

# Some insights on the choice of the future horizon in closed-loop CCA-type Subspace Algorithms

Alessandro Chiuso

**Abstract**—The length of the future horizon is one of the principal “user choices” in subspace identification. It is well known that for the CCA algorithm the asymptotic variance decreases as the future horizon increases when input signals are white (or absent). There are also examples, when the input is not white, in which the asymptotic variance is minimized when the future horizon is equal to the system order.

In this paper we shall discuss how (and why) the choice of the future horizon influences the accuracy. Even though we shall not get to the point of giving a methodology to make this choice, we believe the considerations contained in this paper can be a starting point towards a completely automated choice of user parameters in subspace identification. As an intermediate result we show the a version of the CCA algorithm introduced by Larimore is asymptotically equivalent to a number of recently studied methods, complementing recent results appeared in the literature. The setup we consider includes closed loop identification.

**Index Terms**—Closed Loop Identification, Subspace Methods, Statistical Analysis.

## I. INTRODUCTION

Subspace algorithms for identification of MIMO linear time invariant systems have seen a considerable development in the last two decades (see [31], [21], [32], [2]. Very recently new methods [24], [28], [11], [26], [19], [14], [25] have allowed subspace identification to be applied also in closed loop, making their range of applicability even wider.

All these algorithms are very much related to each other (and to Maximum Likelihood); the main results available to date can be enumerated as follows:

- The CCA algorithm [1], [20], [29] is efficient for time series identification (no inputs) [3] and optimal when the inputs are white [4];
- the algorithms in [19] (SSARX), [14], [15] (PBSID) are asymptotically equivalent<sup>1</sup> [8], [5];
- PBSID and the standard CCA method are asymptotically equivalent (for open loop data) when inputs are white (or absent) [6], [7];
- an optimized version of PBSID (called PBSID<sub>opt</sub>) provides an asymptotic variance which is never larger than that obtained using the CCA method [6], [7];
- PBSID<sub>opt</sub> corresponds exactly to performing VARX estimation followed by the usual steps of subspace

This work has been supported in part by the national project *New methods and algorithms for identification and adaptive control of technological systems* funded by MIUR

A. Chiuso is with the Dipartimento di Tecnica e Gestione dei Sistemi Industriali, Università di Padova (sede di Vicenza), stradella San Nicola, 3 - 36100 Vicenza, Italy. E-mail: chiuso@dei.unipd.it

<sup>1</sup>This means that the estimators obtained have the same asymptotic distribution and hence, in particular, the same asymptotic variance.

identification [10], [5], making it very appealing from the computational point of view. Furthermore PBSID<sub>opt</sub> is a “weighted version” of the algorithm originally discussed in [24] (see [5]).

This shows that most algorithms are essentially equivalent and suggest a quite natural (and computationally attractive) procedure (PBSID<sub>opt</sub>) via VARX modeling followed by model reduction which implements them.

For this class of algorithms (which, having all in common the use of Canonical Correlation Analysis (CCA) [18], can be called of the “CCA-type”) there is yet one parameter<sup>2</sup> to be chosen by the user, i.e. the length of the future horizon ( $\nu$  from now on); this is the number of block rows in the Hankel data matrices containing the future data.

While it has been shown that for the classic CCA algorithm the asymptotic variance of the estimated parameters is a monotonically non-increasing function of  $\nu$  [4] when inputs are white or absent, to the best of the author knowledge there are no results concerning more general cases. Some very preliminary discussion is contained in [27].

In this paper we shall go one step further:

- a) we shall show that also the algorithm in [22] (see also [28]) is asymptotically equivalent to SSARX [19] (see Proposition 3.1);
- b) using the result above we shall study the effect of  $\nu$  in the state construction step. In particular we shall see that the error in the estimated state can be decomposed as the sum of two terms. We shall discuss, to some extent, the role of these errors and their link with  $\nu$  reobtaining the well known results in [4] when inputs are white. An illustrating example is then designed on the basis of these considerations.

Of course, the asymptotic variance formulas derived in [9] provide a natural tool to optimize  $\nu$ ; this is however not very efficient as it would require computing many the variance expressions for various values of  $\nu$ . Moreover we believe it is worth trying to exploit as much as possible the structure rather than performing a (costly) blind optimization.

The structure of the paper is as follows: Section II contains the problem description and some notation; in Section III we describe the two algorithms studied and prove their asymptotic equivalence (Proposition 3.1). In Section IV we shall discuss the effect of the future horizon and Section V

<sup>2</sup>From Assumption 1 the length of the “past” should grow with the number of data. In practice this will have to be estimated from data, typically using standard criterion for ARX model order estimation. See [14], [10], [5] for a discussion.

contains a numerical illustration; finally some conclusions are drawn in Section VI.

## II. STATEMENT OF THE PROBLEM AND NOTATION

Let  $\{\mathbf{y}(t)\}, \{\mathbf{u}(t)\}$  be jointly (weakly) stationary second-order ergodic stochastic processes of dimension  $p$  and  $m$  respectively, which are the output and input signals of a linear stochastic system in innovation form

$$\begin{cases} \mathbf{x}(t+1) &= A\mathbf{x}(t) + B\mathbf{u}(t) + K\mathbf{e}(t) \\ \mathbf{y}(t) &= C\mathbf{x}(t) + D\mathbf{u}(t) + \mathbf{e}(t) \end{cases} \quad t \geq t_0. \quad (\text{II.1})$$

we allow for *feedback* from  $\{\mathbf{y}(t)\}$  to  $\{\mathbf{u}(t)\}$  [17], i.e. we consider “closed loop” identification. Without loss of generality we shall assume that the dimension  $n$  of the state vector  $\mathbf{x}(t)$  is as small as possible, i.e. the representation (II.1) is minimal. For simplicity we assume that  $D = 0$ , i.e. there is no direct feedthrough. For future reference we define  $\bar{A} := A - KC$ . We shall denote the “joint” process as  $\mathbf{z} := [\mathbf{y}^\top \mathbf{u}^\top]^\top$  and assume that its spectral density matrix  $\Phi(z)$  is rational and bounded away from zero on the unit circle  $z = e^{j\omega}$ . Let  $\mu_i$  denote the zeros of the spectral density matrix inside the closed unit disc. We define  $\rho := \max(|\mu_i|)$ . From the assumption  $\Phi(e^{j\omega}) > cI > 0$  it follows that  $\rho < 1$ . Note in particular that  $1 > \rho \geq \max(|\lambda_i(\bar{A})|)$  where  $\lambda_i(\bar{A})$  is the  $i$ -th eigenvalue of  $\bar{A}$ .

