On the Asymptotic Properties of Closed-Loop CCA-Type Subspace Algorithms: Equivalence Results and Role of the Future Horizon

Alessandro Chiuso, *Senior Member, IEEE*

***Abstract*—In this paper, we shall consider a class of subspace algorithms for identification of linear time invariant systems operating in "closed loop."**

In particular we study algorithms based on the so-called "state-sequence" approach; we first show that the ADAPTx algorithm by Larimore is asymptotically equivalent to a number of recently developed algorithms, which we call CCA-type algorithms.

Based on this equivalence result, we then study the effect of the "future horizon," which is one of the principal "user choices" in subspace identification. It is well known that for the CCA algorithm the asymptotic variance of any system invariant is a non increasing function of the future horizon when input signals are white (or absent). In particular we extend this result, valid for white noise input signals to a slightly more general class of input signals, which include proportional (output or state) feedback controllers and LQG controllers, provided the reference input is white. The condition on the input will be expressed in terms of its state space, which we regard as a rather natural condition in this framework.

For the situations not covered by the above result, we shall also describe a computational procedure, based on some recently derived asymptotic variance formulas, which allows to optimize the choice of the future horizon. Some simulation results are included.

***Index Terms*—Closed loop identification, statistical analysis, subspace methods.**

I. Introduction

SUBSPACE algorithms for identification of Multi-Input Multi-Output (MIMO) linear time invariant systems have seen a considerable development in the last two decades (see [4], [6], [11], [15], [29], [44], [45]).

For the reader not familiar with these methods, suffices here to say that subspace methods can be seen as the translation of stochastic realization ideas [14], [33] into algorithms working on data. These algorithms *do not* require parameterizing the model class, are *non-iterative* and *numerically robust* and hence provide a good alternative to standard procedures based on iterative optimization, such as Maximum Likelihood (ML) and Prediction Error Methods (PEM), when it comes to identifying MIMO systems.

Very recently new methods [13], [16], [26], [35], [39], [41] have allowed subspace identification to be applied also in closed loop, making their range of applicability even wider. Of course when many algorithms (or estimators in statistical terms) are available, one has to fix a way to measure their relative performance. It is a standard practice to compare estimators based on their asymptotic distribution, and in particular asymptotic variance, which is the variance of the asymptotic distribution; see later on for a precise definition. Note that, when dealing with state space models, the matrices A, B, C, K are defined up to a similarity transformation. Therefore, when talking about "asymptotic variance of estimated parameters", we either refer to *fixed* basis or consider system invariants such as the transfer function, pole locations etc. In particular, when we say that an algorithm (say Algorithm A) provides estimators whose asymptotic variance is lower than the variance of the estimators using another algorithm (say Algorithm B), we really mean that the asymptotic variance of *any* system invariant estimated using Algorithm A is lower than the asymptotic variance of the same invariant estimated using Algorithm B. Similarly, we say that an algorithm is optimized within a class if it provides the lowest achievable asymptotic variance of any system invariant within that class. We say that an algorithm is asymptotically efficient if the asymptotic variance reaches the Cramér-Rao lower bound assuming Gaussian distribution.

By measuring the relative performance in this way, we can say that most subspace algorithms are very much related to each other (and to Quasi-Maximum Likelihood); the main comparative results available to date can be enumerated as follows:

- The CCA algorithm,[1] which was inspired by pioneering work of Akaike [1], [2], developed in [28] and further refined in [43], is efficient for time series identification (no inputs) [5] and optimal[2] when the inputs are white [7];
- the SSARX [26] and PBSID [16], [17] algorithms are asymptotically equivalent[3][11];
- PBSID and the standard CCA method are asymptotically equivalent (for open loop data) when inputs are white (or absent) [10];
- an optimized version of PBSID (called PBSID_{opt}) provides estimators whose asymptotic variance is never larger than the asymptotic variance of the estimators obtained using the CCA method [10] (for open loop data);

Manuscript received April 17, 2007; revised April 05, 2008. First published January 22, 2010; current version published March 10, 2010. This work was supported in part by MIUR under national project New techniques and applications of identification and adaptive control. Recommended by Associate Editor W. X. Zheng.

The author is with the Dipartimento di Tecnica e Gestione dei Sistemi Industriali, Università di Padova, Vicenza 3-36100, Italy (e-mail: chiuso@dei.unipd.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2009.2039239

[1]Called CVA in [28], but we shall always use the terminology CCA in this paper.

[2]Within the class of algorithms based on the so called "state-sequence" approach [4].

[3]This means that the estimators obtained have the same asymptotic distribution and hence, in particular, the same asymptotic variance.

- PBSID_{opt} corresponds exactly to performing VARX estimation followed by the usual steps of subspace identification [11], making it very appealing from the computational point of view. Furthermore PBSID_{opt} is a "weighted version" of the algorithm originally discussed in [35] (see [11]).

For this class of algorithms (which, having all in common the use of Canonical Correlation Analysis (CCA) [25], can be called of the "CCA-type") three parameters have to be chosen: (i) the system order n, (ii) the length of the "past" horizon p and (iii) the length of the future horizon ν.

As the system order n is concerned, we refer the reader to [3], [4], [6] and references therein; in this paper we shall consider it as *given*.

The two latter parameters are the number of block rows in the block Hankel data matrices containing respectively the "past" (p) and "future" (ν) data. The length of the past horizon p, according to Assumption 1, should grow with the number of data. In practice this will have to be estimated from data, typically using standard criterions for VARX model order estimation. See [11], [16], [27], [38] for a discussion.

In this work we shall mainly be concerned with the length of the future horizon ν. In fact, while it has been shown that for the classical CCA algorithm the asymptotic variance of the estimated parameters is a monotonically non-increasing function of ν [7] when inputs are white or absent, to the best of the author's knowledge there are no results concerning more general cases. Some considerations are reported in [40].

This paper, which is an extension of[4] [12], can be roughly divided into three parts as follows:

a) First we shall show that also the ADAPTx algorithm in [31] (see also [30], [41]) is asymptotically equivalent to SSARX [26] (see Theorem 4.1), and hence to PBSID [11]. This result is instrumental for the second part of the paper for the following two reasons: (i) the derivation will be based on the ADAPTx algorithm, which makes the presentation simpler, and more intuitive in the author's view, than using SSARX or PBSID and (ii) the equivalence result allows to extend the analysis of the role of ν to a whole class of methods rather than a single algorithm.

b) In the second part we shall study the effect of ν in the state construction step in the "CCA" type algorithms discussed above. In particular we shall extend the well known results in [7], showing that the asymptotic variance is non-increasing with ν, to a slightly more general situation (see Condition 1 and Theorem 5.1 for details) which includes closed loop systems with proportional (output or state) feedback controllers and LQG controllers, provided the reference signal is white noise. The author would like to remark that the work in this paper considers a class of algorithms which are consistent also for closed-loop data, while [7] addressed the open loop case; furthermore, even though most of the techniques are adapted from [7], some modifications are introduced which allow to simplify the proofs (e.g., the choice of projectors (V.8)). In addition we stress that the restriction on the set of inputs allowed is given in terms of the state space, which we regard as a quite natural condition in the framework of subspace identification. The analysis contained in this paper, and in particular decomposition (V.1), also suggests a way to understand what happens when things go wrong. We hope also that the qualitative analysis, to be found in Section V-A, of the role of ν on the state estimation step will give inspiration for future research.
As a byproduct of this analysis it is also shown that under the same Assumption 1 the "CCA" weighting in the ADAPTx algorithm is optimal in the same sense as in [7] (see Proposition 7.1).

c) The last part of the paper illustrates, on a few examples, how some recently derived asymptotic variance formulas [9], [18] can be effectively used to optimally choose the length of the future horizon[5]. Optimality here is to be referred to an efficiency index defined in Sections VIII and IX. To the best of the author's knowledge this is also one of the first demonstrations that the expressions derived in [9] can be effectively used for the purpose of optimizing the identification algorithm and are not just of academic interest.

The structure of the paper is as follows: Section II contains the problem description and some notation; in Section III we describe the ADAPTx and SSARX algorithms and in Section IV we prove their asymptotic equivalence (Theorem 4.1). In Section V we shall discuss the effect of the future horizon and state the main result (Theorem 5.1). The particular case of VARX systems is addressed in Section VI, while optimality of the CCA weight is established in Section VII, Proposition 7.1. Section VIII contains a numerical illustration. Section IX will discuss the automated choice of ν, which goes through estimation of the asymptotic variance. Finally some conclusions are drawn in Section X. Most of the proofs and technical lemmas can be found in Appendix.

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 m_y and m_u respectively, which are the output and input signals of a linear stochastic system in innovation form

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

we allow for *feedback* from $\{\mathbf{y}(t)\}$ to $\{\mathbf{u}(t)\}$ [22], i.e., we consider "closed loop" identification; see also Fig. 2. 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 identifiability reasons, see [8], [16], [37] and references therein, we shall 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, of dimension $m_z := m_y + m_u$, as $\mathbf{z} := [\,\mathbf{y}^\top \quad \mathbf{u}^\top\,]^\top$ and assume that its spectral density matrix $\Phi(z)$ is rational, bounded and bounded away from zero on the unit circle $z = e^{j\omega}$. Let μ_i denote the zeros of the spectral density matrix inside the closed unit disc. We define

[4]We warn the reader that the paper [12] contained some minor mistakes and, for reasons of space, an oversimplified analysis.

[5]Note that the conclusion of [7] hinted at this possibility, we believe that actually demonstrating in a few examples that this procedure is viable can be of interest also to practitioners.

$\rho := \max\left(|\mathrm{eig}(\bar{A})|\right)$. From the assumption $\Phi(e^{j\omega}) > cI > 0$ it follows that $\rho \leq \max(|\ \mu_i\ |) < 1$ (see e.g., [16, Section 4]).

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. Let $\Lambda := \mathrm{Var}\{\mathbf{e}(t)\}$ and define 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 use the notation $\mathbf{x}_N = o_P(\mathbf{g}_N)$ if $\mathbf{x}_N/\mathbf{g}_N$ converges to zero in probability, i.e., $\forall\ \delta > 0$, $\lim_{N\to\infty} P[|\ \mathbf{x}_N/\mathbf{g}_N\ | > \delta] = 0$. The symbol $\underline{o}_P(\cdot)$ denotes a random vector whose components are uniformly[6] $o_P(\cdot)$; furthermore $\doteq$ 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 [20]) terms which are $\underline{o}_P(1/\sqrt{N})$ can be neglected when studying the asymptotic statistical properties. We shall also use the same symbol ($\doteq$) when the difference in the equated terms produces nonsingular change of basis $\hat{T}_N$ (up to $\underline{o}_P(1/\sqrt{N})$ and satisfying $\lim_{N\to\infty}\hat{T}_N = I$) in the estimated state sequences. In fact also these differences may be discarded as far as the estimation of system invariants is concerned. For instance, if $\mathbf{x}_1$ and $\mathbf{x}_2$ are two state variables, we shall write $\mathbf{x}_1 \doteq \mathbf{x}_2$ if there exists a non singular $\hat{T}_N$, with $\lim_{N\to\infty}\hat{T}_N = I$, so that $\mathbf{x}_1 - \hat{T}_N\mathbf{x}_2 = \underline{o}_P(1/\sqrt{N})$ (see the [10, Appendix] for details).

