Realization of Stochastic Systems with Exogenous Inputs and Subspace Identification Methods *

Tohru Katayama[†] and Giorgio Picci[‡]

A stochastic realization theory for a discrete-time stationary process with an exogenous input is developed by extending the classical CCA technique. Some stochastic subspace identification methods are derived by adapting the realization procedure to finite input-output data.

Key Words — Stochastic realization; exogenous inputs; canonical correlation analysis; oblique projection; subspace identification.

Abstract

This paper solves the stochastic realization problem for a discrete-time stationary process with an exogenous input. The oblique projection of the future outputs on the space of the past observations along the space of the future inputs is factorized as a product of the extended observability matrix and the state vector. The state vector is chosen by using the canonical correlation analysis (CCA) of past and future conditioned on the future inputs. We then derive the state equations of the optimal predictor of the future outputs in terms of the state vector and the future inputs. These equations lead to a forward innovation model for the output process in the presence of exogenous inputs. The basic step of the realization procedure is a factorization of the conditional covariance matrix of future outputs and past data given future inputs. This factorization is based on CCA and can be easily adapted to finite input-output data. We derive four stochastic subspace identification algorithms which adapt the realization procedure to finite input-output data. Numerical results are also included.

*Received March 13, 1998; revised December 14, 1998; received in final form February , 1999. Part of this paper was presented at the 11th IFAC Symposium on System Identification, Kitakyushu, Japan during July 8-11, 1997. Corresponding author T. Katayama. Tel. +81-75-753-5502; Fax +81-75-761-2437; E-mail katayama@kuamp.kyoto-u.ac.jp.

[†]Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

[‡]Dipartimento di Elettronica e Informatica, Università di Padova and LADSEB-CNR, 35131 Padova, Italy

1 Introduction

Subspace identification methods involve geometric operations on subspaces spanned by the column or row vectors of certain block Hankel matrices formed by the input-output data. These operations are performed numerically in a reliable way based on the singular value decomposition (SVD) and QR decomposition (Golub and Van Loan, 1989); a great advantage being that the problem of local minima and the difficult model selection problem inherent in the classical parameter optimization approaches to MIMO identification (Ljung, 1987) are completely avoided, except for the estimation of the state space dimension.

These methods have attracted much interest in the past years and there is a growing literature on the subject (Jansson and Wahlberg, 1996; Larimore, 1990; Moonen and Vandewalle,1990; Peternell *et al.*, 1996; Van Overschee and De Moor, 1993, 1994, 1996; Verhaegen and Dewilde, 1992; Verhaegen, 1994; Viberg, 1995). Relations among existing subspace identification methods have been discussed by Van Overschee and De Moor (1996) and others. Recently, Lindquist and Picci (1996a, 1996b) have interpreted subspace identification of state-space models for time series from the point of view of realization theory and thereby analyzed some of the possible pitfalls of the method.

A minimal stochastic realization theory based on the canonical correlation analysis (CCA) of the future and the past of a stochastic process has been developed by Akaike in his pioneering work (Akaike, 1974, 1975, 1976). In this approach the state space is defined as a linear space spanned by the best predictors of the future based on the past and a basis in this space is naturally provided by the canonical variates. The computation associated with the CCA technique is performed by the SVD of a normalized block Hankel matrix formed by lagged covariance matrices of the process. Akaike also gave a stochastic interpretation of the basic Ho-Kalman realization method (Ho and Kalman, 1966) and other related realization algorithms.

The idea of using the CCA technique for stochastic state-space modeling and model reduction has been further developed by Desai *et al.*(1985). Aoki (1990) has also proposed a stochastic realization approach to economic time-series identification based on a similar SVD technique. A general geometric theory for solving the stochastic realization problem of time series has been developed by Lindquist and Picci (1985, 1991).

One could say that the relation of stochastic realization theory to subspace identification of stationary time series is now well understood. Instead there is no complete

theory available so far for the realization of stochastic systems excited by an exogenous input. Picci and Katayama (1996) have recently proposed an approach based on separating the "deterministic" and "stochastic" components of the output and shown a possible application to subspace identification of state-space systems with exogenous inputs. But there is still insufficient numerical experience concerning the effectiveness of this decomposition.

Practically all subspace identification methods for systems with inputs reported in the literature use an instrumental variable approach of one sort or another. Sometimes the computational details tend to obscure the main idea. The sound conceptual procedure of first building a state space by projecting the future onto the past and then picking a well-conditioned basis to compute the matrices A, C, $\bar{C}$, etc., which is transparent in the seminal subspace paper of Van Overschee and De Moor (1993) for time series, is somewhat harder to implement when the identification problem involves exogenous inputs. The basic realization ideas get lost in the (sometimes complicated) computations resulting from a variety of possible choices of instruments to compute the observability matrix. We may attribute this state of affairs to the fact that stochastic realization with inputs is, so far, poorly understood.

Motivated by the above, in this paper we consider the realization problem for linear stationary stochastic systems with an exogenous input. The approach to stochastic realization used here is based on oblique projections and conditional CCA which generalize in a natural way the procedure used in realization of time series (without inputs). We propose some subspace identification methods based on the realization procedure which are also meant to shed light on the basic principles underlying some well-known papers on subspace identification (Larimore, 1990; Van Overschee and De Moor, 1994, 1996; Verhaegen and Dewilde, 1992; Verhaegen, 1994) and to allow comparisons.

The construction of a state-space model involves the linear predictor of the future output based on the future inputs and the past observations and is solved by using representation results for two corresponding oblique projection operators. The latter are a basic ingredient in this problem since, as observed by Larimore (1990), the effect of the future inputs is to be removed from the future outputs in order to construct state-space systems with an exogenous input. In the next section it will be clear how oblique projections serve precisely this purpose.

The organization of the paper is as follows. In Section 2, we present some pre-

liminary mathematical facts about oblique projection in Hilbert spaces and consider a linear prediction problem for the future outputs based on the joint past of input and output variables. We derive two discrete Wiener-Hopf type equations satisfied by the oblique projection operators. In Section 3, we define the state vector based on the factorization of a block Hankel matrix using the CCA technique, and construct a forward innovation model for the stochastic system with exogenous inputs. We also develop an idea of stochastic realization based on finite-data. In Section 4, we present two methods of solving the Wiener-Hopf equations based on finite input-output data. Section 5 derives four subspace identification algorithms and some numerical results are shown in Section 6. The conclusion is given in Section 7. The appendix contains some proofs of technical character.

2 Construction of the Predictor Space

2.1 Preliminaries

Consider two discrete-time wide-sense stationary vector processes $\{u(t),\ y(t),\ t = 0, \pm 1, \cdots\}$ of respective dimensions m and p. The first component $u(t)$ models the *input* signal while the second $y(t)$, the *output* of some unknown linear stochastic system which we want to construct from observed input-output data. Throughout this paper we shall assume that the joint process (u, y) is a stationary purely non-deterministic full rank process (Rozanov, 1963). It is also assumed that the two processes are zero-mean and have finite joint covariance matrices.

A basic step in solving this realization problem, which is also the core of the subspace identification algorithms presented later on, is the construction of the *state space* of the system at some chosen instant of time t. In Picci and Katayama (1996) and Picci (1997) a state-space construction was proposed which can be implemented separately for the two orthogonal components, called the "deterministic" and the "stochastic" components of the output process. The decomposition makes sense under a technical condition of absence of feedback (which we shall also assume later on).

In this paper we shall instead construct the minimal state space of the realization without the preliminary decomposition of the output process considered in the two previous references. It will be shown (constructively in Section 3) that under absence of feedback from y to u, the minimal state space leading to minimal causal stationary realizations of the process y involving the exogenous process u as an input variable,

is the so-called forward *predictor space* of y, i.e. a subspace of random variables obtained by projecting the future of the process y obliquely onto the joint past of y and u. This projection onto the joint past space we shall now prepare to analyze.

Let t be a "present" time and let k be a positive integer. We define the stacked vectors of past and future inputs and stacked vectors of past and future outputs as

$$u_-(t) := \begin{bmatrix} u(t-1) \\ u(t-2) \\ \vdots \\ \end{bmatrix}, \quad u_+(t) := \begin{bmatrix} u(t) \\ u(t+1) \\ \vdots \\ u(t+k-1) \end{bmatrix}$$

and

$$y_-(t) := \begin{bmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ \end{bmatrix}, \quad y_+(t) := \begin{bmatrix} y(t) \\ y(t+1) \\ \vdots \\ y(t+k-1) \end{bmatrix}$$

where the present is conventionally included in the future. For notational simplicity we shall drop the subscripts and write

$$\begin{aligned} p(t) &:= \begin{bmatrix} u_-(t) \\ y_-(t) \end{bmatrix} \quad [\infty \times 1 \text{ past observations}] \\ f(t) &:= y_+(t) \quad [kp \times 1 \text{ future outputs}] \end{aligned}$$

It should be noted that the dimension of $p(t)$ is infinite, whereas those of $f(t)$ and $u_+(t)$ are finite.

In the following, $\vee$ denotes (closed) vector sum of subspaces. Let $\mathcal{P}_{t-}$, $\mathcal{Y}_{t-}$ and $\mathcal{U}_{t+}$ be the linear spaces of second order random variables spanned by the past $p(t)$, by $y_-(t)$ and by the future inputs $u_+(t)$, respectively. These spaces are assumed to be closed with respect to the usual root-mean-square norm $\|\xi\| := [E\{\xi^2\}]^{1/2}$, where $E\{\cdot\}$ denotes mathematical expectation. Thus $\mathcal{P}_{t-}$, $\mathcal{Y}_{t-}$ and $\mathcal{U}_{t+}$ are thought of as Hilbert subspaces of an ambient Hilbert space $\mathcal{H} := \mathcal{U} \vee \mathcal{Y}$ containing all linear functionals of the joint process (u, y).

Suppose that $\mathcal{A}$ is a subspace of $\mathcal{H}$. The orthogonal projection of $b \in \mathcal{H}$ onto the subspace $\mathcal{A}$ is denoted by $\hat{E}\{b|\mathcal{A}\}$ or by $b\,|\mathcal{A}$ for short. Let $\mathcal{A}^\perp$ be the orthogonal complement of $\mathcal{A} \subset \mathcal{H}$. Then the orthogonal projection of b onto $\mathcal{A}^\perp$ is defined by $b\,|\mathcal{A}^\perp := b - (b\,|\mathcal{A})$. Suppose further that $\mathcal{A}$ is generated by a random vector a, then the orthogonal projection of $b \in \mathcal{H}$ onto $\mathcal{A}$ can be expressed in terms of a by the well-know formula

$$b\,|\mathcal{A} = E\{ba'\}E\{aa'\}^\dagger a = \Sigma_{ab}\Sigma_{aa}^\dagger a := b\,|a$$

where $\Sigma_{ab} := E\{ab'\}$ is the covariance matrix of the two random vectors a and b, the prime denotes the transpose and $(\cdot)^\dagger$ is the pseudo-inverse (Golub and Van Loan, 1989). The orthogonal projection of b onto $\mathcal{A}^\perp$ can also be expressed in terms of a and is denoted by $b\,|a^\perp$.

2.2 Orthogonal and oblique projections

The following is a basic technical result used in this paper.

