Available online at www.sciencedirect.com
SCIENCE DIRECT®
Automatica 40 (2004) 575–589
automatica
www.elsevier.com/locate/automatica

On the ill-conditioning of subspace identification with inputs☆

Alessandro Chiuso, Giorgio Picci*

Department of Information Engineering, University of Padova, Via Gradenigo 6/a, Padova 35131, Italy

Received 12 March 2001; received in revised form 2 October 2003; accepted 18 November 2003

Abstract

There is experimental evidence that the performance of standard subspace algorithms from the literature (e.g. the N4SID method) may be surprisingly poor in certain experimental conditions. This happens typically when the past signals (past inputs and outputs) and future input spaces are nearly parallel. In this paper we argue that the poor behavior may be attributed to a form of ill-conditioning of the underlying multiple regression problem, which may occur for nearly parallel regressors. An elementary error analysis of the subspace identification problem, shows that there are two main possible causes of ill-conditioning. The first has to do with near collinearity of the state and future input subspaces. The second has to do with the dynamical structure of the input signal and may roughly be attributed to "lack of excitation". Stochastic realization theory constitutes a natural setting for analyzing subspace identification methods. In this setting, we undertake a comparative study of three widely used subspace methods (N4SID, Robust N4SID and PO-MOESP). The last two methods are proven to be essentially equivalent and the relative accuracy, regarding the estimation of the (A, C) parameters, is shown to be the same.

Keywords: Subspace identification; Exogenous inputs; Numerical conditioning; Collinearity; Oblique projections; State-space identification

1. Introduction

Subspace methods for the identification of linear systems have been object of much research in the last 10 years. In particular subspace methods for time series (no observable inputs) have been thoroughly analyzed in the literature (Van Overschee & De Moor, 1993; Bauer, 2002; Lindquist & Picci, 1996) and it seems fair to say that they are perhaps the most efficient and accurate methods available today for multivariable time series identification. The situation is not the same for identification of systems with observable inputs. Although it is generally admitted that subspace methods with inputs offer substantial advantages over traditional PEM identification (Ljung, 1997), especially for identification of multivariable systems, there is experimental evidence that standard subspace methods (e.g., the N4SID method) perform poorly in certain experimental conditions, in particular when the past signals (past inputs and outputs) and future input spaces are nearly parallel (Chiuso & Picci, 1999; Kawauchi, Chiuso, Katayama, & Picci, 1999). Subspace methods operating on joint input–output data involve the solution of a multiple regression problem and the reason for the poor behavior may be attributed to ill-conditioning of the regression problem, which may occur when the regressors are nearly parallel. Although this phenomenon is well-known in multivariate statistics (Stewart, 1987; Belsley, 1991), it does not seem to have been noticed and analyzed in the subspace identification literature. The study of numerical conditioning of the regression problem, besides the effect of numerical roundoff errors, which are clearly irrelevant in the present context, yields information on the sensitivity of the parameter and transfer function estimates to noise in the data, which is instead important to assess the performance of identification methods. It should be intuitively clear, and it will be demonstrated formally in the companion paper (Chiuso & Picci, 2004b), that bad numerical conditioning generally implies a large variance of the estimates. We believe that this aspect should

☆ This work has been supported in part by the ERNSI TMR network *System* Identification and by the national project *Identification* and adaptive control of industrial systems funded by MIUR. Part of this work has been done while the first author (A.C.) was a post-doctoral researcher with the Division of Optimization and Systems Theory, KTH, Stockholm, supported by ERNSI. This paper was recommended for publication in revised form by Associate Editor Brett Niness under the direction of Editor Torsten Söderström.

* Corresponding author. Tel.: +39-049-827-7705; fax: +39-049-827-7699.

E-mail addresses: chiuso@dei.unipd.it (A. Chiuso), picci@dei.unipd.it (G. Picci).

0005-1098/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2003.11.009

be taken into account in the design and comparison of subspace algorithms.

In this paper, which expands on work presented in previous conference papers (Chiuso & Picci, 1999; Kawauchi et al., 1999), we shall provide an elementary error analysis of some well-known subspace methods. We shall see that the sensitivity to random errors of a subspace method can be measured by the condition number of the conditional cross-covariance matrices $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}$, and $\Sigma_{\mathbf{u}^+\mathbf{u}^+|\hat{\mathbf{x}}}$ of the state, given the future inputs, and of the future inputs given the current state, which will be introduced later in the paper. We shall relate the conditioning of these two matrices to the "near parallelism" of the state and future input spaces and thereby see that there are two main possible causes of ill-conditioning. The first is due to the cross-correlation between the state and future inputs, while the second is inherent in the dynamical structure of the input signal (and there is not much one can do about it, if identification has to be based on experiments performed during normal operation of the plant). Here we make contact and extend the analysis of the papers (Jansson & Wahlberg, 1997, 1998); where it has been shown that the singularity of $\Sigma_{\mathbf{xx}|\mathbf{u}^+}$ is a cause of lack of consistency of subspace methods with inputs. In a sense, we study also the effects of near-singularity of this matrix.

A main motivation of this paper is to compare the performance of some widely used subspace methods from the literature. To this purpose, we undertake a comparative analysis of the conditioning of N4SID, "Robust" N4SID and PO-MOESP. The analysis is first directed to recasting the various algorithms into a common setting using ideas from stochastic realization theory. A result of this analysis is that, at least for the estimation of the (A, C) parameters, the "Robust" N4SID and PO-MOESP methods are equivalent. As expected, the original N4SID of (Van Overschee & De Moor, 1994), is generally worse.

The structure of the paper is as follows:

- In Section 2 we review the basic ideas of subspace identification, discuss the finite-interval stochastic realization problem, describe the basic "ideal" Kalman filter model which should be used in identification with inputs and discuss some difficulties which prevent constructing the state from finite input–output data. This explains why there is a multitude of subspace identification methods with inputs and indicates a common background for their analysis.
- In Section 3 we do some error analysis, comparing the conditioning of the identification (regression) problem based on the "ideal model", with the regression problem occurring in the N4SID method.
- In Section 4 we compare the "Robust N4SID" and PO-MOESP methods. We prove that these two methods produce exactly the same estimates of (A, C) and hence the same conditioning analysis holds for these two methods.
- Section 5 contains some conclusions.

As is well-known, numerical conditioning analysis deals in a sense with "worst case" situations and one may wonder what is the practical relevance of the results of this paper in terms of statistical accuracy (e.g., variance) of the estimates. This point is answered in the companion paper (Chiuso & Picci, 2004b), where we introduce asymptotic variance formulas for the (A, C) and (B, D) parameter estimates. From these formulas, the statistical meaning of the analysis of this paper emerges very clearly.

2. A review of subspace identification

Let

$$\{u_{t_0}, \ldots, u_t, \ldots\}, \quad \{y_{t_0}, \ldots, y_t, \ldots\},$$
$$u_t \in \mathbb{R}^p, y_t \in \mathbb{R}^m \tag{2.1}$$

be observed input–output trajectories of an unknown system, which we want to identify. For the moment we shall pretend that the trajectories are infinitely long. We shall assume that the data are sample paths of a pair of zero-mean second-order stationary *true* random processes $\mathbf{y} = \{\mathbf{y}(t)\}$, $\mathbf{u} = \{\mathbf{u}(t)\}$ having a rational spectral density; in other words, data (2.1) are generated by a linear stochastic system of the form

$$\mathbf{x}(t+1) = A\mathbf{x}(t) + B\mathbf{u}(t) + G\mathbf{w}(t),$$
$$\mathbf{y}(t) = C\mathbf{x}(t) + D\mathbf{u}(t) + J\mathbf{w}(t), \quad t \geqslant t_0, \tag{2.2}$$

where $\{\mathbf{x}(t)\}$ is the state process of dimensions n, and $\{\mathbf{w}(t)\}$ is a normalized white noise process uncorrelated with the past history of all other variables, and A, B, G, C, D, J are constant matrices. Here, as in most cases in identification, we are not interested in modelling the exogenous input $\{\mathbf{u}(t)\}$ explicitly. In this paper we shall always make the assumption that *there is no feedback from* $\mathbf{y}$ *to* $\mathbf{u}$. This implies that the processes $\{\mathbf{u}(t)\}$ and $\{\mathbf{w}(t)\}$ are completely uncorrelated. See, e.g. Caines and Chan (1976), Gevers and Anderson (1982), Picci and Katayama (1996) for a discussion of this concept.

The system (2.2) is also called a *stationary stochastic realization* of the output process $\mathbf{y}$ with input $\mathbf{u}$. It is well-known that there are always infinitely many such linear representations of $\mathbf{y}$, which are equivalent up to (conditional) second-order statistics. Without loss of generality we shall only consider realizations which are *stochastically minimal*, in the sense that the state dimension, n, is the smallest possible. This implies in particular (but is not equivalent to) that the triplet $\{C, A, [BG]\}$ is minimal in the usual system-theoretic sense. A realization which is unique up to change of basis, is the so-called "innovation representation"

$$\mathbf{x}(t+1) = A\mathbf{x}(t) + B\mathbf{u}(t) + K\mathbf{e}(t),$$
$$\mathbf{y}(t) = C\mathbf{x}(t) + D\mathbf{u}(t) + \mathbf{e}(t), \tag{2.3}$$

where the white noise $\{\mathbf{e}(t)\}$ has the meaning of (stationary) one step prediction error of $\{\mathbf{y}(t)\}$, given the infinite past history of $\{\mathbf{y}(t)\}$ $\{\mathbf{u}(t)\}$ up to time $t-1$.

Subspace identification is based on the following idea. Since the processes $\{\mathbf{y}(t)\}$, $\{\mathbf{u}(t)\}$, $\{\mathbf{x}(t)\}$ satisfy the equations of the linear innovation model (2.3), it is obvious that the finite "tail" matrices, Y_t, U_t, X_t, constructed at each time t from the observed samples by letting [1]

$$Y_t := [y_t \; y_{t+1} \cdots y_{t+N-1}] \tag{2.4}$$

must also satisfy (2.3), i.e.

$$\begin{aligned} X_{t+1} &= AX_t + BU_t + KE_t \\ Y_t &= CX_t + DU_t + E_t \end{aligned} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X_t \\ U_t \end{bmatrix} + \begin{bmatrix} K \\ I \end{bmatrix} E_t, \tag{2.5}$$

where $E_t := [e_t \; e_{t+1} \cdots e_{t+N-1}]$ is the innovation tail. This equation can be interpreted as a regression model describing X_{t+1}, Y_t in terms of X_t, U_t. Hence, if the tail matrices X_{t+1}, X_t, U_t, Y_t, were given, one could solve (2.5) for the unknown parameters (A, B, C, D), by least squares.

It turns out that, under a generic assumption of invertibility of the joint covariance matrix

$$E\left\{ \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{u}(t) \end{bmatrix} [\mathbf{x}^\top(t) \; \mathbf{u}^\top(t)] \right\}$$

the normal equations obtained by multiplying the last member of (2.5) by $[X_t^\top \; U_t^\top]$, are uniquely solvable in the parameters (A, B, C, D) for N large enough. In the ideal case when infinitely long sample trajectories are available ($N \to \infty$), E_t is orthogonal [2] to the past data, namely $E_t \perp (X_s, U_s)$ for all $s \leqslant t$ by absence of feedback (this is only approximately true for N large but finite). Then, the estimates computed by solving the normal equations coincide, for $N \to \infty$, with the true parameters of the system (consistency). Hence, in an ideal situation where we had available the input–output tail matrices at time t, and also a corresponding pair of *state* tail matrices at the successive time instants t and $t+1$, consistent identification of the parameters (A, B, C, D) of system (2.2) would be a straightforward matter.

In practice the state trajectory is not given to us. However, it is known that the state of certain realizations, in particular the innovation realization (2.3), can be constructed from the input–output processes. In our case we only have available a finite input–output sequence, $\{u_t, y_t\}_{t=0,\ldots,N}$ or equivalently, an input–output tail sequence $\{U_t, Y_t\}_{t=0,\ldots,T}$ (where $T \ll N$) and the state vector at time t needs to be constructed (in general approximately) from the available data. It is seen that the construction of the state becomes a central step in the subspace approach to identification. Most subspace identification methods in the literature can be seen as different ways to implement this step.

The problem of constructing the state and state-space models of stochastic processes is the main concern of stochastic realization theory. The theory provides procedures for state space construction based on geometric operations on certain Hilbert spaces of random variables which are linear functionals of the input and output processes of the system. These spaces will be introduced below.