Our aim is to identify the 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 and the past horizon p go to infinity. We remind that p has to go to infinity at a certain rate depending on the number N of data available. Details can be found, for instance, in [4] where the following assumption is made:

Assumption 1: The past horizon p goes to infinity with N while satisfying

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

See [7], [11] for a discussion on this Assumption.

Boldface (lowercase) letters denote random variables. The symbol $\mathbb{E}\{\cdot\}$ shall denote expectation; given two (zero mean) random vectors $\mathbf{a}$ and $\mathbf{b}$ we shall define $\Sigma_{\mathbf{ab}} := \mathbb{E}\left[\mathbf{a}\mathbf{b}^{\top}\right]$, provided it exists. Lowercase letters denote sample values of a certain random variable. For example we shall denote with $\mathbf{y}(t)$ the output random vector and with y_t the sample value of $\mathbf{y}(t)$. We shall use capitals to denote the tail of length N. For instance $Y_t := [y_t y_{t+1}, \ldots y_{t+N-1}]$, and $Z_t := [\,Y_t^{\top} \quad U_t^{\top}\,]^{\top}$. These are the block rows of the usual *block Hankel data matrices* used in subspace identification. Finite block Hankel data matrices will be denoted using capitals; subscripts shall denote the time interval, e.g., $Y_{[t,s]} := [\,Y_t^{\top} \quad Y_{t+1}^{\top} \quad \ldots \quad Y_s^{\top}\,]^{\top}$ or $Y_{[t,s)} := [\,Y_t^{\top} \quad Y_{t+1}^{\top} \quad \ldots \quad Y_{s-1}^{\top}\,]^{\top}$. A similar notation will be used for random variables, e.g., $\mathbf{y}_{[t,s]} := [\,\mathbf{y}^{\top}(t) \quad \mathbf{y}^{\top}(t+1) \quad \ldots \quad \mathbf{y}^{\top}(s)\,]^{\top}$ and $\mathbf{y}_{[t,s)} := [\,\mathbf{y}^{\top}(t) \quad \mathbf{y}^{\top}(t+1) \quad \ldots \quad \mathbf{y}^{\top}(s-1)\,]^{\top}$.

Row 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 population covariance with a "hat" on top. For example, given finite sequences $A_t := [a_t, a_{t+1} \ldots, a_{t+N-1}]$ and $B_t := [b_t, b_{t+1} \ldots, b_{t+N-1}]$ we shall define the sample covariance matrix $\hat{\Sigma}_{\mathbf{ab}} := (1/N)\sum_{i=0}^{N-1} a_{t+i}b_{t+i}^{\top}$. Under our ergodic assumption $\lim_{N\to\infty}\hat{\Sigma}_{\mathbf{ab}} \overset{a.s.}{=} \Sigma_{\mathbf{ab}}$.

When dealing with tails, e.g., A_t and B_t, containing the sample values a_{t+i}, b_{t+i}, $i = 0, .., N-1$ of the random vectors $\mathbf{a}_N(t)$ and $\mathbf{b}_N(t)$, $A_t \doteq B_t$ really means that $\mathbf{a}(t) \doteq \mathbf{b}(t)$. The orthogonal projection onto the row space of a matrix shall be denoted with the symbol $\hat{E}$; for instance, given a matrix $C_t := [c_t, c_{t+1}, \ldots, c_{t+N-1}]$, $\hat{E}[\cdot|C_t]$ 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[7] $\hat{E}[A_t|C_t] = \hat{\Sigma}_{\mathbf{ac}}\hat{\Sigma}_{\mathbf{cc}}^{\dagger}C_t$. We shall use the notation

$$\Omega(A, C, k) := \begin{bmatrix} C^{\top} & A^{\top}C^{\top} & \ldots & \left(A^{\top}\right)^{k-1}C^{\top} \end{bmatrix}$$

to denote the extended observability matrix with k blocks of the pair A, C. For convenience of notation we denote with $\nu := T - t + 1$ and define the extended observability matrices $\mathcal{O}_{\nu}^{\top} := \Omega(A, C, \nu)$ and $\bar{\mathcal{O}}_{\nu}^{\top} := \Omega(\bar{A}, C, \nu)$. Define also the (reversed extended) reachability matrix

$$\mathcal{C}_p := [\,\bar{A}^{p-1}[K \quad B] \quad \ldots \bar{A}[K \quad B] \quad [K \quad B]\,]. \tag{II.3}$$

Given a finite sequence of matrices of compatible dimensions $M_0, M_1, \ldots, M_{k-1}$ we define the (block) Toeplitz matrix

$$\mathcal{T}(M_0, M_1, \ldots, M_{k-1}) := \begin{bmatrix} M_0 & 0 & \ldots & 0 \\ M_1 & M_0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ M_{k-1} & \ldots & M_1 & M_0 \end{bmatrix}. \tag{II.4}$$

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

Assumption 2: Let $\mathcal{Z}_t^-$ be the σ-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 a $\mathcal{Z}_{t-1}^-$-martingale difference sequence with constant conditional variance, i.e.

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

and has bounded fourth order moments.

Assumptions 1 and 2 will be made throughout the paper without further notice.

ALGORITHMS

As discussed in the introduction it has been shown in [11] that the recently developed SSARX [26] and PBSID [10], [16]

[6] Uniformity is needed when the matrices' sizes increase with N. In this paper uniformity shall be guaranteed by stationarity of the processes involved.

[7] The symbol † denotes Moore-Penrose pseudoinverse.

algorithms are asymptotically equivalent. We take as a representative in this class the SSARX algorithm.

In this section we show that also the ADAPTx algorithm proposed in [31] is asymptotically equivalent to SSARX; this can be seen as a CCA algorithm performed on the data once the "future" has been removed. Taking advantage of this equivalence, we shall stick to ADAPTx for the purpose of analysis. This intermediate result, besides complementing the comparative results already available (see [5], [7], [10], [11], [44] and references therein), turns out to be particularly useful in the analysis as we shall see in Section V.

It is well known that subspace identification can be seen as a 2-step procedure [11], [15], [33], [36], [44]; in a first step the state is estimated and in the second the system matrices are computed from the estimated state. The state is constructed from an estimate $\hat{\mathcal{C}}_p$ as follows:

$$\hat{X}_t := \hat{\mathcal{C}}_p Z_{[t_0,t)} \quad \hat{X}_{t+1} := \hat{\mathcal{C}}_p Z_{[t_0+1,t+1)}. \tag{III.1}$$

Then the state matrices (A, B, C, K) are estimated solving

$$\begin{cases} \hat{X}_{t+1} \simeq A\hat{X}_t + BU_t + K\hat{E}_t \\ Y_t \simeq C\hat{X}_t \end{cases} \tag{III.2}$$

in the least squares sense, where $\hat{E}_t := Y_t - \hat{C}\hat{X}_t$. We shall now discuss the preliminary constructions in SSARX (Section III-A) and ADAPTx (Section III-B) needed to obtain $\hat{\mathcal{C}}_p$ and show that they are asymptotically equivalent (Theorem 4.1). Eventually we shall get into the details of the state construction step in Section IV-B.

A. SSARX Algorithm

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

$$Y_T \simeq \sum_{i=1}^{T-t_0} \left(\Phi_i^y Y_{T-i} + \Phi_i^u U_{T-i}\right) \tag{III.3}$$

where, without loss of generality[8], we have taken the length of the VARX model equal to $T - t_0$. The estimators $\hat{\Phi}_k^u$ and $\hat{\Phi}_k^y$ of Φ_k^u and Φ_k^y are obtained by solving (III.3) in the least squares sense. Then the effect of the future inputs/outputs is removed using $\hat{\Phi}_k^y$, $\hat{\Phi}_k^u$ as[9]

$$Y^S_{[t,T]} := Y_{[t,T]} - \hat{\bar{H}}^u_\nu U_{[t,T]} - \hat{\bar{H}}^y_\nu Y_{[t,T]} \tag{III.4}$$

where $\hat{\bar{H}}^w_\nu := \mathcal{T}(0, \hat{\Phi}^w_1, .., \hat{\Phi}^w_{\nu-1})$. and w is alternatively u or y. Then the state is estimated via SVD decomposition of $\hat{W}_S^{-1}(\nu)\hat{Y}^S_{[t,T]}$ where

$$\hat{Y}^S_{[t,T]} = \hat{E}\left[Y^S_{[t,T]} \mid Z_{[t_0,t)}\right],$$
$$\hat{W}_S(\nu) = \left(\frac{Y^S_{[t,T]}\left(Y^S_{[t,T]}\right)^\top}{N}\right)^{1/2}. \tag{III.5}$$

This is equivalent (as the state construction is concerned) to performing CCA between the corrected future $Y^S_{[t,T]}$ and the past $Z_{[t_0,t)}$.

Let $W_S(\nu) := \lim_{N\to\infty}\hat{W}_S(\nu)$. It is easy to see that $W_S(\nu)W_S^\top(\nu) = \bar{\mathcal{O}}_\nu \Sigma_{\mathbf{xx}} \bar{\mathcal{O}}_\nu^\top + (I \otimes \Lambda)$. Using the same argument as in [7, Lemma 3.3] (see also [10]) it can be proven that $W_S(\nu) = (I \otimes \Lambda)^{1/2}$ gives the same asymptotic distribution of the estimators.

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

$$(I \otimes \Lambda)^{-1/2}\hat{E}\left[Y^S_{[t,T]} \mid Z_{[t_0,t)}\right] \tag{III.6}$$

when constructing the state space.