The white noise process  $\mathbf{e}$ , the innovation of  $\mathbf{y}$  given the joint past of  $\mathbf{y}, \mathbf{u}$ , is defined as the one step ahead (linear) prediction error of  $\mathbf{y}(t)$  given the joint (strict) past of  $\mathbf{u}$  and  $\mathbf{y}$  up to time  $t$ . For future reference we also define the variance of the innovation  $\Lambda := \text{Var}\{\mathbf{e}(t)\}$  and the normalized innovation  $\bar{\mathbf{e}}(t) := \Lambda^{-1/2}\mathbf{e}(t)$ .

Given two sequences of random variables  $\mathbf{x}_N$  and  $\mathbf{g}_N$ , we shall say that  $\mathbf{x}_N = o_P(\mathbf{g}_N)$  if  $\mathbf{x}_N/\mathbf{g}_N$  converges to zero in probability, i.e.  $\forall \delta > 0, \lim_{N \rightarrow \infty} P[|\mathbf{x}_N/\mathbf{g}_N| > \delta] = 0$ .

We shall use the notation  $\underline{o}_P(\cdot)$  to denote a random vector whose components are  $o_P(\cdot)$ .

The symbol  $\stackrel{\cdot}{=}$  shall denote equality in probability up to  $\underline{o}_P(1/\sqrt{N})$  terms, which we shall call *asymptotic equivalence*. In fact, from standard results in asymptotic analysis (see for instance [16]) terms which are  $\underline{o}_P(1/\sqrt{N})$  can be neglected when studying the asymptotic statistical properties.

Our aim is to identify the system parameters  $(A, B, C, K)$ , or equivalently the transfer functions  $F(z) = C(zI - A)^{-1}B$  and  $G(z) = C(sI - A)^{-1}K + I$ , starting from input-output data  $\{y_s, u_s\}, s \in [t_0, T+N]$ , generated by the system (II.1).

The analysis reported in this paper requires that both  $N$ , the length of the finite tails<sup>3</sup> and the past horizon  $t - t_0$ <sup>4</sup> go to infinity. We remind the reader that  $t - t_0$  has to go to infinity at a certain rate depending on the number  $N$  of data available. Details can be found, for instance, in [2] where the following assumption is made:

<sup>3</sup>This is the parameter  $j$  in the notation of Van Overschee and De Moor [30] i.e. the number of columns in the Hankel data matrices used in subspace identification.

<sup>4</sup>The number of block rows in the Hankel data matrix containing the past data.

*Assumption 1:* The past horizon  $t - t_0$  goes to infinity with  $N$  while satisfying:

$$\begin{aligned} t - t_0 &\geq \frac{\log N^{-d/2}}{\log|\rho|} & 1 < d < \infty \\ t - t_0 &= o(\log(N)^\alpha) & \alpha < \infty \end{aligned} \quad (\text{II.2})$$

Under this assumption the effect of terms due to mishandling of the initial condition at time  $t_0$  are  $o(1/\sqrt{N})$  and therefore can be neglected. Moreover, (II.2) ensures that, when regressing onto past data and taking the limit as  $N$  goes to infinity, the computation of sample covariance matrices of increasing size (with  $t - t_0$ ) does not pose any complication in the sense that their limit is well defined and equal to the population counterpart (see the discussion after Lemma 4 in [4]).

We shall use the standard notation of boldface (lowercase) letters to denote random variables. The symbol  $\mathbb{E}\{\cdot\}$  shall denote mathematical expectation; given two (zero mean) random vectors  $\mathbf{a}$  and  $\mathbf{b}$  we shall define  $\Sigma_{\mathbf{ab}} := \mathbb{E}[\mathbf{ab}^\top]$ .

Lowercase letters denote sample values of a certain random variable. For example we shall denote with  $\mathbf{y}(t)$  the random vector denoting the output and with  $y_t$  the sample value of  $\mathbf{y}(t)$ .

We shall use capitals to denote the tail of length  $N$ . For instance  $Y_t := [y_t \ y_{t+1} \ \dots \ y_{t+N-1}]$ , and  $Z_t := [Y_t^\top \ U_t^\top]^\top$ . These are the block rows of the usual *block Hankel data matrices* which appear in subspace identification.

Finite block Hankel data matrices will be denoted using capitals, i.e.  $Y_{[t,s]} := [Y_t^\top \ Y_{t+1}^\top \ \dots \ Y_s^\top]^\top$ .

Spaces generated by finite tails, i.e. spaces generated by the rows of finite block Hankel data matrices, will be denoted with the same symbol used for the matrix itself. Sample covariances will be denoted with the same symbol used for the corresponding random variables with a “hat” on top. For example, given finite sequences  $A_t := [a_t, a_{t+1}, \dots, a_{t+N-1}]$  and  $B_t := [b_t, b_{t+1}, \dots, b_{t+N-1}]$  we shall define the sample covariance matrix

$$\hat{\Sigma}_{\mathbf{ab}} := \frac{1}{N} \sum_{i=0}^{N-1} a_{t+i} b_{t+i}^\top.$$

Under our ergodic assumption  $\lim_{N \rightarrow \infty} \hat{\Sigma}_{\mathbf{ab}} \stackrel{a.s.}{=} \Sigma_{\mathbf{ab}}$ .

The orthogonal projection onto the row space of a matrix shall be denoted with the symbol  $\hat{E}$ ; for instance, given a matrix  $C_t := [c_t, c_{t+1}, \dots, c_{t+N-1}]$ ,  $\hat{E}[C_t]$  will be the orthogonal projection onto the row space of the matrix  $C_t$ ; the symbol  $\hat{E}[A_t|C_t]$  shall denote the orthogonal projection of the rows of the matrix  $A_t$  onto the row space of  $C_t$ , and is given by the formula  $\hat{E}[A_t|C_t] = \hat{\Sigma}_{\mathbf{ac}} \hat{\Sigma}_{\mathbf{cc}}^{-1} C_t$ .

