

# THE ASYMPTOTIC VARIANCE OF SUBSPACE ESTIMATES

ALESSANDRO CHIUSO \* and GIORGIO PICCI†

September 2, 2002

## Abstract

We give new simple general expressions for the asymptotic covariance of the estimated system parameters  $(A, B, C, D)$  in subspace identification. The formulas can be applied to a whole class of subspace methods including N4SID, MOESP, CVA etc. The asymptotic expressions highlight how the conditioning of the estimation problem influences the accuracy of the estimates.

*Journal of Economic Literature Classification:* C49, C59.

**Keywords:** Subspace Identification; Asymptotic Covariance; Stochastic Realization with Inputs.

---

\*Dipartimento di Elettronica e Informatica, University of Padova, 35131 Padua, Italy; Email: [chiuso@dei.unipd.it](mailto:chiuso@dei.unipd.it)

†Dipartimento di Elettronica e Informatica, University of Padova and LADSEB-CNR, 35131 Padua, Italy; Email: [picci@dei.unipd.it](mailto:picci@dei.unipd.it)

# 1 Introduction

In this paper we shall provide new expressions for the asymptotic covariance of the estimated parameters  $(A, B, C, D)$  of a state space model, obtained by some popular subspace identification methods. The expressions are similar but simpler than the asymptotic covariance expressions which have so far been published in the literature (Bauer 1998, Bauer and Jansson 2000, Jansson 2000). The covariance formulas in particular involve the inverses of certain conditional covariance matrices  $(\Sigma_{\hat{x}\hat{x}|\mathbf{u}+})$  which play an important role in measuring the possible ill-conditioning of the identification problem, see (Chiuso and Picci 1999, Chiuso and Picci 2001b), thus providing a direct link of possible ill-conditioning of the estimation problem with the asymptotic variance of the estimates.

The structure of the paper is as follows:

- In Section 2 we shall introduce notations, review a “sample” Hilbert space framework which provides a convenient tool in the analysis, and discuss the basic ideas of stochastic realization involved in subspace identification.
- In Section 3 a *complementary model* is introduced which allows a unified analysis of various estimation algorithms of the  $(A, C)$  parameters of the model. Among these algorithms, common subspace methods like Robust-N4SID, MOESP and CVA methods can be accommodated.
- In Section 4 the complementary model is used to derive error expressions for the  $(A, C)$  parameter estimates and the asymptotic variance formula for the  $(A, C)$  parameters is obtained.
- In Section 5 a Markov estimator of the  $(B, D)$  parameters is discussed and an expression for the asymptotic variance is provided. The estimator is first derived assuming that  $A, C$  are known, but the effect of uncertainty in  $A, C$  can be taken into account at the price of some additional complication.
- Section 6 contains some conclusions.

Unfortunately the proof of the main result (Theorem 4.1 ) requires a good deal of concepts and notations which need to be introduced gradually in the course of the paper. The reader who is only interested in the statement of the result, may just glance at the definition of  $\mathbf{x}^c$  in Sect. 3, understand the (asymptotic) choice of basis defined in Lemma 4.2 and (patiently) keep track of the notations.

## 2 Subspace identification

We shall assume that the observed input-output data

$$\{u_{t_0}, \dots, u_t, \dots\} \quad \{y_{t_0}, \dots, y_t, \dots\} \quad u_t \in \mathbb{R}^p, y_t \in \mathbb{R}^m \quad (2.1)$$

satisfy the zero-average condition

$$\lim_{N \rightarrow \infty} \frac{1}{N+1} \sum_{t=t_0}^{N+t_0} \begin{bmatrix} u_t \\ y_t \end{bmatrix} = 0$$

and that the limit for  $N \rightarrow \infty$  and for all  $\tau \geq 0$ , of the sample correlation

$$\frac{1}{N+1} \sum_{t=t_0}^{N+t_0} \begin{bmatrix} u_{t+\tau} \\ y_{t+\tau} \end{bmatrix} \begin{bmatrix} u_t \\ y_t \end{bmatrix}^\top \quad \tau \geq 0 \quad (2.2)$$

exists, and is independent of the initial time  $t_0$ . A time-series satisfying this assumption is called (*second-order*) *stationary*<sup>1</sup>.

In continuous-time, functions admitting an “ergodic” limit of the sample correlation function (2.2), have been studied by Wiener in his work on Generalized Harmonic Analysis (Wiener 1930, Wiener 1933). It is

---

<sup>1</sup>Also called a *quasi-stationary* signal.

easy to show that Wiener results hold for discrete-time signals as well. In particular, the limits of the time averages (2.2) define a *positive matrix function*, i.e. a bona-fide stationary covariance matrix function,

$$\lim_{N \rightarrow \infty} \frac{1}{N+1} \sum_{t=t_0}^{N+t_0} \begin{bmatrix} u_{t+\tau} \\ y_{t+\tau} \end{bmatrix} \begin{bmatrix} u_t \\ y_t \end{bmatrix}^\top = \Lambda(\tau) = \begin{bmatrix} \Lambda_{uu}(\tau) & \Lambda_{uy}(\tau) \\ \Lambda_{yu}(\tau) & \Lambda_{yy}(\tau) \end{bmatrix} \quad \tau \geq 0, \quad (2.3)$$

The function  $\Lambda(\tau)$  is called the *true covariance* of the data.

Now, given the true covariance  $\Lambda$ , one can formally manufacture a  $\mathbb{R}^{p+m}$ -valued second-order stationary stochastic process say  $\{\mathbf{z}(t)\} = \{[\mathbf{u}(t)^\top \mathbf{y}(t)^\top]^\top\}$ , defined on the sample space  $\Omega = (\mathbb{R}^{p+m})^{\mathbb{Z}_+}$ , having precisely covariance function  $\Lambda$  and zero mean. Actually there is a whole equivalence class of such processes all sharing the same second order statistics; one can fix a representative assuming say a Gaussian probability law. The construction goes back to Kolmogorov and we shall not report it here. The space of elementary random elements  $\Omega$  is the space of all possible sample paths,  $(\mathbb{R}^{p+m})^{\mathbb{Z}_+}$ , of the process.

The observed sample (2.1) is an *ergodic trajectory* of the second-order process  $\mathbf{z}$ , the term meaning that the limit (2.3) determines the covariance function (and hence the probability law in the Gaussian case) of  $\mathbf{z}$  uniquely. Hence, under the assumption of second-order stationarity of the data, an essentially unique pair of second-order stationary stochastic processes (called the “true processes”) exists which produces the observed trajectory (2.1) according to the classical “urn” scheme of probability theory <sup>2</sup>.

In this paper it is assumed that the true processes have a rational spectral density and hence can be described by a linear stochastic system (in innovation form) of the type

$$\begin{cases} \mathbf{x}(t+1) &= A\mathbf{x}(t) + B\mathbf{u}(t) + K\mathbf{e}(t) \\ \mathbf{y}(t) &= C\mathbf{x}(t) + D\mathbf{u}(t) + \mathbf{e}(t) \end{cases} \quad (2.4)$$

where  $\{\mathbf{x}(t)\}$  is the state process of dimensions  $n$ , and  $\{\mathbf{e}(t)\}$  is a white noise process with 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  $t-1$ , and  $A, B, K, C, D$  are constant matrices. Here  $\{\mathbf{u}(t)\}$  is an exogenous input and is not modelled explicitly. In the following we shall make the blanket assumption that *there is no feedback from  $\mathbf{y}$  to  $\mathbf{u}$* . This implies that  $\{\mathbf{u}(t)\}$  and  $\{\mathbf{e}(t)\}$  are uncorrelated at all times. See e.g. (Caines et al. 1976, Gevers and Anderson 1982, Gevers and Anderson 1981, Picci and Katayama 1996) for a discussion of this concept.

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 (2.4), it is obvious that the finite “tail” matrices,  $Y_t, U_t, X_t$ , constructed at each time  $t$  from the observed sample (2.1) by letting

$$\begin{aligned} Y_t &:= \begin{bmatrix} y_t & y_{t+1} & \dots & y_{t+N} \end{bmatrix} \\ U_t &:= \begin{bmatrix} u_t & u_{t+1} & \dots & u_{t+N} \end{bmatrix} \\ X_t &:= \begin{bmatrix} x_t & x_{t+1} & \dots & x_{t+N} \end{bmatrix} \end{aligned}$$

must also satisfy (2.4), i.e.

$$\begin{cases} X_{t+1} &= AX_t + BU_t + KE_t \\ Y_t &= CX_t + DU_t + E_t \end{cases} \quad (2.5)$$

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

In the ideal case when infinitely long sample trajectories are available ( $N \rightarrow \infty$ ),  $E_t$  is orthogonal <sup>3</sup> to the past data, namely  $E_t \perp (X_s, U_s)$  for all  $s \leq t$  by absence of feedback (this is only approximately true for  $N$  large but finite). This implies that, under generic assumptions on the data guaranteeing uniqueness of the solution, the estimates computed by solving the regression equation coincide, for  $N \rightarrow \infty$ , with the true parameters (consistency). Hence, in an ideal situation where we have available the input-output tail matrices at time  $t$ , and also a corresponding pair of *state* tail matrices at the time instants  $t$  and  $t+1$ , consistent identification of the parameters  $(A, B, C, D)$  of the system (2.4) would be a straightforward matter.

<sup>2</sup>A rather artificial stochastic process to be sure, which nevertheless, is mathematically, a perfectly legitimate object.

<sup>3</sup>“Orthogonality” here is with respect to the inner product (2.7) which is defined below.

In practice the state trajectory is not given to us. However, it is known that the state of certain representations, in particular the innovation realization (2.4), can be constructed from the input-output processes. In practice we only have a finite input-output tail sequence  $\{U_t, Y_t\}_{t=0, \dots, T}$  (where  $T \ll N$ ) and the state at time  $t$  needs to be constructed (in general approximately) from these available data. One can see that the construction of the state becomes a central step in subspace identification. In particular, it appears that asymptotic analysis of subspace identification methods needs to be based on an in-depth analysis of the state construction step.

The problem of constructing the state and state-space models of stochastic processes is the main concern of stochastic realization theory. Stochastic realization 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 (2.4). 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  $\Lambda$  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  $-\infty \leq t_0 \leq t \leq T \leq +\infty$  define the Hilbert spaces of scalar second order random variables

$$\begin{aligned} \mathcal{U}_{[t_0, t]} &:= \overline{\text{span}} \{ \mathbf{u}_k(s); k = 1, \dots, p, t_0 \leq s < t \} \\ \mathcal{Y}_{[t_0, t]} &:= \overline{\text{span}} \{ \mathbf{y}_k(s); k = 1, \dots, m, t_0 \leq s < t \} \end{aligned}$$

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. We shall let  $\mathcal{P}_{[t_0, t]} := \mathcal{U}_{[t_0, t]} \vee \mathcal{Y}_{[t_0, t]}$  denote the joint past space of the input and output processes at time  $t$  (the  $\vee$  denotes closed vector sum). Similarly, let  $\mathcal{U}_{[t, T]}$ ,  $\mathcal{Y}_{[t, T]}$  be the respective future spaces up to time  $T$

$$\begin{aligned} \mathcal{U}_{[t, T]} &:= \overline{\text{span}} \{ \mathbf{u}_k(s); k = 1, \dots, p, t \leq s \leq T \} \\ \mathcal{Y}_{[t, T]} &:= \overline{\text{span}} \{ \mathbf{y}_k(s); k = 1, \dots, m, t \leq s \leq T \} \end{aligned}$$

By convention the past spaces do not include the present. When  $t_0 = -\infty$  we shall use the shorthand  $\mathcal{U}_t^-$ ,  $\mathcal{Y}_t^-$  for  $\mathcal{U}_{[-\infty, t)}$ ,  $\mathcal{Y}_{[-\infty, t)}$ , the closed vector sum  $\mathcal{U}_t^- \vee \mathcal{Y}_t^-$  being denoted by  $\mathcal{P}_t^-$  (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.  $\mathcal{U}_{[t, t]}$ ,  $\mathcal{Y}_{[t, t]}$ , etc) are simply denoted  $\mathcal{U}_t$ ,  $\mathcal{Y}_t$ , etc. while for the spaces generated by the whole time history of  $\mathbf{u}$  and  $\mathbf{y}$  we shall use the symbols  $\mathcal{U}$ ,  $\mathcal{Y}$ , respectively.

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

$$\mathcal{U}_{[t_0, T]} = \mathcal{U}_{[t_0, t]} + \mathcal{U}_{[t, T]}, \quad t_0 \leq t \leq T \quad (2.6)$$

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 (or *linearly regular* see, e.g. (Rozanov 1967, p.52)) process  $\mathbf{u}$  to be sufficiently rich it is necessary and sufficient that the determinant of the spectral density matrix  $\Phi_u$  should have no zeros on the unit circle (Hannan and Poskitt 1988).

### The sample-trajectory framework

Under the natural assumption of second-order stationarity, the sequence of semi-infinite tail matrices constructed from the data (2.1), can be looked upon as an object isomorphic to a stationary random process.

The definitions and the basic facts of this isomorphism are shortly reviewed in Appendix A. Then, as it is shown in Appendix A, the vector space  $\overline{\text{span}}\{Y_t, U_t \mid t \geq t_0\}$ , linearly generated by the rows of the semi-infinite tail sequences  $\{Y_t, U_t \mid t \geq t_0\}$  (here  $N = \infty$ !), closed with respect to the norm induced by the inner product of semi-infinite sequences  $\xi, \eta \in \mathbb{R}^{\mathbb{Z}_+}$  defined by the limit <sup>4</sup>

$$\langle \xi, \eta \rangle := \lim_{N \rightarrow \infty} \frac{1}{N+1} \sum_{t=0}^N \xi_t \eta_t \quad (2.7)$$

and the “stochastic” Hilbert space  $\mathcal{Y} \vee \mathcal{U}$  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  $\overline{\text{span}}\{Y_t, U_t \mid t \geq t_0\}$ .

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  $\overline{\text{span}}\{Y_t, U_t \mid t \geq t_0\}$  by boldface lowercase letters, exactly like random quantities in  $\mathcal{Y} \vee \mathcal{U}$ . 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.

From now on we shall denote by **boldface** characters also *finite* data sequences. For a (possibly vector-valued) infinite sequence  $\mathbf{v}$  we shall normally use the subscript “ $\mathbf{v}_N$ ” to denote the tail matrix of  $\mathbf{v}$  obtained by truncation to length  $N$ , and append a superscript “ $N$ ” to the corresponding symbol denoting the subspace spanned by the rows of the (finite) tail matrix  $\mathbf{v}_N$ .

The symbol  $(\mathbf{y}_{[\tau, T]})_N$  will be used to denote the vector (actually a block-Hankel matrix of dimension  $m(T - \tau + 1) \times (N + 1)$ )

$$\begin{bmatrix} \mathbf{y}_N(\tau) \\ \vdots \\ \mathbf{y}_N(T) \end{bmatrix} \equiv \begin{bmatrix} \mathbf{y}(\tau) \\ \vdots \\ \mathbf{y}(T) \end{bmatrix}_N$$

and  $\mathcal{Y}_{[\tau, T]}^N$  the corresponding (finite-dimensional) rowspace. Since for  $N \rightarrow \infty$ ,  $(\mathbf{y}_{[\tau, T]})_N$  “expands” to the  $m(T - \tau + 1) \times \infty$  matrix of semi-infinite tails  $\mathbf{y}_{[\tau, T]}$  (which is the same thing as the  $m(T - \tau + 1)$ -dimensional column random vector  $\mathbf{y}_{[\tau, T]}$ ), we shall agree to say that  $(\mathbf{y}_{[\tau, T]})_N$  “tends” to  $\mathbf{y}_{[\tau, T]}$  as  $N \rightarrow \infty$ . Likewise, for the corresponding rowspaces, we use the notation  $\mathcal{Y}_{[\tau, T]}^N \rightarrow \mathcal{Y}_{[\tau, T]}$  for  $N \rightarrow \infty$ . “Approximating” spaces of random variables by vector spaces spanned by the rows of tail matrices is a standard idea at the heart of subspace identification.