B. ADAPTx Algorithm

Also in this algorithm first a long VARX model is estimated, by solving (III.3) in the least squares sense, obtaining the same coefficients $\hat{\Phi}_k^y$ and $\hat{\Phi}_k^u$ as above. The VARX model described by the coefficients $\hat{\Phi}_k^y$ and $\hat{\Phi}_k^u$ is then realized in state space form obtaining $\hat{A}_{VARX}$, $\hat{B}_{VARX}$, $\hat{C}_{VARX}$, $\hat{K}_{VARX}$; then the Markov parameters[10] $\hat{\Psi}^d_k := \hat{C}_{VARX}\hat{A}^{k-1}_{VARX}\hat{B}_{VARX}$ are computed. Define also $\hat{H}^d_\nu := \mathcal{T}(0, \hat{\Psi}^d_1, .., \hat{\Psi}^d_{\nu-1})$ and observe that $\hat{H}^d_\nu$ is related to $\hat{\Phi}^y_k$ and $\hat{\Phi}^u_k$ as indicated in Lemma A.5 (see Appendix). The ADAPTx algorithm performs CCA between the corrected future $Y^A_{[t,T]} := Y_{[t,T]} - \hat{H}^d_\nu U_{[t,T]}$ and the past $Z_{[t_0,t)}$. This is equivalent, to the purpose of state construction, to performing SVD of $\hat{W}_A^{-1}(\nu)\hat{Y}^A_{[t,T]}$ where

$$\hat{Y}^A_{[t,T]} := \hat{E}\left[Y^A_{[t,T]} \mid Z_{[t_0,t)}\right] \tag{III.7}$$

and

$$\hat{W}_A(\nu) = \left(\frac{Y^A_{[t,T]}\left(Y^A_{[t,T]}\right)^\top}{N}\right)^{1/2}. \tag{III.8}$$

See, in particular, [31, equation (9)] and the discussion following the equation. The asymptotic value $W_A(\nu)$ of the weight $\hat{W}_A(\nu)$ satisfies

$$W_A(\nu)W_A^\top(\nu) = \mathcal{O}_\nu \Sigma_{\mathbf{xx}} \mathcal{O}_\nu^\top + H^s_\nu (I \otimes \Lambda)\left(H^s_\nu\right)^\top$$

where $H^s_\nu := \mathcal{T}(I, CK, CAK, \ldots, CA^{\nu-2}K)$. With an argument completely analogous to that used for $W_S(\nu)$ in Section III-A, $W_A(\nu)$ can be substituted with

$$W_A(\nu) := H^s_\nu (I \otimes \Lambda)^{1/2} \tag{III.9}$$

without altering the asymptotic properties (see [7], [10]). From now on the symbol $W_A(\nu)$ will be used according to definition (III.9). Therefore, for the purpose of analysis, we can assume ADAPTx performs SVD of

$$(I \otimes \Lambda)^{-1/2}\left(H^s_\nu\right)^{-1}\hat{E}\left[Y^A_{[t,T]} \mid Z_{[t_0,t)}\right] \tag{III.10}$$

when constructing the state space.

[8] See, e.g., [11, Lemma 4.2] or a more general result in [27].

[9] We shall use a superscript ^{S} (or subscript $_S$) to denote quantities related to SSARX; similarly ^{A} and $_A$ shall be attached to the ADAPTx algorithm.

[10] The superscript ^{d} stands for "deterministic".

III. CCA-Type Algorithms: Asymptotic Equivalence and State Construction

It has been recently shown, see e.g., [11], that SSARX [26] and PBSID [16] are asymptotically equivalent. In this paper we also show that ADAPTx of [31] is asymptotically equivalent to SSARX; we then refer to this class of algorithms as CCA-type. In this Section we first show the equivalence and then we discuss in more detail the state construction in this class of algorithms.

A. Asymptotic Equivalence

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

Theorem 4.1: The matrices in (III.6) and (III.10) differ only up to a left multiplication for a non-singular matrix which tends to the identity matrix as $N \to \infty$ and therefore the two constructions yield to asymptotically equivalent procedures.

Proof: Note that, using Lemma A.5, $\hat{Y}^S_{[t,T]}$ (III.5) can be rewritten as

$$\begin{aligned}\hat{Y}^S_{[t,T]} &= \hat{E}\left[Y_{[t,T]} - \hat{\bar{H}}^u_\nu U_{[t,T]} - \hat{\bar{H}}^y_\nu Y_{[t,T]} \mid Z_{[t_0,t)}\right] \\ &= \hat{E}\left[\left(I - \hat{\bar{H}}^y_\nu\right) Y_{[t,T]} - \hat{\bar{H}}^u_\nu U_{[t,T]} \mid Z_{[t_0,t)}\right] \\ &= \hat{E}\left[\left(\hat{H}^s_\nu\right)^{-1} Y_{[t,T]} - \hat{\bar{H}}^u_\nu U_{[t,T]} \mid Z_{[t_0,t)}\right] \\ &= \left(\hat{H}^s_\nu\right)^{-1} \hat{E}\left[Y_{[t,T]} - \hat{H}^s_\nu \hat{\bar{H}}^u_\nu U_{[t,T]} \mid Z_{[t_0,t)}\right] \\ &= \left(\hat{H}^s_\nu\right)^{-1} \hat{E}\left[Y_{[t,T]} - \hat{H}^d_\nu U_{[t,T]} \mid Z_{[t_0,t)}\right]\end{aligned}$$

so that (III.6) can be rewritten as

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

It is immediate to recognize that this is exactly the same as in the ADAPTx algorithm (see equation (III.10)), once the asymptotic value of the weight H^s_ν is substituted to its estimator; it is well known (see [4]) that this replacement does not change the asymptotic distribution. ∎

For this reason, when discussing the role of ν, we shall stick to the ADAPTx algorithm described in [31] from now on.

B. State Construction in CCA-Type Subspace Algorithms

The asymptotic equivalence proved above allows to study a whole class of methods which we call of the CCA-type. We now enter a bit more into the details of how the state is usually constructed. The first step is the singular value decomposition[11] (SVD) of (III.10)

$$\hat{U}_n(\nu)\hat{S}_n(\nu)\hat{V}_n^\top(\nu) \simeq W_A^{-1}(\nu)\frac{\hat{E}\left[Y^A_{[t,T]} \mid Z_{[t_0,t)}\right]}{N} \quad \text{(IV.2)}$$

where the subscript $_n$ reminds that only the n largest singular values are retained. The state space is then estimated using

$$\begin{aligned}\hat{X}^\nu_t &:= \hat{S}_n^{-1}(\nu)\hat{U}_n^\top(\nu)W_A^{-1}(\nu)\hat{E}\left[Y^A_{[t,T]} \mid Z_{[t_0,t)}\right] \\ &\doteq S_n^{-1}(\nu)U_n^\top(\nu)W_A^{-1}(\nu)\hat{E}\left[Y^A_{[t,T]} \mid Z_{[t_0,t)}\right] \\ &= \hat{\mathcal{C}}^\nu_p Z_{[t_0,t)} \quad \text{(IV.3)}\end{aligned}$$

where the superscript $^\nu$ reminds that the state space is estimated in a basis which, in general, depends on ν. The second (asymptotic) equivalence stems from the fact that, as thoroughly discussed in [4], only the asymptotic values[12] of $U_n(\nu)$ and $S_n(\nu)$ matter as far as the asymptotic distribution is concerned. The last equation in (IV.3) defines $\hat{\mathcal{C}}^\nu_p$. Note that

$$W_A(\nu)U_n(\nu)S_n(\nu) = \mathcal{O}_\nu T_\nu \quad \text{(IV.4)}$$

where $\mathcal{O}_\nu$ is the "true" observability matrix in a fixed basis such that $\Sigma_{\mathbf{xx}} = I$ and T_ν is a suitable change of basis, depending on ν. Note that, since $\lim_{N\to\infty}\hat{\Sigma}_{\hat{\mathbf{x}}^\nu\hat{\mathbf{x}}^\nu} = \lim_{N\to\infty}(\hat{X}^\nu_t\left(\hat{X}^\nu_t\right)^\top/N) = I$, the matrix T_ν is orthonormal. It is then convenient to define

$$\hat{X}_t := T_\nu \hat{X}^\nu_t \doteq T_\nu \hat{\mathcal{C}}^\nu_p Z_{[t_0,t)} = \hat{\mathcal{C}}_p Z_{[t_0,t)} \quad \text{(IV.5)}$$

where the last equation defines the estimate $\hat{\mathcal{C}}_p$ of $\mathcal{C}_p$, i.e., the controllability matrix expressed in the same fixed basis as $\mathcal{O}_\nu$. This state transformation is convenient to express the variance formulas in a common basis, which is now independent of ν.

IV. Role of the Future Horizon ν

In this section we shall analyze the role of ν in the state construction step. Unfortunately it will be possible to formalize a sharp result only under Condition 1 below, which are slightly more general than those found in [7]. However, besides obtaining a sharp result under suitable assumptions, we shall try to develop some intuition as to which is the role played by the state space of the input process.

Let us define $\tilde{H}^d_\nu := \hat{H}^d_\nu - H^d_\nu$ and recall that $Y_{[t,T]} = \mathcal{O}_\nu X_t + H^d_\nu U_{[t,T]} + H^s_\nu E_{[t,T]}$. Using now the fact that[13] $\hat{E}\left[X_t \mid Z_{[t_0,t)}\right] \doteq X_t$, $\hat{Y}^A_{[t,T]}$ in (III.7) can be written as

$$\hat{Y}^A_{[t,T]} \doteq \mathcal{O}_\nu X_t + \hat{E}\left[H^s_\nu E_{[t,T]} - \tilde{H}^d_\nu U_{[t,T]} \mid Z_{[t_0,t)}\right].$$

Now it is convenient to rewrite (IV.3) and (IV.5) as

$$\begin{aligned}\hat{X}_t &\doteq X_t + T_\nu S_n^{-1}(\nu)\hat{E}\left[U_n^\top(\nu)\bar{E}_{[t,T]} \mid Z_{[t_0,t)}\right] \\ &\quad - T_\nu S_n^{-1}(\nu)U_n^\top(\nu)W_A^{-1}(\nu)\tilde{H}^d_\nu \hat{E}\left[U_{[t,T]} \mid Z_{[t_0,t)}\right] \\ &= X_t + T_\nu S_n^{-1}(\nu)\left(\tilde{X}^e_t + \tilde{X}^u_t\right) \quad \text{(V.1)}\end{aligned}$$

where $\tilde{X}^e_t := \hat{E}\left[U_n^\top(\nu)\bar{E}_{[t,T]} \mid Z_{[t_0,t)}\right]$ and $\tilde{X}^u_t := -U_n^\top(\nu)W_A^{-1}(\nu)\tilde{H}^d_\nu \hat{E}\left[U_{[t,T]} \mid Z_{[t_0,t)}\right]$. Note that the equality $W_A^{-1}(\nu)H^s_\nu E_{[t,T]} = \bar{E}_{[t,T]}$ has been used.

Recall now that the i-th column of $\bar{E}_{[t,T]}$ is a realization of the normalized future innovations $\bar{\mathbf{e}}_{[t+i-1,T+i-1]}$;

[11] In this paper we shall make explicit the dependence on ν of the weights and the matrices involved in the SVD.

[12] To be precise one should also make sure that the state space basis is asymptotically fixed. This can be done in a number of ways. We refer the reader to [7] or to the discussion around formula [9, (26)]. These considerations are inessential when it comes to implementing the algorithms.

[13] 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.