Lemma 1. Let y, a and b be random vectors with components in $\mathcal{H}$, let $\mathcal{A} = \text{span}\{a\}$, $\mathcal{B} = \text{span}\{b\}$ and assume that $\mathcal{A} \cap \mathcal{B} = \{0\}$. Then the vectors $\Pi(y)a$, $\Phi(y)b$ in the orthogonal projection

$$\begin{aligned} \hat{E}\{y \mid \mathcal{A} \vee \mathcal{B}\} &= [E\{ya'\}\ E\{yb'\}] \begin{bmatrix} E\{aa'\} & E\{ab'\} \\ E\{ba'\} & E\{bb'\} \end{bmatrix}^\dagger \begin{bmatrix} a \\ b \end{bmatrix} \\ &:= \Pi(y)a + \Phi(y)b \end{aligned} \tag{1}$$

are the oblique projections of y onto $\mathcal{A}$ along $\mathcal{B}$ and of y onto $\mathcal{B}$ along $\mathcal{A}$ respectively. Moreover define the *conditional* covariance matrices

$$\Sigma_{aa\,|b} := E\{(a\,|b^\perp)(a\,|b^\perp)'\}, \quad \Sigma_{ya\,|b} := E\{(y\,|b^\perp)(a\,|b^\perp)'\}$$

$$\Sigma_{yb\,|a} := E\{(y\,|a^\perp)(b\,|a^\perp)'\}, \quad \Sigma_{bb\,|a} := E\{(b\,|a^\perp)(b\,|a^\perp)'\}.$$

Then the matrices $\Pi(y)$ and $\Phi(y)$ satisfy the discrete Wiener-Hopf type equations

$$\Pi(y)\Sigma_{aa\,|b} = \Sigma_{ya\,|b}, \qquad \Phi(y)\Sigma_{bb\,|a} = \Sigma_{yb\,|a} \tag{2}$$

in which $\Sigma_{aa\,|b}$, $\Sigma_{bb\,|a}$ are nonsingular if so are Σ_{aa}, Σ_{bb}.

Proof: The formulae for Π and Φ are easily proved by using[1]

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} \Delta^{-1} & -\Delta^{-1}BD^{-1} \\ -D^{-1}C\Delta^{-1} & D^{-1} + D^{-1}C\Delta^{-1}BD^{-1} \end{bmatrix}$$

where $\Delta := A - BD^{-1}C$ is the Schur complement. Since $A = E\{aa'\}$, $B = C' = E\{ab'\}$, $D = E\{bb'\}$, we have $\Delta = E\{(a\,|b^\perp)(a\,|b^\perp)'\} = \Sigma_{aa\,|b}$. Thus expanding the first block row of (1) it can be shown that

$$\begin{aligned} \Pi(y) &= E\{ya'\}\Delta^{-1} - E\{yb'\}E\{bb'\}^{-1}E\{ba'\}\Delta^{-1} \\ &= E\{(y\,|b^\perp)(a\,|b^\perp)'\}E\{(a\,|b^\perp)(a\,|b^\perp)'\}^{-1} \\ &= \Sigma_{ya\,|b}\Sigma_{aa\,|b}^{-1} \end{aligned}$$

[1] For simplicity we assume that the indicated inverses exist. Also recall that, if Σ_{bb} is nonsingular, the conditional covariance matrix $\Sigma_{ya\,|b}$ can be expressed as $\Sigma_{ya\,|b} = \Sigma_{ya} - \Sigma_{yb}\Sigma_{bb}^{-1}\Sigma_{ba}$.

Similarly, we have $\Phi(y) = \Sigma_{yb|a}\Sigma_{bb|a}^{-1}$. Since $\Pi(y)a = \Sigma_{ya|b}\Sigma_{aa|b}^{-1}a$, we see that, for $v = \alpha' a$, $z = \beta' b$, $\alpha \in \mathbb{R}^{n_a}$, $\beta \in \mathbb{R}^{n_b}$,

$$\Pi(v)a = \alpha' I a = v, \quad \Pi(z)a = 0$$

Thus $\Pi(\cdot)a$ is idempotent on $\mathcal{A} = \text{span}\{a\}$ and its nullspace is exactly $\mathcal{B} = \text{span}\{b\}$. This shows that $\Pi(y)a$ is the oblique projection of y onto $\mathcal{A}$ along $\mathcal{B}$. A dual argument shows that $\Phi(v)b = 0$ since $\Sigma_{vb|a}$ is zero etc. and $\Phi(y)b$ is the complementary oblique projection. The statement regarding nonsingularity of $\Sigma_{aa|b}$, $\Sigma_{bb|a}$ follows readily from $\mathcal{A} \cap \mathcal{B} = \{0\}$, for if $\alpha'\Sigma_{aa|b} = 0$ with α a nonzero vector, then we would have $\alpha' a \in \mathcal{B}$ which is a contradiction. □

2.3 The optimal predictor

In this section we shall give a representation of the linear predictor of the future output vector $f(t)$ based on the infinite past observations $p(t)$ and on the future inputs $u_+(t)$. We shall need a technical assumption of *sufficient richness* of the input process, which we shall write in the form

$$\mathcal{U}_{t-} \cap \mathcal{U}_{t+} = \{0\} \tag{3}$$

It is shown (Hannan and Poskitt, 1988; Lindquist and Picci, 1996a) that this condition implies that the input process is purely non-deterministic and admits a spectral density matrix without zeros on the unit circle (i.e. coercive). This is too restrictive in many practical situations and we could instead assume just persistently exciting (PE) condition of sufficiently high order and finite dimensionality of an underlying "true" system from the outset. The main reason to choose condition (3) is that it allows a very simple justification of some crucial representation results which would otherwise take a longer and more involved proof and it does not require the assumption of an underlying finite dimensional true system generating the data.

Another crucial assumption in this paper is that the input u is exogenous, i.e. there is *no feedback from the output y to the input u*. This concept is discussed in Caines and Chan (1976), Gevers and Anderson (1982), Picci and Katayama (1996). In Appendix A the geometric definition is recalled briefly.

By a generalization of Lemma 1 we obtain the following representation result.

Theorem 1. Assume (3) and suppose there is no feedback from y to u. Then the optimal predictor $\hat{f}(t)$ of the future output vector $f(t)$ based on $\mathcal{P}_{t-} \vee \mathcal{U}_{t+}$, the space spanned by $p(t)$ and $u_+(t)$, is given by

$$\hat{f}(t) = \hat{E}\{f(t) \mid \mathcal{P}_{t-} \vee \mathcal{U}_{t+}\} = \Pi p(t) + \Phi u_+(t) \tag{4}$$

where $\Pi p(t)$ is the oblique projection of the future $f(t)$ onto the past $\mathcal{P}_{t-}$ along the future $\mathcal{U}_{t+}$, and $\Phi u_+(t)$ is the oblique projection of the future $f(t)$ onto $\mathcal{U}_{t+}$ along $\mathcal{P}_{t-}$ (see Fig. 1). Π and Φ satisfy the discrete Wiener-Hopf type equations

$$\Pi\Sigma_{pp\,|u} = \Sigma_{fp\,|u}, \quad \Phi\Sigma_{uu\,|p} = \Sigma_{fu\,|p} \tag{5}$$

where $\Sigma_{pp\,|u}$, $\Sigma_{uu\,|p}$ are the conditional covariance operators of the past vector $p(t)$ given the future $u_+(t)$ and of the future input $u_+(t)$ given the past $p(t)$.

Proof: The proof consists in showing that $\mathcal{P}_{t-} \cap \mathcal{U}_{t+} = \{0\}$ so that according to Lemma 1, the orthogonal projection $f(t)\,|\mathcal{P}_{t-} \vee \mathcal{U}_{t+}$ splits uniquely into the sum of the oblique projections defined above. This zero intersection property will be proven in Appendix A. □

Denoting, as in Picci and Katayama (1996), by $\hat{E}_{\|\mathcal{B}}\{v|\mathcal{A}\}$ the oblique projection of the random variable v onto the subspace $\mathcal{A}$ along the subspace $\mathcal{B}$, the oblique projections in (4) can be written as

$$\Pi p(t) = \hat{E}_{\|\mathcal{U}_{t+}}\{f(t) \mid \mathcal{P}_{t-}\}, \quad \Phi u_+(t) = \hat{E}_{\|\mathcal{P}_{t-}}\{f(t) \mid \mathcal{U}_{t+}\} \tag{6}$$

where, in matrix representations, Π is a matrix with kp rows and an infinite number of columns and Φ is a $pk \times km$ matrix.

In the next section, we shall show that the first oblique projection in (6) spans the state space of a minimal state-space model of y which is causal with respect to u. Causality has to do with the fact that the second oblique projection in (6) admits, under the stated assumptions, a *causal* representation.

Lemma 2. Suppose there is no feedback from y to u. Then, for all $h \geq 0$,

$$\hat{E}\{y(t+h) \mid \mathcal{P}_{t-} \vee \mathcal{U}_{t+}\} = \hat{E}\{y(t+h) \mid \mathcal{P}_{t-} \vee \mathcal{U}_{[t,t+h]}\} \tag{7}$$

where $\mathcal{U}_{[t,t+h]} = \text{span}\,\{u(t), \cdots, u(t+h)\}$. Consequently Φ has the lower-triangular block Toeplitz structure

$$\Phi = \begin{bmatrix} G_0 & & & & \\ G_1 & G_0 & & & \\ G_2 & G_1 & G_0 & & \\ \vdots & \vdots & & \ddots & \\ G_{k-1} & G_{k-2} & & \cdots & G_0 \end{bmatrix} \tag{8}$$

where $\{G_0,\, G_1, \cdots\}$ can be interpreted as impulse response matrices of a causal system with input u. In other words, Φ is a *causal* operator.

Proof: The proof is deferred to Appendix B. □

3 Canonical Correlation Analysis and Construction of the State Vector

In this section we shall define the state vector for the process y by using a generalization of the method of CCA (Akaike, 1975; Desai *et al.*, 1985; Lindquist and Picci, 1996b) applied to the future outputs and the past observations, with the effect of the future inputs removed.

In the following we shall make an explicit assumption of finite dimensionality. Note that by stationarity $\Sigma_{fp|u}$ is a semi-infinite block Hankel matrix whose rank is a non-decreasing function of k (the future horizon). If $\text{rank}\Sigma_{fp|u}$ stays constant after appending one block row, then a standard Hankel-matrix argument shows that it will stay constant after appending any number of block rows (i.e. after adjoining any number of future output variables after time k). In this case we say that $\Sigma_{fp|u}$ has *finite rank*. This we shall assume from now on.

Let $\text{rank}\Sigma_{fp|u} = n$, and suppose that k is chosen sufficiently large, in particular large enough so that[2] $k > n$. Consider the Cholesky factorizations $\Sigma_{pp|u} = L_p L_p'$ and $\Sigma_{ff|u} = L_f L_f'$. Define

$$\varepsilon_+(t) := L_f^{-1}(f\,|u_+^\perp)(t), \qquad \varepsilon_-(t) := L_p^{-1}(p\,|u_+^\perp)(t)$$

so that

$$E\{\varepsilon_+(t)\varepsilon_-'(t)\} = L_f^{-1}\Sigma_{fp|u}(L_p^{-1})'.$$

Suppose that the SVD of the normalized block Hankel matrix $L_f^{-1}\Sigma_{fp|u}(L_p^{-1})'$ is given by

$$L_f^{-1}\Sigma_{fp|u}(L_p^{-1})' = U\Sigma V' \tag{9}$$

where $U'U = I_n$, $V'V = I_n$ and $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_n)$ is a diagonal matrix with nonzero singular values ($1 \geq \sigma_1 \geq \cdots \geq \sigma_n > 0$).