For convenience of notation we denote with  $\nu := T - t$  and define the extended observability matrices

$$\Gamma_\nu^\top := \begin{bmatrix} C^\top & A^\top C^\top & (A^\top)^2 C^\top & \dots & (A^\top)^{\nu-1} C^\top \end{bmatrix}. \quad (\text{II.3})$$

and

$$\bar{\Gamma}_\nu^\top := \begin{bmatrix} C^\top & \bar{A}^\top C^\top & (\bar{A}^\top)^2 C^\top & \dots & (\bar{A}^\top)^{\nu-1} C^\top \end{bmatrix}. \quad (\text{II.4})$$

We shall also need to make the following assumption on the innovation process  $\mathbf{e}(t)$ :

*Assumption 2:* Let  $\mathcal{F}_t^-$  be the  $\sigma$ -algebra generated by the random variables  $\{\mathbf{y}(s), -\infty < s \leq t\}$  and  $\{\mathbf{u}(s), -\infty < s \leq t\}$  (past outputs and inputs). The innovation process  $\mathbf{e}(t)$  is an  $\mathcal{F}_{t-1}^-$ -martingale difference sequence with constant conditional variance, i.e.

$$\begin{aligned} \mathbb{E}[\mathbf{e}(t) | \mathcal{F}_{t-1}^-] &= 0 \\ \mathbb{E}[\mathbf{e}(t)\mathbf{e}^\top(t) | \mathcal{F}_{t-1}^-] &= \Lambda. \end{aligned} \quad (\text{II.5})$$

### III. ALGORITHMS

As discussed in the introduction it has been shown recently that a whole class of subspace algorithms provide estimators which are asymptotically equivalent. We take as a representative in this class the SSARX algorithm in [19].

For completeness we also consider the algorithm proposed in [22]; this can be seen as a CCA algorithm performed on the data once the ‘‘future’’ has been removed, therefore we shall call it FC-CCA (short for ‘‘future corrected’’ CCA). As an intermediate result we shall argue that also FC-CCA is asymptotically equivalent to SSARX; therefore we can stick to FC-CCA for the purpose of analysis. This intermediate result, besides complementing the results already available, turns out to be particularly useful in the analysis as we shall see in Section IV.

It is well known that subspace identification can be seen as a 2-step procedure; in a first step the state is estimated and in the second the system matrices are computed from the estimated state. Since this second step is common to the algorithms we consider, we only discuss this first step.

#### A. SSARX algorithm

The first step of the SSARX algorithm by Jansson is to estimate one (long) VARX model

$$\begin{aligned} Y_T &\simeq \hat{\Phi}_1^y Y_{T-1} + \hat{\Phi}_2^y Y_{T-2} + \cdots + \hat{\Phi}_{T-t_0}^y Y_{t_0} + \\ &+ \hat{\Phi}_1^u U_{T-1} + \hat{\Phi}_2^u U_{T-2} + \cdots + \hat{\Phi}_{T-t_0}^u U_{t_0} \end{aligned} \quad (\text{III.1})$$

where without loss of generality we have taken the length of the VARX model equal to  $T-t_0$ ; then the effect of the future inputs/outputs is removed using the estimated parameters  $\hat{\Phi}_k^u$ ,  $\hat{\Phi}_k^y$  as<sup>5</sup>:

$$\hat{Y}_{[t,T]}^J := \hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^u U_{[t,T]} - \hat{H}_\nu^y Y_{[t,T]} \mid Z_{[t_0,t]} \right] \quad (\text{III.2})$$

where

$$\hat{H}_\nu^{u,y} := \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \hat{\Phi}_1^{u,y} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \hat{\Phi}_\nu^{u,y} & \cdots & \hat{\Phi}_1^{u,y} & 0 \end{bmatrix}$$

<sup>5</sup>We shall use a superscript  $J$  (or subscript  $J$ ) to denote quantities related to SSARX, as a short for ‘‘Jansson’’; similarly  $L$  and  $L$  shall be attached to the FC-CCA algorithm as a short for ‘‘Larimore’’

Then the state is estimated via SVD decomposition of  $W_J^{-1}(\nu)\hat{Y}_{[t,T]}^J$  where

$$\hat{W}_J(\nu) = \left( \frac{(Y_{[t,T]} - \hat{H}_\nu Z_{[t,T]})(Y_{[t,T]} - \hat{H}_\nu Z_{[t,T]})^\top}{N} \right)^{1/2}. \quad (\text{III.3})$$

This is equivalent (as far as the state construction is concerned) to performing CCA between the corrected future  $Y_{[t,T]} - \hat{H}_\nu Z_{[t,T]}$  and the past  $Z_{[t_0,t]}$ .

Let  $W_J(\nu) := \lim_{N \rightarrow \infty} \hat{W}_J(\nu)$ . It is easy to see that  $W_J(\nu)W_J^\top(\nu) = \Gamma_\nu \Sigma_{\mathbf{xx}} \Gamma_\nu^\top + (I \otimes \Lambda)$ . Using the same argument as in Lemma 3.3 of [4] (see also [7]) it can be proven that  $W_J(\nu) = (I \otimes \Lambda)^{1/2}$  gives the same asymptotic distribution of the estimators.

Therefore, for the purpose of analysis, we can assume SSARX performs SVD of

$$(I \otimes \Lambda)^{-1/2} \hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^u U_{[t,T]} - \hat{H}_\nu^y Y_{[t,T]} \mid Z_{[t_0,t]} \right] \quad (\text{III.4})$$

when constructing the state space.

#### B. FC-CCA algorithm

Also in this algorithm first a long ARX model, as in (III.1), is estimated. Then, using the ARX coefficients  $\hat{\Phi}_i$ , estimates  $\hat{\Phi}_i^d$  of the Markov parameters<sup>6</sup>  $\Psi_i^d = CA^{i-1}B$ ,  $i = 1, \dots, \nu$  are computed.

Defining

$$\hat{H}_\nu^d := \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \hat{\Psi}_1^d & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \hat{\Psi}_\nu^d & \cdots & \hat{\Psi}_1^d & 0 \end{bmatrix},$$

the FC-CCA algorithm performs CCA between the corrected future

$$Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]}$$

and the past  $Z_{[t_0,t]}$ . This is equivalent, to the purpose of state construction, to performing SVD of  $\hat{W}_L^{-1}(\nu)\hat{Y}_{[t,T]}^L$  where

$$\hat{Y}_{[t,T]}^L := \hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]} \mid Z_{[t_0,t]} \right] \quad (\text{III.5})$$

and

$$\hat{W}_L(\nu) = \left( \frac{(Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]})(Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]})^\top}{N} \right)^{1/2}. \quad (\text{III.6})$$

See, in particular, equation (9) in [22] and the discussion following the equation.