In general terms, subspace identification with inputs could be seen as consisting of three basic steps: (i) construction of (a sample estimate of) the state vector of a state-space representation of the process $\mathbf{y}$, (ii) solution of a multiple linear regression problem to determine the system matrices (A, B, C, D) of the deterministic part of the model, (iii) estimation of the stochastic noise parameters K and $\Lambda = E\{\mathbf{e}(t)\mathbf{e}(t)^\top\}$, from the parameters obtained in the previous step.

In this paper we shall not consider the third step at all and concentrate only on the estimation of the "deterministic" parameters (A, B, C, D).

2.1. Notations

For $-\infty \leqslant t_0 \leqslant t \leqslant T \leqslant +\infty$ define the Hilbert space of scalar zero-mean random variables

$$\mathscr{U}_{[t_0,t)} := \overline{\text{span}}\{\mathbf{u}_k(s); k = 1, \ldots, p, t_0 \leqslant s < t\},$$

where the bar denotes closure in mean square, i.e. in the metric defined by the inner product $\langle \xi, \eta \rangle := E\{\xi, \eta\}$, the operator E denoting mathematical expectation. A similar definition holds for $\mathscr{Y}_{[t_0,t)}$. We shall let $\mathscr{P}_{[t_0,t)} := \mathscr{U}_{[t_0,t)} \vee \mathscr{Y}_{[t_0,t)}$ denote the joint past space of the input and output processes at time t (the $\vee$ denotes closed vector sum). Similarly, let $\mathscr{U}_{[t,T]}, \mathscr{Y}_{[t,T]}$ be the respective future spaces up to time T, say:

$$\mathscr{U}_{[t,T]} := \overline{\text{span}}\{\mathbf{u}_k(s); k = 1, \ldots, p, t \leqslant s \leqslant T\}.$$

By convention the past spaces do not include the present. When $t_0 = -\infty$ we shall use the shorthands $\mathscr{U}_t^-, \mathscr{Y}_t^-$ for $\mathscr{U}_{[-\infty,t)}, \mathscr{Y}_{[-\infty,t)}$, the closed vector sum $\mathscr{U}_t^- \vee \mathscr{Y}_t^-$ being denoted by $\mathscr{P}_t^-$ (the infinite joint past at time t). These are the Hilbert spaces of random variables spanned by the infinite past of $\mathbf{u}$ and $\mathbf{y}$ up to time t.

Subspaces spanned by random variables at just one time instant (e.g., $\mathscr{U}_{[t,t]}, \mathscr{Y}_{[t,t]}$, etc.) are simply denoted $\mathscr{U}_t, \mathscr{Y}_t$, etc. while for the spaces generated by the whole time history of $\mathbf{u}$ and $\mathbf{y}$ we shall use the symbols $\mathscr{U}$, $\mathscr{Y}$, respectively.

All through this paper we shall assume that the input process is "sufficiently rich", in the sense that $\mathscr{U}_{[t_0,T]}$ admits the direct sum decomposition

$$\mathscr{U}_{[t_0,T]} = \mathscr{U}_{[t_0,t)} + \mathscr{U}_{[t,T]}, \quad t_0 \leqslant t < T \tag{2.6}$$

the $+$ sign denoting direct sum of subspaces. The symbol $\oplus$ will be reserved for *orthogonal* direct sum. Various conditions ensuring sufficient richness are known. For example, it is well-known that for a full-rank purely non deterministic (p.n.d.) process $\mathbf{u}$ to be sufficiently rich it is necessary and sufficient that the determinant of the spectral density matrix Φ_u should have no zeros on the unit circle (Hannan & Poskitt, 1988).

[1] Similar definitions hold for U_t and X_t.

[2] Orthogonality is with respect to the inner product (2.7) which will be defined later.

2.1.1. The sample-trajectory framework

Under a natural second-order ergodicity assumption, (Lindquist & Picci, 1996), a sequence of semi-infinite tail matrices constructed from a time series, in particular our sample trajectory (2.1), can be looked upon as an object isomorphic to a stationary random process. This isomorphism is defined by the correspondence

$$\mathscr{I}: \begin{cases} a^\top \mathbf{y}(t) \mapsto a^\top Y_t, & a \in \mathbb{R}^m, \\ b^\top \mathbf{u}(t) \mapsto b^\top U_t, & b \in \mathbb{R}^p \end{cases}$$

mapping linear combinations of the components of the random variables at time t of the processes $\{\mathbf{y}\}$ and $\{\mathbf{u}\}$ into the same linear combinations of the rows of the semi-infinite tail matrices at time t of the ergodic trajectory. It is an obvious consequence of ergodicity that if we define an inner product of semi-infinite sequences $\xi, \eta \in \mathbb{R}^{\mathbb{Z}_+}$ by the limit [3]

$$\langle \xi, \eta \rangle := \lim_{N\to\infty} \frac{1}{N+1} \Sigma_{t=0}^{N} \xi_t \eta_t \tag{2.7}$$

then $\mathscr{I}$ is an isometry, i.e. preserves inner products. It follows, (Rozanov, 1967, p. 14), that the Hilbert space $\overline{\text{span}}\{Y_t, U_t | t \geqslant t_0\}$, linearly generated by the rows of the semi-infinite tail sequences $\{Y_t, U_t | t \geqslant t_0\}$ (here $N = \infty$!), closed with respect to the norm induced by the inner product (2.7), and the "stochastic" Hilbert space $\mathscr{Y} \vee \mathscr{U}$ of zero-mean second order random variables introduced above, are isometrically isomorphic Hilbert spaces. This means that for operations concerning computations of second order moments and the relative limits, working with bona-fide random variables as maps defined on a probability space, is equivalent to working with semi-infinite real sequences belonging to the isomorphic Hilbert space $\overline{\text{span}}\{Y_t, U_t | t \geqslant t_0\}$.

Henceforth it will be convenient to regard the two spaces as being the *same* object. We shall therefore denote semi-infinite real or vector-valued sequences in $\overline{\text{span}}\{Y_t, U_t | t \geqslant t_0\}$ by boldface lowercase letters, exactly like random quantities in $\mathscr{Y} \vee \mathscr{U}$. This point of view will turn out to be very convenient later on, since it will allow us to employ in the statistical setup of identification, exactly the same formalism and notations used in the ordinary L^2 setting of stochastic systems.

We shall instead use capitals (e.g. X_t, Y_t, etc.) to denote the *finite* tail matrices (2.4) made of data sequences from t onwards up to time $t + N$. The symbol $Y_{[\tau,T]}$ will be used to denote the Hankel matrix

$$[Y_\tau^\top \cdots Y_T^\top]^\top$$

and $\mathscr{Y}_{[\tau,T]}^N$ the corresponding (finite-dimensional) row-space. Since for $N \to \infty$, $Y_{[\tau,T]}$ becomes the $m(T-\tau+1)$-dimensional column random vector $\mathbf{y}_{[\tau,T]}$ (or equivalently, the $m(T-\tau+1) \times \infty$ matrix of semi-infinite tails), one can say that $\mathscr{Y}_{[\tau,T]}^N \to \mathscr{Y}_{[\tau,T]}$ for $N \to \infty$. "Approximating" spaces of random variables by vector spaces spanned by the rows of tail matrices is a standard device in subspace identification.

In this paper T will denote a fixed terminal time; for simplicity the stochastic vectors $\mathbf{u}_{[t,T-1]}$ and $\mathbf{u}_{[t,T]}$ will occasionally be denoted $\mathbf{u}_t^+$ and $\bar{\mathbf{u}}_t^+$, namely

$$\mathbf{u}_t^+ := [\mathbf{u}^\top(t) \cdots \mathbf{u}^\top(T-1)]^\top, \quad \bar{\mathbf{u}}_t^+ := [(\mathbf{u}_t^+)^\top \mathbf{u}(T)^\top]^\top.$$

Similar notations will be used in the following without further comments.

The matrix of inner products of the finite vector sequences $X = [x_0, x_1, \ldots, x_N], Y = [y_0, y_1, \ldots, y_N]$, generated by sampling the random vectors $\mathbf{x}$ and $\mathbf{y}$, will be denoted

$$E_N[XY^\top] := \frac{1}{N+1} \Sigma_{k=0}^{N} x_k y_k^\top.$$

This is just the *sample covariance* matrix of $\mathbf{x}$ and $\mathbf{y}$, and will also be denoted $\hat{\Sigma}_{\mathbf{xy}}$. In the same spirit we shall write

$$E[\mathbf{x} \,|\, \mathbf{y}] := E[\mathbf{x}\mathbf{y}^\top] E[\mathbf{y}\mathbf{y}^\top]^{-1} \mathbf{y},$$

when $\mathbf{x}$ and $\mathbf{y}$ are random vectors and

$$E_N[X \,|\, Y] := E_N[XY^\top] E_N[YY^\top]^{-1} Y,$$

when X and Y are finite vector sequences. The latter expression is nothing else but the well-known formula solving the (deterministic) least-squares problem

$$\min_{A \in \mathbb{R}^{n\times m}} \|Y - AX\|.$$

Since for $N \to \infty$, $X \to \mathbf{x}$, $Y \to \mathbf{y}$ and $\hat{\Sigma}_{\mathbf{xy}} \to \Sigma_{\mathbf{xy}}$, (the true covariance of $\mathbf{x}$ and $\mathbf{y}$), for infinitely long sequences we have $\lim_{N\to\infty} E_N[X \,|\, Y] = E[\mathbf{x} \,|\, \mathbf{y}]$.

2.2. Constructing the state

The construction of the state space can be based on the prescriptions of stochastic realization theory with inputs (Picci & Katayama, 1996; Katayama & Picci, 1999; Picci, 1997). In particular, we recall that the *state space* at time t of any stationary realization (2.2)

$$\mathscr{X}_t := \overline{\text{span}}\{\mathbf{x}_k(t); k = 1, \ldots, n, \}$$

has the property of being a (*minimal*) *oblique Markovian splitting subspace* for the process $\mathbf{y}$, namely

$$E_{\| \mathscr{U}_t^+}[\mathscr{Y}_t^+ \vee \mathscr{X}_t^+ | \mathscr{X}_t^- \vee \mathscr{P}_t^-] = E_{\| \mathscr{U}_t^+}[\mathscr{Y}_t^+ \vee \mathscr{X}_t^+ | \mathscr{X}_t \vee \mathscr{U}_t] \tag{2.8}$$

where the symbol $E_{\| \mathscr{C}}[\mathscr{A} | \mathscr{B}]$ denotes the oblique projection of the subspace $\mathscr{A}$ onto $\mathscr{B}$ along the subspace $\mathscr{C}$.

The oblique Markovian splitting property is equivalent to the property of state for stochastic systems with inputs. In fact, an arbitrary choice of basis on any such subspace yields a state vector, and leads to a stochastic model of the type (2.2). In this sense, a realization of $\mathbf{y}$ can be seen merely as

[3] Under second-order ergodicity, the sum in (2.7) converges for all sequences whose elements are made of finite linear combinations of the rows of (possibly time-shifted) tails of the given stationary time series.

a particular choice of basis in a minimal oblique Markovian splitting subspace.

Example. It is an important fact, and not difficult to check, that the components of the state vector $\mathbf{x}(t)$ of the innovation model (2.3) form a basis in the *oblique predictor space*

$$\mathscr{X}_t := E_{\|\mathscr{U}_t^+}[\mathscr{Y}_t^+|\mathscr{P}_t^-] \tag{2.9}$$

which is a (minimal) oblique Markovian splitting subspace contained in the past $\mathscr{P}_t^-$. Note that this subspace can in principle be constructed by an oblique projection, using input–output data (on an infinite interval).

The above aims at constructing stationary realizations, assuming that the input and output processes are given on an infinite interval. Unfortunately it is not so simple to construct the state space in the presence of an external input, using input–output data from a *finite* time interval and we shall have to discuss this question in some detail in the next section.

2.3. Finite-interval realizations: the transient Kalman filter

In this section we shall investigate how the state of a stochastic realization of $\mathbf{y}$ may be constructed using the random variables of the input and output processes from a finite time interval $[t_0, T]$. We shall also discuss what kind of state space model of $\mathbf{y}$ can be obtained starting from these data.

To this purpose we need a concept of state space sequence, i.e. subspaces $\{\mathscr{X}_t\}$ of the data space $\mathscr{Y}_{[t_0,T]} \vee \mathscr{U}_{[t_0,T]} (\equiv \mathscr{P}_{[t_0,T]})$, $t = t_0, \dots, T$, which generalizes the properties of an oblique Markovian splitting subspace (2.8) to a possibly non-stationary setting.