We now define

$$\alpha(t) := V'L_p^{-1}(p\,|u_+^\perp)(t), \quad \beta(t) := U'L_f^{-1}(f\,|u_+^\perp)(t)$$

Note that the vectors $\alpha(t)$ and $\beta(t)$ are of dimension n. Since $E\{\alpha(t)\alpha(t)'\} = E\{\beta(t)\beta(t)'\} = I_n$, and $E\{\beta(t)\alpha(t)'\} = \text{diag}(\sigma_1, \cdots, \sigma_n)$, we see that $\alpha(t)$, $\beta(t)$ are canonical variates and the singular values are canonical correlation coefficients

[2]Of course, with real data where n is normally not known, determining whether a chosen k satisfies this condition is a very nontrivial question. We shall not discuss this question in this paper.

between the random vectors $(f\,|u_+^\perp)(t)$ and $(p\,|u_+^\perp)(t)$. In other words, $\alpha(t)$, $\beta(t)$ are *conditional* canonical variates between the past $p(t)$ and the future $f(t)$ *given the future inputs* $u_+(t)$.

Following a standard procedure, see e.g. Desai *et al.* (1985), Aoki (1990), Lindquist and Picci (1996b), we define the extended observability and controllability matrices

$$\mathcal{O} := L_f U \Sigma^{1/2}, \quad \mathcal{C} := \Sigma^{1/2} V' L_p' \tag{10}$$

where $\text{rank}\mathcal{O} = \text{rank}\mathcal{C} = n$. Then from the SVD of (9), the block Hankel matrix $\Sigma_{fp|u}$ has the classical rank factorization

$$\Sigma_{fp|u} = \mathcal{O}\mathcal{C} \tag{11}$$

If the *state vector* is now defined to be the n-dimensional vector

$$x(t) = \mathcal{C}\Sigma_{pp|u}^{-1} p(t) = \Sigma^{1/2} V' L_p^{-1} p(t) \tag{12}$$

it is readily seen that $x(t)$ is a basis for the *stationary oblique predictor space*

$$\mathfrak{X}_t := \hat{E}_{\parallel \mathcal{U}_{t+}}\{\mathcal{Y}_{t+} \mid \mathcal{P}_{t-}\}$$

which, on the basis of general geometric principles (Picci and Katayama, 1996) can be shown to be a minimal state space for the process y. In order to show that $x(t)$ is indeed a basis note, from (5), (12), that the oblique projection of the future outputs onto the past can be expressed as

$$\hat{E}_{\parallel \mathcal{U}_{t+}}\{f(t) \mid \mathcal{P}_{t-}\} = \Pi p(t) = \Sigma_{fp|u}\Sigma_{pp|u}^{-1} p(t) := \mathcal{O}x(t) \tag{13}$$

and since $\text{rank}\mathcal{O} = n$ and the variance matrix of $x(t)$ is nonsingular, the claim follows.

In terms of $x(t)$, the optimal predictor $\hat{f}(t)$ of (4) has the form

$$\hat{f}(t) = \mathcal{O}x(t) + \Phi u_+(t) \tag{14}$$

where Φ is a causal operator (see Lemma 2). It is seen that $x(t)$ is a conditional minimal sufficient statistic carrying exactly all the information contained in $\mathcal{P}_{t-}$ that is necessary for the purpose of estimating the future outputs, *given the future inputs* after time t.

For later reference the state property of $x(t)$, implicit in relation (14), will be stated explicitly below.

Proposition. The process $x = \{x(t)\}$ is conditionally Markov given the future of the input; in fact

$$\hat{E}_{\parallel \mathcal{U}_{t+}}\{x(t+h) \mid \mathcal{P}_{t-}\} = \hat{E}_{\parallel \mathcal{U}_{t+}}\{x(t+h) \mid \mathbf{X}_t\}, \qquad h \geq 0,$$

and an analogous relation holds with $\mathbf{X}_{t-}$ (the past space of x at time t) in place of $\mathcal{P}_{t-}$.

Proof: Rewrite formula (14) shifted in time at time $t+h$,

$$\begin{bmatrix} \hat{y}(t+h) \\ \hat{y}(t+h+1) \\ \vdots \\ \hat{y}(t+h+k-1) \end{bmatrix} = \mathcal{O}_k x(t+h) + \Phi_k u_+(t+h) \tag{15}$$

where $h \geq 0$ is arbitrary. We have introduced the subscript k to denote explicitly the number of block rows in the matrices (i.e. the number of future outputs considered in the future horizon). Similarly, assuming $f(t)$ is made by stacking $k+h$ future output variables, and assuming a consistent choice of bases,

$$\begin{bmatrix} \hat{y}(t) \\ \vdots \\ \hat{y}(t+h-1) \\ \hat{y}(t+h) \\ \vdots \\ \hat{y}(t+h+k-1) \end{bmatrix} = \begin{bmatrix} \mathcal{O}_h \\ \mathcal{O}_{h|k} \end{bmatrix} x(t) + \begin{bmatrix} \Phi_h \\ \Phi_{h|k} \end{bmatrix} u_+(t) \tag{16}$$

where $\mathcal{O}_{h|k}$ and $\Phi_{h|k}$ are matrices made of the last k block rows of $\mathcal{O}_{h+k}$ and Φ_{h+k}, respectively. Now, by construction $\mathcal{O}_k$ has linearly independent columns so that there is an $n \times n$ matrix A_h such that

$$\mathcal{O}_{h|k} = \mathcal{O}_k A_h$$

so that we can write the last k block rows of (16) as

$$\begin{bmatrix} \hat{y}(t+h) \\ \vdots \\ \hat{y}(t+h+k-1) \end{bmatrix} = \mathcal{O}_k A_h x(t) + \Phi_{h|k} u_+(t). \tag{17}$$

Now the claim of the proposition is proven by applying the operator $\hat{E}_{\parallel \mathcal{U}_{t+}}\{\cdot \mid \mathcal{P}_{t-}\}$ on both sides of (15) and (17). The conditional Markov property follows since $\mathbf{X}_{t-} \subset \mathcal{P}_{t-}$. $\square$

Remark 1. Unlike the classical CCA realization procedure, the state vector here is *not* a vector of canonical variates. It may seem more natural to define the state vector as, say, $\Sigma^{1/2}\alpha(t)$, namely

$$\bar{x}(t) = \mathcal{C}\Sigma^{-1}_{pp|u}(p\,|u_+^{\perp})(t) \tag{18}$$

but this definition does not provide a basis in the state space of a causal model. In fact, letting

$$\Pi(p\,|u_+^{\perp})(t) = \mathcal{O}C\Sigma^{-1}_{pp|u}(p\,|u_+^{\perp})(t) = \mathcal{O}\bar{x}(t)$$

we find a different representation of the optimal predictor of the form

$$\hat{f}(t) = \mathcal{O}\bar{x}(t) + Hu_+(t)$$

where

$$Hu_+(t) := (f\,|u_+)(t) = E\{fu_+{}'\}E\{u_+u_+{}'\}^{-1}u_+(t).$$

The operator H, however, no longer has the causal (lower triangular) structure of (8) unless u is a white noise. So the definition (18) would lead to non-causal models. If the input u is a white noise, then $(p\,|u_+^{\perp})(t) = p(t)$ and in this case $\bar{x}(t)$ agrees with $x(t)$. □

3.1 The Stationary Innovation Model

We now derive the forward innovation state-space model for y by using the state vector $x(t)$ of (12) and the optimal predictor $\hat{f}(t)$ of (14).

Let $\mathcal{U}_t$ be the subspace spanned by the present input $u(t)$. Since, from Lemma 2 the first p rows of (14) involve just the one-step predictor of $y(t)$ based on $\mathcal{P}_{t-} \vee \mathcal{U}_t$, we can write

$$\hat{y}(t) := y(t)\,|\mathcal{P}_{t-} \vee \mathcal{U}_t = Cx(t) + Du(t)$$

where C and D are $p \times n$ and $p \times m$ constant matrices.

Define the prediction error by

$$e(t) := y(t) - (y(t)\,|\mathcal{P}_{t-} \vee \mathcal{U}_t) \tag{19}$$

From this definition, we get the output equation

$$y(t) = Cx(t) + Du(t) + e(t)$$

It should be noted that the prediction error $e(t)$ is stationary, since $\{y(t),\ u(t)\}$ are jointly stationary and the projection $y(t)\,|\mathcal{P}_{t-} \vee \mathcal{U}_t$ is based on observations from the infinite past.

The prediction error $e(t)$ is a white noise process uncorrelated with the past outputs $\{y(t-1),\, y(t-2), \cdots\}$ and the present and past inputs $\{u(t),\, u(t-1), \cdots\}$, namely

$$e(t) \;\perp\; \mathcal{Y}_{t-} \;\vee\; \mathcal{U}_{(t+1)-}$$

The orthogonality follows from the definition (19). That $e(t)$ is white also follows readily since $e(t+1) \perp \mathcal{Y}_{(t+1)-} \vee \mathcal{U}_{(t+2)-}$ and the definition (19) implies $e(t) \in \mathcal{P}_{(t+1)-} = \mathcal{Y}_{(t+1)-} \vee \mathcal{U}_{(t+1)-} \subset \mathcal{Y}_{(t+1)-} \vee \mathcal{U}_{(t+2)-}$. Therefore $e(t+1) \perp e(t)$.