The asymptotic value  $W_L(\nu)$  of the weight  $\hat{W}_L(\nu)$  satisfies

$$W_L(\nu)W_L^\top(\nu) = \Gamma_\nu \Sigma_{\mathbf{xx}} \Gamma_\nu^\top + H_\nu^s (I \otimes \Lambda) (H_\nu^s)^\top$$

<sup>6</sup>The superscript  $d$  stands for ‘‘deterministic’’.

where

$$H_\nu^s := \begin{bmatrix} I & 0 & \dots & 0 \\ CK & I & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ CA^{\nu-2}K & \dots & CK & I \end{bmatrix},$$

As above,  $W_L(\nu)$  can be substituted with

$$W_L(\nu) := H_\nu^s(I \otimes \Lambda)^{1/2} \quad (\text{III.7})$$

without altering the asymptotic properties (see [4], [7]). From now on the symbol  $W_L(\nu)$  will be used according to definition (III.7). Therefore, for the purpose of analysis, we can assume FC-CCA performs SVD of

$$(I \otimes \Lambda)^{-1/2}(H_\nu^s)^{-1}\hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]} \mid Z_{[t_0,t]} \right] \quad (\text{III.8})$$

when constructing the state space.

### C. Asymptotic Equivalence of SSARX and FC-CCA

The main result of this section is stated as a proposition:

*Proposition 3.1:* The matrices in (III.4) and (III.8) differ only up to a left multiplication for a non-singular matrix which tends to the identity matrix as  $N \rightarrow \infty$  and therefore the two constructions yield to asymptotically equivalent procedures.

*Proof:* Note that, using Lemma 7.1,  $\hat{Y}_{[t,T]}^J$  equation (III.2) can be rewritten as:

$$\begin{aligned} \hat{Y}_{[t,T]}^J &= \hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^u U_{[t,T]} - \hat{H}_\nu^y Y_{[t,T]} \mid Z_{[t_0,t]} \right] \\ &= \hat{E} \left[ (I - \hat{H}_\nu^y) Y_{[t,T]} - \hat{H}_\nu^u U_{[t,T]} \mid Z_{[t_0,t]} \right] \\ &= \hat{E} \left[ (\hat{H}_\nu^s)^{-1} Y_{[t,T]} - \hat{H}_\nu^u U_{[t,T]} \mid Z_{[t_0,t]} \right] \\ &= (\hat{H}_\nu^s)^{-1} \hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^s \hat{H}_\nu^u U_{[t,T]} \mid Z_{[t_0,t]} \right] \\ &= (\hat{H}_\nu^s)^{-1} \hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]} \mid Z_{[t_0,t]} \right] \end{aligned}$$

so that (III.4) can be rewritten as

$$(I \otimes \Lambda)^{-1/2}(\hat{H}_\nu^s)^{-1} \hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]} \mid Z_{[t_0,t]} \right] \quad (\text{III.9})$$

It is immediate to recognize that this is exactly the same as in the FC-CCA algorithm (see equation (III.8)), once the asymptotic value of the weight  $H_\nu^s$  is substituted to its estimator; it is well known (see [2]) that this replacement does not change the asymptotic distribution. ■

For this reason, when discussing the role of  $\nu$ , we shall stick to the FC-CCA algorithm described in [22] from now on.

### D. State construction

The state is usually constructed from singular value decomposition<sup>7</sup> (SVD) of (III.8)

$$\hat{U}(\nu)\hat{S}(\nu)\hat{V}^\top(\nu) = W_L^{-1}(\nu) \frac{\hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]} \mid Z_{[t_0,t]} \right]}{N}. \quad (\text{III.10})$$

<sup>7</sup>In this paper we shall make explicit the dependence on  $\nu$  of the weights and the matrices involved in the SVD.

The estimator of the observability matrix is given by  $\hat{\Gamma}_\nu := W_L(\nu)\hat{U}_n(\nu)\hat{S}_n(\nu)$ , where the subscript  $n$  reminds that only the  $n - th$  largest singular values are retained. An estimate of the state is given by

$$\hat{X}_t := \hat{S}_n^{-1}(\nu)\hat{U}_n^\top(\nu)W_L^{-1}(\nu)\hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]} \mid Z_{[t_0,t]} \right] \quad (\text{III.11})$$

It is known that, as far as asymptotic distribution is concerned, only the asymptotic values  $U_n(\nu)$  and  $S_n(\nu)$  matter so that only

$$\hat{X}_t := S_n^{-1}(\nu)U_n^\top(\nu)W_L^{-1}(\nu)\hat{E} \left[ Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]} \mid Z_{[t_0,t]} \right] \quad (\text{III.12})$$

will be considered. Note that

$$W_L(\nu)U_n(\nu)S_n(\nu) = \Gamma_\nu, \quad (\text{III.13})$$

where  $\Gamma_\nu$  is the “true” observability matrix in a suitable basis such that  $\Sigma_{\mathbf{x}\mathbf{x}} = \lim_{N \rightarrow \infty} \frac{\hat{X}_t \hat{X}_t^\top}{N} = I$ . This normalization turns out to be convenient for the purpose of this paper and is, one might say, different from the more common choice  $\frac{\hat{X}_t \hat{X}_t^\top}{N} = S_n(\nu)$ . Of course this is just a change of basis and hence does not affect the estimate of any system invariant.

## IV. ROLE OF THE FUTURE HORIZON $\nu$

In this section we shall analyze the role of  $\nu$  in the state construction step. Unfortunately it seems not possible, at the present stage, to formalize a sharp result; it is our opinion, however, that the considerations contained in this section might help developing some intuition.

As mentioned earlier and formally proved in [7][Proposition 3.1] it make sense to compare the state construction without entering into the details of the subsequent steps.

Defining  $\tilde{H}_\nu^d := \hat{H}_\nu^d - H_\nu^d$  and recalling<sup>8</sup> that  $\hat{E} \left[ X_t \mid Z_{[t_0,t]} \right] = X_t$ ,  $\hat{Y}_{[t,T]}^L$  can be written as

$$\hat{Y}_{[t,T]}^L = \Gamma_\nu X_t + \hat{E} \left[ H_\nu^s E_{[t,T]} - \tilde{H}_\nu^d U_{[t,T]} \mid Z_{[t_0,t]} \right].$$

Now it is convenient to rewrite (III.12) as

$$\hat{X}_t = X_t + S_n^{-1}(\nu)\hat{E} \left[ U_n^\top(\nu)\bar{E}_{[t,T]} \mid Z_{[t_0,t]} \right] - S_n^{-1}(\nu)U_n^\top(\nu)W_L^{-1}(\nu)\tilde{H}_\nu^d \hat{E} \left[ U_{[t,T]} \mid Z_{[t_0,t]} \right] \quad (\text{IV.1})$$

Recall now that the  $i$ -th column of  $\bar{E}_{[t,T]}$  is a realization of the normalized future innovations  $\bar{\mathbf{e}}_{[t+i-1, T+i-1]}$ ; using the fact the  $U_n(\nu)$  have orthonormal columns, the  $i$ -th column of  $U_n^\top(\nu)\bar{E}_{[t,T]}$  is a realization of a random vector  $\mathbf{v}(t+i, \nu) := U_n^\top(\nu)\bar{\mathbf{e}}_{[t+i-1, T+i-1]}$  with unit covariance matrix (i.e.  $\mathbb{E} \left[ \mathbf{v}(t+i, \nu)\mathbf{v}^\top(t+i, \nu) \right] = I$ ). Let us define  $V_t(\nu) := [v_t(\nu), v_{t+1}(\nu), \dots, v_{t+N-1}(\nu)]$  where  $v_t(\nu)$  is the sample value of  $\mathbf{v}(t, \nu)$ .

