_j^T y|^2 < \frac{\varepsilon}{2}$. So

$$y^T R_\phi(0)y = (y^T R_\phi(0)y - \frac{1}{p}\sum_{j=k_0}^{k_0+p-1} |\phi_j^T y|^2) + \frac{1}{p}\sum_{j=k_0}^{k_0+p-1} |\phi_j^T y|^2 < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \qquad (3.15)$$

Since ε was arbitrary, $y^T R_\phi(0)y = 0 \Rightarrow y \in \mathrm{Ker}(R_\phi(0)) \Rightarrow x = \|x\| y \in \mathrm{Ker}(R_o(0))$.

$\underline{N_{pe}^\phi \supseteq \mathrm{Ker}(R_\phi(0))}$: Stationarity $\Rightarrow$ $(\forall \varepsilon>0)$ $(\forall x \in \mathbb{R}^{2n})$ $(\exists p_0)$ $(\forall p>p_0)$ $(\forall k)$ $|x^T(R_\phi(0) - \frac{1}{p}\sum_{j=k}^{k+p-1} \phi_j\phi_j^T)x| < \varepsilon \|x\|^2$. Fix $x \in \mathrm{Ker}(R_\phi(0))$. Then for $k=0$ we have $(\forall \varepsilon>0)$ $(\exists p_0)$ $(\forall p>p_0)$ $\sum_{j=0}^{p-1} |\phi_j^T x|^2 < p\varepsilon \|x\|^2$.

claim: $(\forall \alpha>0)$ $(\forall l \in \mathbb{N})$ $(\exists k_0)$ $\sum_{j=k_0}^{k_0+l-1} |\phi_j^T x|^2 < \alpha$. Of course, this implies that $x \in N_{pe}^\phi$.

proof of claim: Fix $\alpha>0$, $l \in \mathbb{N}$ and let $\varepsilon = \frac{\alpha}{l\|x\|^2}$. Then $(\exists p_0)$ $(\forall p>p_0)$ $\sum_{j=0}^{p-1} |\phi_j^T x|^2 < \frac{p\alpha}{l}$. Choose $n \in \mathbb{N}$, $n > \frac{p_0}{l}$ and fix $p = nl$. So $p>p_0$, and

$$\sum_{j=0}^{p-1} |\phi_j^T x|^2 = \sum_{k=0}^{n-1} \left[\sum_{j=kl}^{kl+l-1} |\phi_j^T x|^2 \right] < n\alpha \qquad (3.16)$$

So for some $k \in \{0, \ldots, n-1\}$, $\sum_{j=kl}^{kl+l-1} |\phi_j^T x|^2 < \alpha$. Then $k_0 = kl$ satisfies the claim. ⊠

Conclusion. In comparing persistency of excitation for the two identification schemes, we conclude:

a) If $\{\phi_k\}_{k=0}^\infty$ is persistently exciting then so is $\{\psi_k\}_{k=0}^\infty$; and

b) if $\{\phi_k\}_{k=0}^\infty$ is stationary and persistently exciting of order m (i.e. $\mathrm{rank}(R_\phi(0)) \geq m$), then $\{\psi_k\}_{k=0}^\infty$ is persistently exciting unless a (measure zero) subspace intersection condition is met.

Remark. The stationarity property plays an important role in the analysis of asymptotic properties of identifiers. The matrix function $R_v(m)$ defined in Eq. 3.14 is called the *autocovariance* of $\{v_k\}_{k=0}^\infty$. Similarly, given two staionary vector sequences $\{v_k\}_{k=0}^\infty$ and $\{u_k\}_{k=0}^\infty$, we define their *cross-correlation*

$$R_{uv}(m) := \lim_{p\to\infty} \frac{1}{p}\sum_{j=k}^{k+p-1} u_j v_{j+m}^T$$

This stationarity concept is completely deterministic. For stochastic systems, the results given hold for any stationary *sample paths*.

There is an analogous property of (stochastic) stationarity in the theory of stochastic systems, which is related to the sample path property (3.14). Indeed, for a *wide sense stationary* sequence of random vectors $\{v_k(\omega)\}_{k=0}^\infty$ we define the *stochastic autocovariance*

$$R_v^{stoch}(m) := \mathrm{E}[v_k v_{k+m}^T]$$

which is independent of k. For a wide sense stationary *ergodic* process the sample path autocovariance $R_v(m,\omega)$ as defined in Eq. 3.14 exists and is equal to $R_v^{stoch}(m)$ for almost all ω. Unfortunately, we cannot guarantee that the limit in Eq. 3.14 will be uniform (in k) for almost all ω.