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 $\mathbf{v}(t+i,\nu) := U_n^\top(\nu)\bar{\mathbf{e}}_{[t+i-1,T+i-1]}$ which, regardless of ν, has unit covariance matrix, i.e., $\mathbb{E}\left[\mathbf{v}(t,\nu)\mathbf{v}^\top(t,\nu)\right] = I$; in order for this to hold it is crucial that the CCA weighting in (III.9) is used. For future reference let us define $V_t(\nu) := [v_t(\nu), v_{t+1}(\nu), .., v_{t+N-1}(\nu)] = U_n^\top(\nu)\bar{E}_{[t,T]}$ where $v_t(\nu)$ is the sample value of $\mathbf{v}(t,\nu)$.

In order to study the effect of the error in estimating the state in more detail we introduce the innovation model of the input process; this will include the case in which there is feedback. From the assumption that the joint spectrum is rational, also the input process $\mathbf{u}(t)$ can be thought of as the output of the (of minimal dimension n_u) state-space model[14] in innovation form

$$\begin{cases} \mathbf{x}^u(t+1) = A_u\mathbf{x}^u(t) + B_u\mathbf{y}(t) + K_u\mathbf{n}(t) \\ \mathbf{u}(t) = C_u\mathbf{x}^u(t) + D_u\mathbf{y}(t) + \mathbf{n}(t) \end{cases} \tag{V.2}$$

where $\mathbf{n}(t)$ is the one step ahead linear prediction error of $\mathbf{u}(t)$ given $\mathbf{y}(s)$ up to time t and $\mathbf{u}(s)$ up to time $t-1$. For future reference we also define $\Lambda_{\mathbf{n}} := \mathbb{E}\mathbf{n}(t)\mathbf{n}^\top(t)$. If there is no feedback $B_u = 0$ and $D_u = 0$ hold.

Combining the state space model (V.2) together with (II.1), it is immediate to see that the vector process $\mathbf{q}(t) := [\mathbf{x}^\top(t) \quad (\mathbf{x}^u)^\top(t)]^\top$ is a valid state vector (not necessarily minimal) for the joint process $\mathbf{z}(t)$. In particular $\mathbf{q}(t)$ is the state vector of the innovation model for $\mathbf{z}(t)$, i.e.,

$$\begin{cases} \mathbf{q}(t+1) = \mathcal{A}\mathbf{q}(t) + \mathcal{K}\mathbf{w}(t) \\ \mathbf{y}(t) = \mathcal{C}_y\mathbf{q}(t) + \mathcal{J}_y\mathbf{w}(t) \\ \mathbf{u}(t) = \mathcal{C}_u\mathbf{q}(t) + \mathcal{J}_u\mathbf{w}(t) \end{cases} \tag{V.3}$$

where $\mathbf{w}(t) := [\mathbf{e}^\top(t) \quad \mathbf{n}^\top(t)]^\top$ and $\mathcal{A}, \mathcal{B}, \mathcal{C}_y, \mathcal{J}_y, \mathcal{C}_u, \mathcal{J}_u$ are suitable matrices, see e.g., [16, Section 4] for details.

Let $\mathcal{O}_\nu^u$ be the extended observability matrix of the pair $(\mathcal{A}, \mathcal{C}_u)$, $\mathcal{H}_\nu^u$ the block Toeplitz matrix containing the Markov parameters $\mathcal{J}_u, \mathcal{C}_u\mathcal{K}, \mathcal{C}_u\mathcal{A}\mathcal{K}, \ldots, \mathcal{C}_u\mathcal{A}^{\nu-2}\mathcal{K}$; denote with Q_t the matrix which contains the sample values of $\mathbf{q}(t)$. Then the future inputs $U_{[t,T]}$ can be written in the form

$$U_{[t,T]} = \mathcal{O}_\nu^u Q_t + \mathcal{H}_\nu^u W_{[t,T]}. \tag{V.4}$$

We state now a result which generalizes the behavior with white inputs studied in [7]. This theorem will require the following Condition:

1) Condition 1: The input process $\mathbf{u}(t)$ is such that $\mathbf{x}^u(t) = T_{xu}\mathbf{x}(t)$, for some matrix T_{xu}, i.e., the state space of the input process is a linear function of the state $\mathbf{x}(t)$.

Theorem 5.1: Under Condition 1, the asymptotic variance of any system invariant which depends differentiably on the matrices A, B, C, K, estimated using a CCA-type subspace algorithm is a monotonically non increasing function of ν.

Proof: Let us consider the term $\tilde{X}_t^u := -U_n^\top(\nu)W_A^{-1}(\nu)\tilde{H}_\nu^d\hat{E}\left[U_{[t,T]} \mid Z_{[t_0,t)}\right]$ in (V.1). From equation (V.4) it follows that:

$$\tilde{H}_\nu^d\hat{E}\left[U_{[t,T]} \mid Z_{[t_0,t)}\right] \doteq \tilde{H}_\nu^d\mathcal{O}_\nu^u Q_t$$

where the fact that the elements of $\tilde{H}_\nu^d\hat{E}\left[\mathcal{H}_\nu^u W_{[t,T]} \mid Z_{[t_0,t)}\right]$ are $o_P(1/\sqrt{N})$ has been used. Using now Condition 1 it follows that $Q_t = \begin{bmatrix} I & T_{xu}^\top \end{bmatrix}^\top X_t$. Therefore, there exists a matrix T_u of suitable dimensions such that

$$\tilde{H}_\nu^d\hat{E}\left[U_{[t,T]} \mid Z_{[t_0,t)}\right] \doteq \tilde{H}_\nu^d T_u X_t. \tag{V.5}$$

Plugging (V.5) in (V.1) we obtain

$$\hat{X}_t \doteq \hat{T}(\nu)X_t + T_\nu S_n^{-1}(\nu)\hat{E}\left[U_n^\top(\nu)\bar{E}_{[t,T]} \mid Z_{[t_0,t)}\right] \tag{V.6}$$

where $\hat{T}(\nu) := I - T_\nu S_n^{-1}(\nu)U_n^\top(\nu)W_A^{-1}(\nu)\tilde{H}_\nu^d T_u$. Note that

$$\hat{T}(\nu) \overset{N\to\infty}{\longrightarrow} I. \tag{V.7}$$

Let us define $\Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}} = \mathbb{E}\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}^\top$ and introduce $\tilde{\mathcal{C}}_p^\nu := T_\nu S(\nu)^{-1}V_t(\nu)Z_{[t_0,t)}^\top\left(Z_{[t_0,t)}Z_{[t_0,t)}^\top\right)^{-1}$. Define also the projection matrices[15]

$$\begin{aligned} \Pi &= \Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}\mathcal{C}_p^\top\left(\mathcal{C}_p\Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}\mathcal{C}_p^\top\right)^{-1}\mathcal{C}_p \\ \Pi_c &= I - \Pi. \end{aligned} \tag{V.8}$$

It is now easy to see that, using $X_t \doteq \mathcal{C}_p Z_{[t_0,t)}$, (V.6) can be rewritten as

$$\begin{aligned} \hat{X}_t &\doteq \left[\hat{T}(\nu)\mathcal{C}_p + \tilde{\mathcal{C}}_p^\nu\right]Z_{[t_0,t)} \\ &= \left[\hat{T}(\nu)\mathcal{C}_p + \tilde{\mathcal{C}}_p^\nu\Pi + \tilde{\mathcal{C}}_p^\nu\Pi_c\right]Z_{[t_0,t)} \\ &= \left[\hat{T}_R(\nu)\mathcal{C}_p + \tilde{\mathcal{C}}_p^\nu\Pi_c\right]Z_{[t_0,t)} \\ &\doteq \hat{T}_R(\nu)X_t + \tilde{\mathcal{C}}_p^\nu\Pi_c Z_{[t_0,t)} \end{aligned} \tag{V.9}$$

where

$$\hat{T}_R(\nu) := \hat{T}(\nu) + \tilde{\mathcal{C}}_p^\nu\Pi\mathcal{C}_p^\top\left(\mathcal{C}_p\mathcal{C}_p^\top\right)^{-1}.$$

Note that, from the uniform convergence of sample covariances [24], $\left\|V_t(\nu)Z_{[t_0,t)}^\top/N\right\|_2$ (and hence $\|\tilde{\mathcal{C}}_p^\nu\|_2$) goes to zero in probability as $N \to \infty$; this implies that, for large N, $\hat{T}_R(\nu)$ is non singular, making it a well defined change of basis. In particular also

$$\hat{T}_R(\nu) \overset{N\to\infty}{\longrightarrow} I \tag{V.10}$$

holds true. Note also that, combining (IV.3), (IV.5) and (V.9) we obtain

$$\begin{aligned} \hat{\mathcal{C}}_p Z_{[t_0,t)} &= \hat{\mathcal{C}}_p\Pi Z_{[t_0,t)} + \hat{\mathcal{C}}_p\Pi_c Z_{[t_0,t)} \\ &\doteq \hat{T}_R(\nu)X_t + \tilde{\mathcal{C}}_p^\nu\Pi_c Z_{[t_0,t)} \end{aligned}$$

so that

$$\tilde{X}_t := \tilde{\mathcal{C}}_p^\nu\Pi_c Z_{[t_0,t)} \doteq \hat{\mathcal{C}}_p\Pi_c Z_{[t_0,t)}. \tag{V.11}$$

The choice of projections (V.8) is particularly useful since, as formally proved in Lemma A.1, this guarantees that the sample covariance $(1/N)X_t\tilde{X}_t^\top$ (where $\tilde{X}_t$ has been defined in (V.11))

[14]The subscripts $_u$ and the superscripts ^{u} remind that the corresponding quantities refer to the input model.

[15]To be precise $\Pi^\top$ and $\Pi_c^\top$ are orthogonal projections in the inner product $\langle v, u\rangle_Q := v^\top Q u$ defined by the process $\mathbf{z}$ through its covariance matrix via $Q := \Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}$.

is $\underline{o}_P(1/\sqrt{N})$ and hence can be neglected when computing the asymptotic distribution. This fact is very useful in simplifying the variance expressions[16].

We now observe, see e.g.,[17][7], [9], [10], that the errors in the estimated system matrices can be written, asymptotically, as linear combinations of

$$\tilde{\theta}_1(N) := \operatorname{vec}\left(\hat{\Sigma}_{\mathbf{ez}}\right) \doteq \operatorname{vec}\left(\hat{\Sigma}_{\mathbf{ez}_{[t_0,t)}}\mathcal{C}_p^\top\right)$$
$$\tilde{\theta}_2(N,\nu) := \operatorname{vec}\left(\tilde{\mathcal{C}}_p^\nu \Pi_c \Upsilon_p\right) \tag{V.12}$$

for suitable matrices Υ_p with columns in ℓ^2. This statement is made precise in Lemma A.2.