Definition 1. A sequence of subspaces of the data space, $\mathscr{X}_t \subseteq \mathscr{Y}_{[t_0,T]} \vee \mathscr{U}_{[t_0,T]}$, $t = t_0, \dots, T$, is oblique Markovian splitting in the finite interval $[t_0, T]$, in short, finite-interval oblique Markovian splitting if

$$\begin{aligned} &E[\mathscr{X}_{t+1} \vee \mathscr{Y}_t | \mathscr{X}_{t+1}^- \vee \mathscr{U}_t \vee \mathscr{P}_t^-] \\ &\quad = E[\mathscr{X}_{t+1} \vee \mathscr{Y}_t | \mathscr{X}_t + \mathscr{U}_t]. \end{aligned} \tag{2.10}$$

for all $t = t_0, \dots, T$.

It can be shown that the state space of any finite-interval realization of $\mathbf{y}$ is a finite-interval oblique Markovian splitting subspace and, conversely, given a family of finite-interval oblique Markovian splitting subspaces, $\{\mathscr{X}_t\}$, $t = t_0, \dots, T$, any choice of basis in $\{\mathscr{X}_t\}$ [4] provides the state process of a state space realization of $\mathbf{y}$ with input process $\mathbf{u}$, (Chiuso, 2000). One example will be given shortly. As one can see from (2.10), the state at time t and the present input act as a sufficient statistic also with respect to the information contained in the future inputs space $\mathscr{U}_{(t,T]}$. This guarantees that future inputs will not show up explicitly in the corresponding state-space model, i.e. the model will be causal in $\mathbf{u}$. However the state itself need in general not be a causal function of $\mathbf{u}$.

Let $\mathscr{X}_t$ be a *stationary* oblique Markovian splitting subspace and define the subspaces $\hat{\mathscr{X}}_t$ by

$$\hat{\mathscr{X}}_t := E[\mathscr{X}_t | \mathscr{P}_{[t_0,t)} \vee \mathscr{U}_{[t,T]}], \quad t = t_0, \dots, T. \tag{2.11}$$

Let $\mathbf{x}(t)$ be a basis for $\mathscr{X}_t$ and $\mathbf{x}(t+1)$ its stationary shift. Choose a basis in $\hat{\mathscr{X}}_t$ as

$$\hat{\mathbf{x}}(t) := E[\mathbf{x}(t) | \mathscr{P}_{[t_0,t)} \vee \mathscr{U}_{[t,T]}] \tag{2.12}$$

and let

$$\hat{\mathbf{x}}(t+1) := E[\mathbf{x}(t+1) | \mathscr{P}_{[t_0,t]} \vee \mathscr{U}_{[t+1,T]}] \tag{2.13}$$

(this is a choice of basis coherent with (2.12)). Denote by $\hat{\mathscr{E}}_t$ the *transient innovation space* defined by the orthogonal decomposition

$$\mathscr{P}_{[t_0,t]} \vee \mathscr{U}_{[t+1,T]} = \hat{\mathscr{E}}_t \oplus (\mathscr{P}_{[t_0,t)} \vee \mathscr{U}_{[t,T]}) \tag{2.14}$$

so that $\hat{\mathscr{E}}_t = \operatorname{span}\{\hat{\mathbf{e}}(t)\}$, where $\hat{\mathbf{e}}(t)$ is the *transient* (*conditional*) *innovation process* defined by

$$\hat{\mathbf{e}}(t) = \mathbf{y}(t) - E[\mathbf{y}(t) | \mathscr{P}_{[t_0,t)} \vee \mathscr{U}_{[t,T]}]. \tag{2.15}$$

Let (2.2) be the stationary model associated with the basis $\mathbf{x}(t)$ in the state space $\mathscr{X}_t$. The following result is well-known and has for example been used in Van Overschee and De Moor (1994).

Theorem 1. *The subspaces $\hat{\mathscr{X}}_t$ are finite-interval oblique Markovian splitting. If $\mathscr{X}_t$ is minimal, so are the $\hat{\mathscr{X}}_t$'s. The process* $\mathbf{y}$ *admits the following finite-interval realization, called the* transient conditional Kalman filter realization *on the interval* $[t_0, T]$

$$\begin{aligned} \hat{\mathbf{x}}(t+1) &= A\hat{\mathbf{x}}(t) + B\mathbf{u}(t) + K(t)\hat{\mathbf{e}}(t), \\ \mathbf{y}(t) &= C\hat{\mathbf{x}}(t) + D\mathbf{u}(t) + \hat{\mathbf{e}}(t), \\ \hat{\mathbf{x}}(t_0) &= E[\mathbf{x}(t_0) | \mathscr{U}_{[t_0,T]}], \end{aligned} \tag{2.16}$$

where the matrix gain $K(t)$ is given by

$$K(t) = (AQ(t)C^\top + GJ^\top)(CQ(t)C^\top + JJ^\top)^{-1} \tag{2.17}$$

the matrix $Q(t)$ is the state error covariance $Q(t) = E(\mathbf{x}(t) - \hat{\mathbf{x}}(t))(\mathbf{x}(t) - \hat{\mathbf{x}}(t))^\top$ and can be computed by solving a Riccati difference equation.

We see that $\mathbf{y}$ can be represented by a finite-interval model whose state is a function only of the input–output data on the interval $[t_0, T]$. The deterministic subsystem of the model (2.16) involves the same constant parameters (A, B, C, D) of the stationary model (2.3) that one wants to identify.

[4] In order to preserve the time-invariance of certain parameters of the realization, e.g. the observability matrix, the bases in $\mathscr{X}_{t+1}$ and in $\mathscr{X}_t$ have to be chosen in a suitably "coherent" way, see e.g. Lindquist and Picci (1996, p. 721).

Remark 2. Contrary to the standard Kalman filter, the initial state estimate $\hat{\mathbf{x}}(t_0)$ is not zero and depends on the future inputs $\mathscr{U}_{[t_0T]}$. This fact (which is pointed out formally in Picci and Katayama (1996), Lemma 6.1) happens since the stationary initial state $\mathbf{x}(t_0)$ is a function of the past input history and, unless $\mathbf{u}$ is white noise, $\hat{\mathbf{x}}(t_0)=E[\mathbf{x}(t_0)|\mathscr{U}_{[t_0,T]}]$ is then a nontrivial function of the future inputs. This implies, in spite of the "causal" look of the state Eq. (2.16), that $\hat{\mathbf{x}}(t)$ is also a function of the future inputs on $[t,T]$.

This is not an unfortunate characteristic of the model (2.16), but rather a general fact. For, if we attempt to restrict to *causal* oblique Markovian splitting subspaces, i.e. to state spaces which are contained in $\mathscr{P}_{[t_0,t)}$ for each t, then we end up with a rather restricted class of models.

Proposition 3. *Let* $\mathbf{y}$ *and* $\mathbf{u}$ *be related by a finite dimensional stationary model of type* (2.2). *There are finite-interval oblique Markovian splitting subspace* $\mathscr{X}_t$ *contained in the past* $\mathscr{P}_{[t_0,t)}$ *if and only if the model is of the ARX type of order smaller than* $t-t_0$. *In particular, if common dynamics is not allowed, this holds true if and only if the transfer function* $F(z)=C(zI-A)^{-1}B+D$ *of the deterministic subsystem is a matrix polynomial in* z^{-1}, *i.e.* $F(z)$ *is of the FIR type, and the transfer function* $G(z)=C(zI-A)^{-1}G+J$, *of the stochastic subsystem of* (2.2), *obtained by setting* $B=0, D=0$, *is of the pure AR type.*

Proof. Necessity: Assume there is an oblique Markovian splitting subspace $\mathscr{X}_t$ contained in $\mathscr{P}_{[t_0,t)}$. It follows that there exist a matrix $\Psi=[\Psi_1\,\Psi_2]$ such that $\mathbf{x}(t)=\Psi_1\mathbf{y}_{[t_0,t)}+\Psi_2\mathbf{u}_{[t_0,t)}$. Since $\mathbf{y}(t)=C\mathbf{x}(t)+D\mathbf{u}(t)+\mathbf{e}(t)$ we have that

$$\mathbf{y}(t)=C\Psi_1\mathbf{y}_{[t_0,t)}+C\Psi_2\mathbf{u}_{[t_0,t)}+D\mathbf{u}(t)+\mathbf{e}(t)$$

showing that $\mathbf{y}(t)$ if of the ARX type of order smaller that $t-t_0$.

Sufficiency: Conversely, if the model is ARX of order less that $t-t_0$ then the oblique predictor space

$$\mathscr{X}_t=E_{\|\mathscr{U}_t^+}[\mathscr{Y}_t^+|\mathscr{P}_t^-]=E_{\|\mathscr{U}_t^+}[\mathscr{Y}_t^+|\mathscr{P}_{[t_0,t)}]$$

is contained in the finite (joint) past. It follows that $\hat{\mathscr{X}}_t=\mathscr{X}_t$, showing that there is a finite-interval oblique Markovian splitting subspace $\hat{\mathscr{X}}_t\subseteq\mathscr{P}_{[t_0,t)}$. □

This result states that, excluding a relatively trivial class of stationary processes, finite-interval state space-models with inputs cannot depend causally on the input process. In general the future of $\mathbf{u}$ has to enter in the state of the dynamic equations. We shall have to content ourselves with oblique Markovian splitting subspace which are (*conditionally*) *causal* i.e., such that $\mathscr{X}_t\subseteq\mathscr{Y}_{[t_0,t)}\vee\mathscr{U}_{[t_0,T]}$, like the Kalman filter model (2.16). This concept describes the extent to which the state of a finite time model can be causal in the input signal. Note that the situation is drastically different if $t_0=-\infty$. In this case the predictor space (2.9) is a subspace of the past $\mathscr{P}_t^-$ and defines the well-known stationary Kalman predictor (i.e., innovation) model where the state is a causal function of the past input and output variables.

For the reasons explained in the previous remark, the state space of the Kalman filter model (2.16) is not directly constructible from the input–output data on the interval $[t_0,T]$. We shall show below that it is not possible to extract generators for the state space from the output predictors, as it happens instead in the stationary, infinite past case.

Assume the data are generated by a finite dimensional true system (2.2). Compute the h-step ahead predictor of the output, based on the finite information available at time t

$$\begin{aligned}&E[\mathbf{y}(t+h)|\mathscr{P}_{[t_0,t)}\vee\mathscr{U}_{[t,T]}]\\&\quad=CA^hE[\mathbf{x}(t)|\mathscr{P}_{[t_0,t)}\vee\mathscr{U}_{[t,T]}]+H_{d,h}\mathbf{u}_t^+\\&\quad=CA^h\hat{\mathbf{x}}(t)+H_{d,h}\mathbf{u}_t^+,\\&\quad h=0,1,\dots,v-1,\qquad v:=T-t,\end{aligned}$$

where $H_{d,h}$ is the $h+1$st block-row of the block-Toeplitz matrix

$$H_d:=\begin{bmatrix}D & 0 & \dots & 0 & 0\\ CB & D & \dots & 0 & 0\\ \vdots & & & \ddots & \vdots\\ CA^{v-1}B & CA^{v-2}B & \dots & CB & D\end{bmatrix}. \tag{2.18}$$

Looking at these expressions, it seems that one may compute a basis $\hat{\mathbf{x}}(t)$ for $\hat{\mathscr{X}}_t$ by oblique projections of the output predictors along the future input space $\mathscr{U}_{[t,T]}$, as it is done in the stationary setting (Katayama & Picci, 1999). However, since

$$\begin{aligned}&E[\mathbf{y}(t+h)|\mathscr{P}_{[t_0,t)}\vee\mathscr{U}_{[t,T]}]\\&\quad=E_{\|\mathscr{U}_{[t,T]}}[\mathbf{y}(t+h)|\mathscr{P}_{[t_0,t)}]+E_{\|\mathscr{P}_{[t_0,t)}}[\mathbf{y}(t+h)|\mathscr{U}_{[t,T]}]\\&\quad=CA^hE_{\|\mathscr{U}_{[t,T]}}[\mathbf{x}(t)|\mathscr{P}_{[t_0,t)}]\\&\qquad+CA^hE_{\|\mathscr{P}_{[t_0,t)}}[\mathbf{x}(t)|\mathscr{U}_{[t,T]}]+H_{d,h}\mathbf{u}_t^+,\end{aligned}$$

there is no obvious way to separate the state component $E_{\|\mathscr{P}_{[t_0,t)}}[\mathbf{x}(t)|\mathscr{U}_{[t,T]}]$ from $H_{d,h}\mathbf{u}_t^+$, as both terms have a component along $\mathscr{U}_{[t,T]}$. Hence it is not possible to extract generators for the state space from the output predictors. More generally, there is no known recipe for constructing $\hat{\mathscr{X}}_t$ (or, more generally, any finite-interval oblique Markovian splitting subspace) starting only from the available input–output data $\mathscr{Y}_{[t_0,T]}\vee\mathscr{U}_{[t_0,T]}$. A consequence of this state of the affairs is that we have no practical rule to implement the basic principle on which subspace identification algorithms with inputs should be based.[5] In the literature this difficulty is circumvented either by approximations or by a variety of

[5] This is *not* the case for subspace identification for time-series (no inputs). Here we know how to construct the finite-interval state space from the random output data (infinite tails) observed on that finite interval.

seemingly unrelated ad hoc tricks. This may give the impression that the whole subject hinges on art and trickeries rather than on basic system principles.