Then the state estimation error  $\tilde{X}_t := \hat{X}_t - X_t$  is given by:

$$\tilde{X}_t = S_n^{-1}(\nu)\hat{E} \left[ V_t(\nu) \mid Z_{[t_0,t]} \right] + S_n^{-1}(\nu)U_n^\top(\nu)W_L^{-1}(\nu)\tilde{H}_\nu^d \hat{E} \left[ U_{[t,T]} \mid Z_{[t_0,t]} \right] \quad (\text{IV.2})$$

<sup>8</sup>The asymptotic equivalence stems from the fact that the initial condition at  $t_0$  has an effect which is  $o(1/\sqrt{N})$  thanks to Assumption 1.

We now analyze separately the two contributions to the state estimation error  $\tilde{X}_t$ . Let us define  $\tilde{X}_t^e := S_n^{-1}(\nu)\hat{E}[V_t(\nu) | Z_{[t_0,t]}]$  and  $\tilde{X}_t^u := -S_n^{-1}(\nu)U_n^\top(\nu)W_L^{-1}(\nu)\tilde{H}_\nu^d\hat{E}[U_{[t,T]} | Z_{[t_0,t]}]$ , so that (IV.2) can be written as

$$\tilde{X}_t = \tilde{X}_t^e + \tilde{X}_t^u.$$

- a) The term  $\tilde{X}_t^e$  is a consequence of the fact that the rows of  $\tilde{E}_{[t,T]}$  are only asymptotically orthogonal to the rows of  $Z_{[t_0,t]}$ . Note that, under Assumption 2 the process  $\mathbf{v}(t, \nu) := U_n^\top(\nu)\tilde{\mathbf{e}}_{[t,T]}$  satisfies

$$\begin{aligned} \mathbb{E}[v(t, \nu) | \mathcal{Z}_{t-1}^-] &= 0 \\ \text{Var}\{v(t, \nu) | \mathcal{Z}_{t-1}^-\} &= I \end{aligned} \quad (\text{IV.3})$$

Therefore the asymptotic variance<sup>9</sup> of  $\text{vec}\{V_t(\nu)Z_{[t_0,t]}^\top/N\}$  (and hence of  $\hat{E}[V_t(\nu) | Z_{[t_0,t]}]$ ) does not depend on  $\nu$ .

Instead, the diagonal elements of  $S_n(\nu)$  are monotonically non-decreasing functions of  $\nu$ . They converge to an upper limit and, if the spectrum is rational, this convergence is exponential.

Therefore we can conclude that  $\tilde{X}_t^e$  is monotonically decreasing (in the mean square sense) as a function of  $\nu$ .

- b) The term  $\tilde{X}_t^u$  comes from the errors in estimating the Markov parameters in  $\tilde{H}_\nu^u$ . The perturbation  $\tilde{X}_t^u$  is a linear combination of elements of  $\hat{E}[U_{[t,T]} | Z_{[t_0,t]}]$ . This projection is indeed contained in the predictor space of the joint process  $\mathbf{z}(t)$ :

$$\hat{X}_t^z := \hat{E}[Z_{[t,T]} | Z_{[t_0,t]}].$$

Note that, if the inputs are white (and hence  $\mathbf{u}(\tau)$ ,  $\tau \geq t$  is uncorrelated with the past measurements  $\mathbf{z}(s)$ ,  $s < t$ ), then also  $\hat{E}[U_{[t,T]} | Z_{[t_0,t]}]$  vanishes asymptotically. Hence the product  $\tilde{H}_\nu^d\hat{E}[U_{[t,T]} | Z_{[t_0,t]}]$  has a contribution which is  $o_P(1/\sqrt{N})$  and can be neglected when studying the asymptotic distribution. This fact has been indeed used in [4] where it has been proved that  $\nu$  should increase with  $N$  to achieve optimal accuracy.

#### A. White inputs case

Under this assumption, as discusses in item b) above, the effect of  $\tilde{X}_t^u$  is  $o_P(1/\sqrt{N})$  and therefore can be neglected. Using the fact that:

- the estimated state is normalized so that  $\Sigma_{\mathbf{xx}} = I$  independently on  $\nu$ ;
- the asymptotic variance of  $\tilde{X}_t^e$  is monotonically decreasing with  $\nu$ ;

we essentially re-obtain the well known result of [4] stating that, when the inputs are white, the variance of the estimators is monotonically decreasing as a function of  $\nu$ .

**Remark IV.1** A very intuitive explanation of this fact is that, with an appropriate weighting (indeed the CCA

<sup>9</sup>In a suitable sense as defined in [23] for vectors of increasing size with the number of data.

weight), adding more measurements (i.e. increasing the future horizon) has an ‘‘averaging effect’’ on the noise term  $\hat{E}[\tilde{E}_{[t,T]} | Z_{[t_0,t]}]$  due to projection residuals.  $\diamond$

#### B. Non-white inputs case

In this general situation, which includes the possibility of having feedback (i.e. closed loop identification) the term  $\tilde{X}_t^u$  can no longer be neglected. Its effect essentially depends on the predictor space  $\hat{E}[U_{[t,T]} | Z_{[t_0,t]}]$ .

It seems hard, unfortunately, to give a general recipe; however the optimal choice of  $\nu$  seems to be a trade-off between minimizing the effect of  $\hat{E}[E_{[t,T]} | Z_{[t_0,t]}]$  (increasing  $\nu$ , averaging effect) and minimizing  $\tilde{X}_t^u$ .

It is expected that, the more correlated the input process, the larger the contribution of  $\tilde{X}_t^u$  for increasing  $\nu$ . This is confirmed by the results reported in the Section V which show that, indeed, to wider variations of the input spectrum (we keep fixed the pole while moving the zero location) there corresponds a degradation of performance as  $\nu$  increases.

This fact is related to the observations made (for ‘‘open-loop’’ algorithms) in [13], [12] concerning the maximum eigenvalue<sup>10</sup> of  $\Sigma_{\mathbf{xx}|\mathbf{u}^+}$ . This is in fact a non-decreasing function of the number of future inputs (i.e. length of the future horizon) used in forming  $\mathbf{u}^+ := [\mathbf{u}^\top(t), \dots, \mathbf{u}^\top(t + \nu - 1)]^\top$ .