We now consider the dynamics of the state vector $x(t)$. Let

$$\tilde{x}(t+1) := x(t+1) - (x(t+1)\,|\mathfrak{X}_t \vee \mathcal{U}_t)$$

where $\mathfrak{X}_t$ is the subspace generated by $x(t)$. The last term in the right-hand side is a sum of oblique projections onto finite dimensional spaces. Hence there exist $n \times n$ and $n \times m$ constant matrices A and B such that in the chosen basis we have

$$x(t+1)\,|\mathfrak{X}_t \vee \mathcal{U}_t = Ax(t) + Bu(t).$$

Thus the state equation is given by

$$x(t+1) = Ax(t) + Bu(t) + \tilde{x}(t+1) \tag{20}$$

Lemma 3. The prediction error $\tilde{x}(t+1)$ is a function of $e(t)$, and hence there exists a matrix K such that

$$\tilde{x}(t+1) = Ke(t) \tag{21}$$

Proof: The state (conditional Markov) property implies that $x(t+1) \mid \mathfrak{X}_t \vee \mathcal{U}_t = x(t+1)|\mathcal{P}_{t-} \vee \mathcal{U}_t = x(t+1)|\mathcal{Y}_{t-} \vee \mathcal{U}_{(t+1)-}$. Hence $\tilde{x}(t+1)$ is the projection of $x(t+1)$ onto the orthogonal complement of $\mathcal{Y}_{t-} \vee \mathcal{U}_{(t+1)-}$. Since $x(t+1)$ belongs to

$$\begin{aligned}
\mathcal{P}_{(t+1)-} &:= \overline{\text{span}}\{y(t), y(t-1), \cdots;\, u(t), u(t-1), \cdots\} \\
&= \overline{\text{span}}\{e(t), y(t-1), \cdots;\, u(t), u(t-1), \cdots\} \\
&= \text{span}\{e(t)\} \oplus \mathcal{Y}_{t-} \vee \mathcal{U}_{(t+1)-}
\end{aligned}$$

showing that $e(t)$ spans the orthogonal complement of $\mathcal{Y}_{t-} \vee \mathcal{U}_{(t+1)-}$ in $\mathcal{P}_{(t+1)-}$, we see that $\tilde{x}(t+1) \in \text{span}\{e(t)\}$ and must be a function of $e(t)$. □

Theorem 2. Assume that $\text{rank}\Sigma_{fp|u} = n$ and suppose that there is no feedback from the output y to the input u. Then the output process $y(t)$ admits a minimal

stochastic realization of the form

$$x(t+1) = Ax(t) + Bu(t) + Ke(t) \tag{22}$$

$$y(t) = Cx(t) + Du(t) + e(t) \tag{23}$$

with a state vector $x(t)$ forming a basis in the oblique predictor space $\boldsymbol{\mathcal{X}}_t$. This model is called a (forward) *innovation realization* with exogenous input $u(t)$. □

All other minimal stationary models of y can be written (up to a basis change in the state space) as

$$\xi(t+1) = A\xi(t) + Bu(t) + Gw(t) \tag{24}$$

$$y(t) = C\xi(t) + Du(t) + Hw(t) \tag{25}$$

where w is a normalized white noise process of dimension greater or equal than p, uncorrelated with the input u, and the matrices A, B, C, D are the same as in (22), (23). The state vector $x(t)$ of the innovations model is the *steady-state Kalman filter estimate of* $\xi(t)$ i.e. the orthogonal projection

$$x(t) = \hat{E}\{\xi(t) \mid \mathcal{P}_{t-} \vee \mathcal{U}_{t+}\} = \hat{E}\{\xi(t) \mid \mathcal{P}_{t-}\}$$

The relation between the parameters G, H and $K, \Lambda := E\{e(t)e(t)'\}$ is given by the well-known *Positive Real Lemma* equations which can be found in many references. For these matters one may consult e.g. Faurre *et al.* (1979) and Lindquist and Picci (1991, 1996b).

3.2 Realization based on Finite Data

In practice the computations must be based on finite input-output data. The construction of the predictor space and of the relative innovation model should be based on the predictor of the future outputs $f(t)$, based on the available *finite* past history. Let the past stochastic vectors $u_-(t)$ and $y_-(t)$ be truncated to finite length[3], namely

$$u_-(t) := \begin{bmatrix} u(t-1) \\ u(t-2) \\ \vdots \\ u(t_0) \end{bmatrix}, \quad y_-(t) := \begin{bmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ y(t_0) \end{bmatrix}, \quad t_0 \le t \le T$$

[3]In Section 4, the initial time t_0 will be fixed equal to zero.

and let $\mathcal{P}_{[t_0,t)}$ denote the past data space spanned by the above $(t-t_0-1)(p+m)\times 1$ stochastic vector. The symbol $\mathcal{U}_{t+}$ will denote the (finite) future input history, $\mathcal{U}_{[t,T]}$, after time t.

From these data we can form the *finite-memory* predictor at time t

$$\begin{aligned}\hat{f}_{t_0}(t) &:= \hat{E}\{f(t) \mid \mathcal{P}_{[t_0,t)} \vee \mathcal{U}_{t+}\} \\ &= \hat{E}\{\hat{f}(t) \mid \mathcal{P}_{[t_0,t)} \vee \mathcal{U}_{t+}\} \\ &= \hat{E}_{\parallel \mathcal{U}_{t+}}\{\hat{f}(t) \mid \mathcal{P}_{[t_0,t)}\} + \hat{E}_{\parallel \mathcal{P}_{[t_0,t)}}\{\hat{f}(t) \mid \mathcal{U}_{t+}\}\end{aligned}$$

The following result, which we shall state without proof, explains the role of the transient Kalman filter in finite-data modeling and is in the same spirit of Theorem 6 in Lindquist and Picci (1996b). It relates also to Theorem 3 of Van Overschee and De Moor (1994).

Theorem 3. Assume that $\text{rank}\Sigma_{fp|u} = n$ and suppose that there is no feedback from the output y to the input u. The process $y(t)$ admits a nonstationary realization (called the transient innovation representation) of the form

$$\hat{x}_{t_0}(t+1) = A\hat{x}_{t_0}(t) + Bu(t) + K(t)\hat{e}_{t_0}(t) \tag{26}$$

$$y(t) = C\hat{x}_{t_0}(t) + Du(t) + \hat{e}_{t_0}(t) \tag{27}$$

where the state vector $\hat{x}_{t_0}(t)$ is a basis in the finite-memory predictor space

$$\hat{\mathbf{X}}_t = \hat{E}\{\mathbf{X}_t \mid \mathcal{P}_{[t_0,t)} \vee \mathcal{U}_{t+}\}$$

and the process $\{\hat{e}_{t_0}(t) \mid t \geq t_0\}$ is the transient innovation (i.e. the prediction error) of $\{y(t) \mid t \geq t_0\}$ with respect to the information $\mathcal{P}_{[t_0,t)} \vee \mathcal{U}_{t+}$.

Any basis $\hat{x}_{t_0}(t)$ in $\hat{\mathbf{X}}_t$ has a unique representation as

$$\hat{x}_{t_0}(t) = \hat{E}\{x(t) \mid \mathcal{P}_{[t_0,t)} \vee \mathcal{U}_{t+}\} \tag{28}$$

where $x(t)$ is a basis in the stationary predictor space $\mathbf{X}_t$ and hence $\hat{x}_{t_0}(t)$ is also the transient Kalman filter estimate of $\xi(t)$, the state of any stationary model (24), (25), given the data $\mathcal{P}_{[t_0,t)} \vee \mathcal{U}_{t+}$. The initial condition for (26) is

$$\hat{x}_0 := \hat{x}_{t_0}(t_0) = \hat{E}\{x(t_0) \mid \mathcal{U}_{t_0+}\}$$

The constant matrices A, B, C, D are the same as in (22), (23) while $K(t)$ is the transient Kalman gain given by

$$K(t) = (AP(t)C' + GH')(CP(t)C' + HH')^{-1}$$

where $P(t)$ is the solution of the Riccati equation

$$P(t+1) = AP(t)A' - (AP(t)C' + GH')(CP(t)C' + HH')^{-1}(AP(t)C' + GH')' + GG'$$

with initial condition $P(t_0)$ equal to the covariance matrix of $x(t_0) - \hat{x}_0$. The matrix $P(t)$ can be interpreted as the error covariance matrix

$$P(t) = E\{[\xi(t) - \hat{x}_{t_0}(t)]\,[\xi(t) - \hat{x}_{t_0}(t)]'\}.$$

As $t_0 \to -\infty$ the state of the transient innovation model (26), (27) converges to $x(t)$, $K(t) \to K$ and $P(t)$ to the unique stabilizing solution of the algebraic Riccati equation (ARE)

$$P = APA' - (APC' + GH')(CPC' + HH')^{-1}(APC' + GH')' + GG' \tag{29}$$

Remark 2. The conditional CCA procedure of Section 3 applied to finite past data provides an approximation of the state vector $\hat{x}_{t_0}(t)$ differing by an additive initial condition term which tends to zero as $t_0 \to -\infty$. One has

$$\hat{x}_{t_0}(t) = \hat{E}_{\parallel \mathcal{U}_{t+}}\{x(t) \mid \mathcal{P}_{[t_0,t)}\} + \hat{E}_{\parallel \mathcal{P}_{[t_0,t)}}\{x(t) \mid \mathcal{U}_{t+}\}$$

and the first component on the right-hand side is an oblique projection which can be obtained by conditional CCA of future and finite past data (this is so because $x(t) = \mathcal{O}^{\dagger}\hat{E}_{\parallel \mathcal{U}_{t+}}\{f(t) \mid \mathcal{P}_{t-}\}$, see (13)). The second component tends to zero for $t_0 \to -\infty$ by absence of feedback, since in this case, $x(t) \in \mathcal{P}_{t-}$, and the oblique projection of $x(t)$ along $\mathcal{P}_{[t_0,t)}$ tends to the oblique projection along $\mathcal{P}_{t-}$ which is clearly zero.

Note that the second term can be expressed as a function of the initial condition $x(t_0)$ as

$$\hat{E}_{\parallel \mathcal{P}_{[t_0,t)}}\{x(t) \mid \mathcal{U}_{t+}\} = (A - KC)^{t-t_0}\hat{E}_{\parallel \mathcal{P}_{[t_0,t)}}\{x(t_0) \mid \mathcal{U}_{t+}\},$$

and this quantity tends to zero as $t_0 \to -\infty$, irrespective of absence of feedback, if the (true) system has no zeros on the unit circle (a generic condition).

Hence, by taking a sufficiently large number of lags $t - t_0$, the oblique projection $\hat{E}_{\parallel \mathcal{U}_{t+}}\{f(t) \mid \mathcal{P}_{[t_0,t)}\}$ (which is computed by the CCA procedure) provides, generically, an arbitrarily close approximation of the state of the transient innovation representation (26), (27). □

4 Identification based on Finite Data: I. Computing Oblique Projections

In this section we derive two methods for computing the oblique projections Π and Φ based on given finite input-output data. One is to solve the Wiener-Hopf equations (5) by using the LQ factorization. Since for the finite data case the estimated Φ is generally not lower triangular, we develop another method of estimating Π and Φ by using a constrained least-squares (CLS) technique, where all the elements of the upper right part of Φ are fixed to zero. Once we have estimates of Π and Φ, we can easily derive subspace methods to get the system parameters A, B, C, D and K. This will be discussed in the next section.

Suppose that finite input-output data[4] $u(t)$, $y(t)$ for $t = 0, 1, \cdots, N + 2k - 2$ are given with $k > 0$ and N sufficiently large. We assume that the time series $\{u(t), y(t)\}$ is made of sample values of two jointly stationary processes (u, y) satisfying the assumptions of the previous sections, in particular the finite dimensionality and the feedback-free conditions. In addition, we shall assume throughout this section that the sample averages converge to the "true" expected values as $N \to \infty$.

Define the $km \times N$ and $kp \times N$ block Hankel matrices with N columns

$$U_{0|k-1} := \begin{bmatrix} u(0) & u(1) & \cdots & u(N-1) \\ u(1) & u(2) & \cdots & u(N) \\ \vdots & \vdots & & \vdots \\ u(k-1) & u(k) & \cdots & u(N+k-2) \end{bmatrix}$$

$$Y_{0|k-1} := \begin{bmatrix} y(0) & y(1) & \cdots & y(N-1) \\ y(1) & y(2) & \cdots & y(N) \\ \vdots & \vdots & & \vdots \\ y(k-1) & y(k) & \cdots & y(N+k-2) \end{bmatrix}$$

where the first and the second subscripts denote the indices of the upper-left and the lower-left element of the block Hankel matrix, respectively. In what follows we assume that the integer k is chosen so that $k > n$, where n is the dimension of the underlying stochastic system generating the data.