3. Error analysis of subspace methods

Ideally, the first step of subspace identification should be of constructing the state of a transient Kalman-filter type realization (2.16). Naturally, with only finite data available, the state vector must be approximated by a finite tail matrix and will be affected by random errors. Hence the parameter estimates, obtained by regressing the next state and output variables on the estimated state (and on the observed input) will also be affected by errors. In this section we shall show that the magnitude of these errors depends on the numerical conditioning of the underlying regression problem. In particular *collinearity* of the state and (future) inputs plays an important role. Further, while for time-series (no inputs) a good approximation of the state of the finite-interval Kalman filter can be constructed from a clear-cut stochastic realization procedure (Van Overschee & De Moor, 1993), the situation is unfortunately worse for identification in the presence of exogenous inputs.

3.1. The "ideal" subspace method with inputs

The unknown system matrices (A, B, C, D) parametrize the transient Kalman filter realization (2.16) of $\mathbf{y}$, relative to a sufficiently large time interval $[t_0, T]$. By truncating all random variables to length N, we obtain the analog of the stationary regression model (2.5) of Section 2, now adapted to finite-interval data. Let $\hat{\mathbf{x}}_N(t)$ be the state tail matrix, a basis for $\hat{\mathscr{X}}_t^N$, and $\hat{\mathbf{x}}_N(t+1)$ a coherent basis for $\hat{\mathscr{X}}_{t+1}^N$, and consider the regression model

$$\begin{bmatrix} \hat{\mathbf{x}}_N(t+1) \\ Y_t \end{bmatrix} = \begin{bmatrix} A \\ C \end{bmatrix} \hat{\mathbf{x}}_N(t) + \begin{bmatrix} B \\ D \end{bmatrix} U_t + \begin{bmatrix} K(t)\hat{\mathbf{e}}_N(t) \\ \hat{\mathbf{e}}_N(t) \end{bmatrix}. \tag{3.1}$$

Even if we do not know how to construct a basis $\hat{\mathbf{x}}(t)$ for the Kalman filter realization (2.16) starting from the observable data, we shall for the moment assume that an estimate, $\hat{X}_t$, of the (truncated) Kalman filter state $\hat{\mathbf{x}}_N(t)$ is given to us, such that for $N \to \infty$, $\hat{X}_t \to \hat{\mathbf{x}}(t)$ (the "limit" being understood in the sense explained in Section 2.1). In particular, we shall assume that the covariance $\hat{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, of the finite-sample fluctuation error $\tilde{\mathbf{x}}_N(t) := \hat{\mathbf{x}}_N(t) - \hat{X}_t$ tends to zero for $N \to \infty$. It will be important to keep in mind that the approximation $\hat{X}_t$ is (asymptotically) attached to a certain choice of basis $\hat{\mathbf{x}}(t)$ in $\hat{\mathscr{X}}_t$.

By setting, $\Theta := [\Theta_1 \ \Theta_2] := \left[\begin{smallmatrix} A & C \\ B & D \end{smallmatrix}\right]$ and using the estimate $\hat{X}_t$, we may recast the regression problem (3.1) as a two-blocks regression

$$Z_t := \begin{bmatrix} \hat{X}_{t+1} \\ Y_t \end{bmatrix} = \Theta_1 \hat{X}_t + \Theta_2 U_t + W_t, \tag{3.2}$$

where W_t denotes an error term. This will be called the "ideal" regression model.

We shall now attempt to estimate the errors in the parameters which are caused by the approximation error in the state vector. It will be instructive to do this error analysis for the (unrealistic) ideal model first, since this model is linear in the parameters and the estimates are straightforward. The results will then play the role of a comparison term for the (realistic) methods proposed in the literature which we shall analyze later on.

In the following we shall make the blanket assumption that

$$\hat{\mathscr{X}}_t \cap \mathscr{U}_{[t,T]} = \{0\} \tag{3.3}$$

holds. This condition, which is necessary for making various oblique projection operators well-defined, is equivalent to the non-singularity of the conditional covariance matrix $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}$ which we shall introduce later on, and has been called *consistency condition* by Jansson and Wahlberg (1997,1998). In these references counter-examples are given in which $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}$ may become singular and some sufficient conditions for non-singularity are provided. See e.g. Eq. (25) in Jansson and Wahlberg (1998). The papers Peternell, Scherrer and Deistler, 1996; Bauer and Jansson, 2000 provide generic conditions for the consistency condition to hold, and in Chui, 1997 global consistency conditions for all systems of fixed degree n are discussed.

The least-squares estimate of Θ in (3.2) can be computed using the oblique projection Lemma of Katayama and Picci (1999, Lemma 1, p. 1637), leading to "sample" Wiener–Hopf type equations which can be written in either of the two forms:

$$\hat{\Theta}\hat{\Sigma} = \hat{\Sigma}_1, \quad \hat{\Theta}_1\hat{\Sigma}_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}} = \hat{\Sigma}_{\mathbf{z}\hat{\mathbf{x}}|\mathbf{u}}, \quad \hat{\Theta}_2\hat{\Sigma}_{\mathbf{u}\mathbf{u}|\mathbf{x}} = \hat{\Sigma}_{\mathbf{z}\mathbf{u}|\hat{\mathbf{x}}}, \tag{3.4}$$

where $\hat{\Sigma}$, $\hat{\Sigma}_1$, $\hat{\Sigma}_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}}$, etc. are sample estimates of the joint auto- and cross-covariance matrices of the data and $\hat{\Theta}$ denotes (matrix) parameter estimate. Now the sample covariance matrices can be expressed as the sum of "true" values, say Σ and Σ_1, equal to the limit (i.e. expected) values of $\hat{\Sigma}, \hat{\Sigma}_1$, plus random errors $\tilde{\Sigma}, \tilde{\Sigma}_1$, due to finite-sample effect: $\hat{\Sigma} = \Sigma + \tilde{\Sigma}\hat{\Sigma}_1 = \Sigma_1 + \tilde{\Sigma}_1$. These originate errors in the parameter estimates, namely we shall have $\hat{\Theta} = \Theta + \tilde{\Theta}$, where Θ is the "true" parameter value, solution of the infinite-datalength ($N = \infty$) ideal regression model consisting just of the first two lines of (2.16). For this model, the consistency condition (3.3) trivially implies the direct sum condition $\hat{\mathscr{X}}_t \cap \mathscr{U}_t = \{0\}$, and hence, applying again the oblique projection lemma (Katayama & Picci, 1999; Lemma 1, p. 1637), one gets the limiting expressions ($N = \infty$) of the

parameter estimates as solutions of the Wiener–Hopf type equations which in the present case take the form

$$\begin{bmatrix} A \\ C \end{bmatrix} \Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}} = \begin{bmatrix} \Sigma_{\hat{\mathbf{x}}_1\hat{\mathbf{x}}|\mathbf{u}} \\ \Sigma_{\mathbf{y}\hat{\mathbf{x}}|\mathbf{u}} \end{bmatrix}, \tag{3.5}$$

$$\begin{bmatrix} B \\ D \end{bmatrix} \Sigma_{\mathbf{u}\mathbf{u}|\hat{\mathbf{x}}} = \begin{bmatrix} \Sigma_{\hat{\mathbf{x}}_1\mathbf{u}|\hat{\mathbf{x}}} \\ \Sigma_{\mathbf{y}\mathbf{u}|\hat{\mathbf{x}}} \end{bmatrix}. \tag{3.6}$$

These expressions involve various conditional covariance matrices. Those appearing on the left hand side are

$$\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}} = \mathrm{Var}\{\hat{\mathbf{x}}(t) - E(\hat{\mathbf{x}}(t)|\mathbf{u}(t))\},$$

$$\Sigma_{\mathbf{u}\mathbf{u}|\hat{\mathbf{x}}} = \mathrm{Var}\{\mathbf{u}(t) - E(\mathbf{u}(t)|\hat{\mathbf{x}}(t))\}. \tag{3.7}$$

The formulas for $\Sigma_{\hat{\mathbf{x}}_1\hat{\mathbf{x}}|\mathbf{u}}$, etc. are similar, involving $\hat{\mathbf{x}}_1 \equiv \hat{\mathbf{x}}(t+1)$ and $\mathbf{y} \equiv \mathbf{y}(t)$.

Eqs. (3.5) and (3.6) are exact expressions valid for infinitely long data. Now, if N is large, the fluctuations on the covariance estimates due to the finiteness of the sample used to form the estimate, will be small (namely $O(1/N)$). In particular the perturbations on the singular values will be of the same order of the perturbations on the matrix elements (Stewart & Sun, 1990). Hence, for N large, second-order effects can be neglected and the amplification of the relative errors in the covariance estimate to relative errors in the parameter estimates, will depend only on the *true covariance matrix* Σ of $\hat{\mathbf{x}}, \mathbf{u}$. In particular, the size of the relative errors, say $\|\tilde{\Theta}\|/\|\Theta\|$, can be measured by the *condition number* of Σ. Similarly, we can write (up to second order terms)

$$\tilde{\Theta}_1 \Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}} = \tilde{\Sigma}_{\mathbf{z}\hat{\mathbf{x}}|\mathbf{u}}, \quad \tilde{\Theta}_2 \Sigma_{\mathbf{u}\mathbf{u}|\hat{\mathbf{x}}} = \tilde{\Sigma}_{\mathbf{z}\mathbf{u}|\hat{\mathbf{x}}}$$

and the magnitude of the (relative) errors in the estimates of (A, C) and of (B, D), may be assessed by looking at the condition numbers of the two "true" conditional covariance matrices $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}}$ and $\Sigma_{\mathbf{u}\mathbf{u}|\hat{\mathbf{x}}}$.

Remark 4 (concerning the choice of basis). Clearly, any invertible linear transformation of the $\hat{X}_t$ variable in regression (3.2) leads to a similarity transformation of the estimated model parameters. It is desirable that the measure of relative error on the parameter estimates should be invariant with respect to such similarity transformations, since after all we are only interested in estimating quantities which are *invariant with respect to a change of basis*, such as the transfer function, the system poles, etc.

Unfortunately, the condition number of $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}}$ changes by changing basis (since the condition number of $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$ does) and it may seem that, just by changing basis in the true model (2.16), we could make the worst-case relative norm (or variance) of the parameter estimation error $\tilde{\Theta}_1$, artificially large as we wish. Hence the condition number of the subproblem (3.5) is not a fair measure of error propagation. This is essentially the same phenomenon called *artificial ill-conditioning* by Stewart (1987, p. 7). As argued in this reference, a possible solution to artificial ill-conditioning is to introduce some sort of scaling or normalization of the regressors. In our problem this amounts to normalizing the (error in the) state by fixing a suitable basis. We may for this purpose choose any convenient "canonical" basis in the model (2.16). For example, we may choose a state vector $\hat{\mathbf{x}}(t)$ with orthonormal components, corresponding to taking $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = I$.