Note that the number of columns of $\tilde{\mathcal{C}}_p^\nu \Pi_c$ increases with the sample size N. Therefore the asymptotic analysis has to be performed on linear combinations $\tilde{\mathcal{C}}_p^\nu \Pi_c \Upsilon_p$ where the columns of Υ_p are uniformly bounded (i.e., $\|\Upsilon_p^\top \Upsilon_p\| < M$) and converge in ℓ^2 to a matrix Υ (with columns in ℓ^2), see [7], [10], [32]. The fact that all (column) vectors encountered shall converge in ℓ^2 is guaranteed by the assumption that the spectrum is rational, bounded and bounded away from zero.

As shown in the auxiliary Lemma A.3, $\tilde{\theta}_1(N)$ and $\tilde{\theta}_2(N,\nu)$ are asymptotically uncorrelated for all matrices Υ_p (converging in ℓ^2) introduced above in (V.12) and specified in Lemma A.2. Since ν affects only $\tilde{\theta}_2(N,\nu)$ we can therefore restrict to studying only this term.

It is also shown in Lemma A.4 that, under Condition 1, the asymptotic variance $V_{\theta_2}(\nu) := \operatorname{AsVar}\left\{\sqrt{N}\tilde{\theta}_2(N,\nu)\right\}$ has the form

$$V_{\theta_2}(\nu) = \lim_{N\to\infty}\left(\Upsilon_p^\top \Pi_c^\top \Sigma^{-1}_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}} \Pi_c \Upsilon_p\right) \otimes T_\nu S_n^{-2}(\nu) T_\nu^\top. \tag{V.13}$$

We recall that under the stated assumptions convergence holds; often sample variance matrices are substituted with their population value thanks to their uniform convergence. Also all inverses exist in the limit due to the assumption on the joint spectrum (bounded and bounded away from zero).

In order to show that $V_{\theta_2}(\nu)$ is non increasing as a function of ν, it is sufficient to show that $V_{\theta_2}(\nu-1) \geq V_{\theta_2}(\nu)\ \forall \nu > n$. From the properties of Kronecker products this latter holds if $T_{\nu-1} S_n^{-2}(\nu-1) T_{\nu-1}^\top - T_\nu S_n^{-2}(\nu) T_\nu^\top \geq 0$ or, equivalently, if

$$S_n^{-2}(\nu-1) - T_{\nu-1}^\top T_\nu S_n^{-2}(\nu) T_\nu^\top T_{\nu-1} \tag{V.14}$$

is positive semidefinite. From (IV.4) and recalling that $(H_\nu^s)^{-1}\mathcal{O}_\nu = \bar{\mathcal{O}}_\nu$, the matrices T_ν and $T_{\nu-1}$ are related by:

$$\underline{U}_n(\nu) S_n(\nu) T_\nu^\top = U_n(\nu-1) S_n(\nu-1) T_{\nu-1}^\top \tag{V.15}$$

where $\underline{U}_n(\nu)$ denotes the matrix $U_n(\nu)$ with the last m_y rows removed. Therefore

$$T_{\nu-1}^\top = S_n(\nu-1)^{-1} U_n(\nu-1)^\top \underline{U}_n(\nu) S_n(\nu) T_\nu^\top. \tag{V.16}$$

[16]In this paper, for reasons of space, we shall only address the specific case studied in Theorem 5.1 ; future work will investigate the effect of this simplification in more general cases.

[17]Note that in this paper, for identifiability reasons $D = 0$ is postulated. The same results reported here would hold if, instead, the feedback channel was assumed to have a delay and D was estimated.

Plugging (V.16) in (V.14) we obtain

$$S_n^{-1}(\nu-1)\left[I - \hat{\Sigma}_{U\underline{U}}\hat{\Sigma}_{U\underline{U}}^\top\right] S_n^{-1}(\nu-1) \tag{V.17}$$

where $\hat{\Sigma}_{U\underline{U}} := U_n(\nu-1)^\top \underline{U}_n(\nu) = [U_n(\nu-1)^\top \quad 0]\, U_n(\nu)$. Recalling that $U_n(\nu)^\top U_n(\nu) = I$ and $[U_n(\nu-1)^\top \quad 0][U_n(\nu-1)^\top \quad 0]^\top = I$ the singular values of $\hat{\Sigma}_{U\underline{U}}\hat{\Sigma}_{U\underline{U}}^\top$ are the (squared) canonical correlation coefficients between the column spaces of $U_n(\nu)$ and $[U_n(\nu-1)^\top \quad 0]^\top$, which are (positive and) smaller or equal than 1. This implies that (V.17) is positive semidefinite, which, together with the discussion in [9, [item 4) p. 1305], concludes the proof. ■

Remark V.1: Note that this theorem is an extension of the result found in [7]. In fact, if the input is white CCA is equivalent to PBSID (see [10]), to SSARX (see [11]) and hence to ADAPTx (Theorem 4.1). Also, when the inputs are white, $\mathbf{x}^u(t)$ is the empty vector, so that the theorem applies. ◊

Remark V.2: Using (V.2), the most general input under which Condition 1 is met, has the form $\mathbf{u}(t) = C_u T_{xu}\mathbf{x}(t) + D_u\mathbf{y}(t) + \mathbf{n}(t)$; this includes constant output feedback or (steady-state) LQG controllers[18], provided the reference signal is white.

Note that the assumptions on the input process are expressed in terms of its state space, and in particular studying its relation with the state space of the input-output model (II.1). We believe this is a rather natural condition for subspace methods, which are essentially based on constructing the state space. ◊

A. Some Insights on the Role of ν

Now we would like to discuss the role of ν when Condition 1 does not hold. It is clear that, without this condition, the proofs of Lemmas A.2, A.3 and A.4 do not go through. In this Section we make one step back and only consider the state construction step, see (V.1). In particular we would like to study which is the effect of the length of the future horizon on the error in estimating the state space (which could be measured, for instance, in terms of subspace angles). Note that the perturbation on the estimated state has two sources: $\tilde{X}_t^e$ and $\tilde{X}_t^u$. Let us concentrate on the second term $\tilde{X}_t^u$, which measures how difficult it is to "remove" the part due to future inputs in $Y^A_{[t,T]} = Y_{[t,T]} - \hat{H}_\nu^d U_{[t,T]}$ (see equation (III.7)). It is to be expected that the larger (e.g., in terms of variance) $\tilde{X}_t^u$, the larger the perturbation on the estimated state and, very likely, the worse the estimation performance. We understand this is rather qualitative as we are indeed neglecting correlation effects (e.g., between $\tilde{X}_t^e$ and $\tilde{X}_t^u$, but also in measuring the effect of $\tilde{X}_t^e$ by not studying terms for $i \neq j$ in (A10) (Lemma A.4)). We believe, however, that this discussion does shed some light on the effect of ν.

Having this in mind, note that $\tilde{X}_t^u$ contains two "sources of error": the first, $\tilde{H}_\nu^u$, is related to the error in estimating the input related Markov parameters; the second is the projection of the future inputs onto the joint past. Denote $\mathbf{u}^+ := [\mathbf{u}^\top(t), \ldots, \mathbf{u}^\top(t+\nu-1)]^\top$. It is rather simple to verify that

$$\tilde{H}_\nu^d \hat{E}[U_{[t,T]}|Z_{[t_0,t)}] \doteq \tilde{H}_\nu^d \Sigma_{\mathbf{u}^+\mathbf{x}|\mathbf{x}^u} \Sigma^{-1}_{\mathbf{xx}|\mathbf{x}^u} X_t + \tilde{H}_\nu^d \Sigma_{\mathbf{u}^+\mathbf{x}^u|\mathbf{x}} \Sigma^{-1}_{\mathbf{x}^u\mathbf{x}^u|\mathbf{x}} X_t^u. \tag{V.18}$$

[18]Note that $\mathbf{x}(t)$ is the state space of the innovation model, i.e., of the Kalman filter.

When Condition 1 holds true, as shown in equation (V.5), $\tilde{H}^d_\nu \hat{E}[U_{[t,T]}|Z_{[t_0,t)}]$ is, up to $\varrho_P(1/\sqrt{N})$ terms, a linear combination of the rows of X_t times a matrix which vanishes in probability. This implies that the perturbation $\tilde{X}^u_t$ is really a linear function of X_t itself; it follows that this perturbation just results in a change of basis (see e.g., (V.6)).

Instead, when Condition 1 is not met, $\tilde{X}^u_t$ does not lie in X_t and hence it perturbs the state estimate.

The following lemma makes the dependence of $\tilde{X}^u_t$ on ν, $\mathbf{x}^u$ and $\mathbf{x}$ a bit more precise.

Lemma A.2: Assume, w.l.o.g., that a basis $\mathbf{x}^u$ has been fixed in stochastically balanced form, so that $\Sigma_{\mathbf{x}^u\mathbf{x}^u} = S_{\mathbf{x}^u\mathbf{x}^u} = \mathrm{diag}\{\gamma_i\}$ where γ_i are the canonical correlation coefficients between the past of $\mathbf{z}$ and the future of $\mathbf{u}$. Let $X^\perp_t$ be an orthonormal matrix whose rows span the orthogonal complement of X_t in $\begin{bmatrix} X_t^\top & (X^u_t)^\top \end{bmatrix}^\top$ and Γ^u_ν be the matrix formed by the last n_u columns of $\mathcal{O}^u_\nu$. Then $\tilde{H}^d_\nu \hat{E}[U_{[t,T]}|Z_{[t_0,t)}]$ admits the expression

$$\tilde{H}^d_\nu \hat{E}[U_{[t,T]}|Z_{[t_0,t)}] \doteq \tilde{H}^d_\nu \bar{T}_u X_t + \tilde{H}^d_\nu \Gamma^u_\nu S^{1/2}_{\mathbf{x}^u\mathbf{x}^u} P S_{\mathbf{x}^u\mathbf{x}^\perp} Q X^\perp_t \quad \text{(V.19)}$$

where P and Q are orthonormal matrices and $S_{\mathbf{x}^u\mathbf{x}^\perp} = \mathrm{diag}\{\sqrt{1-\sigma_i^2},\ i = 1,..,\min(n,n_u)\}$, the σ_i^2's being the squared canonical correlation coefficients [21] between the spaces generated by $\mathbf{x}^u(t)$ and $\mathbf{x}(t)$.