As we have seen, in finite-data realization the number of past lags $(t - t_0)$ plays a different role than the extent of the future horizon k. Although it is obviously possible to consider past data on a time interval of different length, in this section we shall take the past and the future intervals of the same length k. This choice

[4] In the following we shall work with the *sample values* of the processes (u, y). We shall denote them by the same symbol as the corresponding random quantities.

conforms with what is commonly assumed in the literature and is not really a loss of generality. With this choice the present time is $t = k$, the past interval is $[0,\ k-1]$, the future interval is $[k,\ 2k-1]$ and notations simplify considerably. Corresponding to this subdivision, the future input and the future output data are arranged in the matrices $U_{k|2k-1} \in \mathbb{R}^{km\times N}$ and $Y_{k|2k-1} \in \mathbb{R}^{kp\times N}$, respectively. Also, the matrix of past input and output data is defined as

$$P_{0|k-1} = \begin{bmatrix} U_{0|k-1} \\ Y_{0|k-1} \end{bmatrix} \in \mathbb{R}^{k(m+p)\times N}$$

The operator Π becomes a $kp \times k(p+m)$ matrix in the following.

4.1 Covariance factorization method

In the following sections it is enough to assume that the input is persistently exciting, so that $U_{0|2k-1}$ has row rank $2km$. Consider the following LQ factorization (Verhaegen and Dewilde, 1992; Van Overschee and De Moor, 1994):

$$\frac{1}{\sqrt{N}} \begin{bmatrix} U_{k|2k-1} \\ P_{0|k-1} \\ Y_{k|2k-1} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix} \begin{bmatrix} Q_1' \\ Q_2' \\ Q_3' \end{bmatrix} := LQ' \tag{30}$$

where L_{11}, L_{22} and L_{33} are $km\times km$, $k(p+m)\times k(p+m)$ and $kp\times kp$ lower triangular matrices, respectively, and Q_i's are orthogonal matrices with $Q_i'Q_j = I\delta_{ij}$.

It follows from (30) that the sample covariance matrix [5]

$$\Sigma_{upf} = \frac{1}{N} \begin{bmatrix} U_{k|2k-1} \\ P_{0|k-1} \\ Y_{k|2k-1} \end{bmatrix} \begin{bmatrix} U_{k|2k-1} \\ P_{0|k-1} \\ Y_{k|2k-1} \end{bmatrix}'$$

can be computed as

$$\Sigma_{upf} := \begin{bmatrix} \Sigma_{uu} & \Sigma_{up} & \Sigma_{uf} \\ \Sigma_{pu} & \Sigma_{pp} & \Sigma_{pf} \\ \Sigma_{fu} & \Sigma_{fp} & \Sigma_{ff} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix} \begin{bmatrix} L_{11}' & L_{21}' & L_{31}' \\ 0 & L_{22}' & L_{32}' \\ 0 & 0 & L_{33}' \end{bmatrix} \tag{31}$$

Conversely, the Cholesky factorization of the sample covariance matrix Σ_{upf} gives L of (30) up to an orthogonal transform. It can be shown from (31) that, since Σ_{uu} is nonsingular,

$$\begin{aligned} \Sigma_{ff|u} &= \Sigma_{ff} - \Sigma_{fu}\Sigma_{uu}^{-1}\Sigma_{uf} \\ &= L_{32}L_{32}' + L_{33}L_{33}' \end{aligned} \tag{32}$$

[5] For simplicity, all sample covariance matrices are written as Σ_{ff}, $\Sigma_{ff|u}$, etc. without using the $\hat{(\cdot)}$ notation. Also, it should be kept in mind that these quantities depend on k. For notational reasons this dependence will not be displayed.

Similarly, we have

$$\Sigma_{pp\,|u} = L_{22}L'_{22}, \quad \Sigma_{fp\,|u} = L_{32}L'_{22} \tag{33}$$

It therefore follows from (5) that

$$\Pi = \Sigma_{fp\,|u}(\Sigma_{pp\,|u})^{\dagger} = L_{32}L_{22}^{\dagger} \tag{34}$$

In order to obtain Φ, we rearrange the covariance matrices as

$$\Sigma_{puf} := \begin{bmatrix} \Sigma_{pp} & \Sigma_{pu} & \Sigma_{pf} \\ \Sigma_{up} & \Sigma_{uu} & \Sigma_{uf} \\ \Sigma_{fp} & \Sigma_{fu} & \Sigma_{ff} \end{bmatrix} = \begin{bmatrix} R_{11} & 0 & 0 \\ R_{21} & R_{22} & 0 \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \begin{bmatrix} R'_{11} & R'_{21} & R'_{31} \\ 0 & R'_{22} & R'_{32} \\ 0 & 0 & R'_{33} \end{bmatrix}$$

Thus, similarly to the above, we have

$$\Sigma_{ff\,|p} = R_{32}R'_{32} + R_{33}R'_{33}$$

$$\Sigma_{uu\,|p} = R_{22}R'_{22}, \quad \Sigma_{fu\,|p} = R_{32}R'_{22}$$

Therefore, from (5) we get

$$\Phi = \Sigma_{fu\,|p}(\Sigma_{uu\,|p})^{\dagger} = R_{32}R_{22}^{\dagger} \tag{35}$$

A problem with the above solution is that Φ may not be lower triangular. This is circumvented by using the CLS method.

4.2 Constrained least-squares method

In this subsection, we develop a method of estimating Π and Φ under the constraint that Φ of (8) has a lower-triangular Toeplitz form. The idea is borrowed from Peternell *et al.* (1996).

We see that (4) results from a regression equation of the form

$$f(t) = [\Pi \;\; \Phi] \begin{bmatrix} p(t) \\ u_+(t) \end{bmatrix} + v(t)$$

where $v(t)$ (the prediction error) is the residual. For the finite-data case, this relation can be written as

$$Y_{k|2k-1} = \Theta \begin{bmatrix} P_{0|k-1} \\ U_{k|2k-1} \end{bmatrix} + V_{k|2k-1}$$

where $\Theta := [\Pi \;\; \Phi] \in \mathbb{R}^{kp \times k(p+2m)}$. By using (30), we get

$$\frac{1}{\sqrt{N}} Y_{k|2k-1} = L_y Q', \quad \frac{1}{\sqrt{N}} \begin{bmatrix} P_{0|k-1} \\ U_{k|2k-1} \end{bmatrix} = L_{pu} Q'$$

so that the regression equation is reduced to

$$L_y = \Theta L_{pu} + V_Q \tag{36}$$

where

$$\begin{aligned} L_y &= [\, L_{31} \;\; L_{32} \;\; L_{33} \,] \in \mathbb{R}^{kp \times 2k(p+m)} \\ L_{pu} &= \begin{bmatrix} L_{21} & L_{22} & 0 \\ L_{11} & 0 & 0 \end{bmatrix} \in \mathbb{R}^{k(p+2m) \times 2k(p+m)} \\ V_Q &= \frac{1}{\sqrt{N}} V_{k|2k-1} Q \in \mathbb{R}^{kp \times 2k(p+m)} \end{aligned}$$

Thus vectorizing (36) yields

$$\mathrm{vec}(L_y) = Z\mathrm{vec}(\Theta) + \mathrm{vec}(V_Q) \tag{37}$$

where $Z = L'_{pu} \otimes I_{kp} \in \mathbb{R}^{2k^2 p(p+m) \times k^2 p(p+2m)}$ and where $\otimes$ denotes the Kronecker product (see Appendix C).

It can be shown that there exists F_{22} with entries equal to zero or one such that

$$F_{22}\mathrm{vec}(G_0, \cdots, G_{k-1}) = \mathrm{vec} \begin{bmatrix} G_0 & & & \\ G_1 & G_0 & & \\ \vdots & \vdots & \ddots & \\ G_{k-1} & G_{k-2} & \cdots & G_0 \end{bmatrix} = \mathrm{vec}(\Phi)$$

Defining $F = \begin{bmatrix} I & \\ 0 & F_{22} \end{bmatrix}$, we get

$$\mathrm{vec}(\Theta) = F \begin{bmatrix} \mathrm{vec}(\Pi) \\ \mathrm{vec}(G_0, \cdots, G_{k-1}) \end{bmatrix} := F\theta$$

in terms of the vector parameter θ. Hence we see from (37) that

$$\mathrm{vec}(L_y) = ZF\theta + \mathrm{vec}(V_Q) \tag{38}$$

Therefore the CLS estimate of Θ is given by a simple multiplication by F of the LS solution of (38).

5 Identification based on Finite Data: II. Stochastic Subspace Methods

By using the estimates Π, Φ and the conditional covariance matrices $\Sigma_{ff|u}$, $\Sigma_{pp|u}$, etc., we can design subspace methods for obtaining estimates of the system parameters A, B, C, D and K. We describe two classes of covariance factorization based subspace methods, in which the first two steps are the same.

Subspace Identification Methods

We assume that the conditional covariance matrices $\Sigma_{ff|u}$, $\Sigma_{pp|u}$, $\Sigma_{fp|u}$ are given by (32) and (33).

Step 1: Compute the SVD of the normalized covariance matrix [see (9)]

$$\Sigma_{ff|u}^{-1/2}\Sigma_{fp|u}(\Sigma_{pp|u}^{-1/2})' = USV' \simeq \hat{U}\hat{S}\hat{V}'$$

to get

$$\Sigma_{fp|u} \simeq \Sigma_{ff|u}^{1/2}\hat{U}\hat{S}\hat{V}'(\Sigma_{pp|u}^{1/2})'$$

where $\hat{S}$ is obtained by neglecting singular values which are "nearly zero", so that the dimension of the state vector equals $\dim \hat{S}$ [6].

Step 2: Set the extended observability and controllability matrices as [see (10)]

$$\mathcal{O}_k = \Sigma_{ff|u}^{1/2}\hat{U}\hat{S}^{1/2}, \quad \mathcal{C}_k = \hat{S}^{1/2}\hat{V}'(\Sigma_{pp|u}^{1/2})'$$

Algorithm A: Regression approach using the state vector

Step A3: The estimate of the state vector is given by [see (12), (28)]

$$\hat{X}_k = \mathcal{C}_k\Sigma_{pp|u}^{-1}P_{0|k-1} = \hat{S}^{1/2}\hat{V}'\Sigma_{pp|u}^{-1/2}P_{0|k-1}$$

Note that this state vector is really an $n \times N$ matrix.

Step A4: The matrices (A, B, C, D) are obtained by applying the least-squares (LS) method to the overdetermined equations

$$\begin{bmatrix} \hat{X}_{k+1} \\ Y_{k|k} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \hat{X}_k \\ U_{k|k} \end{bmatrix} + \begin{bmatrix} W_k \\ E_k \end{bmatrix}$$

where $\hat{X}_{k+1}$ is the CCA state vector computed at time $k+1$. In order to be the time update of the CCA state selected at time k, this vector should be transformed by a change of basis so as to keep the same observability matrix at time k and $k+1$, see e.g. Van Overschee and De Moor (1993).