Consider then the Cholesky factors $L_{\hat{\mathbf{x}}}$ of $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$ and $L_{\mathbf{u}}$ of $\Sigma_{\mathbf{u}\mathbf{u}}$, so that $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = L_{\hat{\mathbf{x}}} L_{\hat{\mathbf{x}}}^\top$, $\Sigma_{\mathbf{u}\mathbf{u}} = L_{\mathbf{u}} L_{\mathbf{u}}^\top$. Using a well-known formula for the conditional covariances, we have

$$\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}} = \Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}} - \Sigma_{\mathbf{x}\mathbf{u}} \Sigma_{\mathbf{u}\mathbf{u}}^{-1} \Sigma_{\mathbf{u}\mathbf{x}} = L_{\hat{\mathbf{x}}}[I - \Pi_{\hat{\mathbf{x}}\mathbf{u}} \Pi_{\hat{\mathbf{x}}\mathbf{u}}^\top] L_{\hat{\mathbf{x}}}^\top,$$

$$\Sigma_{\mathbf{u}\mathbf{u}|\hat{\mathbf{x}}} = \Sigma_{\mathbf{u}\mathbf{u}} - \Sigma_{\mathbf{u}\mathbf{x}} \Sigma_{\mathbf{x}\mathbf{x}}^{-1} \Sigma_{\mathbf{x}\mathbf{u}} = L_{\mathbf{u}}[I - \Pi_{\mathbf{u}\hat{\mathbf{x}}} \Pi_{\mathbf{u}\hat{\mathbf{x}}}^\top] L_{\mathbf{u}}^\top, \tag{3.8}$$

where $\Pi_{\hat{\mathbf{x}}\mathbf{u}}$ is the normalized cross-covariance $\Pi_{\hat{\mathbf{x}}\mathbf{u}} := L_{\mathbf{x}}^{-1} \Sigma_{\mathbf{x}\mathbf{u}} L_{\mathbf{u}}^{-\top} = \Pi_{\mathbf{u}\hat{\mathbf{x}}}^\top$ whose singular values (bounded by one in magnitude) are the well-known *canonical correlation coefficients* of the input and present state. The canonical correlation coefficients are the cosines of the principal angles (Golub & Van Loan, 1989) between the *subspaces* spanned by the two blocks of random variables and are clearly invariant with respect to change of basis. In fact, we can see that the error propagation, in any fixed, in particular orthonormal, basis, is governed by the matrix $I - \Pi_{\hat{\mathbf{x}}\mathbf{u}} \Pi_{\hat{\mathbf{x}}\mathbf{u}}^\top$ which is also basis independent. We may say that error propagation in the regression (3.5) depends on the degree of *collinearity of the regressors*.[6]

Any measure of near singularity of $I - \Pi_{\hat{\mathbf{x}}\mathbf{u}} \Pi_{\hat{\mathbf{x}}\mathbf{u}}^\top$, e.g. the smallest eigenvalue, can be used as an index of collinearity. In fact, except for a pathologically unlikely situation of small but all identical principal angles, the condition number of this matrix can also be used.

Proposition 5. *For the ideal regression* (3.2) *in a fixed basis, the relative accuracy of the estimates of* A, C depends only on the collinearity of $\hat{\mathbf{x}}$ and $\mathbf{u}$, *namely on the size of the smallest singular value of* $I - \Pi_{\hat{\mathbf{x}}\mathbf{u}} \Pi_{\hat{\mathbf{x}}\mathbf{u}}^\top$. *Dually, for a given input covariance, or, in particular*, for a given input signal, *the relative accuracy of the estimates of* B, D *depends only on the size of the smallest singular value of the normalized cross-covariance matrix* $I - \Pi_{\mathbf{u}\hat{\mathbf{x}}} \Pi_{\mathbf{u}\hat{\mathbf{x}}}^\top$.

[6] Regressors are called *collinear* when some principal angle between the relative subspaces is close to zero.

An essentially equivalent statement will be shown to hold also for the N4SID and MOESP methods. See Proposition 7 below.

Remark 6. One may wonder why the standard algorithms in the literature do not suggest simply to choose $\hat{\mathbf{x}}(t)$ with orthonormal components and use instead specific choices of basis which are not orthonormal. The reason seems to be that the choice of basis influences the accuracy of the state estimation step, i.e. the error incurred in the construction of the finite data estimate $\hat{X}_t$. The state estimation is, in fact, a *model reduction step*, implemented in practice by discarding "small" singular values in a SVD of a certain "large" Hankel matrix. This makes some special choices of basis preferable, since they may lead to a better accuracy in the approximation step and eventually to a more accurate transfer function estimate. Consider for example the case when the exogenous input $\mathbf{u}(t)$ is uncorrelated with $\hat{\mathbf{x}}(t)$, in particular when $\mathbf{u}$ is white or absent (e.g. in the situation of time series identification). Then (3.5) is replaced by

$$\begin{bmatrix} A \\ C \end{bmatrix} \Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = \begin{bmatrix} \Sigma_{\mathbf{x}_1\hat{\mathbf{x}}} \\ \Sigma_{\mathbf{y}\hat{\mathbf{x}}} \end{bmatrix}, \tag{3.9}$$

which implies that, in this case, it is the condition number of the *unconditional* state covariance matrix $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$ that determines the accuracy of estimation of (A, C). Since $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}}$ depends only on the particular choice of basis in the state space it could, say, be made equal to the identity matrix by choosing a state vector with orthonormal components, apparently resulting in a perfectly well-conditioned estimation problem.

Actually, for the reasons explained above, this choice of basis is *not* to be recommended; the "best" choice being the so-called *stochastically balanced* basis described (e.g. Desai & Pal, 1984).

Although practical subspace identification methods require to deal with more complicated regression problems, it will be seen in the following sections that the analysis of the ideal case remains valid to a large extent.

3.2. Conditioning of N4SID

We shall quickly review the N4SID algorithm of Van Overschee and De Moor, 1994. The first object which is computed is the output predictor matrix based on joint input–output data. By following the same steps leading to (2.18), but now using tails of finite length, one obtains the relation [7]

$$\begin{aligned} Z_{[t,T]} &:= E_N[Y_{[t,T]}|P_{[t_0,t)} \vee U_{[t,T]}] \\ &= \Gamma_\nu \hat{X}_t + H_d U_{[t,T]} + H_s \tilde{E}_{[t,T]}, \end{aligned} \tag{3.10}$$

where $\Gamma'_\nu \nu = T - t$, is the (extended) observability matrix of model (2.16), $\hat{X}_t$ is the $n \times N$ tail matrix of the Kalman filter state, H_d (see (2.18)) and H_s are the lower triangular block Toeplitz matrices of the deterministic subsystem (A, B, C, D) and of the stochastic subsystem (A, B, K, I) and $\tilde{E}_{[t,T]}$ is an error term given by

$$\tilde{E}_{[t,T]} := E_N[E_{[t,T]}|P_{[t_0,t)} \vee U_{[t,T]}] \tag{3.11}$$

which goes to zero as $N \to \infty$.

Motivated by this expression, an estimate, $\hat{\Gamma}_\nu$, of the observability matrix Γ_ν is obtained by an oblique projection of $Z_{[t,T]}$ along $\mathscr{U}_{[t,T]}$, followed by an SVD factorization and order estimation by neglecting small singular values.

Once $\hat{\Gamma}_\nu$ is computed, the second step of the procedure is to form the matrix $\hat{\Gamma}_\nu^\dagger Z_{[t,T]}$, where † denotes the Moore–Penrose pseudo-inverse. Assuming that the order estimation in the SVD step is consistent, $\hat{\Gamma}_\nu$ converges to the true observability matrix for $N \to \infty$. Now $\hat{\Gamma}_\nu^\dagger Z_{[t,T]}$ approximates the "pseudostate"

$$\Gamma_\nu^\dagger Z_{[t,T]} = \hat{X}_t + \Gamma_\nu^\dagger H_d U_{[t,T]} + \Gamma_\nu^\dagger H_s \tilde{E}_{[t,T]} \tag{3.12}$$

(see Eq. (15) in Van Overschee & De Moor, 1994) which satisfies a linear recursion of the form

$$\begin{bmatrix} \Gamma_{\nu-1}^\dagger Z_{[t+1,T]} \\ Y_t \end{bmatrix} = \begin{bmatrix} A \\ C \end{bmatrix} \Gamma_\nu^\dagger Z_{[t,T]} + \begin{bmatrix} \mathscr{K}_1 \\ \mathscr{K}_2 \end{bmatrix} U_{[t,T]} + E^\perp, \tag{3.13}$$

where $(\mathscr{K}_1, \mathscr{K}_2)$ are known functions of the parameters of the stationary system, depending in particular on the (A, C) parameters, see Eq. (43) in Van Overschee and De Moor (1994). This relation is nevertheless interpreted as a linear regression, as if the unknown parameters (A, C) and $(\mathscr{K}_1, \mathscr{K}_2)$ were independent. With infinitely many data $(N \to \infty)$, the term $E^\perp$ is orthogonal to the row span of $U_{[t_0,T]}$ and to the row span of $\Gamma_\nu^\dagger Z_{[t,T]}$. Hence the least-square solution of the regression problem (3.13) provides consistent estimates of the parameters (A, C) and $(\mathscr{K}_1, \mathscr{K}_2)$. Assuming N is large enough, so that the rowspaces of $\hat{\Gamma}_\nu^\dagger Z_{[t,T]}$ and $U_{[t,T]}$ have only the zero vector in common, the estimates are the oblique projections of the LHS of (3.13) (naturally one uses $\hat{\Gamma}_{\nu-1}^\dagger Z_{[t+1,T]}$ in place of $\Gamma_{\nu-1}^\dagger Z_{[t+1,T]}$), onto the rowspace of $\hat{\Gamma}_\nu^\dagger Z_{[t,T]}$ along the rowspace of $U_{[t,T]}$ and, respectively, onto $U_{[t,T]}$ along the rowspace of $\hat{\Gamma}_\nu^\dagger Z_{[t,T]}$.

Clearly, a crucial issue for assessing the sensitivity of the solutions to noise, is how "parallel" are the rowspaces of $U_{[t,T]}$ and $\hat{\Gamma}_\nu^\dagger Z_{[t,T]}$. For nearly parallel row spaces one expects that the estimation of the parameters (A, C) and $(\mathscr{K}_1, \mathscr{K}_2)$ of the regression will be ill-conditioned, and the parameter estimates be affected by large errors.

To analyze this situation we shall again assume that the sample size $N \to \infty$ and adopt the stochastic setup used in the previous section. Tail matrices will again be

[7] In Van Overschee and De Moor (1994) the notation Z_t is used instead of our $Z_{[t,T]}$.

substituted by their corresponding limit stochastic vectors, denoted by lower case boldface symbols. In particular, the limit pseudo-state vector (3.12) at time t, can be written as the sum of the conditional Kalman filter state, $\hat{\mathbf{x}}(t)$, and the future-input dependent quantity $\Gamma_v^\dagger H_d \mathbf{u}_t^+$ so that, denoting $\hat{\hat{\mathbf{x}}}(t) := \hat{\mathbf{x}}(t) + \gamma_v^\dagger H_d \mathbf{u}_t^+$ and using (2.16), in the limit (3.13) becomes

$$\begin{bmatrix} \hat{\hat{\mathbf{x}}}(t+1) \\ \mathbf{y}(t) \end{bmatrix} = \begin{bmatrix} A \\ C \end{bmatrix} \hat{\hat{\mathbf{x}}}(t) + \begin{bmatrix} \mathscr{K}_1 \\ \mathscr{K}_2 \end{bmatrix} \mathbf{u}_t^+ + \begin{bmatrix} K(t) \\ I \end{bmatrix} \hat{\mathbf{e}}(t), \tag{3.14}$$

where the last term is orthogonal to the space spanned by the random variables $\hat{\hat{\mathbf{x}}}(t), \mathbf{u}_t^+$.

Now, under the consistency condition (3.3), the subspaces spanned by (the components of) $\hat{\hat{\mathbf{x}}}(t)$ and $\mathbf{u}_t^+$ have only the zero random variable in common. In this situation we can use the oblique projection Lemma (Katayama & Picci, 1999; Lemma 1, p. 1637), to express the parameter estimates as solutions of the Wiener–Hopf type equations

$$\begin{bmatrix} A \\ C \end{bmatrix} \Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}|\mathbf{u}^+} = \begin{bmatrix} \Sigma_{\bar{\mathbf{x}}_1\bar{\mathbf{x}}|\mathbf{u}^+} \\ \Sigma_{\mathbf{y}\bar{\mathbf{x}}|\mathbf{u}^+} \end{bmatrix},$$

$$\begin{bmatrix} \mathscr{K}_1 \\ \mathscr{K}_2 \end{bmatrix} \Sigma_{\mathbf{u}^+\mathbf{u}^+|\bar{\mathbf{x}}} = \begin{bmatrix} \Sigma_{\bar{\mathbf{x}}_1\mathbf{u}^+|\bar{\mathbf{x}}} \\ \Sigma_{\mathbf{y}\mathbf{u}^+|\bar{\mathbf{x}}} \end{bmatrix},$$

which also involve various conditional covariance matrices. The expressions appearing in the left hand side are

$$\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}|\mathbf{u}^+} = \mathrm{Var}\{\hat{\hat{\mathbf{x}}}(t) - E(\hat{\hat{\mathbf{x}}}(t)|\mathbf{u}_t^+)\},$$

$$\Sigma_{\mathbf{u}^+\mathbf{u}^+|\bar{\mathbf{x}}} = \mathrm{Var}\{\mathbf{u}_t^+ - E(\mathbf{u}_t^+|\hat{\hat{\mathbf{x}}}(t))\}, \tag{3.15}$$

the formulas for $\Sigma_{\bar{\mathbf{x}}_1\bar{\mathbf{x}}|\mathbf{u}^+}$, etc. being similar, involving $\bar{\mathbf{x}}_1 \equiv \hat{\hat{\mathbf{x}}}(t+1)$ and $\mathbf{y} \equiv \mathbf{y}(t)$. Noting the obvious identity

$$\Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}|\mathbf{u}^+} = \mathrm{Var}\{\hat{\mathbf{x}}(t) - E(\hat{\mathbf{x}}(t)|\mathbf{u}_t^+)\} := \Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}$$

we come to the conclusion that *the conditioning of the problem* (3.13) *is the same of the two linear problems*

$$\begin{bmatrix} A \\ C \end{bmatrix} \Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+} = \begin{bmatrix} \Sigma_{\hat{\mathbf{x}}_1\bar{\mathbf{x}}|\mathbf{u}^+} \\ \Sigma_{\mathbf{y}\hat{\mathbf{x}}|\mathbf{u}^+} \end{bmatrix}, \tag{3.16}$$

$$\begin{bmatrix} \mathscr{K}_1 \\ \mathscr{K}_2 \end{bmatrix} \Sigma_{\mathbf{u}^+\mathbf{u}^+|\bar{\mathbf{x}}} = \begin{bmatrix} \Sigma_{\hat{\mathbf{x}}_1\mathbf{u}^+|\bar{\mathbf{x}}} \\ \Sigma_{\mathbf{y}\mathbf{u}^+|\bar{\mathbf{x}}} \end{bmatrix}. \tag{3.17}$$