#### C. Identification of ARX systems

In the particular case of identification of ARX systems (i.e. systems such that  $\bar{A}^k = 0$  for some  $k \leq n$ ), it is possible to conclude that, increasing  $\nu \geq k$  does not affect the performance of the algorithm. This we state in the following proposition, the proof can be found in the Appendix.

Let us call *index of nilpotency*  $k$  the smallest integer such that  $\bar{A}^k = 0$ .

*Proposition 4.1:* Let us assume that the system is of the ARX type with index of nilpotency  $k$ . Then the state estimator (III.12) (and hence any estimator based on the state sequence) is not affected by the choice of  $\nu$  provided  $\nu \geq k$ .

We agree that this result is not surprising at all but, to the best of the author’s knowledge, it has not been proved before. This result was stated in a weaker form in [4] (see the discussion after Corollary 2) where only white inputs were allowed for.

With a similar argument it can be verified that, in general, the variance depends on  $\nu$  only through  $\bar{A}^\nu$  and hence it converges (to its asymptotic value) exponentially, the rate being determined by the noise zeros.

This is confirmed by the experimental results which show that, indeed, the efficiency index  $Eff(\nu)$  in (V.1) does not change when  $\nu$  is increased over a certain threshold. In the examples reported in the next section this threshold is roughly  $\nu \simeq 10$ . Note that, for this example  $\bar{A} = 0.5$ , so that  $\bar{A}^{10} \simeq, 10^{-3}$ , which is of the same order of magnitude of the change observed in  $Eff(\nu)$  for  $\nu > 10$ .

<sup>10</sup>In a normalized basis s.t.  $\Sigma_{\mathbf{xx}} = I$  for instance.

## V. EXPERIMENTAL RESULTS

In order to illustrate the results of this paper we consider the first order ARMAX model

$$\mathbf{y}(t) - 0.5\mathbf{y}(t-1) = \mathbf{u}(t-1) + \mathbf{e}(t) + 0.5\mathbf{e}(t-1).$$

The input  $\mathbf{u}(t)$  is generated using the difference equation

$$\mathbf{u}(t) - 0.9\mathbf{u}(t) = \mathbf{n}(t) - \alpha\mathbf{n}(t-1)$$

where  $\mathbf{n}(t)$  is unit variance white noise uncorrelated from  $\mathbf{e}(t)$  (i.e. open loop operation is assumed<sup>11</sup>.)

The input spectrum is controlled through the zero location  $\alpha$ ; three possible values of  $\alpha$  are considered:  $\alpha_1 = 0.5$  (slightly correlated input),  $\alpha_2 = -0.2$  (moderately correlated input) and  $\alpha_3 = -0.9$  (highly correlated input).

Let us denote with  $\hat{F}_\nu(e^{j\omega})$  the estimator of the deterministic transfer function  $F(e^{j\omega}) := \frac{1}{e^{j\omega}-0.5}$  using the FC-CCA algorithms as a function of the future horizon  $\nu$ . Denote also with  $CRLB_{\hat{F}}(j\omega)$  the Cramér-Rao lower bound for any unbiased estimator of  $F(e^{j\omega})$  as a function of the normalized frequency  $\omega \in [0, 2\pi]$ .

In Figures 1, 2 and 3 we show, respectively for the three values of  $\alpha$ , the input spectrum and the behavior of the efficiency index

$$Eff(\nu) := \frac{\int_0^{2\pi} \text{AsVar}\{\hat{F}_\nu(e^{j\omega})\} d\omega}{\int_0^{2\pi} CRLB_{\hat{F}}(j\omega) d\omega} \quad (\text{V.1})$$

as a function of  $\nu$ .

The asymptotic variance is computed using the formulas in [9] and assuming  $t - t_0 = 30$  (which makes the effect of transients due to mishandling of the initial condition of the order of  $10^{-9}$ ).

As summarized in table I, it should be observed that the best performance (in terms of the index  $Eff(\nu)$ ) does not degrade as the input spectrum varies more widely. Indeed, denoting with

$$\nu_{opt} := \arg \min_{\nu} Eff(\nu)$$

we have that  $Eff(\nu_{opt}) = Eff(1)$  for  $\alpha_3 = -0.9$  is roughly 1.0380 while for  $\alpha_2 = -0.2$ ,  $Eff(\nu_{opt}) = Eff(2) \simeq 1.0720$ . Note also that for  $\alpha_1 = 0.5$ ,  $Eff(\nu)$  decreases monotonically reaching the asymptotic value  $Eff(\nu_{opt}) = Eff(\infty) \simeq 1.0005$ .

As a side, note that for  $\alpha = 0.9$  (i.e. with white input, in which case  $Eff(\nu)$  decreases monotonically in  $\nu$ ) the computed value of  $Eff(\infty) \simeq 1 + 10^{-9}$ ; In this case the algorithm is expected to be efficient and the  $10^{-9}$  error should be attributed to the numerical computation of the integral in (V.1) and to the fact the asymptotic variance formulas are computed assuming  $t - t_0 = 30$  which, as mentioned after formula (V.1), yields errors of the order of  $10^{-9}$  with respect to  $t - t_0 = \infty$ .

Note, as a side, that the FC-CCA is not efficient in all the examples considered for any choice of  $\nu$ .

<sup>11</sup>Note that the algorithms described in this paper yield consistent estimators also for closed-loop operating conditions [14].

$\alpha$	$\nu_{opt}$	$Eff(\nu_{opt})$
0.9	$\infty$	1
0.5	$\infty$	1.0005
-0.2	2	1.0720
-0.9	1	1.0380

TABLE I

OPTIMAL FUTURE HORIZON  $\nu_{opt}$  AND RELATIVE EFFICIENCY

## VI. CONCLUSIONS

In this paper it has been shown that the method presented in [22] (called here FC-CCA) and the SSARX method in [19] are asymptotically equivalent. Using this fact and the results in [6], [10], [7], [5] it has been possible to study the effect of the “future horizon” in whole class of algorithms, which we call of CCA-type (as explained in Section I). Even though it has not been possible to give a recipe for choosing this parameter optimally, we believe our results help developing some intuition, hopefully leading to more considerable progresses in the future.