Since the only difference between operations on finite and infinite sequences is in the inner product, we shall use a different notation for the inner product. Namely, we shall denote by  $E[\mathbf{xy}^\top]$  the mathematical expectation (true covariance matrix) when  $\mathbf{x}$  and  $\mathbf{y}$  are random vectors (or infinitely long vector-valued data sequences) and by  $\hat{E}_N[\mathbf{xy}^\top]$  the sample covariance matrix of the finite (vector) sequences  $\mathbf{x}, \mathbf{y}$ ,

$$\hat{E}_N[\mathbf{xy}^\top] := \frac{1}{N+1} \sum_{t=0}^N x_t y_t^\top.$$

so that  $\lim_{N \rightarrow \infty} \hat{E}_N[\mathbf{xy}^\top] = E[\mathbf{xy}^\top]$ . In the same spirit we shall understand that  $E[\mathbf{x} \mid \mathbf{y}]$  is the wide-sense conditional expectation

$$E[\mathbf{x} \mid \mathbf{y}] := E[\mathbf{xy}^\top] E[\mathbf{yy}^\top]^{-1} \mathbf{y}$$

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

$$\hat{E}_N[\mathbf{x} \mid \mathbf{y}] := \hat{E}_N[\mathbf{xy}^\top] \hat{E}_N[\mathbf{yy}^\top]^{-1} \mathbf{y}$$

---

<sup>4</sup>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.

will denote the corresponding object when  $\mathbf{x}$  and  $\mathbf{y}$  are finite sequences. This is nothing else but the well-known formula solving the (deterministic) Least-Squares problem

$$\min_{A \in \mathbb{R}^{n \times m}} \|\mathbf{y} - A\mathbf{x}\|$$

Clearly, under second-order stationarity,  $\lim_{N \rightarrow \infty} \hat{E}_N[\mathbf{x} | \mathbf{y}] = E[\mathbf{x} | \mathbf{y}]$ . This simple fact will be used quite frequently in this paper.

## 2.2 Constructing the state

The construction of the state should be based on the prescriptions of stochastic realization theory with inputs (Picci and Katayama 1996, Katayama and Picci 1999). In particular, we recall that the *state space* at time  $t$  of the stationary realization (2.4)

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

is the so-called *oblique predictor space*

$$\mathcal{X}_t := E_{\|\mathcal{U}_t^+\} [\mathcal{Y}_t^+ | \mathcal{P}_t^-] \quad (2.8)$$

where the symbol  $E_{\|\mathcal{C}} [A | \mathcal{B}]$  denotes oblique projection of the subspace  $\mathcal{A}$  onto  $\mathcal{B}$  along the subspace  $\mathcal{C}$  (Picci and Katayama 1996, Katayama and Picci 1999). Note that this subspace can in principle be constructed using input-output data, although the complete *infinite* past,  $\mathcal{P}_t^-$ , is needed in (2.8).

In identification the infinite past is never available and the state construction must be done starting from input-output tails  $\{\mathbf{y}(t), \mathbf{u}(t)\}$  on a *finite* interval,  $[t_0, T]$ . Obviously, since in practice the data are finite, we should consider finite length tails  $\{\mathbf{y}_N(t), \mathbf{u}_N(t); t \in [t_0, T]\}$ . However since the discussion here is intended primarily to clarify conceptual issues, we shall pretend below that  $N = \infty$ , in other words that  $\{\mathbf{y}(t), \mathbf{u}(t)\}$  are true random variables. Dealing with data of finite length will be the main objective of the following sections.

Henceforth we shall consider the problem of constructing state-space representations of the process  $\mathbf{y}$  where the state is a function of the input and output variables on a finite interval  $[t_0, T]$  only. These models will be called *finite-interval* realizations. In general they involve non-stationary parameters.

In principle  $\mathbf{y}$  can be represented by a finite-interval realization involving the same constant parameters  $(A, B, C, D)$  of the stationary model (2.4) that one wants to identify. This is the *transient conditional Kalman filter realization* on the interval  $[t_0, T]$  first used in (Van Overschee and De Moor 1994)

$$\begin{cases} \hat{\mathbf{x}}(t+1) &= A\hat{\mathbf{x}}(t) + B\mathbf{u}(t) + K(t)\hat{\mathbf{e}}(t) \\ \mathbf{y}(t) &= C\hat{\mathbf{x}}(t) + D\mathbf{u}(t) + \hat{\mathbf{e}}(t) \\ \hat{\mathbf{x}}(t_0) &= E[\mathbf{x}(t_0) | \mathcal{U}_{[t_0, T]}] \end{cases} \quad (2.9)$$

where

$$\hat{\mathbf{x}}(t) := E[\mathbf{x}(t) | \mathcal{P}_{[t_0, t-1]} \vee \mathcal{U}_{[t, T]}] \quad (2.10)$$

$\mathbf{x}(t)$  being a basis for a stationary state space  $\mathcal{X}_t$  and  $\hat{\mathbf{e}}(t)$  the *transient (conditional) innovation process* defined by

$$\hat{\mathbf{e}}(t) = \mathbf{y}(t) - E[\mathbf{y}(t) | \mathcal{P}_{[t_0, t-1]} \vee \mathcal{U}_{[t, T]}] \quad (2.11)$$

Note that the *transient innovation space*  $\hat{\mathcal{E}}_t$  defined by the orthogonal decomposition

$$\mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t+1, T]} = \hat{\mathcal{E}}_t \oplus (\mathcal{P}_{[t_0, t-1]} \vee \mathcal{U}_{[t, T]}) \quad (2.12)$$

is precisely spanned by the components of  $\hat{\mathbf{e}}(t)$ , i.e.  $\hat{\mathcal{E}}_t = \text{span} \{\hat{\mathbf{e}}(t)\}$ .

**Remark 2.1** Contrary to the standard Kalman filter, the initial state estimate  $\hat{\mathbf{x}}(t_0)$  is not zero and depends on the future inputs  $\mathcal{U}_{[t_0, T]}$ . This implies that  $\hat{\mathbf{x}}(t)$  is also influenced by future inputs on  $[t, T]$ , in spite of the “causal” look of the state equation (2.9). For this reason, the construction of (a basis,  $\hat{\mathbf{x}}(t)$ , in) the state space  $\hat{\mathcal{X}}_t := E[\mathcal{X}_t | \mathcal{P}_{[t_0, t-1]} \vee \mathcal{U}_{[t, T]}]$  of the model (2.9), using only finite input-output data, is apparently

an impossible task. See (Chiuso and Picci 2001b) for a discussion of this point. The simple strategy based on solving a linear regression problem, which was alluded at in the beginning of the previous section, cannot be implemented if we work with the model (2.9).  $\diamond$

Ideally we would like to construct state subspaces of  $\mathcal{Y}_{[t_0, T]} \vee \mathcal{U}_{[t_0, T]}$ , leading to state-space models which are *causal* in  $\mathbf{u}$ , as well as in the driving noise (the “transient innovation”  $\hat{\mathbf{e}}(t)$ ). In the next section we shall obtain models of this type.

### 3 The complementary model

In this section we shall build a special finite interval realization which permits a unified analysis of most subspace methods with inputs. This model will also be instrumental for the derivation of the asymptotic variances expressions to be given in the following sections.

The basic idea, inspired by a preliminary orthogonal projection step first introduced in the MOESP-type algorithms (Verhaegen and Dewilde 1992, Verhaegen 1994), and then also used in the “Robust N4SID, and CCA (with finite data) algorithms, is to form the orthogonal complement in  $\mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t, T]}$  of the future input space (i.e. to “subtract off” the effect of future inputs) and to study the dynamics of the system on this subspace. As we shall see, this will lead to a stochastic realization which is constructible from finite input-output data, and involves (modulo a change of basis) the same parameters ( $A, C$ ) of the steady-state model.

Introduce the orthogonal complement of  $\mathcal{U}_{[t, T]}$  in  $\mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t, T]}$

$$\mathcal{F}_{[t_0, t-1]} := (\mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t, T]}) \ominus \mathcal{U}_{[t, T]}$$

and similarly

$$\mathcal{F}_{[t_0, t]} := (\mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t+1, T]}) \ominus \mathcal{U}_{[t+1, T]}$$

Note that at different times we are taking orthogonal complements in a different ambient space. For later reference we point out the following “constructive” formula whose proof will be left to the reader.

**Lemma 3.1**

$$\mathcal{F}_{[t_0, t-1]} = E\{\mathcal{P}_{[t_0, t]} \mid \mathcal{U}_{[t, T]}^\perp\} := \text{span}\{\mathbf{p} - E[\mathbf{p} \mid \mathcal{U}_{[t, T]}] \mid \mathbf{p} \in \mathcal{P}_{[t_0, t]}\} \quad (3.1)$$

Sometimes we shall use  $(\mathcal{P}_{[t_0, t]} \mid \mathcal{U}_{[t, T]}^\perp)$  as a shorthand for  $E\{\mathcal{P}_{[t_0, t]} \mid \mathcal{U}_{[t, T]}^\perp\}$ .

Now assume the true process  $\mathbf{y}$  admits a (minimal) stationary realization, not necessarily of the innovation type,

$$\begin{cases} \mathbf{x}(t+1) &= A\mathbf{x}(t) + B\mathbf{u}(t) + G\mathbf{w}(t) \\ \mathbf{y}(t) &= C\mathbf{x}(t) + D\mathbf{u}(t) + J\mathbf{w}(t) \end{cases} \quad (3.2)$$

with state space  $\mathcal{X}_t := \text{span}\{\mathbf{x}(t)\}$  and let  $\mathcal{W}_t := \text{span}\{\mathbf{w}(t)\}$ .

Define the “complementary” process

$$\mathbf{y}^c(t) := E[\mathbf{y}(t) \mid \mathcal{F}_{[t_0, t]}] \quad (3.3)$$

which by construction is orthogonal to the future input space, and introduce the “complementary state space”  $\hat{\mathcal{X}}_t^c$  as follows:

$$\hat{\mathcal{X}}_t^c := E[\mathcal{X}_t \mid \mathcal{F}_{[t_0, t-1]}] = E\left[\mathcal{X}_t \mid (\mathcal{P}_{[t_0, t]} \mid \mathcal{U}_{[t, T]}^\perp)\right] \quad , \quad t = t_0, \dots, T \quad (3.4)$$

A basis in  $\hat{\mathcal{X}}_t^c$  can be constructed by choosing

$$\hat{\mathbf{x}}^c(t) := E[\mathbf{x}(t) \mid \mathcal{F}_{[t_0, t-1]}] \quad (3.5)$$

The following Lemma will be used below to show that the complementary state space can indeed give origin to state space models.

**Lemma 3.2** Let  $\mathcal{E}_t$  be the transient innovation space (2.12) and let

$$\hat{\mathcal{V}}_t := \mathcal{U}_{[t, T]} \ominus \mathcal{U}_{[t+1, T]}, \quad (3.6)$$

be the transient backward innovation space of the input process  $\mathbf{u}$ . Then the following inclusions hold

$$\hat{\mathcal{X}}_{t+1}^c \subseteq \hat{\mathcal{X}}_t^c \oplus \hat{\mathcal{V}}_t \oplus \hat{\mathcal{E}}_t \quad (3.7)$$

and

$$E\{\mathcal{Y}_t \mid \mathcal{F}_{[t_0, t]}\} \subseteq \hat{\mathcal{X}}_t^c \oplus \hat{\mathcal{V}}_t \oplus \hat{\mathcal{E}}_t \quad (3.8)$$

*Proof.* From the definition of backward transient innovation of  $\mathbf{u}$  we have that

$$\begin{aligned} \mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t, T]} &= \mathcal{U}_{[t, T]} \oplus \mathcal{F}_{[t_0, t-1]} \\ &= \mathcal{U}_{[t+1, T]} \oplus \hat{\mathcal{V}}_t \oplus \mathcal{F}_{[t_0, t-1]} \end{aligned}$$

and since

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

we obtain the decomposition

$$\mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t+1, T]} = \mathcal{U}_{[t+1, T]} \oplus \left( \hat{\mathcal{V}}_t \oplus \mathcal{F}_{[t_0, t-1]} \oplus \hat{\mathcal{E}}_t \right)$$

and hence

$$\mathcal{F}_{[t_0, t]} = \mathcal{F}_{[t_0, t-1]} \oplus \hat{\mathcal{V}}_t \oplus \hat{\mathcal{E}}_t \quad (3.9)$$

The first statement of the Proposition follows by projecting the subspace inclusion

$$\mathcal{X}_{t+1} \subseteq (\mathcal{X}_t \vee \mathcal{U}_t) \oplus \mathcal{W}_t$$

(which is a coordinate free version of the state equation in (3.2)) onto  $\mathcal{F}_{[t_0, t]}$ . The second also follows by projecting

$$\mathcal{Y}_t \subseteq (\mathcal{X}_t \vee \mathcal{U}_t) \oplus \mathcal{W}_t$$

onto  $\mathcal{F}_{[t_0, t]}$ .

□

Obtaining state-space models is now just a matter of choosing bases, a particularly convenient choice being that given in (3.5).

**Proposition 3.1** Let  $\hat{\mathbf{x}}^c(t)$  be the basis defined in (3.5). Let also  $\hat{\mathbf{v}}(t)$  be the backward innovation process,  $\hat{\mathbf{v}}(t) = \mathbf{u}(t) - E[\mathbf{u}(t) \mid \mathcal{U}_{[t+1, T]}]$  (a basis for the space  $\hat{\mathcal{V}}_t$ ). Then the following representation holds

$$\begin{cases} \hat{\mathbf{x}}^c(t+1) &= A\hat{\mathbf{x}}^c(t) + \bar{B}(t)\hat{\mathbf{v}}(t) + K(t)\hat{\mathbf{e}}(t) \\ \mathbf{y}^c(t) &= C\hat{\mathbf{x}}^c(t) + \bar{D}(t)\hat{\mathbf{v}}(t) + \hat{\mathbf{e}}(t) \end{cases} \quad (3.10)$$

for all  $t_0 \leq t \leq T$ . All the terms on the right hand side are mutually uncorrelated. The matrix coefficients are given by  $\bar{B}(t) = (AK_u(t) + B)$ ,  $\bar{D}(t) = (CK_u(t) + D)$ , where

$$K_u(t) := E[\mathbf{x}(t)\hat{\mathbf{v}}^\top(t)] \left( E[\hat{\mathbf{v}}(t)\hat{\mathbf{v}}^\top(t)] \right)^{-1} \quad (3.11)$$

and  $K(t)$  is the transient Kalman filter gain in (2.9).

*Proof.* Since  $\mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t, T]} \subset \mathcal{F}_{[t_0, t-1]}$ , it is the same to use the Kalman filter state  $\hat{\mathbf{x}}(t)$  in place of  $\mathbf{x}(t)$  in formula (3.5). Projecting term by term the conditional Kalman filter equations (2.9) onto  $\mathcal{F}_{[t_0, t]}$ , using the decomposition (3.9) and noting that  $E\{\hat{\mathbf{x}}(t) | \hat{\mathcal{E}}_t\} = 0$ ,  $E\{\mathbf{u}(t) | \hat{\mathcal{E}}_t\} = 0$  we get

$$\begin{aligned} E\{\mathbf{x}(t) | \mathcal{F}_{[t_0, t]}\} &= E\{\mathbf{x}(t) | \mathcal{F}_{[t_0, t-1]}\} + E\{\mathbf{x}(t) | \hat{\mathcal{V}}_t\} + E\{\mathbf{x}(t) | \hat{\mathcal{E}}_t\} \\ &= \hat{\mathbf{x}}^c(t) + K_u(t)\hat{\mathbf{v}}(t) \\ E\{\mathbf{u}(t) | \mathcal{F}_{[t_0, t]}\} &= E\{\mathbf{u}(t) | \mathcal{F}_{[t_0, t-1]}\} + E\{\mathbf{u}(t) | \hat{\mathcal{V}}_t\} + E\{\mathbf{u}(t) | \hat{\mathcal{E}}_t\} \\ &= K_u(t)\hat{\mathbf{v}}(t) \end{aligned}$$

the realization (3.10) follows after some rearrangement of terms. The formulas for  $\bar{B}(t)$ ,  $\bar{D}(t)$  also follow after some simple algebra.

□

We shall now show that a basis in the complementary state space  $\hat{\mathcal{X}}_t^c$  can be constructed starting from the observed data. We shall do this here for infinite length sequences (random variables) and postpone the discussion of the finite length case to the next section.