Because of the transient Kalman Filter realization (see Theorem 3), even if the past is finite and $\hat{X}_k$ is non-stationary, this LS procedure provides the correct[7] estimates of the parameters (A, B, C, D) of the stationary model (22), (23). If the

[6] This empirical procedure of order estimation is not very satisfactory and better justified statistical methods are currently under study. On this point we shall however have to refer the reader to the literature.

[7] In fact, consistent for $N \to \infty$ for fixed k, provided k is chosen suitably large.

integer k is large enough, the state vector is approximately stationary. Then $\hat{X}_{k+1}$ may be approximated by the one-step shifted version of the vector $\hat{X}_k$ obtained in Step A3.

The matrices W_k, E_k are residuals which represent LS estimates of model and observation noise terms.

Step A5: Define the covariance matrices of residuals as

$$\begin{bmatrix} \Sigma_{ww} & \Sigma_{we} \\ \Sigma_{ew} & \Sigma_{ee} \end{bmatrix} = \frac{1}{N} \begin{bmatrix} W_k W_k' & W_k E_k' \\ E_k W_k' & E_k E_k' \end{bmatrix}$$

and solve the ARE (compare (29))

$$P = APA' - (APC' + \Sigma_{we})(CPC' + \Sigma_{ee})^{-1}(APC' + \Sigma_{we})' + \Sigma_{ww} \qquad (39)$$

to get the stabilizing solution $P \geq 0$ and the corresponding Kalman gain

$$K = (APC' + \Sigma_{we})(CPC' + \Sigma_{ee})^{-1}$$

making $A - KC$ is stable.

This procedure is correct for infinitely long past data and $N \to \infty$. In this case, it follows from Lemma 3 that the exact relations $\Sigma_{ee} = \Lambda = E\{e(t)e(t)'\}$, $\Sigma_{ww} = K\Sigma_{ee}K'$, $\Sigma_{we} = K\Sigma_{ee}$ (constant with k) should hold, so that the unique stabilizing solution of the ARE (39) exists and is actually zero.

For the finite data case these exact relations do not hold and the sample covariance matrices computed in Step A5 vary with k. However, under the standing assumption that the data are generated by a true system of order n, if N and k are chosen large enough with $N \gg k$ (i.e. so that in particular the order is estimated correctly), then the procedure provides consistent estimates. Note that the ARE (39) has a unique stabilizing solution $P \geq 0$ from which we can estimate K. This is so, since by construction of the extended observability matrix $\mathcal{O}_k$, the pair (C, A) is observable and we have $W_k W_k' > 0$ generically.

For further discussion of this point and the related positivity issue we refer the reader to Section 7 of Lindquist and Picci (1996a).

Algorithm B: Realization approach

Step B3: Compute the estimates of C and A by (Verhaegen and Dewilde, 1992)

$$C = \mathcal{O}_k(1:p, 1:n)$$

$$A = \mathcal{O}_k(p+1:kp, 1:n)^{\dagger}\mathcal{O}_k(1:(k-1)p, 1:n)$$

Step B4: For given (A, C), and Φ of (35), it follows from (8) that the estimates of B and D are given by solving

$$\left[\begin{array}{c|c} I_p & 0 \\ 0 & \mathcal{O}(1:p(k-1),1:n) \\ \hline I_p & 0 \\ 0 & \mathcal{O}(1:p(k-2),1:n) \\ \hline \vdots & \vdots \\ \hline I_p & 0 \end{array}\right] \left[\begin{array}{c} D \\ \hline B \end{array}\right] = \left[\begin{array}{c} \Phi(1:pk,1:m) \\ \hline \Phi(p+1:pk,m+1,2m) \\ \hline \vdots \\ \hline \Phi(p(k-1):pk,m(k-1):km) \end{array}\right]$$

In the same way, we get another two subspace identification methods by combining Algorithms A and B above and the CLS-based method of solving Wiener-Hopf equations. Hence, we get four subspace identification methods based on our realization procedures for stochastic systems with exogenous inputs.

6 Simulation Results

Some results of computer simulations are presented to show the performance of the stochastic subspace identification algorithms developed in Section 5. We consider a fifth-order SISO model shown in Fig. 2, proposed in Viberg (1995), where u is the input, and w and v are additive white noises with mean zeros. The transfer function is given by $G(z) = B(z) \,/\, A(z)$, where

$$B(z) = 0.0275z^{-4} + 0.0551z^{-5}$$

$$A(z) = 1 - 2.3443z^{-1} + 3.081z^{-2} - 2.5274z^{-3} + 1.2415z^{-4} - 0.3686z^{-5}$$

Thus $G(z)$ has a zero at $z = -2$ and poles at $z = 0.9,\ 0.8e^{\pm j},\ 0.8e^{\pm 1.2j}$.

The performance of the algorithms is measured by the norm of the estimation error

$$I_N = \frac{1}{M}\sum_{l=1}^{M}\left(\sum_{j=1}^{10}[\theta_j - \hat{\theta}_j(l,N)]^2\right), \quad N = 200, 500, 1000, 2000$$

where θ_j denotes the true parameter and $\hat{\theta}_j(l,N)$ is the estimate of θ_j at l-th run with the number of data N, and M denotes the number of simulation runs.

The following algorithms are used for simulations.

COV-a: Compute Π, Φ by the covariance factorization and use Algorithm A.

COV-b: Compute Π, Φ as above and use Algorithm B.

CLS-a: Compute Π, Φ by the CLS method and use Algorithm A.

CLS-b: Compute Π, Φ as above and use Algorithm B.

Example 1 (White noise input)

In this case, the input u is a zero-mean white Gaussian noise with variance $\sigma_u^2 = 1$, where the variances of noises w and v are $\sigma_w^2 = (0.05)^2$, $\sigma_v^2 = (0.05)^2$. Fig. 3 shows the performance of COV-a and COV-b with $k = 8$, $M = 100$, together with results due to Basic 4SID (Verhaegen and Dewilde, 1992) and PO-MOESP (Verhaegen, 1994), where the performance of COV-b and PO-MOESP is indistinguishable in this experiment. The performance of CLS-a and CLS-b, not shown here, is pretty close to that of COV-a and CLS-a, respectively. Figs. 4 and 5 respectively depict the pole estimates by COV-a and PO-MOESP for $k = 8$, $M = 100$, $N = 1000$; they are quite close to the true pole positions (denoted by +).

Example 2 (Sinusoidal input)

To compare the effect of input signals, the input is now chosen as

$$u(t) = U_0 \sum_{i=1}^{20} \sin(\omega_i t)$$

where the frequencies ω_i's are uniformly spaced in the interval $(0.1, 3)$(rad) and where U_0 is adjusted to yield $\sigma_u^2 = 1$. The noise variances are chosen as $\sigma_w^2 = \sigma_v^2 = (0.05)^2$. Since the sinusoidal input above is PE of order 40, the number of rows k must be less than 20 [8]. Fig. 6 depicts the performance of five algorithms including PO-MOESP, where $k = 8$, $M = 100$. In this case, COV-a, COV-b and PO-MOESP show similar performance, but the performance of CLS-a and CLS-b is rather different from that of others. In order to analyze this fact, we have simulated CLS-a and CLS-b for several different k's, where $M = 50$, $N = 1000$. We see from Fig. 7 that both methods give similar performance for k greater than 10, but for the smaller k, CLS-a shows better performance. In Figs. 8 and 9, the pole estimates by COV-a and CLS-a are depicted for $k = 8$, $N = 1000$. □

We see from Figs. 3 and 6 that the white noise inputs give better performance over the sinusoidal inputs in that even the worst Basic 4SID performance for the white noise inputs is better than the best CLS-a performance for sinusoidal inputs. Moreover, the simulation results for the sinusoidal inputs widely vary with algorithms. In fact, COV-a gives a rather scattered pole estimates as shown in Fig. 8, but CLS-a yields better pole estimates with a smaller variability as shown in Fig. 9. It should be noted that the simulation results also crucially depend on the S/N ratio; in fact, although not shown here, if the noise variances are reduced, then the

[8] It is apparent that this input does not satisfy the condition (3). But it is enough to consider the input with a PE condition of sufficiently high order to apply the subspace methods.

Table 1: The number of flops required for 50 simulation runs

	COV-a	COV-b	CLS-a	CLS-b	PO-MOESP	
Flops	1.56	2.34	10.9	7.18	1.33	$(\times 10^8)$

performance and pole estimates by COV-a and CLS-a become much better than the results shown in Figs. 6, 8 and 9.

Table 1 shows the number of flops of five algorithms, where it includes all the computations for the whole simulations by each algorithm for $k = 8$, $M = 50$, $N = 1000$. Also, we see that CLS-a, CLS-b are computationally more expensive than COV-a, COV-b, respectively. This may be due to the fact that the regression equation of (38) for CLS method contains matrices F and Z of large dimensions.

7 Conclusions

In this paper we have presented a realization theory for stochastic system with exogenous inputs under the assumption that there is no feedback from the output to the control input. A generalized CCA technique is employed for defining the state vector and a forward innovation representation of a stochastic system with exogenous input is derived. From this we have developed four stochastic subspace identifications methods based on approximate solutions of two discrete Wiener-Hopf equations by covariance factorization and by CLS method. In simulation studies CLS-a yields the best performance among the four subspace identification algorithms presented here.

The present approach is meant to shed some light on well-known procedures of subspace identification (Larimore, 1990; Peternell *et al.*, 1996; Van Overschee and De Moor, 1994, 1996; Verhaegen and Dewilde, 1992; Verhaegen, 1994) and to allow comparisons. It seems to us that in the presence of stochastic dynamics, the subspace method based on the approximate solutions of the discrete Wiener-Hopf equations is a more natural approach than the procedures reported in the literature which are based on instrumental variable approaches.

Of course, numerical results may be different from those presented here, under different simulation conditions. In fact, for generic data, there may be cases in which our subspace methods fail. This is a known common feature of all subspace meth-

ods where the noise parameters are estimated by realization techniques (Dahlén, Lindquist and Mari, 1998), and may be regarded as a symptom that the model is "very far" from the data being modeled. Often this problem can be fixed by just choosing k large enough.

Acknowledgments

We would like to express our thanks to S. Omori and D. Huang for their assistance in simulation studies.

References

[1] Akaike, H. (1974). Stochastic theory of minimal realization. *IEEE Trans. Automat. Contr.*, **AC-19**, 667–674.

[2] Akaike, H. (1975). Markovian representation of stochastic processes by canonical variables. *SIAM J. Control*, **13**, 162–173.

[3] Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion. In *System Identification: Advances and Case Studies* (R. Mehra and D. Lainiotis, Eds.), Academic, 27–96.

[4] Aoki, M. (1990). *State Space Modeling of Time Series* (2nd ed.), Springer.

[5] Caines, P. E. and C. W. Chan (1976). Estimation, identification and feedback. In *System Identification: Advances and Case Studies* (R. Mehra and D. Lainiotis, Eds.), Academic, 349–405.

[6] Dahlén, A., A. Lindquist and J. Mari (1998). Experimental evidence showing that stochastic subspace identification methods fail. *Systems & Contr. Lett.*, **34**, 303–312.