The concept of (deterministic) stationarity was one of the tools developed by Wiener in [14]. A very brief survey of some of the results with respect to their application for adaptive systems is included in Boyd and Sastry [5].

3.4 Parameter Error Comparisons

We now derive a direct relation between $\underset{\sim}{\theta}_k^e$ and $\underset{\wedge}{\theta}_k^e$, the reduced- and full-order parameter errors due to "noise." (Recall that $\underset{\wedge}{\theta}_k^e \to 0$ as $1/\sqrt{n}$, while $\underset{\wedge}{\theta}_k^{ic} \to 0$ as $1/n$. So $\underset{\wedge}{\theta}_k^e$ is the asymptotically dominant term.) We have

$$\underset{\wedge}{\theta}_k^e = M_k^{-1} \sum_{s=1}^{k} \phi_s e_s \tag{3.17}$$

$$\underset{\wedge}{\gamma}_k^e = N_k^{-1} \sum_{s=1}^{k} \psi_s e_s \tag{3.18}$$

$$\underset{\sim}{\theta}_k^e = D \underset{\wedge}{\gamma}_k^e \tag{3.19}$$

Since $\psi_s = D^T \phi_s$ and $N_0 = D^T M_0 D$ we have $N_k = D^T M_k D$ and

$$\underset{\sim}{\theta}_k^e = D (D^T M_k D)^{-1} D^T \sum_{s=1}^{k} \phi_s e_s \tag{3.20}$$

But, from Eq. 2.5,

$$\sum_{s=1}^{k} \phi_s e_s = M_k \underset{\wedge}{\theta}_k^e \tag{3.21}$$

So

$$\underset{\sim}{\theta}_k^e = D (D^T M_k D)^{-1} D^T M_k \underset{\wedge}{\theta}_k^e \tag{3.22}$$

and the connection between the two parameter errors is established. To make more sense of Eq. 3.22, we let M_k have a singular value decomposition

$$M_k = U \Sigma U^T \tag{3.23}$$

and define

$$L := \Sigma^{1/2} U^T D$$

$$T := U \Sigma^{-1/2}$$

Then Eq. 3.22 can be written

$$\underset{\sim}{\theta}_k^{\ell} = (U\Sigma^{-1/2})L(L^TL)^{-1}L^T(\Sigma^{1/2}U^T)\underset{\wedge}{\theta}_k^{\ell} \tag{3.24}$$

$$= T\,Pr\,T^{-1}\underset{\wedge}{\theta}_k^{\ell}$$

where Pr is the orthogonal projection onto $\mathrm{Im}(L)$. So the map $\underset{\wedge}{\theta}_k^{\ell} \mapsto \underset{\sim}{\theta}_k^{\ell}$ is a similarity transformation of an orthogonal projection. The result is a projection onto $\mathrm{Im}(T) = \mathrm{Im}(D)$ as expected. Of course, similarity transformations do not preserve orthogonality, so we may not conclude that $\|\underset{\sim}{\theta}_k^{\ell}\| \leq \|\underset{\wedge}{\theta}_k^{\ell}\|$ in general. None-the-less, $\|\underset{\sim}{\theta}_k^{\ell}\| / \|\underset{\wedge}{\theta}_k^{\ell}\|$ can be bounded above.

$$\|\underset{\wedge}{\theta}_k^{\ell}\| \leq \sigma_{\max}(T)\,\sigma_{\max}(T^{-1})\,\|\underset{\wedge}{\theta}_k^{\ell}\| \tag{3.25}$$

$$= \sigma_{\min}(M_k)^{-1/2}\sigma_{\max}(M_k)^{1/2}\,\|\underset{\wedge}{\theta}_k^{\ell}\|$$

Finally, applying the persistent excitation condition (Eq. 3.6) yields

$$\frac{\|\underset{\sim}{\theta}_k^{\ell}\|}{\|\underset{\wedge}{\theta}_k^{\ell}\|} \leq \sqrt{\frac{\alpha_2}{\alpha_1}}\sqrt{\frac{k+l+\sigma_0/\alpha_2}{k-l}} \underset{k\to\infty}{\to} \sqrt{\frac{\alpha_2}{\alpha_1}} \tag{3.26}$$

where $\sigma_0 = \sigma_{\max}(M_0)$.