Consider the output predictors based on the complementary past information up to time  $t-1$

$$\hat{\mathbf{y}}(t+k | t-1) := E[\mathbf{y}(t+k) | \mathcal{F}_{[t_0, t-1]}] \quad k \geq 0 \quad (3.12)$$

Note that we can decompose the output string at time  $t+k$  as  $\mathbf{y}(t+k) = \mathbf{y}^c(t+k) + \tilde{\mathbf{y}}^c(t+k)$  where  $\mathbf{y}^c(t+k) = E[\mathbf{y}(t+k) | \mathcal{F}_{[t_0, t+k]}]$  is the complementary output at time  $t+k$  while  $\tilde{\mathbf{y}}^c(t+k) = E[\mathbf{y}(t+k) | \mathcal{U}_{[t+k+1, T]}]$  is the part of  $\mathbf{y}(t+k)$  which is predictable based on future inputs after time  $t+k$ . Since  $\mathcal{U}_{[t+k+1, T]} \subset \mathcal{U}_{[t, T]} \perp \mathcal{F}_{[t_0, t-1]}$ , we have

$$\hat{\mathbf{y}}(t+k | t-1) = E[\mathbf{y}^c(t+k) | \mathcal{F}_{[t_0, t-1]}], \quad k \geq 0$$

Note that since  $\mathcal{F}_{[t_0, t-1]}$  can be computed from the data, this quantity is also computable from the data. Assume that the integer  $\nu := T-t$  is greater or equal than the order  $n$  of the true model (2.4), and let

$$\hat{\mathbf{y}}_t^+ := \begin{bmatrix} \hat{\mathbf{y}}(t | t-1) \\ \hat{\mathbf{y}}(t+1 | t-1) \\ \vdots \\ \hat{\mathbf{y}}(T-1 | t-1) \end{bmatrix} \quad \hat{\mathbf{y}}_{t+1}^+ := \begin{bmatrix} \hat{\mathbf{y}}(t+1 | t) \\ \hat{\mathbf{y}}(t+2 | t) \\ \vdots \\ \hat{\mathbf{y}}(T | t) \end{bmatrix} \quad (3.13)$$

the vector  $\hat{\mathbf{y}}_{t+1}^+$  is called the (one step ahead) *conditional shift* of  $\hat{\mathbf{y}}_t^+$ . By minimality and by the orthogonality property of the complementary state equations (3.10) we have

$$\text{row-span}\{\hat{\mathbf{y}}_t^+\} = \text{span}\{\hat{\mathbf{x}}_k^c(t); k = 1, \dots, n\} = \hat{\mathcal{X}}_t^c \quad (3.14)$$

and changing  $t$  into  $t+1$ , the analogous relation results for  $\hat{\mathcal{X}}_{t+1}^c$ .

It also follows from (3.10) that the matrices  $(A, C)$  are uniquely determined by the chosen (state vector) bases  $\hat{\mathbf{x}}^c(t+1)$ ,  $\hat{\mathbf{x}}^c(t)$ , and by  $\mathbf{y}(t)$ , by the formulas

$$A = E\left[\hat{\mathbf{x}}^c(t+1) (\hat{\mathbf{x}}^c(t))^\top\right] \left(E\left[\hat{\mathbf{x}}^c(t) (\hat{\mathbf{x}}^c(t))^\top\right]\right)^{-1} \quad (3.15)$$

and

$$C = E\left[\mathbf{y}(t) (\hat{\mathbf{x}}^c(t))^\top\right] \left(E\left[\hat{\mathbf{x}}^c(t) (\hat{\mathbf{x}}^c(t))^\top\right]\right)^{-1}. \quad (3.16)$$

where we have assumed invertibility of  $E\left[\hat{\mathbf{x}}^c(t) (\hat{\mathbf{x}}^c(t))^\top\right]$ . Clearly in (3.16) we can substitute  $\mathbf{y}^c(t)$  with  $\mathbf{y}(t)$ , as this does not change the covariance.

One can show (Chiuso and Picci 2001b), that certain particular choices of basis in  $\hat{\mathcal{X}}_t^c$  give origin to the (theoretical) state underlying well-known subspace identification methods like the robust N4SID algorithm

(vanOverschee and De Moor 1994), Verhaegen’s MOESP algorithm with “shift invariance” (Verhaegen 1994), and also the canonical correlation analysis (CCA) method based on a *finite* data window. Following early observations of (Van Overschee and De Moor 1995) the different bases can all be seen as “canonical” variates obtained by SVD of a correlation matrix between suitably weighted future outputs and past input-output data. It follows that, provided the state  $\hat{\mathbf{x}}^c$  is chosen in the appropriate coordinate system, the same formulas (3.15), (3.16) describe asymptotically the estimates of  $(A, C)$  by the various subspace methods mentioned above.

The original N4SID method can also be related to a particular choice of basis in the complementary state space  $\hat{\mathcal{X}}_t^c$  since, as observed in (Chiuso and Picci 2001b), the asymptotic estimates of the  $A$  and  $C$  matrices by N4SID, can be written

$$A = \Sigma_{\hat{\mathbf{x}}' \hat{\mathbf{x}} | \mathbf{u}^+} \Sigma_{\hat{\mathbf{x}} \hat{\mathbf{x}} | \mathbf{u}^+}^{-1} \quad C = \Sigma_{\hat{\mathbf{y}} \hat{\mathbf{x}} | \mathbf{u}^+} \Sigma_{\hat{\mathbf{x}} \hat{\mathbf{x}} | \mathbf{u}^+}^{-1} \quad (3.17)$$

where  $\hat{\mathbf{x}} \equiv \hat{\mathbf{x}}(t)$  is the state of the Kalman filter realization (2.9),  $\hat{\mathbf{x}}' \equiv \hat{\mathbf{x}}(t+1)$  and

$$\begin{aligned} \Sigma_{\hat{\mathbf{x}}' \hat{\mathbf{x}} | \mathbf{u}^+} &:= E \left[ \left( \hat{\mathbf{x}}(t+1) - E[\hat{\mathbf{x}}(t+1) | \mathcal{U}_{[t, T]}] \right) \left( \hat{\mathbf{x}}(t) - E[\hat{\mathbf{x}}(t) | \mathcal{U}_{[t, T]}] \right)^\top \right] \\ \Sigma_{\hat{\mathbf{x}} \hat{\mathbf{x}} | \mathbf{u}^+} &:= E \left[ \left( \hat{\mathbf{x}}(t) - E[\hat{\mathbf{x}}(t) | \mathcal{U}_{[t, T]}] \right) \left( \hat{\mathbf{x}}(t) - E[\hat{\mathbf{x}}(t) | \mathcal{U}_{[t, T]}] \right)^\top \right] \end{aligned}$$

Clearly formulas (3.17) are the same as (3.15) (3.16), as it follows from the equality

$$\hat{\mathbf{x}}^c(t) := E[\mathbf{x}(t) | \mathcal{F}_{[t_0, t-1]}] = E[\hat{\mathbf{x}}(t) | \mathcal{F}_{[t_0, t-1]}] = E[\hat{\mathbf{x}}(t) | \mathcal{U}_{[t, T]}^\perp] \quad (3.18)$$

We shall take up the state construction step again in more detail in section 4 below.

The discussion above concerns estimation of  $(A, C)$ . Unfortunately a unified treatment regarding estimation of the  $(B, D)$  parameters seems not to be possible. In Section 5 we shall do some asymptotic analysis of the  $(B, D)$  estimates obtained by the so-called “linear regression” method.

### 3.1 A finite data length complementary model

In this section we shall construct the “finite-data” version of the complementary model (3.10) starting from real data of finite duration.

**WARNING:** in this and following subsection, boldface quantities will denote tail sequences of *finite* length  $N$  (tail matrices with  $N+1$  columns). In particular, the matrices of future and past input and output data, like all other tail matrices we shall form with the data, are of finite length  $N$ . For an infinite sequence  $\mathbf{v}$  (possibly vector-valued) we shall normally use the subscript “ $\mathbf{v}_N$ ” to denote truncation to length  $N$ . To be completely consistent one should append an “ $N$ ” to all corresponding symbols denoting subspaces, but, in order to keep notations simple, we shall refrain from doing that. The reader should keep in mind that the same symbols  $\mathcal{U}_{[t, T]}$ ,  $\mathcal{Y}_{[t, T]}$ ,  $\mathcal{U}_{[t_0, t]}$ ,  $\mathcal{P}_{[t_0, t]}$  etc. which were used for the “theoretical” subspaces made of infinite-length sequences will now be used for the corresponding subspaces spanned by the rows of the relevant *finite* tail matrices.

Let

$$\mathbf{y}_N^+(t) := \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{y}(t+1) \\ \vdots \\ \mathbf{y}(T-1) \end{bmatrix}_N \quad \mathbf{u}_N^+(t) := \begin{bmatrix} \mathbf{u}(t) \\ \mathbf{u}(t+1) \\ \vdots \\ \mathbf{u}(T-1) \end{bmatrix}_N \quad (3.19)$$

$$\mathbf{u}_N^-(t) := \begin{bmatrix} \mathbf{u}(t_0) \\ \mathbf{u}(t_0+1) \\ \vdots \\ \mathbf{u}(t-2) \\ \mathbf{u}(t-1) \end{bmatrix}_N \quad \mathbf{p}_N^-(t) := \begin{bmatrix} \mathbf{y}(t_0) \\ \mathbf{u}(t_0) \\ \vdots \\ \mathbf{y}(t-2) \\ \mathbf{u}(t-2) \\ \mathbf{y}(t-1) \\ \mathbf{u}(t-1) \end{bmatrix}_N \quad (3.20)$$

be the finite-data tail matrices with  $N + 1$  columns.

It is not difficult to see that all the subspace manipulations introduced in the previous section make sense also in the present setting. Everything we did can actually be repeated verbatim for finite-length subspaces provided the finite expectation symbol  $\hat{E}_N$  is substituted in place of the ordinary expectation  $E$ . In particular, the (finite) orthogonal complement of  $\mathcal{U}_{[t, T]}$  in  $\mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t, T]}$ , is defined as

$$\mathcal{F}_{[t_0, t-1]} := (\mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t, T]}) \ominus \mathcal{U}_{[t, T]}$$

and similarly for  $\mathcal{F}_{[t_0, t]}$ .

The *finite-length innovation* at time  $t$  is defined by

$$\hat{\mathbf{e}}_N(t) := \mathbf{y}_N(t) - \hat{E}_N [\mathbf{y}_N(t) | \mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t, T]}] \quad (3.21)$$

and is the finite-length counterpart of (2.12) with the finite projection operator  $\hat{E}_N[\cdot | \cdot]$  replacing the stochastic (infinite-length) projection  $E[\cdot | \cdot]$ . Note that, since the truncation and projection operators do not commute,  $\hat{\mathbf{e}}_N(t)$  is not equal to  $\hat{\mathbf{e}}_N(t)$ . The *finite-length innovation space* is

$$\hat{\mathcal{E}}_t := \text{span}\{\hat{\mathbf{e}}_N(t)\}$$

The basic recursion

$$\mathcal{F}_{[t_0, t]} = \mathcal{F}_{[t_0, t-1]} \oplus \hat{\mathcal{V}}_t \oplus \hat{\mathcal{E}}_t \quad (3.22)$$

still holds, the finite-length backward innovation space  $\hat{\mathcal{V}}_t$  being the orthogonal complement of  $\mathcal{U}_{[t+1, T]}$  in  $\mathcal{U}_{[t, T]}$ . The generator of  $\hat{\mathcal{V}}_t$ , i.e. the *finite-length backward innovation* is

$$\hat{\mathbf{v}}_N(t) := \mathbf{u}_N(t) - \hat{E}_N [\mathbf{u}_N(t) | \mathcal{U}_{[t+1, T]}] \quad (3.23)$$

Consider now the “truncated” stationary innovation model obtained by truncating to length  $N$  all stochastic variables in the model (2.4)

$$\begin{cases} \mathbf{x}_N(t+1) &= A\mathbf{x}_N(t) + B\mathbf{u}_N(t) + K\mathbf{e}_N(t) \\ \mathbf{y}_N(t) &= C\mathbf{x}_N(t) + D\mathbf{u}_N(t) + \mathbf{e}_N(t) \end{cases} \quad (3.24)$$

Let us introduce the finite-length complementary process

$$\mathbf{y}_N^c(t) := \hat{E}_N [\mathbf{y}_N(t) | \mathcal{F}_{[t_0, t]}] \quad (3.25)$$

which is orthogonal to the (finite-length) future input space  $\mathcal{U}_{[t+1, T]}$ . Consider also the projection of the state,  $\mathbf{x}_N(t)$ , of the model (3.24) onto  $\mathcal{F}_{[t_0, t-1]}$

$$\hat{\mathbf{z}}_N^c(t) := \hat{E}_N [\mathbf{x}_N(t) | \mathcal{F}_{[t_0, t-1]}] \quad (3.26)$$

here, again, we warn the reader that  $\hat{\mathbf{z}}_N^c(t) \neq \hat{\mathbf{x}}_N^c(t)$ .

Introduce the finite-length complementary state space  $\hat{\mathcal{X}}_t^c$  as follows:

$$\hat{\mathcal{X}}_t^c := \text{span}\{\hat{\mathbf{z}}_N^c(t)\} = \hat{E}_N [\mathcal{X}_t | \mathcal{F}_{[t_0, t-1]}] \quad (3.27)$$

where  $\mathcal{X}_t = \text{span}\{\mathbf{x}_N(t)\}$ . Likewise one should remember that all subspaces in this subsection are of finite length  $N$ . Again, for later reference, we rewrite this as

$$\hat{\mathcal{X}}_t^c = E \left[ \mathcal{X}_t | \left( \mathcal{P}_{[t_0, t]} | \mathcal{U}_{[t, T]}^\perp \right) \right] \quad , \quad t = t_0, \dots, T \quad (3.28)$$

**Proposition 3.2** *Let  $\hat{\mathbf{z}}_N^c(t)$  be the basis in the finite-length complementary state space  $\hat{\mathcal{X}}_t^c$  defined in (3.26) and let  $\hat{\mathbf{z}}_N^c(t+1)$  be its conditional shift. Let*

$$\hat{\mathbf{z}}_N^c(t) := \hat{E}_N [\mathbf{e}_N(t) | \mathcal{F}_{[t_0, t-1]}] \quad (3.29)$$

Then the finite length complementary process  $\mathbf{y}_N^c$  admits the following representation

$$\begin{cases} \hat{\mathbf{z}}_N^c(t+1) &= A\hat{\mathbf{z}}_N^c(t) + \hat{B}(t)\hat{\mathbf{v}}_N(t) + \hat{K}(t)\hat{\mathbf{e}}_N(t) + K\hat{\boldsymbol{\zeta}}_N(t) \\ \mathbf{y}_N^c(t) &= C\hat{\mathbf{z}}_N^c(t) + \hat{D}(t)\hat{\mathbf{v}}_N(t) + \hat{\mathbf{e}}_N(t) + \hat{\boldsymbol{\zeta}}_N(t) \end{cases} \quad (3.30)$$

for all  $t_0 \leq t \leq T$ . The matrix coefficients are given by the expressions  $\hat{B}(t) = \left( A\hat{K}_u(t) + B + E(t) \right)$ ,  $\hat{D}(t) = \left( C\hat{K}_u(t) + D + E(t) \right)$ , where

$$\hat{K}_u(t) := \hat{E}_N [\mathbf{x}_N(t)\hat{\mathbf{v}}_N(t)^\top] \left( \hat{E}_N [\hat{\mathbf{v}}_N(t)\hat{\mathbf{v}}_N(t)^\top] \right)^{-1} \quad (3.31)$$

$$\hat{K}(t) := \hat{E}_N [\mathbf{x}_N(t+1)\hat{\mathbf{e}}_N(t)^\top] \left( \hat{E}_N [\hat{\mathbf{e}}_N(t)\hat{\mathbf{e}}_N(t)^\top] \right)^{-1} \quad (3.32)$$

$$E(t) := \hat{E}_N [\mathbf{e}_N(t)\hat{\mathbf{v}}_N(t)^\top] \left( \hat{E}_N [\hat{\mathbf{v}}_N(t)\hat{\mathbf{v}}_N(t)^\top] \right)^{-1} \quad (3.33)$$

For  $N \rightarrow \infty$ ,

$$\hat{\boldsymbol{\zeta}}_N(t) \rightarrow 0 \quad \hat{\mathbf{v}}_N(t) \rightarrow \hat{\mathbf{v}}(t) \quad \hat{\mathbf{e}}_N(t) \rightarrow \hat{\mathbf{e}}(t) \quad (3.34)$$

and the matrix coefficients of (3.30) converge to those of the complementary model (3.10).