[7] Desai, U. B., D. Pal and R. D. Kirkpatrick (1985). A realization approach to stochastic model reduction. *Int. J. Control*, **42**, 821–838.

[8] Faurre, P., M. Clerget and F. Germain (1979). *Opérateurs Rationnels Positifs*, Dunod.

[9] Gevers, M. R. and B. D. O. Anderson (1982). On jointly stationary feedback-free stochastic processes. *IEEE Trans. Automat. Contr.*, **AC-27**, 431–436.

[10] Golub, G. H. and C. R. Van Loan (1989). *Matrix Computations* (2nd ed.). The Johns Hopkins Univ. Press.

[11] Hannan, E. J. and D. S. Poskit (1988). Unit canonical correlations between future and past. *Annals Statistics*, **16**, 784–790.

[12] Jansson, M. and B. Wahlberg (1996). A linear regression approach to state-space subspace system identification. *Signal Processing*, **52**, 103–129.

[13] Ho, B. L. and R. E. Kalman (1966). Effective construction of linear state-variable models from input/output functions. *Regelungstechnik*, **14**, 545–548.

[14] Larimore, W. E. (1990). Canonical variate analysis in identification, filtering, and adaptive control. In *Proc. 29th IEEE Conf. Decision & Control*, Honolulu, 596–604.

[15] Lindquist, A. and G. Picci (1985). Realization theory for multivariable stationary Gaussian processes. *SIAM J. Control Optim.*, **23**, 809–857.

[16] Lindquist, A. and G. Picci (1991). A geometric approach to modelling and estimation of linear stochastic systems. *J. Math. Systems, Estimation & Control*, **1**, 241–333.

[17] Lindquist, A. and G. Picci (1996a). Geometric methods for state space identification. In *Identification, Adaptation, Learning* (Lectures given at the NATO-ASI School, *From Identification to Learning*, Como, Italy, August 1994), Springer, 1–69.

[18] Lindquist, A. and G. Picci (1996b). Canonical correlation analysis, approximate covariance extension, and identification of stationary time series. *Automatica*, **32**, 709–733.

[19] Ljung, L. (1987). *System Identification – Theory for the User*. Prentice-Hall.

[20] Moonen, M. and J. Vandewalle (1990). QSVD approach to on- and off-line state-space identification. *Int. J. Control*, **51**, 1133–1146.

[21] Peternell, K., W. Scherrer and M. Deistler (1996). Statistical analysis of novel subspace identification methods. *Signal Processing*, **52**, 161–177.

[22] Picci, G. (1997). Stochastic realization and system identification. In *Statistical Methods in Control and Signal Processing* (T. Katayama and S. Sugimoto, Eds.), Dekker, pp. 1–35.

[23] Picci, G. and T. Katayama (1996). Stochastic realization with exogenous inputs and "subspace methods" identification. *Signal Processing*, **52**, 145–160.

[24] Rozanov, N. I. (1963). *Stationary Random Processes*. Holden-Day.

[25] Van Overschee, P. and B. De Moor (1993). Subspace algorithms for the stochastic identification problem. *Automatica*, **29**, 649–660.

[26] Van Overschee, P. and B. De Moor (1994). N4SID: Subspace algorithms for the identification of combined deterministic - stochastic systems. *Automatica*, **30**, 75–93.

[27] Van Overschee, P. and B. De Moor (1996). *Subspace Identification for Linear Systems*. Kluwer Academic Publications.

[28] Verhaegen, M. (1994). Identification of the deterministic part of MIMO state space models given in innovations form from input–output data. *Automatica*, **30**, 61–74.

[29] Verhaegen, M. and P. Dewilde (1992). Subspace model identification, Part 1. The output-error state-space model identification class of algorithms; Part 2. Analysis of the elementary output-error state-space model identification algorithm. *Int. J. Control*, **56**, 1187–1210 & 1211–1241.

[30] Viberg, M. (1995). Subspace-based methods for the identification of linear time-invariant systems. *Automatica*, **31**, 1835–1851.

A Proof of Theorem 1

Recall that the feedback-free property

$$\hat{E}\{\mathcal{U}_{t+} \mid \mathcal{P}_{t-}\} = \hat{E}\{\mathcal{U}_{t+} \mid \mathcal{U}_{t-}\} \tag{A.1}$$

is equivalent to *conditional orthogonality* of the past of y and the future of u, given the past of u, which is denoted

$$\mathcal{Y}_{t-} \perp \mathcal{U}_{t+} \mid \mathcal{U}_{t-}. \tag{A.2}$$

It is explained in detail in Section 2 of Picci and Katayama (1996) that another equivalent way of writing this condition is

$$\hat{E}\{\mathcal{Y}_{t-} \mid \mathcal{U}\} = \hat{E}\{\mathcal{Y}_{t-} \mid \mathcal{U}_{t-}\} \tag{A.3}$$

(causality). Now, decomposing the past of y orthogonally into the past spaces of the "deterministic" ($y^d(t) = y(t)|\mathcal{U}$) and "stochastic" ($y^s(t) = y(t)|\mathcal{U}^\perp$) components of y (Picci and Katayama, 1996), it follows that $\mathcal{Y}_{t-} \subset \mathcal{Y}^d_{t-} \oplus \mathcal{Y}^s_{t-}$, where $\mathcal{Y}^d_{t-} \subset \mathcal{U}_{t-}$. Therefore, by elementary vector space geometry

$$\begin{aligned} \mathcal{U}_{t+} \cap \mathcal{P}_{t-} &= \mathcal{U}_{t+} \cap (\mathcal{U}_{t-} \vee \mathcal{Y}_{t-}) \\ &\subset \mathcal{U}_{t+} \cap (\mathcal{U}_{t-} \vee (\mathcal{Y}^d_{t-} \oplus \mathcal{Y}^s_{t-})) \\ &\subset \mathcal{U}_{t+} \cap (\mathcal{U}_{t-} \oplus \mathcal{Y}^s_{t-}) \end{aligned}$$

and since $\mathcal{Y}^s_{t-} \perp \mathcal{U}_{t+}$, the last intersection is equal to $\mathcal{U}_{t+} \cap \mathcal{U}_{t-}$, which is zero by assumption (3). □

B Proof of Lemma 2

We recall a known characterization of conditional orthogonality of two subspaces $\mathcal{A}$, $\mathcal{B}$ given a third subspace $\mathcal{X}$, see Proposition 1.1 in Picci and Katayama (1996) and Lindquist and Picci (1985).

$$\mathcal{A} \perp \mathcal{B} \mid \mathcal{X} \Leftrightarrow \mathcal{A} \perp \mathcal{B} \mid \mathcal{X} \vee \mathcal{A}_0, \quad \forall \mathcal{A}_0 \subset \mathcal{A} \tag{B.1}$$

The feedback-free condition is equivalent to the conditional orthogonality $\mathcal{Y}_{(t+h+1)-} \perp \mathcal{U}_{(t+h+1)+} \mid \mathcal{U}_{(t+h+1)-}$ for $h \geq 0$. Since $\mathcal{Y}_{t-} \subset \mathcal{Y}_{(t+h+1)-}$, we have also $\mathcal{Y}_{(t+h+1)-} \perp \mathcal{U}_{(t+h+1)+} \mid \mathcal{Y}_{t-} \vee \mathcal{U}_{(t+h+1)-}$ which is equivalent to

$$\hat{E}\{\mathcal{Y}_{(t+h+1)-} \mid \mathcal{Y}_{t-} \vee \mathcal{U}\} = \hat{E}\{\mathcal{Y}_{(t+h+1)-} \mid \mathcal{Y}_{t-} \vee \mathcal{U}_{(t+h+1)-}\} \tag{B.2}$$

Now $y(t+h) \in \mathcal{Y}_{(t+h+1)-}$ so that (B.2) implies

$$\hat{E}\{y(t+h) \mid \mathcal{Y}_{t-} \vee \mathcal{U}\} = \hat{E}\{y(t+h) \mid \mathcal{Y}_{t-} \vee \mathcal{U}_{(t+h+1)-}\} = \hat{E}\{y(t+h) \mid \mathcal{P}_{t-} \vee \mathcal{U}_{[t,t+h]}\}$$

which is (7). The last statement of the lemma follows readily by decomposing $\hat{E}\{y(t+h) \mid \mathcal{P}_{t-} \vee \mathcal{U}_{[t,t+h]}\}$ for $h = 0, 1, \cdots, k-1$ into oblique projections. Using (7) we see that

$$\Phi u_+(t) = \hat{E}_{\parallel \mathcal{P}_{t-}}\{f(t) \mid \mathcal{U}_{[t,t+k-1]}\}$$

which is in fact a function only of $\{u(t), \cdots, u(t+k-1)\}$. That Φ is a time-invariant operator follows by joint stationarity. □

C Kronecker Products

Let A be an $m \times n$ matrix with $A = [a_1, \cdots, a_n]$, where $a_i \in \mathbb{R}^n$, $i = 1, \cdots, n$. Then the $mn \times 1$ stacked vector $\mathrm{vec}(A)$ is defined by

$$\mathrm{vec}(A) = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$$

Moreover, we define Kronecker product of $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{q \times l}$ as

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{mq \times nl}$$

Then, for $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times l}$, we have

$$\mathrm{vec}(AB) = (I_l \otimes A)\mathrm{vec}(B) = (B' \otimes I_m)\mathrm{vec}(A) \in \mathbb{R}^{ml}$$

Figures 1 to 9

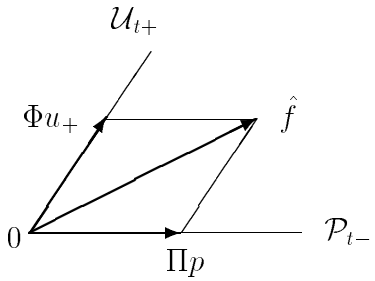

Figure 1: Oblique projections.

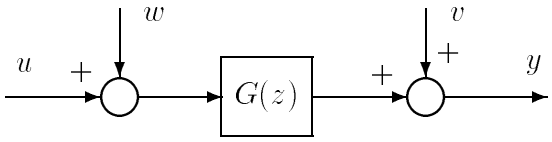

Figure 2: The plant model.

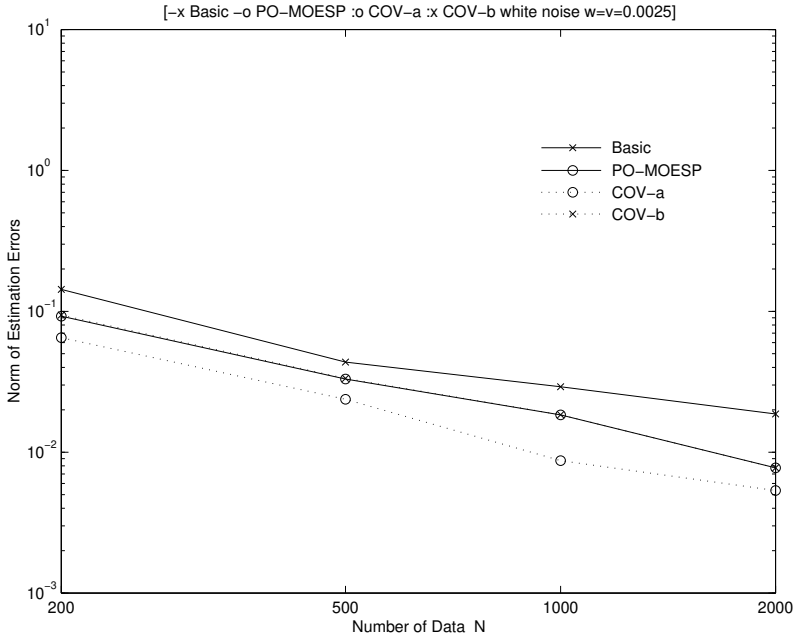

Figure 3: Performance of four algorithms where the control input is white noise, and $\sigma_w^2 = \sigma_v^2 = (0.05)^2$.