4. ROBUSTNESS OF REDUCED ORDER IDENTIFICATION

4.1 Robustness to Improper Parameterization

In the reduced order identification scheme, we restrict the set of allowable parameters to $\Theta = \theta_0 + \mathrm{Im}(D)$. We now ask what to expect if this re-parameterization is not *proper*; i.e. if the true parameter θ is *not* in Θ.

First we consider the deterministic exact least squares algorithm. Assume that $\theta \notin \Theta$ and define

$$d := \min_{v \in \mathbf{R}^m} \{ \|\theta - \theta_0 - Dv\| \} \tag{4.1}$$

the distance from θ to Θ. Let v' be the minimizing argument of Eq. 1, and let $\theta' = \theta_0 + Dv'$. So $d = \|\theta - \theta'\|$, and θ' is the nearest element of Θ to θ.

Theorem 4.1 Suppose $\{\phi_k\}_{k=0}^{\infty}$ is persistently exciting, with l, α_1, and α_2 defined as in Eq. 3.6. Then

$$\|\theta - \ddot{\theta}_k\| \le \sqrt{\frac{k+l}{k-l}}\sqrt{\frac{\alpha_2}{\alpha_1}}\, d \underset{k\to\infty}{\to} \sqrt{\frac{\alpha_2}{\alpha_1}}\, d \tag{4.2}$$

Proof: $\ddot{\theta}_k$ minimizes $\|Y_k - \Phi_k^T\ddot{\theta}_k\|$ over Θ. Since $\theta' \in \Theta$, we have

$$\|Y_k - \Phi_k^T\ddot{\theta}_k\| \le \|Y_k - \Phi_k^T\theta'\| \tag{4.3}$$

Since $Y_k = \Phi_k^T\theta$ and $\|\theta - \theta'\| = d$,

$$\|\theta - \ddot{\theta}_k\| \le \frac{\sigma_{\max}(\Phi_k^T)}{\sigma_{\min}(\Phi_k^T)}\, d \tag{4.4}$$

Applying the persistent excitation bounds on $\Phi_k^T\Phi_k = \sum_{j=1}^{k}\phi_j\phi_j^T$ from Eq. 3.6 gives the result. ⊠

Note that when $\sum_{j=1}^{k}\phi_j\phi_j^T$ is "perfectly conditioned" (i.e. $\sigma_{\min} = \sigma_{\max}$), Theorem 4.1 implies that $\ddot{\theta}_k = \theta'$. In any case, Theorem 4.1 and Pythagoras' Theorem tell us that $\ddot{\theta}_k$ will converge to the set

$$\left\{\theta \in \Theta : \|\theta - \theta'\| \le \sqrt{\frac{\alpha_2}{\alpha_1} - 1}\; d\right\}$$

In order to guarantee that $\ddot{\theta}_k$ actually converges we assume that $\{\phi_k\}_{k=0}^{\infty}$ is stationary as well as persistently exciting and denote the limit θ^*. (Proof of convergence: $\{\phi_k\}_{k=0}^{\infty}$ stat. and p.e. $\Rightarrow$ $\{\psi_k\}_{k=0}^{\infty}$ stat. and p.e. and $\{z_k\}_{k=0}^{\infty}$ stat. $\Rightarrow$ $\hat{v}_k$ converges to $v^* = R_{\psi}(0)^{-1}R_{\psi z}(0) \Rightarrow \ddot{\theta}_k$ converges

to $\theta^* = \theta_0 + D\nu^*$.) Of course, we have:

$$\|\theta^* - \theta\| \le \sqrt{\frac{\alpha_2}{\alpha_1}}\, d \tag{4.5}$$

$$\|\theta^* - \theta'\| \le \sqrt{\frac{\alpha_2}{\alpha_1} - 1}\, d \tag{4.6}$$

$$\|\nu^* - \nu'\| \le \frac{1}{\sigma_{\min}(D)} \sqrt{\frac{\alpha_2}{\alpha_1} - 1}\, d \tag{4.7}$$

where ν' is specified by $\theta' = \theta_0 + D\nu'$. In fact, since $\{\phi_k\}_{k=0}^{\infty}$ is stationary, we can obtain tighter bounds in Eq. 4.5-4.7 by replacing $\sqrt{\alpha_2/\alpha_1}$ with the condition number of $R_\phi(0)$.