*Proof.* Denote  $\mathcal{S}_{[t_0, t]} := \mathcal{F}_{[t_0, t-1]} \oplus \hat{\mathcal{V}}_t$ . Projecting the truncated innovation model (3.24) onto the subspace  $\mathcal{F}_{[t_0, t]} = \mathcal{S}_{[t_0, t]} \oplus \hat{\mathcal{E}}_t$  one obtains

$$\begin{aligned} \hat{E}_N \{ \mathbf{x}_N(t+1) \mid \mathcal{F}_{[t_0, t]} \} &= A\hat{E}_N \{ \mathbf{x}_N(t) \mid \mathcal{S}_{[t_0, t]} \} + B\hat{E}_N \{ \mathbf{u}_N(t) \mid \mathcal{S}_{[t_0, t]} \} + \\ &+ K\hat{E}_N \{ \mathbf{e}_N(t) \mid \mathcal{S}_{[t_0, t]} \} + \hat{E}_N \{ \mathbf{x}_N(t+1) \mid \hat{\mathcal{E}}_t \} \\ \mathbf{y}_N^c(t) &= C\hat{E}_N \{ \mathbf{x}_N(t) \mid \mathcal{S}_{[t_0, t]} \} + D\hat{E}_N \{ \mathbf{u}_N(t) \mid \mathcal{S}_{[t_0, t]} \} + \\ &+ \hat{E}_N \{ \mathbf{e}_N(t) \mid \mathcal{S}_{[t_0, t]} \} + \hat{\mathbf{e}}_N(t) \end{aligned}$$

and the representation (3.30) follows from the equalities

$$\begin{aligned} \hat{E}_N \{ \mathbf{x}_N(t) \mid \mathcal{S}_{[t_0, t]} \} &= \hat{E}_N \{ \mathbf{x}_N(t) \mid \mathcal{F}_{[t_0, t-1]} \} + \hat{E}_N \{ \mathbf{x}_N(t) \mid \hat{\mathcal{V}}_t \} = \hat{\mathbf{z}}_N^c(t) + \hat{K}_u(t)\hat{\mathbf{v}}_N(t) \\ \hat{E}_N \{ \mathbf{u}_N(t) \mid \mathcal{S}_{[t_0, t]} \} &= \hat{E}_N \{ \mathbf{u}_N(t) \mid \mathcal{F}_{[t_0, t-1]} \} + \hat{E}_N \{ \mathbf{u}_N(t) \mid \hat{\mathcal{V}}_t \} = 0 + \hat{\mathbf{v}}_N(t) \\ \hat{E}_N \{ \mathbf{e}_N(t) \mid \mathcal{S}_{[t_0, t]} \} &= \hat{E}_N \{ \mathbf{e}_N(t) \mid \mathcal{F}_{[t_0, t-1]} \} + \hat{E}_N \{ \mathbf{e}_N(t) \mid \hat{\mathcal{V}}_t \} = \hat{\boldsymbol{\zeta}}_N(t) + E(t)\hat{\mathbf{v}}_N(t) \end{aligned}$$

of which only the second requires some justification. Namely, the first term in the right is zero since  $\mathcal{F}_{[t_0, t-1]} \subset \mathcal{U}_{[t, T]}^\perp$  and, by definition,  $\hat{E}_N \{ \mathbf{u}_N(t) \mid \hat{\mathcal{V}}_t \} = \hat{\mathbf{v}}_N(t)$ . Moreover, from the definition of finite-length innovation space,  $\hat{\mathcal{E}}_t \perp \mathcal{U}_{[t_0, T]}$ , so that  $\hat{E}_N \{ \mathbf{u}_N(t) \mid \hat{\mathcal{E}}_t \} = 0$ .

The statements in (3.34) concerning the limit of  $\hat{\mathbf{v}}_N(t)$ ,  $\hat{\mathbf{e}}_N(t)$ , are obvious. That  $\hat{\boldsymbol{\zeta}}_N(t) \rightarrow 0$  follows since the finite truncation  $\mathbf{e}_N(t)$  of the stationary innovation process will, in the limit for  $N \rightarrow \infty$ , become orthogonal to  $\mathcal{U} \supset \mathcal{U}_{[t, T]}$ , because of the feedback-free hypothesis, and to  $\mathcal{P}_t^- \supset \mathcal{P}_{[t_0, t]}$ , by definition of stationary innovation. Since  $\mathcal{F}_{[t_0, t-1]} = \text{span}\{ \mathbf{p}_N - \hat{E}_N [\mathbf{p}_N \mid \mathcal{U}_{[t, T]}] \mid \mathbf{p}_N \in \mathcal{P}_{[t_0, t]} \} \subset \mathcal{P}_{[t_0, t]} \vee \mathcal{U}_{[t, T]}$ , we see that

$$\hat{E}_N [\mathbf{e}_N(t) \mid \mathcal{F}_{[t_0, t-1]}] \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

Hence  $E(t) \rightarrow 0$  and  $\hat{B}(t) \rightarrow AK_u(t) + B = \bar{B}(t)$ ,  $\hat{D}(t) \rightarrow CK_u(t) + D = \bar{D}(t)$ ,  $\hat{K}(t) \rightarrow K(t)$ .

□

## 4 The state approximation step

From the finite-data complementary model (3.30) of the previous section, one can naturally write subspace estimates of  $(A, C)$  which are the finite-length counterpart of formulas (3.15), (3.16) providing, under a non-singularity assumption of the covariance matrix  $E \left[ \hat{\mathbf{x}}^c(t) (\hat{\mathbf{x}}^c(t))^\top \right]$ , consistent estimates. The only difficulty

with these estimates is that the finite-length state variable  $\hat{\mathbf{z}}_N^c(t)$  is not directly computable since it involves the (truncated) unmeasurable state of the stationary innovation model. In practice, an estimate of  $\hat{\mathbf{z}}_N^c(t)$  must be constructed from the available input-output data.

Assume for the moment that the order  $n$  of the model (3.10) is known. Consider an estimate,  $\boldsymbol{\xi}(t)$ , of the state at time  $t$  of the model (3.10) based on input-output data of length  $N + 1$ . This estimate will be a certain  $n \times (N + 1)$  tail matrix which we shall construct later on as an ‘‘approximation’’ of the state  $\hat{\mathbf{z}}_N^c(t)$  of the finite-length model (3.30). Since  $\boldsymbol{\xi}(t)$  will approximate  $\hat{\mathbf{z}}_N^c(t)$  only in a certain basis, which will in general be different from the particular basis chosen for the model (3.30), we shall write

$$\boldsymbol{\xi}(t) = T_N \hat{\mathbf{z}}_N^c(t) + \tilde{\boldsymbol{\xi}}(t) \quad (4.1)$$

where  $T_N$  is a  $n \times n$  (data-dependent) nonsingular matrix, and  $\tilde{\boldsymbol{\xi}}(t)$  is an error term.

We shall assume that state estimate at time  $t + 1$ , obeys an analogous relation

$$\boldsymbol{\xi}(t + 1) = T_N \hat{\mathbf{z}}_N^c(t + 1) + \tilde{\boldsymbol{\xi}}(t + 1)$$

which means that  $\boldsymbol{\xi}(t + 1)$  is obtained by updating in a suitably ‘‘coherent’’ way the construction of  $\boldsymbol{\xi}(t)$  implemented at time  $t$ . We shall be more precise on this point in the next section. Then, using the complementary model (3.30) we can formally write a state-space representation of  $\mathbf{y}_N^c(t)$  in terms of the finite-length state estimate as

$$\begin{cases} \boldsymbol{\xi}(t + 1) &= A_N \boldsymbol{\xi}(t) + \hat{B}_N(t) \hat{\mathbf{v}}_N(t) + \boldsymbol{\varepsilon}_x(t) \\ \mathbf{y}_N^c(t) &= C_N \boldsymbol{\xi}(t) + \hat{D}(t) \hat{\mathbf{v}}_N(t) + \boldsymbol{\varepsilon}_y(t) \end{cases} \quad (4.2)$$

where

$$A_N = T_N A T_N^{-1} \quad C_N = C T_N^{-1} \quad B_N = T_N B \quad (4.3)$$

$\boldsymbol{\xi}(t_0) = 0$ , and the ‘‘error terms’’ are given by

$$\boldsymbol{\varepsilon}_x(t) := \left( \tilde{\boldsymbol{\xi}}(t + 1) - A_N \tilde{\boldsymbol{\xi}}(t) + \hat{K}_N(t) \hat{\boldsymbol{\varepsilon}}_N(t) + K_N \hat{\boldsymbol{\zeta}}_N(t) \right) \quad \boldsymbol{\varepsilon}_y(t) := \left( \hat{\boldsymbol{\varepsilon}}_N(t) - C_N \tilde{\boldsymbol{\xi}}(t) + \hat{\boldsymbol{\zeta}}_N(t) \right) \quad (4.4)$$

where  $K_N := T_N K$  and  $\hat{K}_N(t) := T_N K(t)$ . Naturally, we shall look for state estimates  $\boldsymbol{\xi}(t)$  which also belong to the finite-length subspace  $\mathcal{F}_{[t_0, t-1]}$ . In this case the first two terms on the right of (4.2) are orthogonal (i.e. finite-time uncorrelated).

The estimates of the model parameters,  $A_N, C_N$ , obtained by using the approximate states  $\boldsymbol{\xi}(t)$  and  $\boldsymbol{\xi}(t + 1)$  are defined as the matrices  $\hat{A}_N, \hat{C}_N$  solving the regression problem (4.2) in the least-squares sense.

**Lemma 4.1** *Assume that the state estimate  $\boldsymbol{\xi}(t) \in \mathcal{F}_{[t_0, t-1]}$  and that for  $N$  large enough, the covariance  $\hat{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}$  is non-singular. Then, the least-squares estimates of the parameters  $A_N, C_N$  in the regression model (4.2), are given by the formulas*

$$\hat{A}_N := \hat{\Sigma}_{\boldsymbol{\xi}'\boldsymbol{\xi}} \hat{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} = A_N + \left( \hat{\Sigma}_{\tilde{\boldsymbol{\xi}}'\boldsymbol{\xi}} - A_N \hat{\Sigma}_{\tilde{\boldsymbol{\xi}}\boldsymbol{\xi}} + K_N \hat{\Sigma}_{\hat{\boldsymbol{\zeta}}\boldsymbol{\xi}} \right) \hat{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} \quad (4.5)$$

$$\hat{C}_N := \hat{\Sigma}_{\mathbf{y}\boldsymbol{\xi}} \hat{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} = C_N + \left( \hat{\Sigma}_{\hat{\boldsymbol{\zeta}}\boldsymbol{\xi}} - C_N \hat{\Sigma}_{\tilde{\boldsymbol{\xi}}\boldsymbol{\xi}} \right) \hat{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} \quad (4.6)$$

where

$$\begin{aligned} \hat{\Sigma}_{\boldsymbol{\xi}'\boldsymbol{\xi}} &:= \hat{E}_N \{ \boldsymbol{\xi}(t + 1) \boldsymbol{\xi}(t)^\top \} & \hat{\Sigma}_{\tilde{\boldsymbol{\xi}}'\boldsymbol{\xi}} &:= \hat{E}_N \{ \tilde{\boldsymbol{\xi}}(t + 1) \boldsymbol{\xi}(t)^\top \} \\ \hat{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}} &:= \hat{E}_N \{ \boldsymbol{\xi}(t) \boldsymbol{\xi}(t)^\top \} & \hat{\Sigma}_{\tilde{\boldsymbol{\xi}}\boldsymbol{\xi}} &:= \hat{E}_N \{ \tilde{\boldsymbol{\xi}}(t) \boldsymbol{\xi}(t)^\top \} \\ \hat{\Sigma}_{\mathbf{y}\boldsymbol{\xi}} &:= \hat{E}_N \{ \mathbf{y}_N(t) \boldsymbol{\xi}(t)^\top \} & \hat{\Sigma}_{\hat{\boldsymbol{\zeta}}\boldsymbol{\xi}} &:= \hat{E}_N \{ \hat{\boldsymbol{\zeta}}_N(t) \boldsymbol{\xi}(t)^\top \} \end{aligned} \quad (4.7)$$

*Proof.* Project both members of (4.2) onto  $\mathcal{F}_{[t_0, t-1]}$ . Since the terms containing  $\hat{\mathbf{v}}_N(t)$  and  $\hat{\mathbf{e}}_N(t)$  are orthogonal to  $\mathcal{F}_{[t_0, t-1]}$ , we are left with

$$\begin{cases} \boldsymbol{\xi}(t+1) &= A_N \boldsymbol{\xi}(t) + \left( \tilde{\boldsymbol{\xi}}(t+1) - A_N \tilde{\boldsymbol{\xi}}(t) + K_N \hat{E}_N \left[ \hat{\boldsymbol{\zeta}}_N(t) \mid \mathcal{F}_{[t_0, t-1]} \right] \right) \\ \mathbf{y}_N^c(t) &= C_N \boldsymbol{\xi}(t) + \left( -C_N \tilde{\boldsymbol{\xi}}(t) + \hat{E}_N \left[ \hat{\boldsymbol{\zeta}}_N(t) \mid \mathcal{F}_{[t_0, t-1]} \right] \right) \end{cases}$$

Formulas (4.5), (4.6) are obtained by right multiplying the above formulas by  $(\boldsymbol{\xi}(t))^\top$  which amounts to taking (finite) expectations on both members.

□

#### 4.1 Construction of the state from measured data

Formulas (4.5), (4.6) provide a rather explicit expression for the estimation errors of the  $A, C$  matrices, but to proceed further in our analysis we need to introduce a specific state estimate. In this subsection we shall review a general state estimation procedure based on (weighted) canonical correlation analysis, a well-known concept in subspace identification (vanOverschee and De Moor 1995). We shall just introduce a slight variation in the standard procedure which, as we shall argue, yields a lower error covariance as it eliminates one source of error.

Consider the prediction of future outputs based on the (finite-length) “complementary past” subspace  $\mathcal{F}_{[t_0, t-1]}$

$$\hat{\mathbf{y}}_N(t+k \mid t-1) := \hat{E}_N [\mathbf{y}_N(t+k) \mid \mathcal{F}_{[t_0, t-1]}] = \hat{E}_N \left[ \mathbf{y}_N(t+k) \mid \left( \mathbf{p}_N^-(t) \mid (\mathbf{u}_N^+(t))^\perp \right) \right] \quad k \geq 0. \quad (4.8)$$

Note that we can decompose the output string at time  $t+k$  as  $\mathbf{y}_N(t+k) = \mathbf{y}_N^c(t+k) + \tilde{\mathbf{y}}_N^c(t+k)$  where  $\mathbf{y}_N^c(t+k) = \hat{E}_N [\mathbf{y}_N(t+k) \mid \mathcal{F}_{[t_0, t+k]}]$  is the complementary output at time  $t+k$  while  $\tilde{\mathbf{y}}_N^c(t+k) = \hat{E}_N [\mathbf{y}_N(t+k) \mid \mathcal{U}_{[t+k+1, T]}]$  is the part of  $\mathbf{y}_N(t+k)$ , predictable based on future inputs after time  $t+k$ . Since  $\mathcal{U}_{[t+k+1, T]} \subset \mathcal{U}_{[t, T]} \perp \mathcal{F}_{[t_0, t-1]}$ , it follows readily from (3.30) that

$$\begin{aligned} \hat{\mathbf{y}}_N(t+k \mid t-1) &= \hat{E}_N [\mathbf{y}_N^c(t+k) \mid \mathcal{F}_{[t_0, t-1]}] = CA^k \hat{\mathbf{z}}_N^c(t) + \\ &+ \hat{E}_N \left[ CA^{k-1} K \hat{\boldsymbol{\zeta}}_N(t) + CA^{k-2} K \hat{\boldsymbol{\zeta}}_N(t+1) + \dots \right. \\ &+ \left. CK \hat{\boldsymbol{\zeta}}_N(t+k-1) + \hat{\boldsymbol{\zeta}}_N(t+k) \mid \mathcal{F}_{[t_0, t-1]} \right] \end{aligned} \quad (4.9)$$

Now, introduce

$$(\hat{\mathbf{y}}_t^+)_N := \begin{bmatrix} \hat{\mathbf{y}}_N(t \mid t-1) \\ \hat{\mathbf{y}}_N(t+1 \mid t-1) \\ \vdots \\ \hat{\mathbf{y}}_N(T-1 \mid t-1) \end{bmatrix}, \quad (\hat{\boldsymbol{\zeta}}_t^+)_N := \hat{E}_N \left\{ \begin{bmatrix} \mathbf{e}_N(t) \\ \mathbf{e}_N(t+1) \\ \vdots \\ \mathbf{e}_N(T-1) \end{bmatrix} \mid \mathcal{F}_{[t_0, t-1]} \right\} \quad (4.10)$$

(note that here we consider a future history up to time  $T-1$ ) and let

$$\hat{\mathbf{w}}_t^+ := H_s (\hat{\boldsymbol{\zeta}}_t^+)_N \quad (4.11)$$

where  $H_s$  is the block-lower triangular Toeplitz matrix of the stochastic subsystem,

$$H_s = \begin{bmatrix} I & 0 & \dots & 0 & 0 \\ CK & I & \dots & 0 & 0 \\ \vdots & & & \ddots & \vdots \\ CA^{\nu-2}K & CA^{\nu-3}K & \dots & & I \end{bmatrix}, \quad \nu := T-t \quad (4.12)$$

so that we can write

$$(\hat{\mathbf{y}}_t^+)_N = \Gamma \hat{\mathbf{z}}_N^c(t) + \hat{\mathbf{w}}_t^+ \quad (4.13)$$

where  $\Gamma$  is the observability matrix  $\Gamma = [C^\top (CA)^\top \dots ((CA)^{\nu-1})^\top]^\top$ . Note that the “noise” vector  $\hat{\mathbf{w}}_t^+$  is zero for infinite data length.