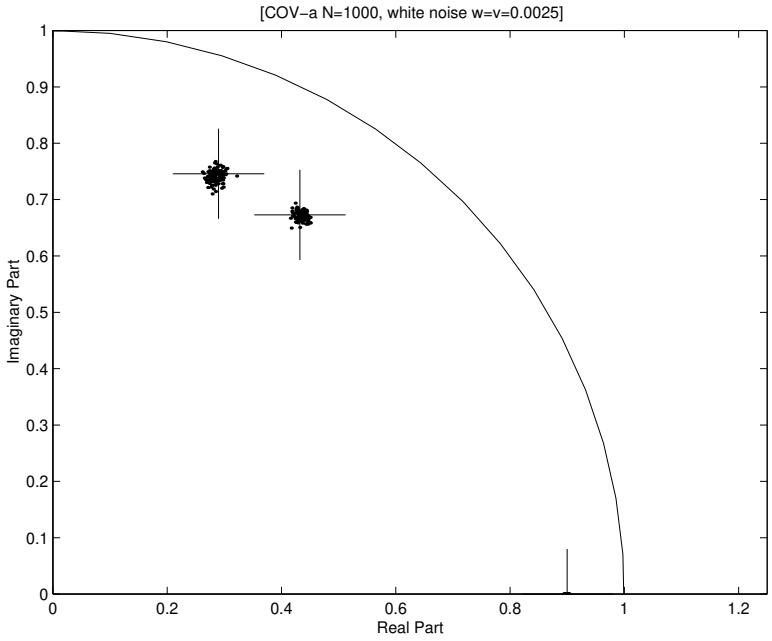

Figure 4: Pole estimates by COV-a where the control input is white noise, and $\sigma_w^2 = \sigma_v^2 = (0.05)^2$.

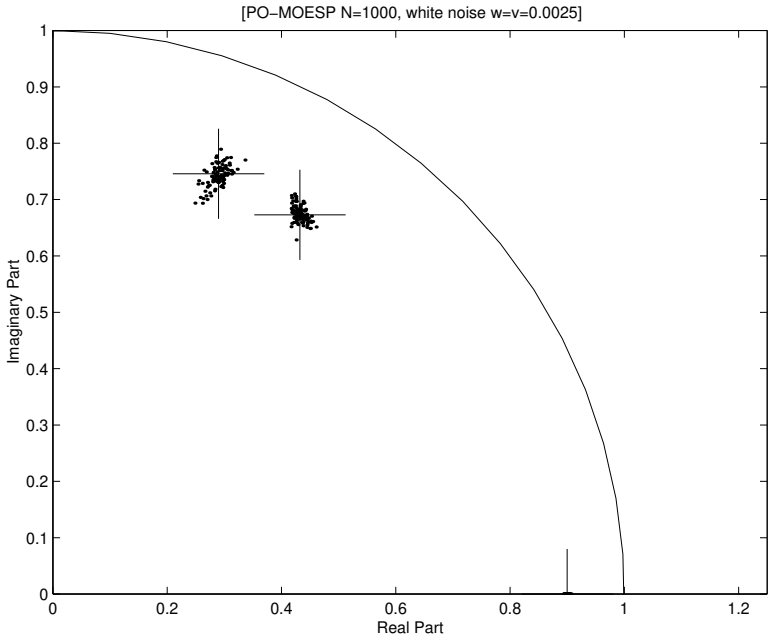

Figure 5: Pole estimates by PO-MOESP where the control input is white noise, and $\sigma_w^2 = \sigma_v^2 = (0.05)^2$.

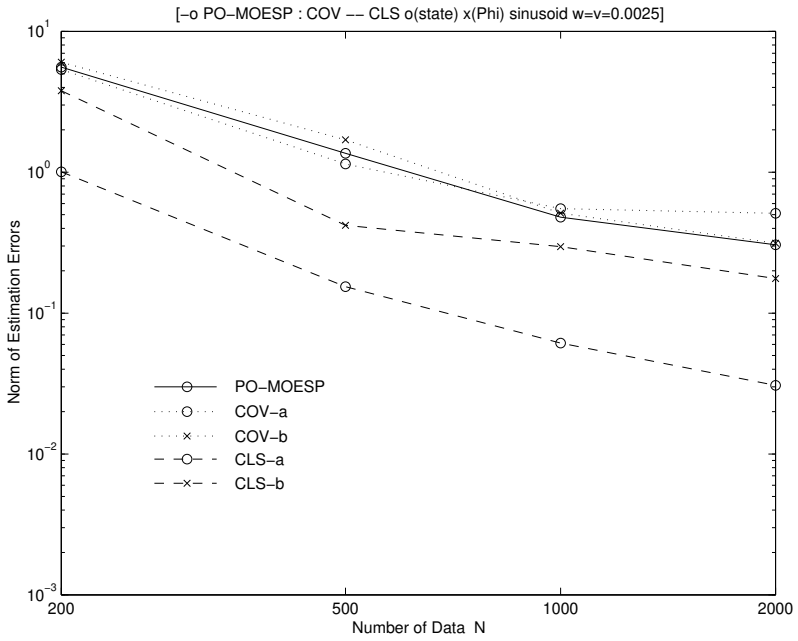

Figure 6: Performance of five algorithms where the control input is a sum of sinusoids, and $\sigma_w^2 = \sigma_v^2 = (0.05)^2$.

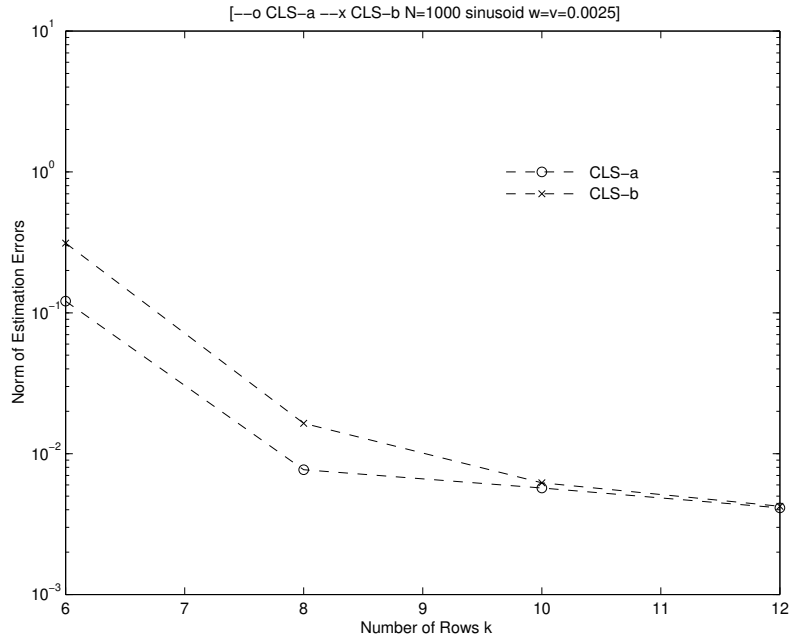

Figure 7: Performance in function of number of rows k.

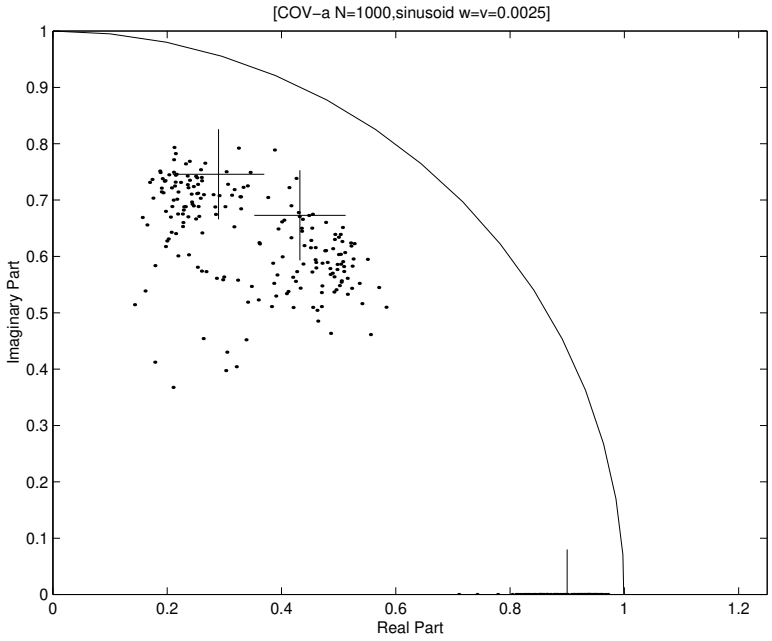

Figure 8: Pole estimates by COV-a where the control input is a sum of sinusoids, and $\sigma_w^2 = \sigma_v^2 = (0.05)^2$.

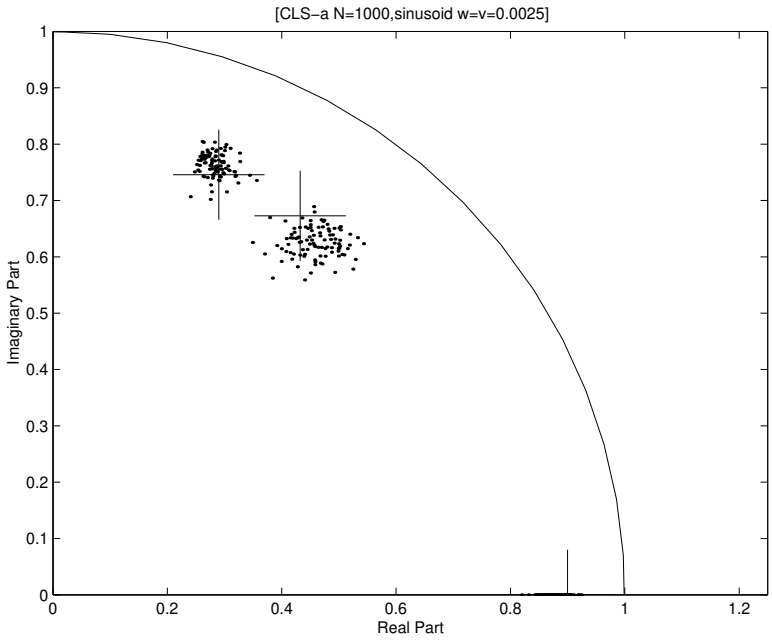

Figure 9: Pole estimates by CLS-a where the control input is a sum of sinusoids, and $\sigma_w^2 = \sigma_v^2 = (0.05)^2$.