## REFERENCES

- [1] H. Akaike. Markovian representation of stochastic processes by canonical variables. *SIAM J. Control*, 13:162–173, 1975.
- [2] D. Bauer. Asymptotic properties of subspace estimators. *Automatica*, 41:359–376, 2005.
- [3] D. Bauer. Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs. *Journal of Time Series Analysis*, 26:631–668, 2005.
- [4] D. Bauer and L. Ljung. Some facts about the choice of the weighting matrices in Larimore type of subspace algorithm. *Automatica*, 38:763–773, 2002.
- [5] A. Chiuso. On the role of Vector AutoRegressive modeling in subspace identification. Submitted to *Automatica*, available at <http://www.dei.unipd.it/~chiuso>.
- [6] A. Chiuso. On the relation between CCA and predictor-based subspace identification. In *Proc. of the 44rd IEEE Conf. on Dec. and Control*, Sevilla, Spain, 2005.
- [7] A. Chiuso. On the relation between CCA and predictor based subspace identification. Submitted to *IEEE Trans. on Aut. Control*, 2005. available at <http://www.dei.unipd.it/~chiuso>.
- [8] A. Chiuso. Asymptotic equivalence of certain closed-loop subspace identification methods. In *Proc. of SYSID 2006*, Newcastle, Australia, 2006.
- [9] A. Chiuso. Asymptotic variance of closed-loop subspace identification algorithms. *IEEE Trans. on Aut. Control*, 51(8):1299–1314, 2006.
- [10] A. Chiuso. On the role of Vector AutoRegressive modeling in subspace identification. In *Proc. of CDC 2006 (to appear)*, San Diego (USA), Dec. 2006.
- [11] A. Chiuso and G. Picci. Constructing the state of random processes with feedback. In *Proc. of the IFAC Int. Symposium on System Identification (SYSID)*, Rotterdam, August 2003.
- [12] A. Chiuso and G. Picci. Numerical conditioning and asymptotic variance of subspace estimates. *Automatica*, 40(4):677–683, 2004.
- [13] A. Chiuso and G. Picci. On the ill-conditioning of subspace identification with inputs. *Automatica*, 40(4):575–589, 2004.
- [14] A. Chiuso and G. Picci. Consistency analysis of some closed-loop subspace identification methods. *Automatica*, 41(3):377–391, 2005.
- [15] A. Chiuso and G. Picci. Prediction error vs. subspace methods in closed-loop identification. In *Proc. of the 16th IFAC World Congress*, Prague, July 2005.
- [16] T. Ferguson. *A Course in Large Sample Theory*. Chapman and Hall, 1996.
- [17] C.W.J. Granger. Economic processes involving feedback. *Information and Control*, 6:28–48, 1963.
- [18] H. Hotelling. Relations between two set of variables. *Biometrika*, 28:321–377, 1936.

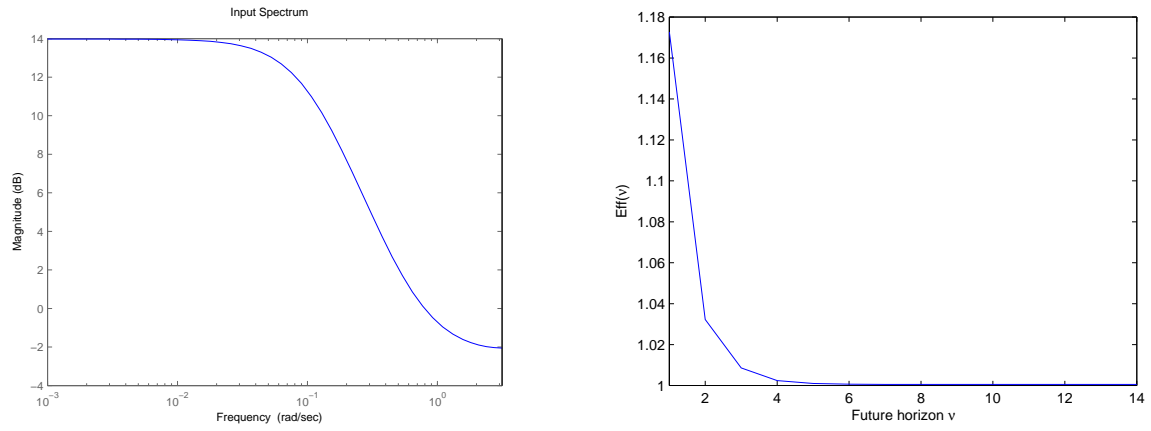


Fig. 1.  $\alpha = 0.5$ . Left: Input spectrum, right:  $Eff(v)$ .

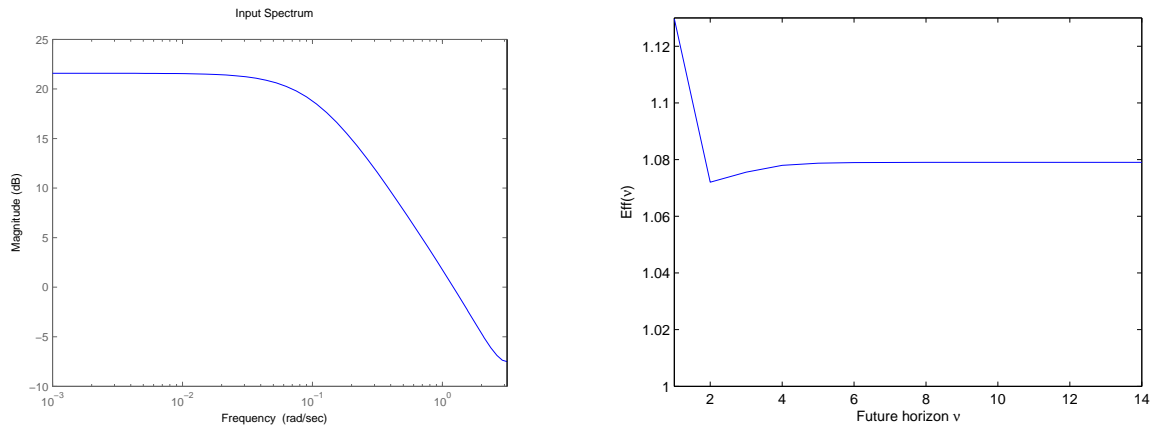


Fig. 2.  $\alpha = -0.2$ . Left: Input spectrum, right:  $Eff(v)$ .