Similarly, letting  $\hat{\mathbf{y}}_N(t+k | t) := \hat{E}_N [\mathbf{y}_N(t+k) | \mathcal{F}_{[t_0, t]}] = \hat{E}_N [\mathbf{y}_N(t+k) | (\mathbf{p}_N^-(t+1) | (\mathbf{u}_N^+(t+1))^\perp)]$  and bringing in the conditional shifts

$$(\hat{\mathbf{y}}_{t+1}^+)_N := \begin{bmatrix} \hat{\mathbf{y}}_N(t+1 | t) \\ \hat{\mathbf{y}}_N(t+2 | t) \\ \vdots \\ \hat{\mathbf{y}}_N(T | t) \end{bmatrix}, \quad (\hat{\boldsymbol{\zeta}}_{t+1}^+)_N := \hat{E}_N \left\{ \begin{bmatrix} \mathbf{e}_N(t+1) \\ \mathbf{e}_N(t+2) \\ \vdots \\ \mathbf{e}_N(T) \end{bmatrix} \middle| \mathcal{F}_{[t_0, t]} \right\} \quad (4.14)$$

we can write

$$(\hat{\mathbf{y}}_{t+1}^+)_N = \Gamma \hat{\mathbf{z}}_N^c(t+1) + \hat{\mathbf{w}}_{t+1}^+ \quad (4.15)$$

here, in accordance with (4.11), we have set  $\hat{\mathbf{w}}_{t+1}^+ := H_s(\hat{\boldsymbol{\zeta}}_{t+1}^+)_N$ .

Now, since we are operating with data of finite length, due to the additive noise term  $\hat{\mathbf{w}}_t^+$ , we no longer have equality between (finite-length) predictor space and state space as in (3.14). In fact  $\text{row-span}\{(\hat{\mathbf{y}}_t^+)_N\}$  will in general be of full dimension  $m\nu = m(T-t)$ . One has to construct a suitable subspace of the (finite-length) predictor space  $\text{row-span}\{(\hat{\mathbf{y}}_t^+)_N\}$ , i.e. an “approximate” state space built from the available input-output data, which is a “best” approximation of the (unknown) theoretical finite-length state space  $\text{row-span}\{\hat{\mathbf{z}}_N^c(t)\}$ .

A standard way to solve this problem, is to consider the singular value decomposition of the (weighted) covariance matrix<sup>5</sup>

$$W \hat{E}_N \{(\hat{\mathbf{y}}_t^+)_N ((\hat{\mathbf{y}}_t^+)_N)^\top\} W^\top = [\hat{U}_N \hat{V}_N] \text{diag}\{\hat{S}_N^2, \tilde{S}_N^2\} [\hat{U}_N \hat{V}_N]^\top = \hat{U}_N \hat{S}_N^2 \hat{U}_N^\top + \hat{V}_N \tilde{S}_N^2 \hat{V}_N^\top; \quad (4.16)$$

where  $W$  is a nonsingular weighting matrix<sup>6</sup>,  $[\hat{U}_N \hat{V}_N]$  is an orthogonal matrix and  $\tilde{S}_N^2$  is the diagonal matrix of “small” squared singular values which are declared to be noise. Deciding how many singular values are declared to be zero and how many are retained in the first piece of formula (4.16) is the order estimation problem which we shall not discuss in this paper. We shall just assume that the order estimator is *consistent* in the sense that the correct number  $n$  of (nonzero) singular values is retained for  $N$  large enough.

After separating “signal” from “noise”, the approximate basis in the state space at time  $t$  is taken to be

$$\boldsymbol{\xi}(t) := \hat{S}_N^{-1/2} \hat{U}_N^\top W (\hat{\mathbf{y}}_t^+)_N \quad (4.17)$$

while, at time  $t+1$  we choose the conditional shift

$$\boldsymbol{\xi}(t+1) := \hat{S}_N^{-1/2} \hat{U}_N^\top W (\hat{\mathbf{y}}_{t+1}^+)_N. \quad (4.18)$$

From the viewpoint frequently taken in subspace identification, one might, equivalently, say that  $\hat{\Gamma}_N := W^{-1} \hat{U}_N \hat{S}_N^{1/2}$  is the estimate of the observability matrix  $\Gamma$  (in the chosen basis). This is clearly the same thing as saying that  $\boldsymbol{\xi}(t) = \hat{\Gamma}_N^{-L} (\hat{\mathbf{y}}_t^+)_N$  is the chosen basis for the approximate state space. Here we shall always use the left inverse given by

$$\hat{\Gamma}_N^{-L} := \hat{S}_N^{-1/2} \hat{U}_N^\top W. \quad (4.19)$$

**Lemma 4.2** Consider the vector of infinite-length predictors (3.13) and let

$$WE\{\hat{\mathbf{y}}_t^+ (\hat{\mathbf{y}}_t^+)^\top\} W^\top = US^2U^\top \quad S^2 = \text{diag}\{\sigma_1^2, \dots, \sigma_n^2\} \quad (4.20)$$

<sup>5</sup>To avoid ambiguities we shall henceforth assume that the singular values are distinct and  $U$  is taken with a positive first non zero element in each column.

<sup>6</sup>Which for simplicity will be assumed constant; in reality  $W = W_N$  is data dependent, but converges to a fixed nonsingular matrix as  $N \rightarrow \infty$ , so the more complicated asymptotics which applies to the data dependent weighting will in the end give the same results as for a constant weighting.

where the singular values are ordered in decreasing magnitude. Assume that the  $n$ -th singular value of (4.20) is positive<sup>7</sup> and that the following canonical basis

$$\hat{\mathbf{x}}^c(t) := S^{-1/2}U^\top W \hat{\mathbf{y}}_t^+ \quad (\text{here } N = \infty) \quad (4.21)$$

is chosen in the state space (3.14) of the true complementary model (3.10). Then, for data of infinite length, the approximate basis  $\boldsymbol{\xi}(t)$  coincides with (4.21), which we shall write as

$$\lim_{N \rightarrow \infty} \boldsymbol{\xi}(t) = \hat{\mathbf{x}}^c(t) \quad (4.22)$$

*Proof.* By consistency of order estimation, for  $N \rightarrow \infty$  the term  $\hat{V}_N \tilde{S}_N^2 \hat{V}_N^\top$  tends to zero and the factorization in (4.16) converges to (4.20). Hence  $S^{-1/2}U^\top W$  is the asymptotic value of  $\hat{S}_N^{-1/2} \hat{U}_N^\top W$  for data of infinite length (i.e. for  $N \rightarrow \infty$ ).

□

Having fixed the basis in the true model, it is clear that the  $n \times n$  matrix

$$\hat{T}_N := \hat{S}_N^{-1/2} \hat{U}_N^\top W \Gamma = \hat{S}_N^{-1/2} \hat{U}_N^\top U S^{1/2}, \quad (4.23)$$

defines the change of basis which was alluded to in (4.1). Note that  $\hat{T}_N$  asymptotically tends to the identity matrix and hence it may be assumed of full rank provided  $N$  is taken sufficiently large. From equations (4.13), (4.14), (4.17), and (4.18) we obtain

$$\boldsymbol{\xi}(t) = \hat{T}_N \hat{\mathbf{z}}_N^c(t) + \hat{S}_N^{-1/2} \hat{U}_N^\top W \hat{\mathbf{w}}_t^+ \quad (4.24)$$

and

$$\boldsymbol{\xi}(t+1) = \hat{T}_N \hat{\mathbf{z}}_N^c(t+1) + \hat{S}_N^{-1/2} \hat{U}_N^\top W \hat{\mathbf{w}}_{t+1}^+ \quad (4.25)$$

where, as required in the preceding discussion, the same matrix  $\hat{T}_N$  appears in both equations.

**Remark 4.1** Introduce the “augmented” noise vector

$$\hat{\mathbf{w}}_t^+ := \bar{H}_s \hat{E}_N \left\{ \begin{array}{c} \mathbf{e}_N(t) \\ \mathbf{e}_N(t+1) \\ \vdots \\ \mathbf{e}_N(T-1) \\ \mathbf{e}_N(T) \end{array} \middle| \mathcal{F}_{[t_0, t-1]} \right\} \quad (4.26)$$

and let  $\bar{H}_s$  be the block Toeplitz matrix  $H_s$  of (4.12) bordered with one more block row and column.

Most standard procedures in subspace identification use an “augmented” predictor vector  $(\hat{\mathbf{y}}_t^+)_N := [\hat{\mathbf{y}}_N(t | t-1)^\top \dots \hat{\mathbf{y}}_N(T-1 | t-1)^\top \hat{\mathbf{y}}_N(T | t-1)^\top]^\top$  with  $T-t+1$  block rows and row-span $\{(\hat{\mathbf{y}}_t^+)_N\}$  in lieu of row-span $\{\hat{\mathbf{y}}_t^+\}$ . This leads to an extended observability matrix  $\hat{\hat{\Gamma}}_N$  with one extra block in the formula for the state space at time  $t$  while  $\hat{\Gamma}_N$  is used instead at time  $t+1$ . With this choice, we have that  $\hat{\hat{T}}_N = \hat{\hat{\Gamma}}_N^{-L} \Gamma \neq \hat{\Gamma}_N^{-L} \Gamma = \hat{T}_N$ , due to errors in the estimation of the observability matrix. Therefore a further source of errors may be introduced due to the fact that equations (4.24) and (4.25) now read

$$\boldsymbol{\xi}(t) = \hat{\hat{T}}_N \hat{\mathbf{z}}_N^c(t) + \hat{\hat{\Gamma}}_N^{-L} \hat{\mathbf{w}}_t^+ \quad (4.27)$$

and

$$\boldsymbol{\xi}(t+1) = \hat{\hat{T}}_N \hat{\mathbf{z}}_N^c(t+1) + \hat{\hat{\Gamma}}_N^{-L} \hat{\mathbf{w}}_{t+1}^+. \quad (4.28)$$

Here the difference between  $\hat{T}_N$  and  $\hat{\hat{T}}_N$  should be accounted for in the computation of the error covariance.

◇

<sup>7</sup>This is generically true. See however (Chui 1997, Jansson and Wahlberg 1997) for precise conditions on the underlying processes.

Using the formulas (4.24), (4.25) and comparing with (4.1), we get explicit expressions of the error terms  $\tilde{\xi}(t)$ ,  $\tilde{\xi}(t+1)$  and  $\hat{e}_x(t)$ ,  $\hat{e}_y(t)$ . After substituting in (4.4) we obtain

$$\hat{E}_N \left[ \left( \tilde{\xi}(t+1) - A_N \tilde{\xi}(t) + \hat{K}_N(t) \hat{e}_N(t) + K_N \hat{\zeta}_N(t) \right) \mid \mathcal{F}_{[t_0, t-1]} \right] = [(K_N \quad \hat{\Gamma}_N^{-L}) - A_N (\hat{\Gamma}_N^{-L} \quad 0_{n \times m})] \hat{\mathbf{w}}_t^+$$

Similarly

$$\begin{aligned} \hat{E}_N \left[ \left( \hat{e}_N(t) - C_N \tilde{\xi}(t) + \hat{\zeta}_N(t) \right) \mid \mathcal{F}_{[t_0, t-1]} \right] &= \hat{\zeta}_N(t) - C_N \hat{\Gamma}_N^{-L} \hat{\mathbf{w}}_t^+ \\ &= [(I_m \quad 0_{m \times m(\nu-1)}) - C_N \hat{\Gamma}_N^{-L}] \hat{\mathbf{w}}_t^+ \end{aligned}$$

In order to work with more compact formulas we introduce the matrices

$$\hat{M}_N := [(K_N \quad \hat{\Gamma}_N^{-L}) - A_N (\hat{\Gamma}_N^{-L} \quad 0_{n \times m})] \quad \hat{R}_N := [(I_m \quad 0_{m \times m(\nu-1)}) - C_N \hat{\Gamma}_N^{-L}] \quad (4.29)$$

which, for  $N \rightarrow \infty$ , tend to the limits  $M$  and  $R$  given by

$$M := [(K \quad \Gamma^{-L}) - A (\Gamma^{-L} \quad 0_{n \times m})] \quad R := [(I_m \quad 0_{m \times m(\nu-1)}) - C \Gamma^{-L}]$$

**Proposition 4.1** *The errors on the system matrix estimates with data of length  $N$ , can be expressed as*

$$\tilde{A}_N = \hat{A}_N - A_N = \hat{M}_N \bar{H}_s \hat{E}_N [\bar{\mathbf{e}}_t^+ \xi(t)^\top] \hat{\Sigma}_{\xi\xi}^{-1} := \hat{M}_N \bar{H}_s \hat{\Sigma}_{\bar{\mathbf{e}}^+ \xi} \hat{\Sigma}_{\xi\xi}^{-1} \quad (4.30)$$

$$\tilde{C}_N = \hat{C}_N - C_N = \hat{R}_N H_s \hat{E}_N \{ \mathbf{e}_t^+ \xi(t)^\top \} \hat{\Sigma}_{\xi\xi}^{-1} := \hat{R}_N H_s \hat{\Sigma}_{\mathbf{e}^+ \xi} \hat{\Sigma}_{\xi\xi}^{-1} \quad (4.31)$$

where  $\bar{\mathbf{e}}_t^+ := [\mathbf{e}_N(t)^\top \quad \mathbf{e}_N(t+1)^\top \quad \dots \quad \mathbf{e}_N(T-1)^\top \quad \mathbf{e}_N(T)^\top]^\top$  is the augmented truncated stationary innovation vector of the true model.

*Proof.* It follows from (4.5) and (4.6) that

$$\begin{aligned} \tilde{A}_N &= \hat{A}_N - A_N := \hat{M}_N \hat{E}_N \{ \hat{\mathbf{w}}_t^+ \xi(t)^\top \} \hat{\Sigma}_{\xi\xi}^{-1} \\ \tilde{C}_N &= \hat{C}_N - C_N := \hat{R}_N \hat{E}_N \{ \hat{\mathbf{w}}_t^+ \xi(t)^\top \} \hat{\Sigma}_{\xi\xi}^{-1}. \end{aligned}$$

and since  $\xi(t) \in \mathcal{F}_{[t_0, t-1]}$  we may write the finite expectation term in the first formula as

$$\hat{E}_N \{ \hat{\mathbf{w}}_t^+ \xi(t)^\top \} = \bar{H}_s \hat{E}_N [\bar{\mathbf{e}}_t^+ \xi(t)^\top] = \bar{H}_s \hat{E}_N \left\{ \begin{bmatrix} \mathbf{e}_N(t) \\ \mathbf{e}_N(t+1) \\ \vdots \\ \mathbf{e}_N(T-1) \\ \mathbf{e}_N(T) \end{bmatrix} \xi(t)^\top \right\} = \bar{H}_s \hat{\Sigma}_{\bar{\mathbf{e}}^+ \xi}$$

Likewise, we can write  $\hat{E}_N \{ \hat{\mathbf{w}}_t^+ \xi(t)^\top \} = H_s \hat{\Sigma}_{\mathbf{e}^+ \xi}$ .