Let us now introduce the cross-covariance matrices $\bar{\Pi} := E[\mathbf{u}_t^+ \hat{\hat{\mathbf{x}}}(t)^\top]$, $\Pi := E[\mathbf{u}_t^+ \hat{\mathbf{x}}(t)^\top]$, and let $L_{\mathbf{u}^+}$ denote the Cholesky factor of $\Sigma_{\mathbf{u}^+\mathbf{u}^+}$ (the covariance matrix of $\mathbf{u}_t^+$), so that $L_{\mathbf{u}^+} L_{\mathbf{u}^+}^\top = \Sigma_{\mathbf{u}^+\mathbf{u}^+}$. Further, let $\hat{\Pi}$ and $\hat{\bar{\Pi}}$ be the normalized cross-covariances

$$\hat{\bar{\Pi}} := L_{\mathbf{u}^+}^{-1} \bar{\Pi} L_{\bar{\mathbf{x}}}^{-\top}, \quad \hat{\Pi} := L_{\mathbf{u}^+}^{-1} \Pi L_{\hat{\mathbf{x}}}^{-\top},$$

whose singular values are the canonical correlation coefficients of $\mathbf{u}_t^+$ and $\hat{\mathbf{x}}(t)$, and of $\mathbf{u}_t^+$ and $\hat{\hat{\mathbf{x}}}(t)$, respectively. These quantities are interpreted as the cosines of the principal angles between the subspaces generated by $\mathbf{u}_t^+$ and $\hat{\mathbf{x}}(t)$ and, respectively, between the subspaces generated by $\mathbf{u}_t^+$ and $\hat{\hat{\mathbf{x}}}(t)$. Evidently these angles do not depend on the particular generators, but only on the spanned subspaces.

Proposition 7. *In any fixed basis* $\hat{\mathbf{x}}(t)$, *the condition number* $\kappa\left(\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}\right)$ *depends only on the principal angles between the future input subspace generated by* $\mathbf{u}_t^+$ *and the state space* $\hat{\mathscr{X}}_t$. *In fact, we have the following bound*:

$$\kappa\left(\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}\right) \leqslant \kappa\left(\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\right) \frac{1 - \sigma_{\min}^2(\hat{\Pi})}{1 - \sigma_{\max}^2(\hat{\Pi})}, \tag{3.18}$$

where $\sigma_{\max}, \sigma_{\min}$ *denote the largest and smallest singular values. In particular, if* $\hat{\mathbf{x}}(t)$ *is an orthonormal basis for the state space* $\hat{\mathscr{X}}_t$ *of the model* (2.16), *the maximal and minimal singular values of* $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}$ *are* $1 - \sigma_{\min}^2(\hat{\Pi})$ *and* $1 - \sigma_{\max}^2(\hat{\Pi})$, *respectively and* $\kappa\left(\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}\right)$ *is equal to the right factor in* (3.18).

Similarly, for a fixed input covariance, $\Sigma_{\mathbf{u}^+\mathbf{u}^+}$, *the condition number* $\kappa\left(\Sigma_{\mathbf{u}^+\mathbf{u}^+|\bar{\mathbf{x}}}\right)$ *depends only on the principal angles between the subspaces generated by* $\mathbf{u}_t^+$ *and* $\hat{\hat{\mathbf{x}}}(t)$ *and can be estimated by*

$$\kappa\left(\Sigma_{\mathbf{u}^+\mathbf{u}^+|\bar{\mathbf{x}}}\right) \leqslant \kappa\left(\Sigma_{\mathbf{u}^+\mathbf{u}^+}\right) \frac{1}{1 - \sigma_{\max}^2(\hat{\bar{\Pi}})}. \tag{3.19}$$

These bounds are sharp (*i.e. there are situations in which the inequalities* (3.18) *and* (3.19) *become equalities*).

Proof. The formulas are based on the factorizations $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+} = L_{\hat{\mathbf{x}}}(I - \hat{\Pi}^\top \hat{\Pi}) L_{\hat{\mathbf{x}}}^\top$ and $\Sigma_{\mathbf{u}^+\mathbf{u}^+|\bar{\mathbf{x}}} = L_{\mathbf{u}^+}(I - \hat{\bar{\Pi}}\hat{\bar{\Pi}}^\top) L_{\mathbf{u}^+}^\top$, which follow from well-known expressions for the conditional covariances $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}$ and $\Sigma_{\mathbf{u}^+\mathbf{u}^+|\bar{\mathbf{x}}}$. From these formulas it is readily seen that

$$\lambda_{\max}\left(\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}\right) \leqslant \lambda_{\max}\left(L_{\hat{\mathbf{x}}} L_{\hat{\mathbf{x}}}^\top\right)\left(1 - \lambda_{\min}\left(\hat{\Pi}^\top \hat{\Pi}\right)\right),$$

$$\lambda_{\min}\left(\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}\right) \geqslant \lambda_{\min}\left(L_{\hat{\mathbf{x}}} L_{\hat{\mathbf{x}}}^\top\right)\left(1 - \lambda_{\max}\left(\hat{\Pi}^\top \hat{\Pi}\right)\right),$$

so that taking the quotient one gets (3.18). The proof of (3.19) is similar; just note that $\lambda_{\min}\left(\hat{\bar{\Pi}}\hat{\bar{\Pi}}^\top\right) = 0$ since in all

subspace methods the dimension of $\mathbf{u}_t^+$ is larger than that of $\hat{\hat{\mathbf{x}}}(t)$ so that $\hat{\hat{\Pi}}\hat{\hat{\Pi}}^\top$ is singular. □

As the state space $\hat{\mathscr{X}}_t$ and the future input space $\mathscr{U}_{[t,T]}$ become "closer" (i.e. the smallest canonical angle approaches zero), $\sigma_{\max}\left(\hat{\Pi}\right)$ and $\sigma_{\max}\left(\hat{\hat{\Pi}}\right)$ tend to one and the problem becomes more ill-conditioned. In fact, when, say, the smallest canonical angle between the subspaces generated by $\mathbf{u}_t^+$ and $\hat{\mathbf{x}}(t)$ tends to zero, the two spaces tend to have a non-zero intersection and $\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}$ becomes singular, i.e. $\sigma_{\min}(\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}) \to 0$ in which case one has, except for very degenerate cases, $\kappa(\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+}) \to \infty$. Similarly one may show that $\kappa\left(\Sigma_{\mathbf{u}^+\mathbf{u}^+|\bar{\mathbf{x}}}\right) \to \infty$.

As we can see from (3.19), the computation of $(\mathscr{K}_1, \mathscr{K}_2)$ can be ill-conditioned also when $\kappa(\Sigma_{\mathbf{u}^+\mathbf{u}^+})$ is large. This possible cause of ill-conditioning has to do with wide variations in the amplitude of the input spectrum (Grenander & Szegö, 1958; Söderström & Stoica, 1989) and is observed in particular when there are frequency bands where the spectrum is nearly zero, causing "insufficient excitation". Although there is a general diffuse understanding of this phenomenon, a precise characterization of its role in subspace identification does not seem to have been pointed out before. The situation is relatively safe if $\mathbf{u}$ is a white noise process, in which case $\kappa(\Sigma_{\mathbf{u}^+\mathbf{u}^+}) = 1$. However, even when $\mathbf{u}$ is nearly white, the problem (3.17) could still be badly conditioned due to small canonical angles between the state and future input spaces as it can be seen from (3.19).

One also sees that the condition numbers of the N4SID regression are always *larger than those occurring for the ideal regression problem* (3.1), i.e. for the linear Eqs. (3.5) and (3.6). The more so, the larger the future horizon $v = T - t$. This is so, since the bounds (3.18) (3.19) involve angles with the whole future input space, while for the ideal regression (3.13) only angles with the *present* input space are involved.

4. Conditioning of the robust N4SID and of MOESP-type methods

In this section we shall see that the computation of the asymptotic estimates of (A, C) in the "robust" N4SID method of Van Overschee and De Moor (1996) and in the so-called PO-MOESP method of Verhaegen, 1994, is described, except for a change of basis in the state space, by the same formulas found for N4SID. Therefore the same conditioning analysis which applies to N4SID (in particular Proposition 7) also applies to these two methods.

In the process of doing this, we shall actually establish that the algorithm PO-MOESP and the "robustified" N4SID method, produce the same estimates of A and C even for finite sample size N. This fact seems to have been noticed experimentally before, but, to the best of our knowledge, has never been formally proven.

Both methods are based on a predictor matrix $Z^{\mathrm{c}}_{[t,T]}$, defined as the orthogonal projection of the future outputs $Y_{[t,T]}$ onto the "complementary" data space [8] spanned by the rows of the matrix $U^{\perp}_{[t,T]} := P_{[t_0,t)} - E_N[P_{[t_0,t)}|U_{[t,T]}]$.

Consider the singular value decomposition

$$Z^{\mathrm{c}}_{[t,T]} = USV^\top = [U_1\ U_2] \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix}$$

having say n_{c} "most significant" singular values, where U_1 are the first n_{c} columns of U, V_1 the first n_{c} rows of V and S_1 the upper-left $n_{\mathrm{c}} \times n_{\mathrm{c}}$ corner of S. By neglecting the "small" singular values, we obtain a full-rank factorization

$$Z^{\mathrm{c}}_{[t,T]} = \hat{\Gamma}^{\mathrm{c}}_v \hat{X}^{\mathrm{c}}_t, \tag{4.1}$$

where

$$\hat{\Gamma}^{\mathrm{c}}_v = U_1 S_1^{1/2}, \quad \hat{X}^{\mathrm{c}}_t = S_1^{1/2} V_1^T. \tag{4.2}$$

Since as $N \to \infty$, the last term in the expression

$$\begin{aligned} Z^{\mathrm{c}}_{[t,T]} &= E_N[Z_{[t,T]}|U^{\perp}_{[t,T]}] \\ &= \Gamma_v E_N[\hat{X}_t|U^{\perp}_{[t,T]}] + E_N[W^{\perp}|U^{\perp}_{[t,T]}] \end{aligned} \tag{4.3}$$

tends to zero and, in force of the consistency condition (3.3), the rank of the projected matrix $E_N[\hat{X}_t|U^{\perp}_{[t,T]}]$ is equal to n (the true state dimension), the oblique projection of $Z_{[t,T]}$ along the rowspace of $U_{[t,T]}$ and $Z^{\mathrm{c}}_{[t,T]}$ asymptotically have the same column spaces and the same rank n.

Hence if the rank determination step in the factorization (4.1) is statistically consistent (i.e., asymptotically $n_{\mathrm{c}} = n$) the two factors in (4.1) both admit a limit, in a sense made precise in the following proposition.