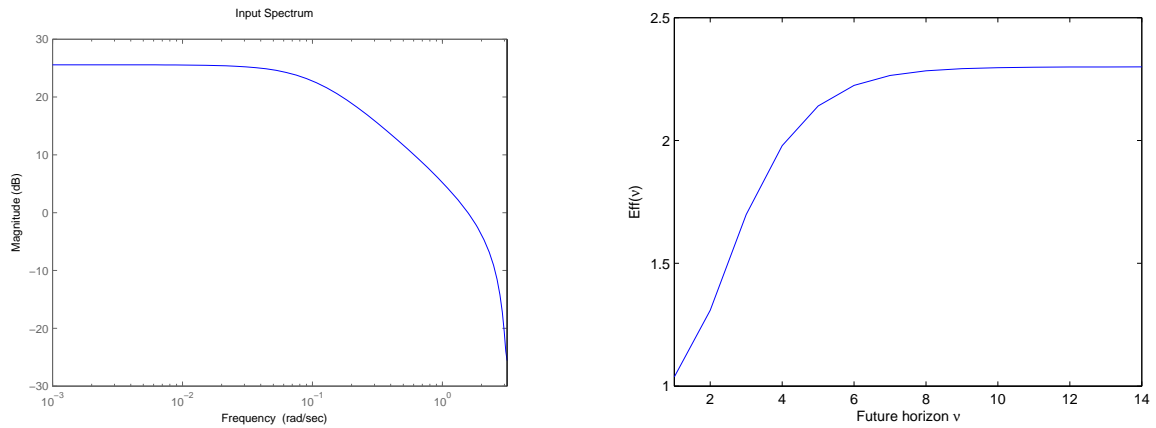


Fig. 3.  $\alpha = -0.9$ . Left: Input spectrum, right:  $Eff(v)$ .

[19] M. Jansson. Subspace identification and ARX modeling. In *Proceedings of SYSID 2003*, Rotterdam, 2003.

[20] W.E. Larimore. System identification, reduced-order filtering and modeling via canonical variate analysis. In *Proc. American Control Conference*, pages 445–451, 1983.

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

[22] W.E. Larimore. Large sample efficiency for ADAPTX subspace identification with unknown feedback. In *Proc. of IFAC DYCOPS'04*,

- Boston, MA, USA, 2004.
- [23] R. Lewis and G.C. Reinsel. Prediction of multivariate time series by autoregressive model fitting. *J. of Multivariate Analysis*, 16:393–411, 1985.
  - [24] L. Ljung and T. McKelvey. Subspace identification from closed loop data. *Signal Processing*, 52(2):209–216, 1996.
  - [25] K. Onodera, G. Emoto, and S.J. Qin. A new subspace identification method for closed loop systems. In *Proceedings of SYSID 2006*, Newcastle, Australia, 2006.
  - [26] S.J. Qin and L. Ljung. Closed-loop subspace identification with innovation estimation. In *Proceedings of SYSID 2003*, Rotterdam, 2003.
  - [27] S.J. Qin and L. Ljung. On the role of future horizon in closed-loop subspace identification. In *Proceedings of SYSID 2006*, Newcastle, Australia, 2006.
  - [28] F. Shi and J.F. MacGregor. A framework for subspace identification. In *Proc. of IEEE ACC*, Arlington, VA, 2001.
  - [29] P. Van Overschee and B. De Moor. Subspace algorithms for the stochastic identification problem. *Automatica*, 29:649–660, 1993.
  - [30] P. Van Overschee and B. De Moor. N4SID: Subspace algorithms for the identification of combined deterministic– stochastic systems. *Automatica*, 30:75–93, 1994.
  - [31] P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems*. Kluwer Academic Publications, 1996.
  - [32] M. Verhaegen. Identification of the deterministic part of MIMO state space models given in innovations form from input-output data. *Automatica*, 30:61–74, 1994.

## VII. APPENDIX

*Lemma 7.1:* Let us denote with  $\hat{H}_\nu^s$  the estimate of the matrix  $H_\nu^s$  obtained from the ARX estimators  $\hat{\Psi}_k^y, \hat{\Psi}_k^u$ . Then

$$\hat{H}_\nu^s = (I - \hat{H}_\nu^y)^{-1}$$

and

$$\hat{H}_\nu^d = \hat{H}_\nu^s \hat{H}_\nu^u$$

*Proof:* The proof is a simple matrix manipulation with the state space parameters of the estimated ARX model and is omitted for reasons of space ■

*Proof of Proposition 4.1.* First of all note under the assumption  $\bar{A}^k = 0$  the observability matrix  $\bar{\Gamma}_\nu$  has the structure

$$\bar{\Gamma}_\nu^\top = [C^\top \bar{A}^\top C^\top \dots (\bar{A}^\top)^{k-1} 0 \dots 0]$$

It is also easy to verify that  $(H_\nu^s)^{-1} \Gamma_\nu = \bar{\Gamma}_\nu$ . Therefore, from (III.13)  $(H_\nu^s)^{-1} \Gamma_\nu = (I \otimes \Lambda)^{1/2} U_n(\nu) S_n(\nu) = \bar{\Gamma}_\nu$  and therefore also  $(I \otimes \Lambda)^{1/2} U_n(\nu) S_n(\nu)$  has the structure

$$((I \otimes \Lambda)^{1/2} U_n(\nu) S_n(\nu))^\top = \underbrace{[* \dots *]}_k \underbrace{[0 \dots 0]}_{\nu - k}$$

and the elements denoted with stars do not change as a function of  $\nu \geq k$ . From the block diagonal structure of  $(I \otimes \Lambda)$  also  $S_n(\nu) U_n^\top(\nu) (I \otimes \Lambda)^{-1/2}$  is of the form

$$S_n(\nu) U_n^\top(\nu) (I \otimes \Lambda)^{-1/2} = \underbrace{[* \dots *]}_k \underbrace{[0 \dots 0]}_{\nu - k}$$

with non-zero elements independent of  $\nu \geq k$ . From the lower triangular structure of  $(H_\nu^s)^{-1}$  also  $S_n(\nu) U_n^\top(\nu) (I \otimes \Lambda)^{-1/2} (H_\nu^s)^{-1} = S_n(\nu) U_n^\top(\nu) W_L^{-1}(\nu)$  satisfies

$$S_n(\nu) U_n^\top(\nu) W_L^{-1}(\nu) = \underbrace{[* \dots *]}_k \underbrace{[0 \dots 0]}_{\nu - k}$$

where again the non-zero elements do not change as a function of  $\nu \geq k$ . In particular also the diagonal elements

of  $S_n(\nu)$  do not increase for  $\nu$  larger than  $k$ , i.e.  $S_n(\nu) = S_n(k)$ ,  $\nu \geq k$ .

Using these considerations it follows that:

- a) The variance of  $\tilde{X}_t^e$  does not depend on  $\nu \geq k$ .
- b) Using the lower triangular structure of  $\tilde{H}_\nu^d$  it follows that  $\tilde{X}_t^u$  is invariant for  $\nu \geq k$ .

Therefore the estimation error  $\tilde{X}_t$  does not change (in distribution) as a function of  $\nu \geq k$ , proving the proposition. □