□

**Remark 4.2** We have chosen to express the estimation errors  $\tilde{A}_N, \tilde{C}_N$  in the current, data-dependent, basis determined by the SVD step of the estimation algorithm. In other words, both the estimates and the true values  $A, C$  are expressed in the basis (4.17) determined by the SVD (4.16). It is however immediate to express the estimation errors in the asymptotic canonical basis of the true system, defined by (4.21). In fact, since the estimates expressed in the asymptotic basis are, respectively, given by the formulas  $\hat{T}_N^{-1} \tilde{A}_N \hat{T}_N$  and  $\hat{C}_N \hat{T}_N$  the errors in the asymptotic basis are just

$$\tilde{\tilde{A}}_N := \hat{T}_N^{-1} \tilde{A}_N \hat{T}_N - A = \hat{T}_N^{-1} (\hat{A}_N - A_N) \hat{T}_N = \hat{T}_N^{-1} \tilde{A}_N \hat{T}_N \quad (4.32)$$

$$\tilde{\tilde{C}}_N := \hat{C}_N \hat{T}_N - C = (\hat{C}_N - C_N) \hat{T}_N = \tilde{\tilde{C}}_N \hat{T}_N \quad (4.33)$$

◇

## 4.2 Main Result

We now come to the main result of this section. We shall assume that in the model (2.4) of the true system generating the data, the innovation process  $\{\mathbf{e}(t)\}$  is a martingale difference with respect to the  $\sigma$ -algebra  $\mathcal{E}_t \vee \mathcal{U}$  generated by the random variables  $\{\mathbf{e}(s); s < t\}$  and  $\{\mathbf{u}(t); t \in \mathbb{Z}\}$ , more precisely, assume for  $j, k \geq 0$ , that

$$E \{\mathbf{e}(t+k) \mid \mathcal{E}_t \vee \mathcal{U}\} = 0 \quad k \geq 0 \quad (4.34a)$$

$$E \{\mathbf{e}(t+j)\mathbf{e}(t+k)^\top \mid \mathcal{E}_t \vee \mathcal{U}\} = E \{\mathbf{e}(t+j)\mathbf{e}(t+k)^\top\} = \Lambda \delta_{jk} \quad (4.34b)$$

for a positive definite matrix  $\Lambda$ . We shall also need boundedness of the fourth moment of  $\{\mathbf{e}(t)\}$ . These “noise conditions” are often found in the statistical literature, see e. g. (Hannan and Deistler 1988); they hold, for example, if  $\{\mathbf{e}(t)\}$  is a i.i.d. process (strict sense white noise) with finite fourth order moments, independent of  $\mathbf{u}$ , or if  $\{\mathbf{e}(t)\}$  is Gaussian, independent of  $\mathbf{u}$ . In the first situation we shall also assume that the observed trajectory (2.1) is an ergodic trajectory of the joint input-output process. For Gaussian process, second-order ergodicity suffices since it is the same as ergodicity.

The Gaussian distribution with mean  $\mu$  and covariance matrix  $\Sigma$  is denoted  $\mathcal{N}(\mu, \Sigma)$ . If a sequence of random vectors  $\{\mathbf{z}_N\}$  converges almost surely to a constant  $z_0$  and is *asymptotically normal*, i.e.  $\sqrt{N}(\mathbf{z}_N - z_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ , where  $\xrightarrow{d}$  denotes convergence in distribution, one says that  $\Sigma$  is the *asymptotic variance* of  $\{\sqrt{N} \mathbf{z}_N\}$ . Notation:  $\Sigma = \text{AsVar}(\sqrt{N} \mathbf{z}_N)$ . The *asymptotic covariance* of two, asymptotically jointly Gaussian, sequences is defined in a similar way.

**Theorem 4.1** *Assume that the stationary innovation process,  $\{\mathbf{e}(t)\}$ , in the model (2.4) of the true system generating the data, satisfies the conditions (4.34) and has finite fourth order moments. Then the vectorized parameter estimates  $[\text{vec}(\hat{A}_N)^\top \text{vec}(\hat{C}_N)^\top]^\top$  form an asymptotically Gaussian sequence with*

$$\text{AsVar} \left( \sqrt{N} \text{vec}(\hat{A}_N) \right) = \bar{F} \left\{ \sum_{|\tau| \leq \nu} \Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}(\tau) \otimes \Sigma_{\bar{\mathbf{e}}^+ \bar{\mathbf{e}}^+}(\tau) \right\} \bar{F}^\top \quad (4.35)$$

$$\text{AsVar} \left( \sqrt{N} \text{vec}(\hat{C}_N) \right) = F \left\{ \sum_{|\tau| < \nu} \Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}(\tau) \otimes \Sigma_{\mathbf{e}^+ \mathbf{e}^+}(\tau) \right\} F^\top \quad (4.36)$$

$$\text{AsCov} \left( \sqrt{N} \text{vec}(\hat{A}_N), \sqrt{N} \text{vec}(\hat{C}_N) \right) = \bar{F} \left\{ \sum_{\tau=-\nu}^{\tau=\nu-1} \Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}(\tau) \otimes \Sigma_{\bar{\mathbf{e}}^+ \mathbf{e}^+}(\tau) \right\} F^\top \quad (4.37)$$

where  $F := \Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}^{-1} \otimes [R H_s]$ ,  $\bar{F} := \Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}^{-1} \otimes [M \bar{H}_s]$  and

$$\Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}(\tau) := E \{ \sigma^\tau \hat{\mathbf{x}}^c(t) \hat{\mathbf{x}}^c(t)^\top \}, \quad \Sigma_{\bar{\mathbf{e}}^+ \mathbf{e}^+}(\tau) = E \{ \mathbf{e}_{t+\tau}^+ (\mathbf{e}_t^+)^\top \} \quad (4.38)$$

the operator  $\sigma^\tau$  being the  $\tau$ -steps ahead stationary shift of the processes  $\mathbf{y}, \mathbf{u}$ , whereby

$$\sigma^\tau \hat{\mathbf{x}}^c(t) = \sigma^\tau E [\mathbf{x}(t) \mid \mathcal{F}_{[t_0, t-1]}] = E [\mathbf{x}(t+\tau) \mid \mathcal{F}_{[t_0+\tau, t+\tau-1]}] \quad (4.39)$$

The understanding here is that  $\mathcal{F}_{[t_0+\tau, t+\tau-1]} := (\mathcal{P}_{[t_0+\tau, t+\tau]} \vee \mathcal{U}_{[t+\tau, T+\tau]}) \ominus \mathcal{U}_{[t+\tau, T+\tau]}$ .

*Proof.* The proof follows a standard line of arguments (Hannan and Deistler 1988, Viberg et al. 1997, Jansson 2000). By elementary manipulations of Kronecker products we can write the errors given in Proposition 4.1 as, say,

$$\text{vec} \left( \tilde{A}_N \right) = \hat{\Sigma}_{\xi\xi}^{-1} \otimes \left[ \hat{M}_N \bar{H}_s \right] \cdot \text{vec} \left( \hat{E}_N \{ \bar{\mathbf{e}}_t^+ \xi(t)^\top \} \right) := \hat{F}_N \mathbf{w}_N(t)$$

where the matrix  $\hat{F}_N$  and the vector  $\mathbf{w}_N(t)$  are

$$\hat{F}_N := \hat{\Sigma}_{\xi\xi}^{-1} \otimes \left[ \hat{M}_N \bar{H}_s \right], \quad \mathbf{w}_N(t) := \text{vec} \left( \hat{E}_N \{ \bar{\mathbf{e}}_t^+ \xi(t)^\top \} \right) = \hat{E}_N \{ \xi(t) \otimes \bar{\mathbf{e}}_t^+ \}$$

Consider now the quantity

$$\begin{aligned}\hat{E}_N [\boldsymbol{\xi}(t) \otimes \bar{\mathbf{e}}_t^+] \hat{E}_N [\boldsymbol{\xi}(t) \otimes \bar{\mathbf{e}}_t^+]^\top &= \frac{1}{(N+1)^2} \sum_{i=0}^N \sum_{j=0}^N [\boldsymbol{\xi}_{t+i} \otimes \bar{\mathbf{e}}_{t+i}^+] [\boldsymbol{\xi}_{t+j}^\top \otimes (\bar{\mathbf{e}}_{t+j}^+)^\top] \\ &= \frac{1}{(N+1)^2} \sum_{i=0}^N \sum_{j=0}^N [\boldsymbol{\xi}_{t+i} \boldsymbol{\xi}_{t+j}^\top] \otimes [\bar{\mathbf{e}}_{t+i}^+ (\bar{\mathbf{e}}_{t+j}^+)^\top]\end{aligned}$$

This quantity of course changes randomly depending on the particular sample trajectory (2.1) chosen by “nature”, so we can (and shall) think of it as a realization of a *bona-fide* random variable whose expected value we want to compute. To this purpose we may also think of the column vectors  $\boldsymbol{\xi}_{t+i}, \boldsymbol{\xi}_{t+j}, \bar{\mathbf{e}}_{t+i}^+, \dots$  as particular sample values of random variables (which we shall here denote by the same symbols), each of which has been properly defined in the course of the preceding sections, as a function of the sample trajectory (2.1).

**Lemma 4.3** *Under the stated assumptions the following limit relation holds*

$$\lim_{N \rightarrow \infty} NE\{\hat{E}_N [\boldsymbol{\xi}(t) \otimes \bar{\mathbf{e}}_t^+] \hat{E}_N [\boldsymbol{\xi}(t) \otimes \bar{\mathbf{e}}_t^+]^\top\} = \sum_{|\tau| \leq \nu} \Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}(\tau) \otimes \Sigma_{\bar{\mathbf{e}}^+ \bar{\mathbf{e}}^+}(\tau) \quad (\dagger)$$

where  $\Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}(\tau)$  is defined in (4.38).

*Proof.* Note that  $\boldsymbol{\xi}_{t+i}$  is measurable with respect to  $\mathcal{E}_{t+i} \vee \mathcal{U}$  so that, using (4.34), after some rearrangements and using ergodicity, one obtains

$$E\{\hat{E}_N [\boldsymbol{\xi}(t) \otimes \bar{\mathbf{e}}_t^+] \hat{E}_N [\boldsymbol{\xi}(t) \otimes \bar{\mathbf{e}}_t^+]^\top\} = \frac{1}{N+1} \sum_{\tau=-\nu}^{\nu} \left(1 - \frac{|\tau|}{N+1}\right) \Sigma_{\boldsymbol{\xi} \boldsymbol{\xi}}(\tau) \otimes \Sigma_{\bar{\mathbf{e}}^+ \bar{\mathbf{e}}^+}(\tau)$$

where  $\Sigma_{\boldsymbol{\xi} \boldsymbol{\xi}}(\tau) = E[\sigma^\tau \boldsymbol{\xi}(t) \boldsymbol{\xi}(t)^\top]$ . The limits in the sum can be taken to be  $\pm\nu$  since  $\Sigma_{\bar{\mathbf{e}}^+ \bar{\mathbf{e}}^+}(\tau)$  is zero for  $|\tau| > \nu$ . Note that (4.22) implies that  $\hat{\Sigma}_{\boldsymbol{\xi} \boldsymbol{\xi}} \rightarrow \Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}$ .  
□

We can now invoke a version of the central limit theorem, see e.g. (? , p. 550), to conclude that  $\sqrt{N} \mathbf{w}_N(t) \xrightarrow{d} \mathcal{N}(0, P)$  where the asymptotic variance  $P$  is the matrix in the last member of (†).

Finally, since  $\hat{F}_N := \hat{\Sigma}_{\boldsymbol{\xi} \boldsymbol{\xi}}^{-1} \otimes [\hat{M}_N \hat{H}_s]$  converges almost surely (and hence in probability) to the constant matrix  $\bar{F} := \Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}^{-1} \otimes [M \bar{H}_s]$ , we easily conclude that

$$\sqrt{N} \tilde{A}_N = \sqrt{N} \hat{F}_N \mathbf{w}_N(t) \xrightarrow{d} \mathcal{N}(0, \bar{F} P \bar{F}^\top)$$

which is (4.35).

The proof of (4.36) and of (4.37) is analogous.  
□

**Remark 4.3** Formulas (4.35), (4.36) and (4.37) should be compared with the asymptotic variance expressions in the literature, notably with those obtained in (Bauer and Jansson 2000, Jansson 2000). In this respect we highlight the following points

1. The state covariance matrix  $\Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}(\tau)$  appears in place of the joint data covariance matrix  $R_{\zeta\zeta}(\tau)$  of formula (37) in (Jansson 2000). Moreover the identity (3.18) already discussed in section 3, relates the variance of the estimates with the possible ill-conditioning of the subspace estimation problem, see (Chiuso and Picci 2001b).
2. The expressions (4.35), (4.36), (4.37) can be used for parameter estimates obtained by many subspace methods, namely MOESP, Robust N4SID, and finite-interval CCA, by specializing the choice of the weighting matrix  $W$ , see (vanOverschee and De Moor 1995, Van Overschee and De Moor 1996) for the particular expression of  $W$  which applies in each case. This may allow to compare the accuracy of different methods, given that both  $\Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}$  and the matrices  $M$  and  $R$  in general depend on the choice

of the weighting matrix  $W$ . The dependence can be seen for instance from the formula  $\Gamma \Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c} \Gamma^\top = E\{\hat{\mathbf{y}}_t^+ (\hat{\mathbf{y}}_t^+)^{\top}\}$  (Lemma 4.2), where  $\Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c} \equiv S_s$  depends on  $W$  (unless  $W$  is chosen to be an orthogonal matrix). A comparison of these methods from the point of view of relative asymptotic efficiency is however outside the scope of this paper.

◇

**Remark 4.4** For the practical computation of  $\Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}(\tau)$  one should be careful not to confuse the stationary shift  $\sigma^\tau \hat{\mathbf{x}}^c(t)$  with the conditional shift  $\hat{\mathbf{x}}^c(t + \tau)$  (which is not stationarily correlated with  $\hat{\mathbf{x}}^c(t)$ ). The stationary shift makes it easy to approximate  $\Sigma_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}(\tau)$  from finite I/O data. As seen in the course of the proof, a natural sample estimate can be obtained by just using the state approximation  $\xi(t)$ , computed at time  $t$ , using the formula

$$\hat{\Sigma}_{\hat{\mathbf{x}}^c \hat{\mathbf{x}}^c}(\tau) \simeq \frac{1}{N+1} \sum_{i=0}^{N-\tau} \xi_{t+i+\tau} \xi_{t+i}^\top \quad N \rightarrow \infty$$

of course we should make sure that the estimate is a positive function, but we shall not insist on this point here.

◇

**Remark 4.5** The asymptotic variance expressions (4.35), (4.36), (4.37), describe the errors in a data-dependent basis. However the same formulas provide also the asymptotic variances of the estimation errors expressed in the asymptotic canonical basis (4.21) of the true system. We state this formally in the following Corollary.

**Corollary 4.1** *Exactly the same asymptotic variance expressions (4.35), (4.36), (4.37), hold for the errors  $\tilde{A}_N, \tilde{C}_N$ , expressed in the asymptotic canonical basis of the true system.*

*Proof.* This follows from formulas (4.32) (4.33), the first of which can be rewritten

$$\tilde{A}_N = \left( I - (\hat{T}_N - I) \right)^{-1} \tilde{A}_N \left( I - (\hat{T}_N - I) \right) = \tilde{A}_N + (\hat{T}_N - I) \tilde{A}_N + \tilde{A}_N (\hat{T}_N - I) + O(\|\hat{T}_N - I\|^2)$$

and the fact that, from (4.23),  $\hat{T}_N - I = [\hat{S}_N^{-1/2} \hat{U}_N - S^{-1/2} U]^\top U S^{1/2} = O(\frac{1}{\sqrt{N}})$  for  $N \rightarrow \infty$ . In other words,  $\sqrt{N} [\tilde{A}_N - \tilde{A}_N] \rightarrow 0$  almost surely (and in probability) which implies that  $\sqrt{N} \tilde{A}_N$  and  $\sqrt{N} \tilde{A}_N$  have the same asymptotic distribution, see e.g. (? , Theorem 6, p. 39). An analogous expansion holds for  $\tilde{C}_N$ .

□

More generally, it may be worth stressing that, provided of course the estimates are consistent and asymptotically expressed in the *same basis* chosen for the true parameters, knowing the asymptotic variance of the estimates ( $\hat{A}_N, \hat{C}_N, \hat{B}_N, \hat{D}_N$ ), permits to compute the asymptotic variance of any smooth function of the true parameters, in particular of any system invariant. More precisely, let

$$\theta := [\text{vec}(A)^\top \text{vec}(C)^\top \text{vec}(B)^\top \text{vec}(D)^\top]^\top$$

denote the true system matrices and let  $\hat{\theta}_N$  denote the estimate of the (vectorized) system matrices based on  $N$  data points. Assuming that  $\sqrt{N} \hat{\theta}_N$  is consistent and asymptotically normal, i.e. that  $\hat{B}_N, \hat{D}_N$  have the same kind of asymptotic behavior as  $\hat{A}_N, \hat{C}_N$ , the asymptotic variance of the estimate,  $g(\hat{\theta}_N)$ , of any smooth function  $g(\theta)$ , can be computed by a well-known linearization technique, see (? , Thm. 7, p.45),

$$\text{AsVar} \left[ \sqrt{N} g(\hat{\theta}_N) \right] = \frac{\partial g}{\partial \theta} \Big|_{\theta} \text{AsVar} \{ \sqrt{N} \hat{\theta}_N \} \frac{\partial g}{\partial \theta} \Big|_{\theta}^\top \quad (4.40)$$

We will see an application of this formula to the estimate of the transfer function, at the end of the next section.

As suggested by a reviewer, we shall demonstrate the use of the expression of the asymptotic variance (4.35) for computing the asymptotic variance of certain system invariants, in particular the eigenvalues of the system.