Proposition 8. *Assume that the rank determination step in the factorization* (4.1) *is statistically consistent, then, in the limit for* $N \to \infty$, *the factors* (4.1) *of the complementary predictor* $Z^{\mathrm{c}}_{[t,T]}$ *converge, in the following sense. There is a* $n \times n$ *nonsingular matrix* T *such that*,

$$\hat{\Gamma}^{\mathrm{c}}_v \to \Gamma_v T^{-1} \tag{4.4}$$

and the tail matrix $\hat{X}^{\mathrm{c}}_t$ *becomes the random vector*

$$\hat{\mathbf{x}}^{\mathrm{c}}(t) := TE[\hat{\mathbf{x}}(t)|\mathscr{U}^{\perp}_{[t,T]}] = T(\hat{\mathbf{x}}(t) - E[\hat{\mathbf{x}}(t)|\mathscr{U}_{[t,T)}]) \tag{4.5}$$

called the complementary state *of the system. The complementary state satisfies the recursion*

$$\begin{bmatrix} \hat{\mathbf{x}}^{\mathrm{c}}(t+1) \\ \mathbf{y}(t) \end{bmatrix} = \begin{bmatrix} A^{\mathrm{c}} \\ C^{\mathrm{c}} \end{bmatrix} \hat{\mathbf{x}}^{\mathrm{c}}(t) + \begin{bmatrix} \mathscr{B}_1 \\ \mathscr{B}_2 \end{bmatrix} \mathbf{u}_t^+ + \tilde{\mathbf{e}}^{\perp}, \tag{4.6}$$

where

$$A^{\mathrm{c}} = TAT^{-1} \quad C^{\mathrm{c}} = CT^{-1}. \tag{4.7}$$

[8] "Complementary", since it is the orthogonal complement of $\mathscr{U}^N_{[t,T]}$ in the data space $\mathscr{P}^N_{[t_0,t)} \vee \mathscr{U}^N_{[t,T]}$. The notation $U^{\perp}_{[t,T]}$ is not completely consistent since the ambient space of the complement varies with t.

$\mathscr{B}_1, \mathscr{B}_2$ are suitable matrix functions of the parameters of the stationary system generating the data, and $\tilde{\mathbf{e}}^\perp$, a random vector orthogonal to the data space $\mathscr{P}_{[t_0,t)} \vee \mathscr{U}_{[t,T]}$.

The first two terms in the right member of (4.6) are uncorrelated. Hence the parameters (A^c, C^c) are the solution of the Wiener–Hopf type equations

$$A^c(E[\hat{\mathbf{x}}^c(t)(\hat{\mathbf{x}}^c(t))^\top]) = E[\hat{\mathbf{x}}^c(t+1)(\hat{\mathbf{x}}^c(t))^\top], \tag{4.8}$$

$$C^c(E[\hat{\mathbf{x}}^c(t)(\hat{\mathbf{x}}^c(t))^\top]) = E[\mathbf{y}(t)(\hat{\mathbf{x}}^c(t))^\top]. \tag{4.9}$$

Proof. That for $N \to \infty$, $Z^c_{[t,T]}$ tends to $\Gamma_v \hat{\mathbf{x}}^c(t)$ follows from (4.3) and from

$$\hat{X}^c_t := E_N[\hat{X}_t | U^\perp_{[t,T]}] \to \hat{\mathbf{x}}^c(t)$$

(second-order ergodicity). Then (4.4) follows from the consistency assumption and the discussion preceding the statement of the proposition. Consider next the N4SID pseudostate $\hat{\tilde{\mathbf{x}}}(t)$, introduced in the previous section. It follows from (4.5) that

$$T\hat{\tilde{\mathbf{x}}}(t) = T\hat{\mathbf{x}}(t) + T\gamma_v^\dagger H_d \mathbf{u}_t^+ = \hat{\mathbf{x}}^c(t) + \phi_t \mathbf{u}_t^+,$$

where $\Phi_t \mathbf{u}_t^+ := TE\left[\hat{\mathbf{x}}(t)|\mathbf{u}_t^+\right] + T\Gamma_v^\dagger H_d \mathbf{u}_t^+$. It is clear that Φ_t depends only on the parameters of the system. Substituting the last expression into (3.14) and collecting the terms which depend on $\mathbf{u}_t^+$, one obtains (4.6), with $\tilde{\mathbf{e}}^\perp = T\mathbf{e}^\perp$.

By construction $\hat{\mathbf{x}}^c(t)$ and $\mathbf{u}_t^+$ are uncorrelated, therefore right-multiplying (4.6) by $\hat{\mathbf{x}}^c(t)^\top$ and taking expectations one obtains (4.8), (4.9). □

Note that the covariance of the complementary state is, modulo a change of basis, just the conditional covariance of the Kalman state $\hat{\mathbf{x}}(t)$ given the future outputs $\mathbf{u}_t^+$,

$$\Sigma_{\hat{\mathbf{x}}^c\hat{\mathbf{x}}^c} = (E[\hat{\mathbf{x}}^c(t)(\hat{\mathbf{x}}^c(t))^\top]) = T\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+} T^\top. \tag{4.10}$$

This is a direct consequence of (4.5), which will be used later.

4.1. The "Robust" N4SID

The new estimate (4.2) of the observability matrix can be used instead of $\hat{\Gamma}_v$ in order to get a more robust estimate of the N4SID pseudostate (3.12). This leads to the "robust" N4SID method, where the estimates of A and C are obtained by solving in the least square sense the regression

$$\begin{bmatrix} (\hat{\Gamma}^c_{v-1})^\dagger Z_{[t+1,T]} \\ Y_t \end{bmatrix} = \begin{bmatrix} A^c \\ C^c \end{bmatrix} (\hat{\Gamma}^c_v)^\dagger Z_{[t,T]} + \begin{bmatrix} \mathscr{K}^c_1 \\ \mathscr{K}^c_2 \end{bmatrix} U_{[t,T]} + W^\perp. \tag{4.11}$$

The parameters now carry a superscript c since they are not exactly the same of the regression model (3.13) but are instead related by the change of basis

$$A_c = TAT^{-1}, \quad C_c = CT^{-1}, \quad \mathscr{K}^c_i = T\mathscr{K}_i, \quad i = 1, 2, \tag{4.12}$$

where T is the asymptotic change of basis corresponding to the SVD factorization (4.1).

It follows from the asymptotic analysis of the previous paragraphs that solving these equations leads asymptotically to Wiener–Hopf equations of the same type as (3.16) (3.17). In fact, we see that the estimates of the standard and robustified N4SID methods asymptotically differ only by a nonsingular change of basis, and therefore the same conditioning analysis of the previous subsection applies to the estimation of A, C by the robustified method.

It should be noted that although the change of basis induced by the matrix $T = (\Gamma^c_v)^\dagger \Gamma_v$ leads to a change of the eigenvalues of $T\Sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}|\mathbf{u}^+} T^\top$, and seemingly to a possibly different conditioning of the two methods, the change of basis does not affect the canonical angles and hence the effect of a near parallelism of the state and future input spaces in the accuracy of computation of A and C is exactly the same in the two methods.

Remark 9. It may be objected that in practice one observes an improvement in the robust N4SID with respect to the standard N4SID method. The improvement is apparently due to avoiding the computation of the observability matrix by an oblique projection as it was instead suggested in the early version of the algorithm. The crucial modification to this effect is the right-multiplication of the predictor matrix $Z_{[t,T]}$ by the projection matrix onto the orthogonal complement of the future input space, $U^\perp_{[t,T]}$, to form $Z^c_{[t,T]}$. See formula (4.3). This leads to a different SVD factorization and in general to a better estimate of the observability matrix. We have just shown however, that, asymptotically, the regression determining A, C is exactly the same as that occurring in N4SID, except for a change of basis in the state space. One can see from the formulas given in the companion paper Chiuso & Picci, 2004b (see also Chiuso & Picci, 2004a), that the asymptotic variances of the two estimates are in fact the same up to a change of basis in the state space. This of course has no influence on the variance of the estimated transfer function.

To add some more evidence supporting this conclusion, one may check with various expressions for the asymptotic variance derived in Bauer, Deistler and Scherrer (2000) and Bauer and Ljung (2001), where it is shown that, provided the basis has been fixed, the "right" weighting matrix in the SVD calculation, has asymptotically no influence on the accuracy of estimates obtained by algorithms based on the so called "state sequence" approach, which is precisely the one pursued by N4SID and its robustified version.

Of course asymptotic variances are only part of the story and it is well known that procedures which yield

asymptotically equivalent estimates may lead to noticeable differences in finite time computations. As shown in Chiuso and Picci, 2000, the (finite-sample) accuracy of the estimates in the standard and robustified N4SID methods may indeed differ dramatically. It is clear (and also shown in Chiuso & Picci, 2000) that poor accuracy in computing the N4SID estimate of Γ_ν with a finite sample, can also be attributed to near collinearity of the state and future inputs, since there is an oblique projection involved. However the errors due to collinearity enter in this analysis only as second order terms in the computation of the perturbation in A and C and so they do not influence the asymptotic variance. Clearly this does not prevent the finite sample accuracies to be significantly different as noted experimentally for instance looking at the poles of the estimates. The improvement in robust N4SID is essentially a finite-data phenomenon which is not visible from the asymptotic analysis performed in this section.

There are other improvements in the robust version regarding the estimation of (B, D) which we shall not discuss at this stage.

4.2. Equivalence of Robust N4SID and MOESP

We now show that the estimates of A and C obtained from the so-called PO-MOESP method (Verhaegen, 1994), are identical to those obtained from the "robust" N4SID method. Given a matrix Φ we shall denote by $\bar{\Phi}$ the matrix obtained from Φ removing the first m rows and by $\underline{\Phi}$ the one removing the last m rows.

Lemma 10. *For finite N, the right factor $\hat{X}_t^{\mathrm{c}}$ in (4.1) satisfies the recursion*

$$\begin{bmatrix} \hat{X}_{t+1}^{\mathrm{c}} \\ Y_t \end{bmatrix} = \begin{bmatrix} A_{\mathrm{c}} \\ C_{\mathrm{c}} \end{bmatrix} \hat{X}_t^{\mathrm{c}} + \begin{bmatrix} \mathscr{B}_1 \\ \mathscr{B}_2 \end{bmatrix} U_{[t,T]} + \tilde{W}^{\perp}, \tag{4.13}$$

where $\lim_{N\to\infty} \tilde{W}^{\perp} := \tilde{\mathbf{e}}^{\perp}$, a random vector orthogonal to the data space $\mathscr{P}_{[t_0,t)} \vee \mathscr{U}_{[t,T]}$. The first two terms in the sum in the right-hand side are orthogonal.

The least-squares solutions $(\hat{A}, \hat{C})$ of the regressions (4.11) and (4.13) are the same.

Proof. Since

$$\hat{X}_t^{\mathrm{c}} = (\hat{\Gamma}_\nu^{\mathrm{c}})^{\dagger} E_N[Z_{[t,T]} | U_{[t,T]}^{\perp}],$$

we can write

$$\hat{X}_t^{\mathrm{c}} = (\hat{\Gamma}_\nu^{\mathrm{c}})^{\dagger} Z_{[t,T]} + \Phi_t U_{[t,T]}.$$

for some suitable matrix Φ_t. In analogy with (4.1) we also define

$$\begin{aligned} Z_{t+1}^{\mathrm{c}} &= E_N[Y_{[t+1,T]} | U_{[t+1,T]}^{\perp}] \\ &= E_N[Z_{[t+1,T]} | U_{[t+1,T]}^{\perp}] \end{aligned} \tag{4.14}$$

and note that $\hat{X}_{t+1}^{\mathrm{c}} = (\hat{\Gamma}_{\nu-1}^{\mathrm{c}})^{\dagger} Z_{t+1}^{\mathrm{c}}$. Similarly, one can see that

$$\hat{X}_{t+1}^{\mathrm{c}} = (\hat{\Gamma}_{\nu-1}^{\mathrm{c}})^{\dagger} Z_{[t+1,T]} + \Phi_{t+1} U_{[t+1,T]}.$$

Substituting these expressions into (4.11) we get (4.13). The statement about $\tilde{\mathbf{e}}^{\perp}$ follows from Proposition 8. □

Proposition 11. *Let the estimate of the observability matrix $\hat{\Gamma}_\nu^{\mathrm{c}}$ be given by formula (4.2). Let $\hat{A} = \underline{(\hat{\Gamma}_\nu^{\mathrm{c}})}^{\dagger} \overline{(\hat{\Gamma}_\nu^{\mathrm{c}})}$ and let $\hat{C}$ be given by the first m rows of $\hat{\Gamma}_\nu^{\mathrm{c}}$, i.e. $\hat{C} = (\hat{\Gamma}_\nu^{\mathrm{c}})_{[1:m]}$. Then $(\hat{A}, \hat{C})$ solve (4.13) (and therefore (4.11)) in the least squares sense.*

Proof. The least-squares solutions, say $(\hat{A}^{\mathrm{c}}, \hat{C}^{\mathrm{c}})$, of (4.13), are given by

$$\hat{A}^{\mathrm{c}} = \hat{X}_{t+1}^{\mathrm{c}} (\hat{X}_t^{\mathrm{c}})^{\top} [\hat{X}_t^{\mathrm{c}} (\hat{X}_t^{\mathrm{c}})^{\top}]^{-1},$$

$$\hat{C}^{\mathrm{c}} = Y_t (\hat{X}_t^{\mathrm{c}})^{\top} [\hat{X}_t^{\mathrm{c}} (\hat{X}_t^{\mathrm{c}})^{\top}]^{-1}.$$