Proof: Equation (V.18) can be rewritten as

$$\tilde{H}^d_\nu \hat{E}[U_{[t,T]}|Z_{[t_0,t)}] \doteq \tilde{H}^d_\nu \bar{T}_u X_t + \tilde{H}^d_\nu \Sigma_{\mathbf{u}^+\mathbf{x}^\perp} X^\perp_t \quad \text{(V.20)}$$

for a suitable choice of the matrix $\bar{T}_u$. From (V.3) it is easy to compute

$$\Sigma_{\mathbf{u}^+\mathbf{x}^\perp} = \mathcal{O}^u_\nu \Sigma_{\mathbf{q}\mathbf{x}^\perp} = \Gamma^u_\nu \Sigma_{\mathbf{x}^u\mathbf{x}^\perp}$$

where Γ^u_ν is the matrix formed by the last n_u columns of $\mathcal{O}^u_\nu$. Since $\mathbf{x}^u$ has been normalized so that $\Sigma_{\mathbf{x}^u\mathbf{x}^u} = S_{\mathbf{x}^u\mathbf{x}^u}$, the singular values of $S^{-1/2}_{\mathbf{x}^u\mathbf{x}^u}\Sigma_{\mathbf{x}^u\mathbf{x}^\perp}$ are $\sqrt{1-\sigma_i^2}$, $i = 1,..,\min(n,n_u)$, where the σ_i^2's are the squared canonical correlation coefficients [21] between the spaces generated by $\mathbf{x}^u(t)$ and $\mathbf{x}(t)$, which concludes the proof. ∎

It follows from (V.19) and (V.1) that the state error $\hat{E}\left[\tilde{X}^u_t | X^\perp_t\right]$, i.e., the contribution of $\tilde{X}^u_t$ orthogonal to X_t, can be written in the form

$$-U_n^\top(\nu) W_A^{-1}(\nu) \tilde{H}^d_\nu \Gamma^u_\nu S^{1/2}_{\mathbf{x}^u\mathbf{x}^u} P S_{\mathbf{x}^u\mathbf{x}^\perp} Q X^\perp_t. \quad \text{(V.21)}$$

We would like to understand when this perturbation term becomes large; this depends on

(i) how far, in terms of (squared) canonical correlations σ_i^2's, is the space generated by $\mathbf{x}^u(t)$ from the space generated by $\mathbf{x}(t)$;
(ii) how close are the spaces generated by past and future inputs, measured by the canonical correlations γ_i's;
(iii) how large is the variance of the elements $\hat{\Psi}^d_k, k = 1,..,\nu-1$ of $\tilde{H}^d_\nu$ and how it depends upon k (and hence ν).

The closer the σ_i's to 1 and the closer the γ_i's to zero, the more the results based on Condition 1 can be considered a good approximation of what really happens.

Instead, when (V.21) cannot be neglected, anything can happen and, indeed, it is possible to tailor examples in which the asymptotic variance is monotonically decreasing, monotonically increasing or show a local minima as ν varies.

TABLE I
DETAILS OF THE SIMULATION SETUP

	example 1	example 2	example 3
$F(z)$	$\frac{1}{z-0.5}$	$\frac{z}{z^2-1.5z+0.7}$	$\frac{z}{z^2-1.5z+0.7}$
$G(z)$	$\frac{z+0.5}{z-0.5}$	$\frac{z^2-0.5z+0.7}{z^2-1.5z+0.7}$	$\frac{z^2-0.5z+0.7}{z^2-1.5z+0.7}$
$H(z)$	0	$-0.05\frac{z-0.2}{z-0.5}$	$-0.05\frac{z-0.2}{z-0.5}$
$K(z)$	$\frac{z-\alpha}{z-0.9}$	$\frac{z^2-0.5z+0.7}{z^2-1.5z+0.7}$	$\frac{z^2-1.4z+0.7}{z^2-1.5z+0.7}$

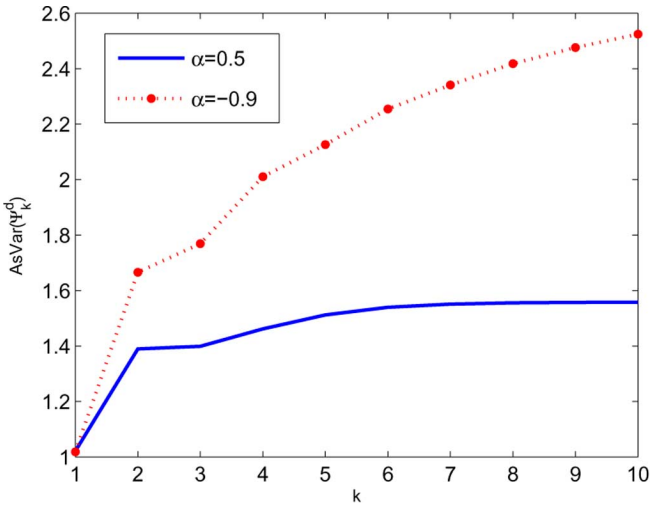

Fig. 1. Asymptotic variance of $\hat{\Psi}^d_k$, $k \in [1, 10]$.

Let us consider example 1 in Section VIII (see also Table I). For $\alpha = 0.5$ the canonical correlation between past and future of $\mathbf{u}$ is $\gamma_{0.5} \simeq 0.7273$ while, for $\alpha = -0.9$, $\gamma_{-0.9} \simeq 0.9945$ holds. Similarly it is also possible to compute the canonical correlations between $\mathbf{x}^u$ and $\mathbf{x}$, which are, respectively, $\sigma_{0.5} \simeq 0.8932$ and $\sigma_{-0.9} \simeq 0.9225$.

Since in this example both $\mathbf{x}$ and $\mathbf{x}^u$ have dimension 1, also P and Q in (V.19) are equal to 1. Therefore, for $\alpha = 0.5$, $S^{1/2}_{\mathbf{x}^u\mathbf{x}^u} P S_{\mathbf{x}^u\mathbf{x}^\perp} Q \simeq \sqrt{0.7273\,(1-0.8932^2)} = 0.1471$ holds. Instead, for $\alpha = -0.9$, $S^{1/2}_{\mathbf{x}^u\mathbf{x}^u} P S_{\mathbf{x}^u\mathbf{x}^\perp} Q \simeq \sqrt{0.9945\,(1-0.9225^2)} = 0.1482$.

Thus there are no major differences as far as items (i) and (ii) above are concerned, and in both cases the term due to $\tilde{X}^u_t$ cannot be neglected. However, in the first situation, i.e., $\alpha = 0.5$ the variance is monotonically decreasing in ν while the opposite happens for $\alpha = -0.9$.

As far as item (iii) above, the asymptotic variance of $\hat{\Psi}^d_k$ in $\hat{H}^d_\nu$ is proportional, for a suitable H which does not depend on the input spectrum, to $H\Sigma^{-1}_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}} H^\top$; this quantity may become large (depending upon H) when the joint spectrum of $\mathbf{z}$ varies widely (see e.g., [23]). Fig. 1 reports the asymptotic variance of $\hat{\Psi}^d_k$ for $k = 1,\ldots,10$; as expected to wider variations of the input spectrum ($\alpha = -0.9$) there corresponds larger variance of $\hat{\Psi}^d_k$.

V. IDENTIFICATION OF VARX SYSTEMS

In the particular case of identification of VARX 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.

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

Proposition 6.1: Let us assume that the system is of the VARX type with index of nilpotency k. Then the state estimator (IV.3), (IV.5) (and hence any estimator based on the state sequence) is not affected by the choice of ν provided $\nu \geq k$.

Proof: See Appendix. ■

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 [7] (see the discussion after Corollary 2) where only white inputs were allowed for.

With a similar argument it can be verified that, when $\bar{A}$ is not nilpotent, the asymptotic variance $V_\theta(\nu) := \text{AsVar}\left\{\sqrt{N}\,[\tilde{\theta}_1^\top(N) \quad \tilde{\theta}_2^\top(N,\nu)]^\top\right\}$ converges, as $\nu \to \infty$, to a limit $V_\theta(\infty)$ at least as fast as[19] ρ^ν goes to zero, where ρ is the maximum (in modulus) eigenvalue of $\bar{A}$, i.e., $(V_\theta(\nu) - V_\theta(\infty))/\rho^\nu = O(1)$. This is rather simple, but tedious, to show and relies on the fact that the matrix $\bar{\mathcal{O}}_\infty^\top$, has the structure $[\bar{\mathcal{O}}_\nu^\top \quad \rho^\nu \tilde{\mathcal{O}}_\infty^\top]$ with $\|\tilde{O}_\infty\|_2$ bounded. With an argument similar to the one used in the proof of lemma 6.1, it can be shown that $V_\theta(\nu)$ differs from $V_\theta(\infty)$ by terms which go to zero (at least) as fast as ρ^ν. Details are omitted for reasons of space.

This is confirmed by the experimental results which show that, indeed, the efficiency index $Eff(\nu)$ in (VIII.2) does not change when ν is increased over a certain threshold. In the examples reported in Section VII this threshold is roughly $\nu \simeq 10$. Note that, for this example $\rho = |\bar{A}| = 0.5$, so that $\rho^{10} = \bar{A}^{10} \simeq 10^{-3}$ which is of the same order of magnitude of the change observed in $Eff(\nu)$ for $\nu > 10$.

Remark VI.3: The reader may argue that Proposition 6.1 has a purely theoretical interest since (i) if one knew the system is VARX there would be no need to use subspace methods for identification and (ii) VARX models (of "low" order) are hardly found in practice. However the fact that $(V_\theta(\nu) - V_\theta(\infty))/\rho^\nu = O(1)$ suggests that the search for the "optimal" ν can be limited to values of ν which are smaller than a certain value $\bar{\nu}$ such that $\rho^{\bar{\nu}}$ becomes negligible. A suitable $\bar{\nu}$ can be estimated using standard order estimation techniques for long autoregressions [24], [27], [34], [42]. In practice $\bar{\nu} = p$ can be taken, as for instance it is done in [19]. ◊

VI. Optimality of the CCA Weighting

Note that, even though in this paper we restrict our attention to methods employing the CCA weight, the ADAPTx method described in Section III-B, and all CCA-type algorithms, could be implemented with any other (left) weight, i.e., with other choices of $\hat{W}_A(\nu)$ in (III.8). It is natural to ask whether this weighting has some optimality properties. The following proposition partially answer this question, extending the results in [7].

Proposition 7.1: Under Condition 1 the CCA weighting used in the ADAPTx method is optimal, in the sense of yielding, for any value of ν, lower (or equal) asymptotic variance of the estimators of any system invariant which depends differentiably on the matrices A, B, C, K, than any other weighting scheme.

Proof: Under the assumption of Theorem 5.1 and using Lemmas A.2, A.3 and A.4 the asymptotic variance expression has the same form as that found in [7]; hence the proof follows the same lines as in [7], and is therefore omitted. ■

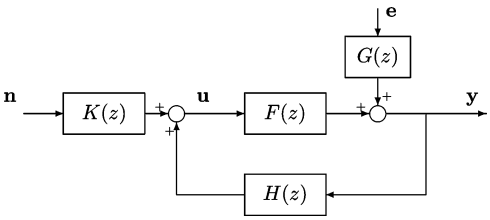

Fig. 2. Closed-loop system.

VII. Role of ν : Experimental Results

We shall consider an experimental setup as described in Fig. 2. The white noises $\mathbf{e}(t)$ and $\mathbf{n}(t)$ are independent, zero mean and unit variance.