Assume for simplicity that the “true” matrix  $A$  has simple eigenvalues. According to (Stewart and Sun 1990, Thm 2.3, p. 183), there is an eigenvalue  $\lambda^i$  of  $A$  such that the difference between the  $i$ -th eigenvalue of  $(\hat{T}_N)^{-1}\hat{A}_N\hat{T}_N$ ,  $\hat{\lambda}_N^i$ , and  $\lambda^i$ , satisfies

$$\hat{\lambda}_N^i - \lambda^i = \frac{v_i^\top \tilde{A}_N u_i}{v_i^\top u_i} + O(\|\tilde{A}_N\|^2) \quad (4.41)$$

where  $v_i$  and  $u_i$  are the normalized left and right eigenvectors of  $A$  corresponding to  $\lambda^i$ . From this it is immediate to see that  $\sqrt{N}(\hat{\lambda}_N^i - \lambda^i)$  is also asymptotically normal with asymptotic variance

$$\text{AsVar}[\sqrt{N}(\hat{\lambda}_N^i - \lambda^i)] = \frac{1}{(v_i^\top u_i)^2} (u_i^\top \otimes v_i^\top) \text{AsVar} \left\{ \sqrt{N} \text{vec}(\tilde{A}_N) \right\} (u_i \otimes v_i) \quad (4.42)$$

which in particular implies

$$\text{AsVar}[\sqrt{N}(\hat{\lambda}_N^i - \lambda^i)] \leq \frac{1}{(v_i^\top u_i)^2} \lambda_{\max} \left\{ \text{AsVar}[\sqrt{N} \text{vec}(\tilde{A}_N)] \right\} \quad (4.43)$$

where  $\lambda_{\max}[\cdot]$  means maximum eigenvalue. Note that  $(v_i^\top u_i)^2$  is the square of the cosine of the angle between the two eigenvectors. This is less or equal to one and equal to one just in case the matrix  $A$  is symmetric (in which case  $v_i = u_i$ ).

Formula (4.42) provides a simple and useful estimate for the asymptotic variance of the eigenvalues of the system.  $\diamond$

**Remark 4.6** Under slightly more stringent assumptions guaranteeing that  $\hat{F}_N := \hat{\Sigma}_{\xi\xi}^{-1} \otimes [\hat{M}_N \bar{H}_s]$  (or the companion sequence of estimates  $\hat{F}_N$ ) converges to  $\bar{F}$  ( $F$ ) also in  $L^2$ , it is possible to show that the asymptotic variances (4.35) are actual limits of the finite sample variances, e.g.

$$\lim_{N \rightarrow \infty} NE \left\{ \text{vec}(\tilde{A}_N) \text{vec}(\tilde{A}_N)^\top \right\} = \text{AsVar} \left( \sqrt{N} \text{vec}(\tilde{A}_N) \right) \quad (4.44)$$

$$\lim_{N \rightarrow \infty} NE \left\{ \text{vec}(\tilde{C}_N) \text{vec}(\tilde{C}_N)^\top \right\} = \text{AsVar} \left( \sqrt{N} \text{vec}(\tilde{C}_N) \right) \quad (4.45)$$

$$\lim_{N \rightarrow \infty} NE \left\{ \text{vec}(\tilde{A}_N) \text{vec}(\tilde{C}_N)^\top \right\} = \text{AsCov} \left( \sqrt{N} \text{vec}(\tilde{A}_N) \sqrt{N} \text{vec}(\tilde{C}_N) \right) \quad (4.46)$$

This means that the asymptotic variance formulas of Theorem 4.1 describe the outcome of a Monte-Carlo simulation where the data are generated by a known true system, whose  $(A, C)$  parameters may always, by standard computations, be brought to the (asymptotic) canonical basis (4.21). Of course the sample variance of the results of say  $M$  Monte Carlo runs should be computed subtracting the “true” known mean values. For example, the right hand side of (4.35) is the limit for  $M \rightarrow \infty$  of the average

$$\frac{N}{M} \sum_{i=1}^M \text{vec} \left[ (\hat{T}_N^i)^{-1} \hat{A}_N^i \hat{T}_N^i - A_0 \right] \text{vec} \left[ (\hat{T}_N^i)^{-1} \hat{A}_N^i \hat{T}_N^i - A_0 \right]^\top$$

where  $A_0$  is the true known  $A$  matrix of the simulated system, expressed in the canonical basis (4.21).  $\diamond$

## 5 The asymptotic variance of $(B, D)$

Several algorithms have been proposed in the literature for the estimation of the matrices  $(B, D)$ , see e.g. (vanOverschee and De Moor 1996, Verhaegen and Dewilde 1992, Verhaegen 1994). In this section we shall generalize slightly a standard procedure which is based on “linear regression on  $B, D$ ”. The algorithm of (Verhaegen and Dewilde 1992) is a special case of the one described below.

We shall derive the minimum variance (Markov) estimate of  $(B, D)$  and the relative expression for the error covariance, assuming first that  $A, C$  are known. An expression for the asymptotic variance which takes into account also the sample variations in the estimates of  $A, C$  can be obtained from these expressions using a linearization technique similar to that employed in (Jansson 2000). The calculations are easy but tedious and since do not add anything conceptually new we shall omit most of the details.

Let us consider the equations obtained by substituting all infinite-length (random) variables in the model (2.4) by the corresponding tail matrices of length  $N$ . In the following, unless otherwise stated, all **bold symbols** will represent tail matrices with  $N$  columns and for simplicity we shall not use subscripts.

Let  $\hat{E}_N [\mathbf{y}_t^+ | \mathbf{u}_t^+]$  be the projection of the future outputs onto future inputs at time  $t$ . Here there is no need to chose the same “present” time  $t$  as in the previous sections (in fact, it may be reasonable to pick  $t = t_0$ ), but, just in order to avoid having to introduce further notations, we shall keep the same meaning of  $t$ . The vectors  $\mathbf{y}_t^+, \mathbf{u}_t^+$  are defined in (3.19) where they carry a subscript  $N$  which has now been dropped. This projection can be written as

$$\hat{E}_N [\mathbf{y}_t^+ | \mathbf{u}_t^+] = \Gamma \hat{E}_N [\mathbf{x}(t) | \mathbf{u}_t^+] + H_d(B, D) \mathbf{u}_t^+ + H_s \hat{E}_N [\mathbf{e}_t^+ | \mathbf{u}_t^+]. \quad (5.1)$$

where  $H_d$  is the lower triangular block-Toeplitz matrix of the Markov parameters of the “deterministic” subsystem, namely

$$H_d = H_d(B, D) = \begin{bmatrix} D & 0 & \dots & 0 & 0 \\ CB & D & \dots & 0 & 0 \\ \vdots & & & \ddots & \vdots \\ CA^{\nu-2}B & CA^{\nu-3}B & \dots & & D \end{bmatrix},$$

and where  $H_s$  is the lower triangular block-Toeplitz matrix of the Markov parameters of the “stochastic” subsystem defined as in formula (4.12). The third term in (5.1) is the regression of future innovations  $\{\mathbf{e}(t), \dots, \mathbf{e}(T-1)\}$  on future inputs at time  $t$ . By the feedback-free assumption, it should ideally be zero; in practice, due to finite sample length effects, it is not. It can formally be expressed by the formula

$$\hat{E}_N [\mathbf{e}_t^+ | \mathbf{u}_t^+] = \hat{\Sigma}_{\mathbf{e}+\mathbf{u}^+} \hat{\Sigma}_{\mathbf{u}^+\mathbf{u}^+}^{-1} \mathbf{u}_t^+.$$

The left-hand side of (5.1) has the form  $\hat{\mathbf{y}}_t^+ = \hat{E}_N [\mathbf{y}_t^+ | \mathbf{u}_t^+] := \hat{\Phi}_y \mathbf{u}_t^+$  where the regression matrix  $\hat{\Phi}_y$  can be computed by solving a least-squares problem. Hence equation (5.1) is rewritten as

$$\hat{\Phi}_y \mathbf{u}_t^+ = \left( \Gamma \hat{\Phi}_x + H_d(B, D) + H_s \hat{\Sigma}_{\mathbf{e}+\mathbf{u}^+} \hat{\Sigma}_{\mathbf{u}^+\mathbf{u}^+}^{-1} \right) \mathbf{u}_t^+, \quad (5.2)$$

where  $\hat{\Phi}_x$  is the regression matrix of the (unknown) state on  $\mathbf{u}^+$ . Due to the “sufficient richness” (or persistence of excitation) of  $\mathbf{u}$ , (5.2) is immediately seen to be equivalent (for  $N$  large enough) to the “dual” equation for the coefficients

$$\hat{\Phi}_y = \Gamma \hat{\Phi}_x + H_d(B, D) + H_s \hat{\Sigma}_{\mathbf{e}+\mathbf{u}^+} \hat{\Sigma}_{\mathbf{u}^+\mathbf{u}^+}^{-1}. \quad (5.3)$$

Now,  $H_d(B, D)$  is linear in the parameters  $(B, D)$  so that it can be written in vectorized form as

$$\text{vec } H_d(B, D) = L \begin{bmatrix} \text{vec}(B) \\ \text{vec}(D) \end{bmatrix} \quad (5.4)$$

for a suitable matrix  $L$  depending on  $(A, C)$ , and equation (5.3) re-stated in vectorized form is rewritten

$$\text{vec} \left( \hat{\Phi}_y \right) = [I_{km} \otimes \Gamma] \text{vec} \left( \hat{\Phi}_x \right) + L \begin{bmatrix} \text{vec}(B) \\ \text{vec}(D) \end{bmatrix} + \left( \hat{\Sigma}_{\mathbf{u}^+\mathbf{u}^+}^{-1} \otimes H_s \right) \text{vec} \left( \hat{\Sigma}_{\mathbf{e}+\mathbf{u}^+} \right) \quad (5.5)$$

Assuming  $(A, C)$  are known, this relation can be interpreted as a linear regression of the (known) vector  $\text{vec } \hat{\Phi}_y$  on the (known) quantities  $([I_{km} \otimes \Gamma], L)$ , with unknown parameters  $(\hat{\Phi}_x, B, D)$ .

The additive term  $\text{vec} \left( \hat{\Sigma}_{\mathbf{e}+\mathbf{u}^+} \right)$  is regarded as a random perturbation vector whose covariance matrix  $\Sigma_0$ , can be computed exactly for finite  $N$ . Under the same assumptions (4.34) of the previous section, we have

$$\Sigma_0 = E \left[ \text{vec} \left( \hat{\Sigma}_{\mathbf{e}+\mathbf{u}^+} \right) \text{vec} \left( \hat{\Sigma}_{\mathbf{e}+\mathbf{u}^+} \right)^\top \right] = \frac{1}{N+1} \sum_{|\tau| < \nu} \left( 1 - \frac{|\tau|}{N+1} \right) \Sigma_{\mathbf{u}^+\mathbf{u}^+}(\tau) \otimes \Sigma_{\mathbf{e}+\mathbf{e}^+}(\tau) \quad (5.6)$$

where

$$\Sigma_{\mathbf{u}+\mathbf{u}^+}(\tau) = E\{\mathbf{u}_{t+\tau}^+(\mathbf{u}_t^+)^{\top}\} \quad (5.7)$$

Assuming  $N$  is large enough, so that  $\hat{\Sigma}_{\mathbf{u}+\mathbf{u}^+}(\tau) \simeq \Sigma_{\mathbf{u}+\mathbf{u}^+}(\tau)$  (the population covariance) we get a good approximation,  $W_0$ , of the variance of the additive noise term in (5.5) by the formula

$$W_0 := (\Sigma_{\mathbf{u}+\mathbf{u}^+}^{-1} \otimes H_s) \Sigma_0 (\Sigma_{\mathbf{u}+\mathbf{u}^+}^{-1} \otimes H_s)^{\top}$$

The following statement easily follows.

**Theorem 5.1** *Assuming  $(A, C)$  are known, the formula,*

$$\begin{bmatrix} \text{vec}(\hat{B}) \\ \text{vec}(\hat{D}) \end{bmatrix} = \left( L^{\top} W_0^{-T/2} \Delta^{\perp} W_0^{-1/2} L \right)^{-\sharp} L^{\top} W_0^{-T/2} \Delta^{\perp} W_0^{-1/2} \text{vec}(\hat{\Phi}_y) \quad (5.8)$$

provides the minimum variance linear (Markov) estimate of  $B, D$  from the “dual” regression equation (5.5). Here  $W_0^{-T/2} := (W_0^{-1/2})^{\top}$ ,  $\Delta := W_0^{-1/2} [I_{km} \otimes \Gamma]$  and  $\Delta^{\perp} = I - \Delta(\Delta^{\top} \Delta)^{-1} \Delta^{\top}$  is the orthogonal projection onto  $[\text{col-span}(\Delta)]^{\perp}$ , and  $^{-\sharp}$  denotes Moore-Penrose pseudoinverse. The variance of the estimates of the  $B, D$  parameters is

$$\text{Var} \left\{ \begin{bmatrix} \text{vec}(\hat{B}) \\ \text{vec}(\hat{D}) \end{bmatrix} \right\} = \left( L^{\top} W_0^{-T/2} \Delta^{\perp} W_0^{-1/2} L \right)^{-\sharp} \quad (5.9)$$

To obtain realistic expressions for the asymptotic variance one needs to account for the uncertainty in the parameters  $A$  and  $C$ . In order to streamline notation, let us define

$$\Pi := \left( L^{\top} W_0^{-T/2} \Delta^{\perp} W_0^{-1/2} L \right)^{-\sharp} L^{\top} W_0^{-T/2} \Delta^{\perp} W_0^{-1/2}.$$

and denote by  $\hat{\Pi}_N, \hat{L}_N, \hat{\Gamma}_N$  and  $\Pi_N, L_N, \Gamma_N$  the matrices  $\Pi, L, \Gamma$  computed using respectively the estimates  $\hat{A}_N$  and  $\hat{C}_N$  and the true values,  $A_N$  and  $C_N$ , converted to the basis defined by (4.23). In the same basis the  $B$  and  $K$  matrices of the system are given by  $B_N := T_N B, K_N := T_N K$ . The estimates provided by formula (5.8) will be denoted by  $\hat{B}_N, \hat{D}_N$ . In the following, the subscript  $N$  will be used to denote estimates computed with a data set of length  $N$ , expressed with respect to the basis defined by (4.23).

Recall that, by construction,  $\hat{\Pi}_N \hat{L}_N = I$  and  $\hat{\Pi}_N (I \otimes \hat{\Gamma}_N) = 0$ , so that we can write the estimate  $\text{vec} \begin{pmatrix} \hat{B}_N \\ \hat{D}_N \end{pmatrix} := \hat{\Pi}_N \text{vec}(\hat{\Phi}_y)_N$  as

$$\begin{aligned} \begin{bmatrix} \text{vec}(\hat{B}_N) \\ \text{vec}(\hat{D}_N) \end{bmatrix} &= \begin{bmatrix} \text{vec}(\hat{B}_N) \\ \text{vec}(\hat{D}) \end{bmatrix} + \hat{\Pi}_N (I \otimes (\Gamma_N - \hat{\Gamma}_N)) (\hat{\Phi}_x)_N + \hat{\Pi}_N (L_N - \hat{L}_N) \begin{bmatrix} \text{vec}(\hat{B}_N) \\ \text{vec}(\hat{D}_N) \end{bmatrix} + \\ &+ \hat{\Pi}_N \left( \hat{\Sigma}_{\mathbf{u}+\mathbf{u}^+}^{-1} \otimes (\hat{H}_s)_N \right) \text{vec}(\hat{\Sigma}_{\mathbf{e}+\mathbf{u}^+}) \end{aligned}$$