4.2 Robustness to Unmodelled Dynamics

We demonstrate robustness to higher order dynamics as a special case of improper parameterization. We assume our system is actually n^{th} order but we identify it as an $\bar{n}^{th}$ order system with $\bar{n} < n$. Θ is the $2\bar{n}$ dimensional linear subspace of $\mathbb{R}^{2n}$ given by

$$a_{\bar{n}+1} = \cdots = a_n = b_{\bar{n}+1} = \cdots = b_n = 0 \tag{4.8}$$

The reduced parameter space $\mathbb{R}^{2\bar{n}}$ can therefore be identified directly with Θ, and D is just the inclusion map ($\theta_0 = 0$). Having made the identification, we say $\hat{\nu}_k = \ddot{\theta}_k$, $\nu' = \theta'$, and $\nu^* = \theta^*$. (Since $\sigma_{\min}(D) = 1$, Eqs. 4.6 and 4.7 are still consistent.) Theorem 4.1 can now be applied directly to give the desired robustness result:

$$\hat{\nu}_k - \nu' \le \sqrt{\frac{k+l}{k-l}} \sqrt{1 - \frac{\alpha_2}{\alpha_1}}\, d \tag{4.9}$$

$$d^2 = a_{\bar{n}+1}^2 + \cdots + a_n^2 + b_{\bar{n}+1}^2 + \cdots + b_n^2 \tag{4.10}$$

It is clear that this concept can be extended to combined robustness to unmodelled dynamics *and* improper parameterization. Instead of identifying Θ directly with $\mathbb{R}^{2\bar{n}}$, we identify Θ with an affine subspace of $\mathbb{R}^{2\bar{n}}$.

Finally, we remark that ν^* has been called the 'tuned value' for the modelled system in the presence of unmodelled dynamics for the case $\Theta = \mathbb{R}^{2\bar{n}}$. The tuned value depends not only on the system and modelling assumptions, but on the actual signals (i.e. on the choice of input) as well. This dependence is demonstrated explicitly in the equation

$$\nu^* = R_\psi(0)^{-1} R_{\psi z}(0) \tag{4.11}$$

5. REFERENCES

[1] Åström, K. J., "Theory and Applications of Adaptive Control - A Survey," *Automatica*, Vol. 19, no. 5, 1983, pp. 471-486.

[2] Bai, E. W. and Sastry, S. S., "Parameter Identification Using Prior Information," *International Journal of Control*, Vol. 44, 1986, no. 2, pp. 455-473.

[3] Bai, E. W. and Sastry, S. S., "Persistency of Excitation, Sufficient Richness and Parameter Convergence in Discrete Time Adaptive Control," *System and Control Letters*, Vol. 6, 1985, pp. 153-163.

[4] Bodson, M and Sastry, S. S., *Stability, Convergence and Robustness of Adaptive Systems*, Memorandum No. UCB/ERL M86/66, Electronic Research Laboratory, University of California, Berkeley, 1986.

[5] Boyd, S. and Sastry, S. S., "Necessary and Sufficient Conditions for Parameter Convergence in Adaptive Control," *Proc. of the Ames-Berkeley Conf. on Non-Linear Dynamics and Control*, Math-Sci Press, Brookline, Massachusetts, 1984, pp. 81-101.

[6] Clary, J. P., *Robust Algorithms in Adaptive Control*, Ph. D. Thesis, Information Systems Laboratory, Stanford University, 1984.

[7] Dasgupta, S., *Adaptive Identification and Control*, Ph. D. Thesis, Australia National University, Canberra Australia, 1984.

[8] Dasgupta, S., Anderson, B. D. O. and Kaye, R. J., "Robust Identification of Partially Known Systems," *Proc. of the 22nd IEEE Conference on Decision and Control*, pp. 1510-1514, 1983.

[9] Goodwin, G. C. and Payne, R. L., *Dynamic System Identification: Experiment Design and Data Analysis*, Academic Press, 1977.

[10] Goodwin, G. C. and Sin, K. S., *Adaptive Filtering, Prediction and Control*, Prentice-Hall, 1984.

[11] Kumar, P. R. and Varaiya, P., *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, 1986.

[12] Ljung, L. and Söderström, T., *Theory and Practice of Recursive Identification*, MIT Press, 1983.

[13] Mason, J. E. and Sastry, S. S., *Use of Specific Prior Information in the Adaptive Identification of Continuous Time Systems*, Memorandum no. UCB/ERL M86/60, Electronic Research Laboratory, University of California, Berkeley, 1986.

[14] Wiener, N., "Generalized Harmonic Analysis," *Acta Mathematica*, vol. 55, 1930, pp. 117-258.