We shall focus first on the equation for $\hat{A}^{\mathrm{c}}$. From (4.1), and (4.14) we see that

$$\begin{aligned} &\hat{X}_{t+1}^{\mathrm{c}} (\hat{X}_t^{\mathrm{c}})^{\top} \\ &\quad = (\hat{\Gamma}_{\nu-1}^{\mathrm{c}})^{\dagger} E_N[Y_{[t+1,T]} | U_{[t+1,T]}^{\perp}] [(\hat{\Gamma}_\nu^{\mathrm{c}})^{\dagger} E_N[Y_{[t,T]} | U_{[t,T]}^{\perp}]]^{\top}. \end{aligned}$$

Since $\operatorname{span} U_{[t,T]}^{\perp} \subseteq \operatorname{span} U_{[t+1,T]}^{\perp}$, the formula above can be rewritten as

$$\begin{aligned} &\hat{X}_{t+1}^{\mathrm{c}} (\hat{X}_t^{\mathrm{c}})^{\top} \\ &\quad = (\hat{\Gamma}_{\nu-1}^{\mathrm{c}})^{\dagger} E_N[Y_{[t+1,T]} | U_{[t,T]}^{\perp}] [(\hat{\Gamma}_\nu^{\mathrm{c}})^{\dagger} E_N[Y_{[t,T]} | U_{[t,T]}^{\perp}]]^{\top} \\ &\quad = (\hat{\Gamma}_{\nu-1}^{\mathrm{c}})^{\dagger} \bar{Z}_{[t,T]}^{\mathrm{c}} (Z_{[t,T]}^{\mathrm{c}})^{\top} ((\hat{\Gamma}_\nu^{\mathrm{c}})^{\dagger})^{\top}. \end{aligned}$$

Similarly, we have

$$[\hat{X}_t^{\mathrm{c}} (\hat{X}_t^{\mathrm{c}})^{\top}] = (\hat{\Gamma}_\nu^{\mathrm{c}})^{\dagger} Z_{[t,T]}^{\mathrm{c}} (Z_{[t,T]}^{\mathrm{c}})^{\top} ((\hat{\Gamma}_\nu^{\mathrm{c}})^{\dagger})^{\top}.$$

Now it follows from (4.2) that the pseudo-inverses have the following expressions:

$$(\hat{\Gamma}_\nu^{\mathrm{c}})^{\dagger} = S_1^{-1/2} U_1^{\top}, \quad (\hat{\Gamma}_{\nu-1}^{\mathrm{c}})^{\dagger} = \underline{(U_1 S_1^{1/2})}^{\dagger}$$

and hence the least-squares estimate of A^{c} by the "robust" N4SID method can be rewritten as

$$\begin{aligned} \hat{A}^{\mathrm{c}} &= \underline{(\hat{\Gamma}_\nu^{\mathrm{c}})}^{\dagger} \overline{USV^{\top}} (VSU^{\top}) (S_1^{-1/2} U_1^{\top})^{\top} \\ &\quad \times [(S_1^{-1/2} U_1^{\top}) (USV^{\top}) (USV^{\top})^{\top} (S_1^{-1/2} U_1^{\top})^{\top}]^{-1} \\ &= \underline{(\hat{\Gamma}_\nu^{\mathrm{c}})}^{\dagger} \overline{U_1 S_1^{1/2}} = \underline{(\hat{\Gamma}_\nu^{\mathrm{c}})}^{\dagger} \overline{(\hat{\Gamma}_\nu^{\mathrm{c}})}, \end{aligned} \tag{4.15}$$

which is exactly the estimate of the PO-MOESP method.

Next, we shall deal with $\hat{C}^{\mathrm{c}}$. Denote by Z_t^{c} the matrix obtained selecting the first m rows of $Z_{[t,T]}^{\mathrm{c}}$. The estimate of C^{c} is computed by an orthogonal projection of Y_t onto $\hat{X}_t^{\mathrm{c}}$ (see equation (4.13). However, from (4.1), $Y_t = Z_t^{\mathrm{c}} \oplus \tilde{Y}_t$ where $\tilde{Y}_t$ is orthogonal to $Z_{[t,T]}^{\mathrm{c}}$ and therefore to $\hat{X}_t^{\mathrm{c}}$.

This implies that

$$\hat{C}^{c} = Y_t(\hat{X}_t^{c})^{\top}[\hat{X}_t^{c}(\hat{X}_t^{c})^{\top}]^{-1}$$
$$= Z_t^{c}(\hat{X}_t^{c})^{\top}[\hat{X}_t^{c}(\hat{X}_t^{c})^{\top}]^{-1} = (\hat{\Gamma}_v^{c})_{[1:m]} \quad (4.16)$$

which, once again, coincides with the PO-MOESP estimate. □

From what we have shown above we can assert that PO-MOESP can also be seen as a realization-based method which uses as (theoretical) pseudo-state the complementary state vector $\mathbf{x}^{c}$ of (4.6). The estimates of A and C are also expressible asymptotically as solutions to the Wiener–Hopf equation (4.8). In fact, as far as the estimation of (A, C) is concerned, *the conditioning of the PO-MOESP and Robust N4SID methods are the same*.

Remark 12. We would like to warn the reader that the last statement is not true if a different estimate of the observability matrix than (4.2) is used. For instance estimates obtained by enforcing shift invariance of the estimate $\hat{\Gamma}_v$ in the standard N4SID algorithm, are not similar to the robust N4SID estimates.

Sometimes the estimate of the observability matrix is taken to be $\hat{\Gamma}_v^{c} := U_1$ without the diagonal factor $S_1^{1/2}$. It is straightforward to check that the above calculations hold verbatim in this case also, showing that even with this choice of the observability matrix, PO-MOESP and "robust" N4SID give identical estimates.

5. Conclusions

In this paper we have presented an error analysis which applies to some commonly used subspace identification methods with inputs. We have shown that in presence of collinearity of the regressors these methods may lead to inaccurate estimates of the system parameters (and of the relative transfer function). We have also demonstrated that some of the most well-known algorithms in the literature (in particular robust N4SID and PO-MOESP) are equivalent as far as estimation of the matrices (A, C) is concerned.

References

Bauer, D. (2002). Comparing the cca subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs. *Journal of Time Series Analysis*, submitted for publication.

Bauer, D., & Jansson, M. (2000). Analysis of the asymptotic properties of the moesp type of subspace algorithms. *Automatica*, *36*, 497–509.

Bauer, D., & Ljung, L. (2001). Some facts about the choice of the weighting matrices in larimore type of subspace algorithm. *Automatica*, *36*, 763–773.

Bauer, D., Deistler, M., & Scherrer, W. (2000). On the impact of weighting matrices in subspace algorithms. In *Proceedings of IFAC international symposium on system identification*, Santa Barbara.

Belsley, D. A. (1991). *Conditioning diagnostics, collinearity and weak data regression*. New York: Wiley.

Caines, P. E., & Chan, C. W. (1976). Estimation, identification and feedback. In: R. Mehra, D. Lainiotis (Eds.), System identification: advances and case studies (pp. 349–405). New York, Academic.

Chiuso, A. (2000). Geometric methods for subspace identification. Ph.D. thesis. Dept. of Electronics and informatics, University of Padova. Padova, Italy.

Chiuso, A., & Picci, G. (1999). Subspace identification by orthogonal decomposition. In *Proceedings of the 14th IFAC world congress*, Vol. I (pp. 241–246).

Chiuso, A., & Picci, G. (2000). Error analysis of certain subspace methods. In *Proceedings of IFAC international symposium on system identification*, Santa Barbara (pp. 85–90).

Chiuso, A., & Picci, G. (2004a). The asymptotic variance of subspace estimates. *Journal of Econometrics*, *118*(1–2), 257–291.

Chiuso, A., & Picci, G. (2004b). Numerical conditioning and asymptotic variance of susbspace estimates. *Automatica*, this issue.

Chui, N. L. C. (1997). *Subspace methods and informative experiments for subspace identification*. Ph.D. thesis. Pembroke College, Cambridge.

Desai, U. B., & Pal, D. (1984). A realization approach to stochastic model reduction. *IEEE Transactions Automatic Control*, *29*, 1097–1100.

Gevers, M. R., & Anderson, B. D. O. (1982). On jointly stationary feedback-free stochastic processes. *IEEE Transactions Automatic Control*, *27*, 431–436.

Golub, G. H., & Van Loan, C. R. (1989). *Matrix computation* (2nd ed.). The Johns Hopkins Univ. Press.

Grenander, U., & Szegö, G. (1958). *Toeplitz forms and their applications*. New York: Chelsea.

Hannan, E. J., & Poskitt, D. S. (1988). Unit canonical correlations between future and past. *The Annals of Statistics*, *16*, 784–790.

Jansson, M., & Wahlberg, B. (1997). Counterexample to general consistency of subspace system identification methods. In *Proceedings of SYSID97*, Fukoka, Japan (pp. 1677–1682).

Jansson, M., & Wahlberg, B. (1998). On consistency of subspace methods for system identification. *Automatica*, *34*, 1507–1519.

Katayama, T., & Picci, G. (1999). Realization of stochastic systems with exogenous inputs and subspace system identification methods. *Automatica*, *35*(10), 1635–1652.

Kawauchi, H., Chiuso, A., Katayama, T., & Picci, G. (1999). Comparison of two subspace identification methods for combined deterministic-stochastic systems. In *Proceedings of the 31st ISCIE international symposium on stochastic systems theory and its applications*. Yokohama, Japan.

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

Ljung, L. (1997). *System identification; theory for the user*. Englewood Cliffs, NJ: Prentice-Hall.

Peternell, K., Scherrer, W., & Deistler, M. (1996). Statistical analysis of novel subspace identification methods. *Signal Processing*, *52*, 161–178.

Picci, G. (1997). Oblique splitting susbspaces and stochastic realization with inputs. In: D. Prätzel-Wolters, U. Helmke, & E. Zerz (Eds.), *Operators, systems and linear algebra* (pp. 157–174). Stuttgart: Teubner.

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

Rozanov, Y. A. (1967). *Stationary random processes*. San Francisco: Holden-Day.

Söderström, T., & Stoica, P. (1989). *System identification*. Englewood Cliffs, NJ: Prentice-Hall.

Stewart, G. W. (1987). Collinearity and least squares regression. *Statistical Science*, *2*, 68–100.

Stewart, G. W., & Sun, J. G. (1990). *Matrix perturbation theory*. New York: Academic Press.

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

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

Van Overschee, P., & De Moor, B. (1996). *Subspace identification for linear systems*. Dordrecht: Kluwer Academic Publications.

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.

Alessandro Chiuso received his D.Ing. degree Cum Laude in 1996, and the Ph.D. degree in Systems Engineering in 2000 both from the University of Padova. In 1998/99 he was a Visiting Research Scholar with the Electronic Signal and Systems Research Laboratory (ESSRL) at Washington University, St. Louis. From March 2000 to July 2000 he has been Visiting Post-Doctoral (EU-TMR) fellow with the Division of Optimization and System Theory, Department of Mathematics, KTH, Stockholm, Sweden. Since March 2001 he is Research Faculty ("Ricercatore") with the Dept. of Information Engineering, University of Padova. In the summer 2001 he has been visiting researcher with the Dept. of Computer Science, University of California Los Angeles. His research interests are mainly in Identification and Estimation Theory, System Theory and Computer Vision.

Giorgio Picci holds a full professorship with the University of Padova, Italy, Department of Information Engineering, since 1980. He graduated (cum laude) from the University of Padova in 1967 and since then has held several long-term visiting appointments with various American and European universities among which Brown University, M.I.T., the University of Kentucky, Arizona State University, the Center for Mathematics and Computer Sciences (C.W.I.) in Amsterdam, the Royal Institute of Technology, Stockholm Sweden, Kyoto University and Washington University in St. Louis, Mo.

He has been contributing to Systems and Control theory mostly in the area of modeling, estimation and identification of stochastic systems and published over 100 papers and edited three books in this area. Since 1992 he has been active also in the field of Dynamic Vision and scene and motion reconstruction from monocular vision.

He has been involved in various joint research projects with industry and state agencies. He is currently general coordinator of the italian national project *New techniques for identification and adaptive control of industrial systems*, funded by MIUR (the Italian ministery for higher education), has been project manager of the italian team for the Commission of the European Communities Network of Excellence *System Identification* (ERNSI) and is currently general project manager of the Commission of European Communities IST project RECSYS, in the fifth Framework Program.

Giorgio Picci is a Fellow of the IEEE, past chairman of the IFAC Technical Committee on Stochastic Systems and a member of the EUCA council.