Note that if  $A_N$  and  $C_N$  were known exactly, the two terms containing  $(\Gamma_N - \hat{\Gamma}_N)$  and  $(L_N - \hat{L}_N)$  would vanish and from this expression one would get back the variance formula (5.9). Moreover, since these terms are linear functions of the errors  $\text{vec}(\tilde{A}_N)$  and  $\text{vec}(\tilde{C}_N)$ , linearizing  $\hat{\Pi}_N$  and  $(\hat{\Phi}_x)_N$  around the true values  $A_N$  and  $C_N$ , i.e. substituting  $\hat{\Pi}_N = (\hat{\Pi}_N - \Pi_N) + \Pi_N$ ,  $(\hat{\Phi}_x)_N = [(\hat{\Phi}_x)_N - (\Phi_x)_N] + (\Phi_x)_N$  etc. and neglecting higher order terms whose variance goes to zero faster than  $1/N$  for  $N \rightarrow \infty$ , the error can be expressed as

$$\begin{aligned} \begin{bmatrix} \text{vec}(\tilde{B}_N) \\ \text{vec}(\tilde{D}_N) \end{bmatrix} &= \Pi_N (I \otimes (\Gamma_N - \hat{\Gamma}_N)) \text{vec}(\Phi_x)_N + \Pi_N (L_N - \hat{L}_N) \begin{bmatrix} \text{vec}(B_N) \\ \text{vec}(D) \end{bmatrix} + \\ &+ \Pi_N \left( \hat{\Sigma}_{\mathbf{u}+\mathbf{u}^+}^{-1} \otimes H_{s,N} \right) \text{vec}(\hat{\Sigma}_{\mathbf{e}+\mathbf{u}^+}) + o\left(\frac{1}{\sqrt{N}}\right) \end{aligned}$$

and hence as a linear function of  $\text{vec}(\tilde{A}_N)$  and  $\text{vec}(\tilde{C}_N)$  as follows:

$$\begin{bmatrix} \text{vec}(\tilde{B}_N) \\ \text{vec}(\tilde{D}_N) \end{bmatrix} = L_{A_N} \text{vec}(\tilde{A}_N) + L_{C_N} \text{vec}(\tilde{C}_N) + L_{H_N} \text{vec}(\hat{\Sigma}_{\mathbf{e}+\mathbf{u}^+}) + o\left(\frac{1}{\sqrt{N}}\right) \quad (5.10)$$

where  $L_{A_N}, L_{C_N}, L_{H_N}$  are the evaluation at  $A = A_N, C = C_N, B = B_N$  of

$$\begin{aligned} L_A &:= \Pi_N \left[ \left( I \otimes \left( \frac{\partial}{\partial \text{vec}(A)} \Gamma \right) \right) \otimes_b \text{vec}(\Phi_x) + \left( \frac{\partial}{\partial \text{vec}(A)} L \right) \otimes_b \text{vec} \begin{pmatrix} B \\ D \end{pmatrix} \right] \\ L_C &:= \Pi_N \left[ \left( I \otimes \left( \frac{\partial}{\partial \text{vec}(C)} \Gamma \right) \right) \otimes_b \text{vec}(\Phi_x) + \left( \frac{\partial}{\partial \text{vec}(C)} L \right) \otimes_b \text{vec} \begin{pmatrix} B \\ D \end{pmatrix} \right] \\ L_H &:= \Pi_N (\Sigma_{\mathbf{u}^+ \mathbf{u}^+}^{-1} \otimes H_s) \end{aligned}$$

Here  $\otimes_b$  denotes block Kronecker product. Introducing compact symbols for the various asymptotic covariances

$$\begin{aligned} \Sigma(A, C) &:= \text{AsVar} \left\{ \begin{bmatrix} \text{vec}(\hat{A}_N) \\ \text{vec}(\hat{C}_N) \end{bmatrix} \right\} := \begin{bmatrix} (4.35) & (4.37) \\ (4.37)^\top & (4.36) \end{bmatrix} \\ \Sigma(B, D) &:= \text{AsVar} \left\{ \begin{bmatrix} \text{vec}(\hat{B}_N) \\ \text{vec}(\hat{D}_N) \end{bmatrix} \mid \hat{A}_N = A, \hat{C}_N = C \right\} = N \left( L^\top W_0^{-T/2} \Delta^\perp W_0^{-1/2} L \right)^{-\sharp} \end{aligned}$$

we obtain the following expressions for the covariance matrices,

$$\begin{aligned} \text{AsCov} \left\{ \begin{bmatrix} \text{vec}(\hat{A}_N) \\ \text{vec}(\hat{C}_N) \end{bmatrix} \begin{bmatrix} \text{vec}(\hat{B}_N) \\ \text{vec}(\hat{D}_N) \end{bmatrix}^\top \right\} &= \Sigma(A, C) \begin{bmatrix} L_A^\top \\ L_C^\top \end{bmatrix} + \\ &+ \begin{bmatrix} (\Sigma_{\hat{x}^c \hat{x}^c}^{-1} \otimes M \bar{H}_s) (\sum_{\tau=-\nu}^{\nu} \Sigma_{\hat{x}^c \mathbf{u}^+}(\tau) \otimes \Sigma_{\bar{\mathbf{e}}^+ \bar{\mathbf{e}}^+}(\tau)) \\ (\Sigma_{\hat{x}^c \hat{x}^c}^{-1} \otimes R H_s) (\sum_{\tau=-\nu+1}^{\nu} \Sigma_{\hat{x}^c \mathbf{u}^+}(\tau) \otimes \Sigma_{\mathbf{e}^+ \bar{\mathbf{e}}^+}(\tau)) \end{bmatrix} L_H^\top, \\ \text{AsVar} \left\{ \begin{bmatrix} \text{vec}(\hat{B}_N) \\ \text{vec}(\hat{D}_N) \end{bmatrix} \right\} &= \Sigma(BD) + [L_A \ L_C] \Sigma(A, C) \begin{bmatrix} L_A^\top \\ L_C^\top \end{bmatrix} \\ &+ [L_A \ L_C] \begin{bmatrix} (\Sigma_{\hat{x}^c \hat{x}^c}^{-1} \otimes M \bar{H}_s) (\sum_{\tau=-\nu}^{\nu} \Sigma_{\hat{x}^c \mathbf{u}^+}(\tau) \otimes \Sigma_{\bar{\mathbf{e}}^+ \bar{\mathbf{e}}^+}(\tau)) \\ (\Sigma_{\hat{x}^c \hat{x}^c}^{-1} \otimes R H_s) (\sum_{\tau=-\nu+1}^{\nu} \Sigma_{\hat{x}^c \mathbf{u}^+}(\tau) \otimes \Sigma_{\mathbf{e}^+ \bar{\mathbf{e}}^+}(\tau)) \end{bmatrix} L_H^\top \\ &+ \left( [L_A \ L_C] \begin{bmatrix} (\Sigma_{\hat{x}^c \hat{x}^c}^{-1} \otimes M \bar{H}_s) (\sum_{\tau=-\nu}^{\nu} \Sigma_{\hat{x}^c \mathbf{u}^+}(\tau) \otimes \Sigma_{\bar{\mathbf{e}}^+ \bar{\mathbf{e}}^+}(\tau)) \\ (\Sigma_{\hat{x}^c \hat{x}^c}^{-1} \otimes R H_s) (\sum_{\tau=-\nu+1}^{\nu} \Sigma_{\hat{x}^c \mathbf{u}^+}(\tau) \otimes \Sigma_{\mathbf{e}^+ \bar{\mathbf{e}}^+}(\tau)) \end{bmatrix} L_H^\top \right)^\top \end{aligned}$$

The formulas above provide complete expressions for the overall asymptotic covariance matrix of the parameter estimates. Similar (although a bit less explicit) expressions have been obtained by (Jansson 2000) based on an unweighted least squares estimator of  $B, D$ .

**Remark 5.1** Naturally, for assessing the overall quality of the estimates, the most interesting quantity to consider is just the system transfer function. According to the general principle discussed in Remark ??, the asymptotic variance of the transfer function can be computed by using the previous expressions for the asymptotic covariance of the estimates  $(\hat{A}_N, \hat{C}_N, \hat{B}_N, \hat{D}_N)$ . The result follows by a straightforward linearization,

$$\begin{aligned} \text{vec}(\hat{W}(z) - W(z)) &\simeq \\ &\simeq [W_1 W_2 \ W_1 \ W_2 \ I] \text{vec}([\hat{A}_N \ \hat{C}_N \ \hat{B}_N \ \hat{D}_N]). \end{aligned} \quad (5.11)$$

where

$$W_1 := \left( (zI - A_N)^{-1} B_N \right)^\top \otimes I, \quad W_2 := I \otimes C_N (zI - A_N)^{-1}$$

as explained in the paper (Jansson 2000), to which the reader is referred for the details.  $\diamond$

## 6 Conclusions

Using ideas of stochastic realization we have derived asymptotic expressions for the covariance matrix of subspace estimates of the matrices  $(A, B, C, D,)$  of a state-space realization. These expressions provide new insight in the estimation problem. In particular

1. The variance of the estimates of  $A, C$  is seen to be roughly ‘‘proportional’’ to the inverse of the conditional covariance  $\Sigma_{\hat{x}\hat{x}|\mathbf{u}^+}$ . This relates the statistical accuracy to the possible ill-conditioning of the computation of the estimates.

2. The inverse of the covariance of the input process appears in the expression of the variance of the  $B, D$  parameters (5.9). This describes the influence of the conditioning of the input process on the estimates of  $B, D$ . A poorly conditioned input Toeplitz matrix  $\Sigma_{\mathbf{u}+\mathbf{u}^+}$  is seen to correspond to a “large” additive noise variance  $W_0$  and to poor estimates.
3. The formulas can be used for several estimation algorithms (CVA, N4SID, MOESP) by specializing the choice of the weighting matrix  $W$  as described in (vanOverschee and De Moor 1995).

## Acknowledgments

This work has been supported by the national project *Identification and Adaptive Control of Industrial Systems* funded by MURST. The European Commission is herewith acknowledged for its financial support in part to the research reported on in this paper. The support was provided via the Program Training and Mobility of Researchers (TMR) and Project System Identification (ERB FMRX CT98 0206) to the European Research Network System Identification (ERNSI).

## Appendix A: The sample-trajectory framework

Let (2.1) be a second-order ergodic<sup>8</sup> trajectory of a bona-fide  $(m+p)$ -dimensional second-order stationary process  $\mathbf{z} = [\mathbf{y}^\top, \mathbf{u}^\top]^\top$ . Consider the correspondence

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

associating a generic linear combination of the components of the  $t$ -th random variables of the processes  $\{\mathbf{y}\}$  and  $\{\mathbf{u}\}$  to the same linear combination of the rows of the “tail” matrices made with the present and future after time  $t$  of the ergodic trajectory.

This map can be extended by linearity to all combinations of random variables of the processes  $\{\mathbf{y}\}$  and  $\{\mathbf{u}\}$ . In fact, the correspondence  $\mathcal{T}$  seen as a map from the “stochastic” Hilbert space  $\mathcal{Y} \vee \mathcal{U}$  of zero-mean second order random variables to the vector space  $\text{span}\{Y_t, U_t, | t \in \mathbb{Z}_+\}$  closed with respect to the inner product (2.7), is an *isometry*, i.e. it maps random variables into semi-infinite sequences, preserving their inner product. It follows from a general theorem on isometric maps on Hilbert spaces (Rozanov 1967), that  $\mathcal{T}$  can be extended to an isometric map from the Hilbert space generated by zero mean second order random variables of the process  $\{\mathbf{z}(t)\}$ , into the Hilbert space  $\overline{\text{span}}\{Y_t, U_t, | t \in \mathbb{Z}_+\}$  generated by the tails constructed with the ergodic trajectory. We can actually make this map *unitary* by identifying stationary trajectories which give rise to the same true covariance.

Hence the “stochastic” Hilbert space of zero mean second order random variables and the Hilbert space of a stationary sample function (2.1) of the underlying stochastic process are *isometrically isomorphic*. 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 Hilbert space  $\overline{\text{span}}\{Y_t, U_t, | t \in \mathbb{Z}_+\}$ . For this reason we shall denote this latter space by the same symbol introduced for subspaces of random variables, and denote also the corresponding elements (semi-infinite tail sequences) by boldface letters as done for random quantities. This useful correspondence was introduced and used in (Lindquist and Picci 1996a), (Lindquist and Picci 1996b).

## References

- Bauer, D., 1998, Some Asymptotic Theory for the Estimation of Linear Systems Using Maximum Likelihood Methods or Subspace Algorithms, Ph. D. Thesis, TU Wien.
- Bauer, D. and M. Jansson, 2000, Analysis of the asymptotic properties of the moesp type of subspace algorithms, *Automatica* 36, No. 4, 497-509.

<sup>8</sup>This is the same thing as a “second-order stationary” or “quasi-stationary” signal, as defined in Section 1.

- Bauer, D., 2000, Asymptotic efficiency of the CCA subspace method in the case of no exogenous inputs, submitted to Journal of Time Series Analysis.
- Bauer, D. and L. Ljung, 2001, Some facts about the choice of the weighting matrices in Larimore type of subspace algorithm, to appear in Automatica.
- Caines, P. E. and C. W. Chan, 1976, Estimation, identification and feedback, in R. Mehra and D. Lainiotis, eds., System Identification: Advances and Case Studies (Academic), 349-405.
- Chiuso, A. and G. Picci, 1999, Subspace Identification by Orthogonal Decomposition, in Proceedings 14th IFAC World Congress, Vol. I, 241-246
- Chiuso, A. and G. Picci, 2000, Error Analysis of Certain Subspace Methods, in Proceedings of IFAC International Symposium on System Identification, Santa Barbara, June 2000, pp. 85-90.
- Chiuso, A. and G. Picci, 2001a, Some Algorithmic aspects of Subspace Identification with Inputs, Applied Mathematics and Computer Sciences, Vol. 11, No. 1, 55-75.
- Chiuso, A. and G. Picci, 2001b, On the Ill-Conditioning of Subspace Identification with Inputs, Tech Rep. TRITA/MATH-01-OS5, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden (submitted for publication).
- Chui, N.L.C. 1997, Subspace methods and informative experiments for subspace identification. Ph.D. Thesis, Pembroke College Cambridge. Obtained at <http://www-control.eng.com.ac.uk/nlcc/report/thesis.ps>.
- Gevers, M. R. and B. D. O. Anderson, 1982, On jointly stationary feedback-free stochastic processes, IEEE Transactions on Automatic Control, 27, 431-436.
- Gevers, M.R. and B.D.O. Anderson, 1981, Representation of jointly stationary feedback free processes, International Journal of Control, 33, pp.777-809.
- Hannan, E. J. and D. S. Poskitt, 1988, Unit canonical correlations between future and past, The Annals of Statistics, 16, 784-790.
- Hannan, E.J. and M. Deistler, 1988, The Statistical Theory of Linear Systems (Wiley).
- Jansson, M. and B. Wahlberg, 1996, A linear regression approach to state-space subspace system identification, Signal Processing, 52, 103-129.
- Jansson, M. and B. Wahlberg, 1997, Counterexample to general consistency of subspace system identification methods, Automatica, 34, No. 12, 1507-1519
- Jansson, M., 2000, Asymptotic Variance Analysis of Subspace Identification Methods, in Proceedings of IFAC International Symposium on System Identification, S. Barbara (CA).
- Katayama T. and G. Picci, 1999, Realization of Stochastic Systems with Exogenous Inputs and Subspace System Identification Methods, Automatica 35, No. 10, 1635-1652.
- Larimore, W. E., 1983, System identification, reduced-order filtering and modeling via canonical variate analysis, Proceedings of the American Control Conference, 445-451.
- Lindquist A. and G. Picci, 1991, A geometric approach to modelling and estimation of linear stochastic systems, Journal of Mathematical Systems, Estimation and Control 1, 241-333.
- Lindquist A. and G. Picci, 1996a, Canonical correlation analysis approximate covariance extension and identification of stationary time series, Automatica 32, pp. 709-733.
- Lindquist A. and G. Picci, 1996b, Geometric Methods for State-Space Identification, in S. Bittanti and G. Picci. eds, Identification, Adaptation, Learning, (Springer Verlag), 1-69.

- Picci G. and T. Katayama, 1996, Stochastic realization with exogenous inputs and “Subspace Methods” Identification, *Signal Processing* 52, 145-160.
- Rozanov, Y.A., 1967, *Stationary Random Processes* (Holden-Day, San Francisco).
- Stewart G. W. and J. Sun, 1990, *Matrix Perturbation Theory* (Academic Press).
- Van Overschee, P. and B. De Moor, 1993, Subspace algorithms for the stochastic identification problem, *Automatica* 29, 649-660.
- Van Overschee, P. and B. De Moor, 1994, N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems, *Automatica* 30 75-93.
- Van Overschee, P. and B. De Moor, 1994, A unifying theorem for subspace system identification algorithms and its interpretation. *Proceedings of the 10th IFAC Symposium on System Identification* 2, 145-156.
- Van Overschee, P. and B. De Moor, 1995, Choice of State-space Basis in Combined Deterministic-Stochastic Subspace Identification, *Automatica* 31, 1877-1883.
- Van Overschee, P. and B. De Moor, 1996, *Subspace Identification for Linear Systems* (Kluwer Academic Publications).
- Viberg, M., B. Wahlberg and B. Ottersten, 1997, Analysis of State Space System Identification Methods Based on Instrumental Variables and Subspace Fitting, *Automatica* 33, 1603-1616.
- 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. *International Journal of Control* 56, 1187-1210 & 1211-1241.
- 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.
- Wiener, N., 1930, Generalized Harmonic Analysis. *Acta Mathematica* 55, 117-258.
- Wiener, N., 1933, Generalized Harmonic Analysis, in *The Fourier Integral and Certain of its Applications* (Cambridge U.P.).