In a first example we choose $H_1(z) = 0$ (i.e., open loop[20]),

$$F_1(z) = \frac{1}{z - 0.5}, \quad G_1(z) = \frac{z + 0.5}{z - 0.5}, \tag{VIII.1}$$

which corresponds to a first order ARMAX model

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

and $K_1(z) = (z - \alpha/z - 0.9)$.

In this first example the input spectrum (note that in open loop $\mathbf{u} = K_1(z)\mathbf{n}$) is controlled through the zero location α; three possible values of α are considered: $\alpha_1 = 0.5$ (slightly correlated input), $\alpha_2 = -0.2$ (moderately correlated input) and $\alpha_3 = -0.9$ (highly correlated input). While very simple, this example captures the fact that, for a given system, changing the input spectrum can drastically change the effect of ν.

We shall also consider two examples (example 2 and example 3) in a closed loop configuration. The details are found in Table I. All simulation results in this paper shall refer to one of these three configurations, named respectively *example 1, example 2* and *example 3*.

Let us denote with $\hat{F}(e^{j\omega}, \nu)$ the estimator of the deterministic transfer function $F(e^{j\omega}) := F(z)_{|z=e^{j\omega}}$ using the ADAPTx algorithm as a function of the future horizon ν. Denote also with $CRLB_F(j\omega)$ 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 Fig. 3 we show, respectively for the three values of α, the input spectrum and the behavior of the efficiency index[21]

$$Eff(\nu) := \frac{\int_0^{2\pi} Tr\left[\text{AsVar}\left\{\hat{F}(e^{j\omega}, \nu)\right\}\right] d\omega}{\int_0^{2\pi} Tr\left[CRLB_{\hat{F}}(j\omega)\right] d\omega} \tag{VIII.2}$$

[19]In fact, finer estimates of the convergence rate are possible. For instance, in Example 1 it can actually be verified that the rate of convergence is $\max\{\rho|\alpha|, \rho^2\}$ where $\rho = |-0.5|$ is the absolute value of the zero of $G(z)$ and α is the zero of $K(z)$.

[20]Note that the algorithms described in this paper yield consistent estimators also for closed-loop operating conditions [16]. We shall also consider a closed loop setup in Section IX.

[21]Note that the transfer function is a complete invariant; if the transfer function estimator is efficient then the estimator of any invariant will be efficient. Note also that here we only consider $F(z)$ in the definition of $Eff(\nu)$. Of course also the "stochastic" model $G(z)$ could be taken into account.

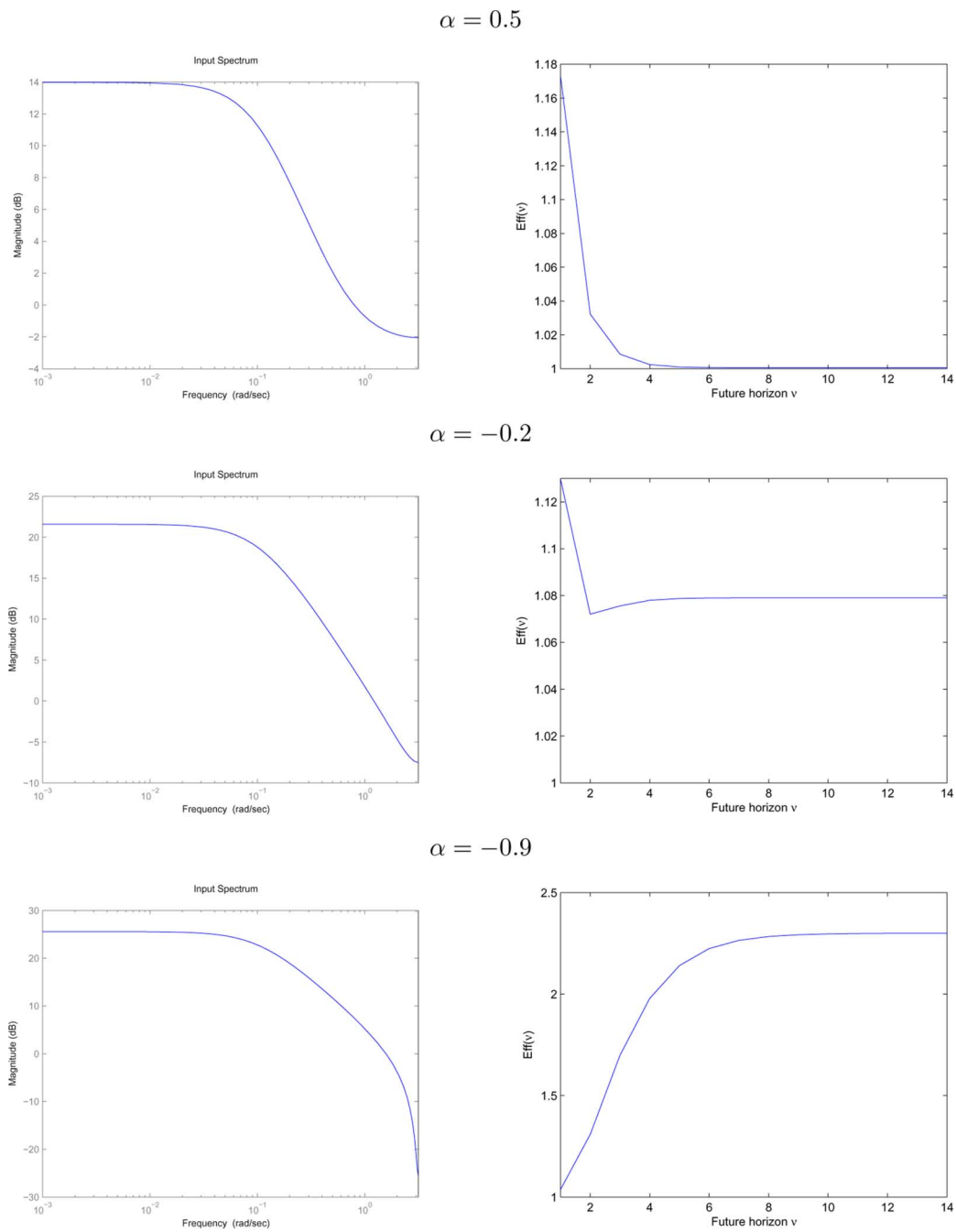

Fig. 3. Example 1. Left: Input spectrum, right: $Eff(\nu)$. Top to bottom: $\alpha = 0.5$, $\alpha = -0.2$, $\alpha = -0.9$.

as a function of ν. When performing optimization of ν shall consider the non-normalized index[22]

$$\overline{Eff}(\nu) := \int_0^{2\pi} Tr\left[\text{AsVar}\left\{\hat{F}(e^{j\omega}, \nu)\right\}\right] d\omega. \qquad \text{(VIII.3)}$$

Of course, when it will come to optimize ν, minimizing $Eff(\nu)$ or $\overline{Eff}(\nu)$ is the same.

The asymptotic variance is computed using the formulas in [9] and assuming $p = 100$.

As the zero location changes (see Fig. 3), $Eff(\nu)$ can be monotonically decreasing ($\alpha = 0.5$) monotonically increasing ($\alpha = -0.9$), or show a minimum for an intermediate value of ν ($\alpha = -0.2$).

TABLE II
OPTIMAL FUTURE HORIZON ν_{opt} AND RELATIVE EFFICIENCY

α	ν_{opt}	$Eff(\nu_{opt})$
0.9	∞	1
0.5	∞	1.0005
-0.2	2	1.0720
-0.9	1	1.0380

As summarized in Table II, 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

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

[22]Actually an estimate of $\overline{Eff}(\nu)$, see Section IX.

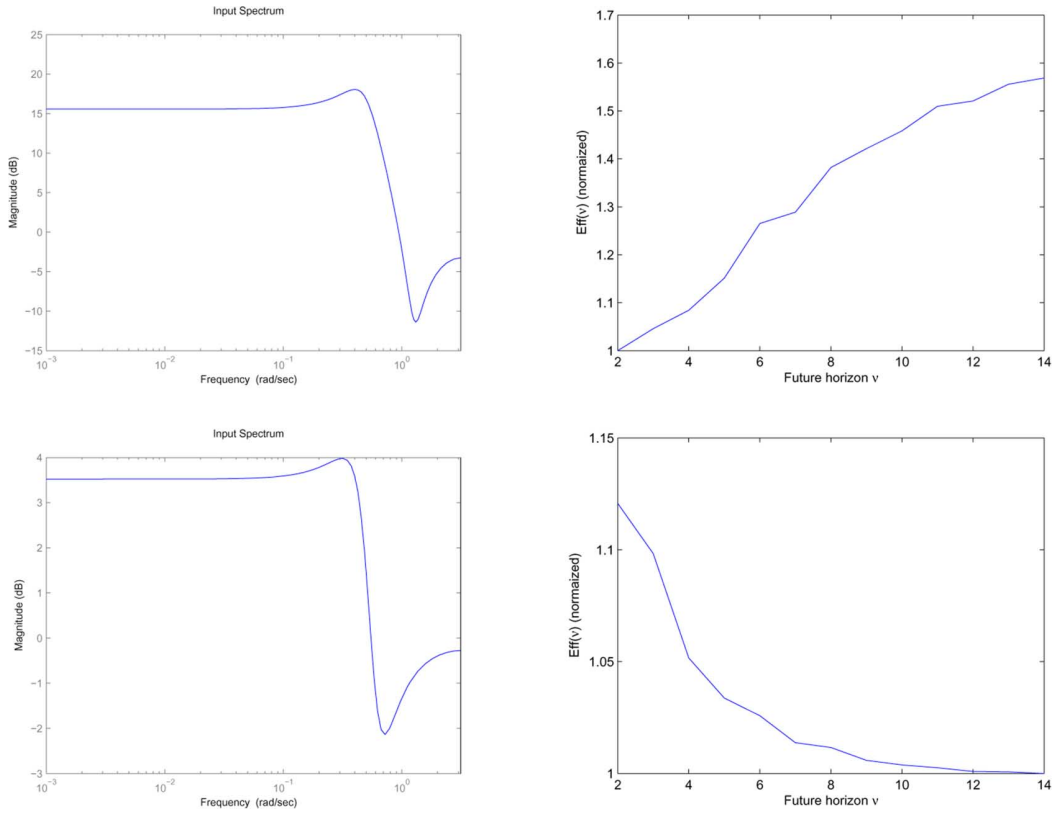

Fig. 4. Left: Reference signal spectrum. Right: (normalized) $Eff(\nu)$. Top row: example 2, bottom row: example 3.

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 ν) the computed value of $Eff(\infty)$ is 1 (within numerical precision) as expected; in fact in this case these algorithms are expected to be efficient. Note also that ADAPTx is not efficient in all the examples considered regardless of choice of ν. This provides a counterexample to the statements in [31], claiming efficiency of ADAPTx.

The reader may argue that in some of these examples the influence of ν is marginal; e.g., in Fig. 3, $\alpha = -0.2$, $Eff(\nu)$ ranges (roughly) in the interval $[1.075 \quad 1.125]$, suggesting that any choice of ν would do a good job. That this is not always the case can be seen in Fig. 3, $\alpha = -0.9$ in which $\nu = 14$ gives a performance roughly 2.3 times worse than $\nu = 1$.

In Fig. 4 we depict the index[23] $Eff(\nu)$ respectively for example 2 and example 3; the only difference is the zero location of the reference signal. Again changing the zero location $Eff(\nu)$ changes from monotonically decreasing to monotonically increasing.

VIII. Automatic Selection of ν

Unfortunately the results presented so far, while providing useful insights on the effect of the future horizon in the state construction step, do not lead to general guidelines as to how ν should be chosen when Condition 1 is not met; the value of ν depends upon the specific experimental conditions.

The purpose of this Section is to provide a general computational recipe, based on the results found in [9], [18] and the consideration reported in this paper, to optimize[24] the choice of ν. We also validate, through some Monte-Carlo simulations, the proposed procedure.

The results presented show that, indeed, optimally choosing the length of the future horizon, can considerably improve the quality of the estimators.

Remark IX.4: As pointed out by an anonymous reviewer, the reader may argue that for each value of ν there is an optimal choice of $W_A(\nu)$ (in general different from the "CCA weight"); therefore, in order to optimize performance one should first optimize $W_A(\nu)$ for each ν and then chose the optimal ν. Of course this is possible. Unfortunately, to the best of the author knowledge, there are no closed form expressions for this optimal weight, making it necessary to perform a non-linear search over the space of (non-singular) matrices $W_A(\nu)$ of dimension $m_y\nu \times m_y\nu$. Instead ν is a scalar integer parameter; making its choice a rather straightforward task.

To the author's experience, the optimal value of ν allows to obtain estimators whose (asymptotic) variance is very close to the Cramèr-Rao lower bound even for fixed $\hat{W}_A(\nu)$ as in (III.8) (or (III.9)). $\Diamond$

The procedure we suggest goes through the following steps:

a) Obtain an initial estimate of the system parameters A, B, C, K and of the innovation variance Λ. This of course

[23]Normalized so that its minimum is one since the Cramér-Rao Lower bound has not been computed in this case.

[24]W.r.t. the cost function $Eff(\nu)$. Of course other choices of cost functions are possible, but we regard this choice as irrelevant as far as this paper is concerned.

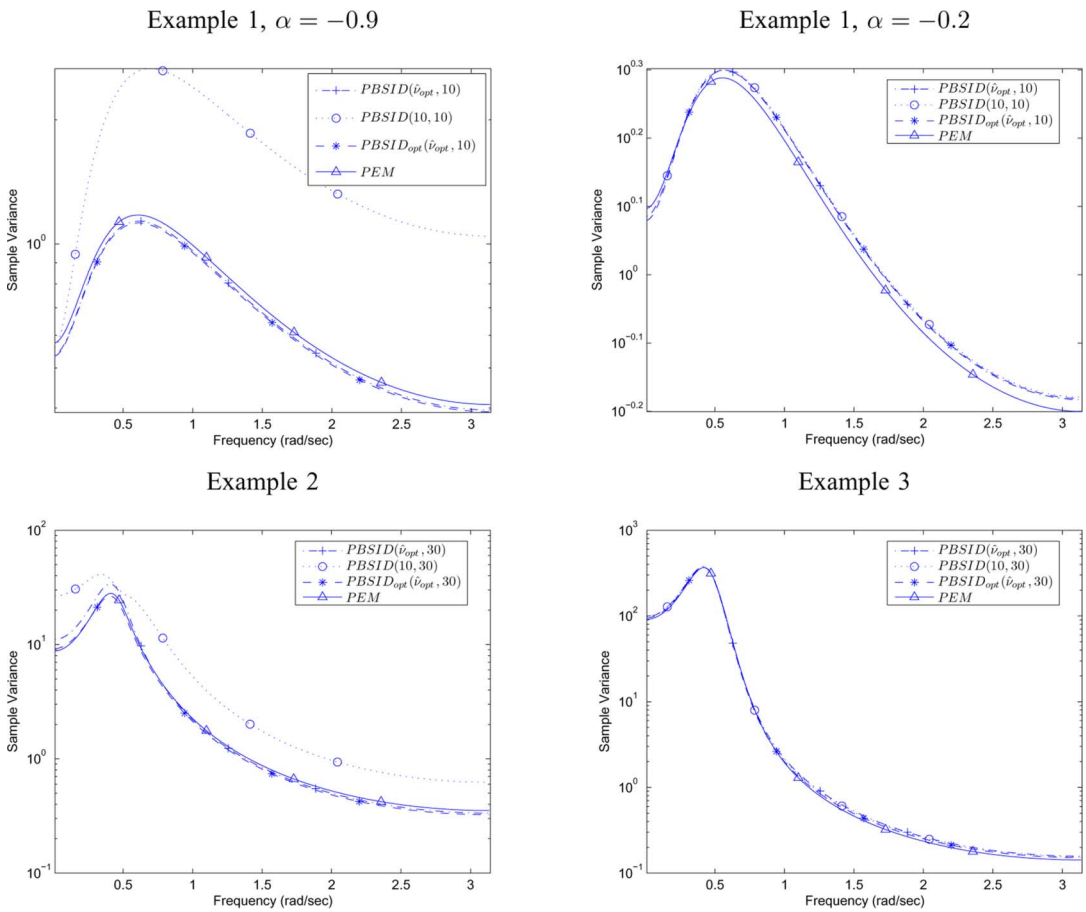

Fig. 5. Sample variance of $\hat{F}(e^{j\omega})$ (100 Monte Carlo runs). In the legend the two indexes between brackets denote, respectively the values of ν and p. The Cramér-Rao lower bound is indistinguishable from the sample variance for PEM and hence not shown.

requires choosing a candidate value of ν, which we shall denote with ν_0 from now on. As we shall argue later on the specific value of ν_0 will not play a very delicate role.

b) Using the pre-estimated parameters, the asymptotic variance $\mathrm{AsVar}\ \{\hat{F}(e^{j\omega}, \nu)\}$ can be estimated as described in [18]. We denote this estimator by $\widehat{\mathrm{AsVar}}\ \{\hat{F}(e^{j\omega}, \nu)\}$. Then the optimal value $\hat{\nu}_{opt}$ is estimated as

$$\hat{\nu}_{opt} := \arg\min_{\nu} \widehat{\overline{Eff}}(\nu) \tag{IX.1}$$

where $\widehat{\overline{Eff}}(\nu) := \int_0^{2\pi} \widehat{\mathrm{AsVar}}\ \{\hat{F}(e^{j\omega}, \nu)\} d\omega$, and the integration is performed numerically. Note that, when $Eff(\nu)$ is monotonically decreasing in ν, the optimal value would be infinity. However, as discussed in Section VI and Remark VI.4, the search for the optimal value in equation (IX.1) can be limited to "small" values of ν. In particular $\nu \leq p$ will always be used.

c) Estimate the parameters using the estimated "optimal" value $\hat{\nu}_{opt}$.

Remark IX.5: Note that this procedure is an instance of a commonly used two-step approach in statistical inference which prescribes using preliminary estimated quantities when constructing an estimator. A simple yet important example is the construction of an (asymptotic) Markov estimator [42] using an estimate of the noise variance rather than the true (but unknown) one. ◊

In Section IX-A, together with some practical considerations, we shall illustrate this procedure through some Monte-Carlo experiments. We stress that in performing these Monte-Carlo experiments (and in particular estimating ν_{opt}) *no prior information* is used concerning the true system except for the system order which is assumed to be known.

A. Simulation Results and Practical Considerations

Since, as mentioned in the introduction, a whole family of algorithms are asymptotically equivalent, we report the simulation results for the PBSID algorithm [10], [16]; we shall use the notation $\mathrm{PBSID}(\nu, p)$ meaning that the PBSID algorithm has been run with the indicated user parameters ν and p. We shall also report results using the "optimized" version of PBSID_{opt} (see [10], [11]); this algorithm improves in terms of asymptotic variance over PBSID while also being computationally advantageous [11].

We compare (via the sample variance through 100 Monte-Carlo simulations) $\mathrm{PBSID}(10, p)$, $\mathrm{PBSID}(\hat{\nu}_{opt}, p)$ and $\mathrm{PBSID}_{opt}(\hat{\nu}_{opt}, p)$. The value $\hat{\nu}_{opt}$ is computed by the procedure described above (see (IX.1)) *for each* Monte Carlo run. It is clear that in Fig. 5 ($\alpha = -0.9$) a considerable improvement (about a factor 2) is obtained by optimally choosing ν. On the contrary, in Fig. 5 (top right) ($\alpha = -0.2$, $Eff(\nu)$ almost flat) there is essentially no gain in performing the "optimal" choice of ν.

TABLE III
EXAMPLE 1. EFFECT OF THE ESTIMATION OF ν_{opt}: AVERAGE AND STANDARD DEVIATION (STD.) OF THE EFFICIENCY NUMBERS $Eff(\hat{\nu}_{opt})$ AND OF THE ESTIMATED OPTIMAL HORIZON $\hat{\nu}_{opt}$ OVER ALL TRIALS: 10 MONTE CARLO RUNS FOR EACH PAIR (ν_0, N), $\nu_0 \in [1,9]$ AND $N \in [500,1000,1500,2000]$ (TOTAL 360 TRIALS)

Input	mean(Eff)	std(Eff)	mean($\hat{\nu}_{opt}$)	std($\hat{\nu}_{opt}$)
$\alpha = -0.2$	1.0740	0.0022	2.6056	1.3226
$\alpha = -0.9$	1.0372	0	1	0

Some considerations on the procedure above are now in order. First of all one might ask whether the procedure is particularly sensitive to the value of ν_0 in the first step. Since the estimators are consistent regardless of the value of ν_0, it is reasonable to assume that the estimators will provide a decent approximation of the "true" system. It should also be observed that $Eff(\nu)$ is a continuous function of the system parameters; therefore, if we denote by $\hat{\nu}_{opt}$ the estimated optimal value obtained by optimizing the function $\widehat{Eff}(\nu)$, it is to be expected that $Eff(\nu_{opt}) \simeq Eff(\hat{\nu}_{opt})$. Note that, depending upon the sensitivity of $Eff(\nu)$ upon ν, the actual value of $\hat{\nu}_{opt}$ may vary considerably.

These, we have to admit rather qualitative, statements are indeed supported by the experimental results. Consider example 1 (see Table I) with $\alpha = -0.2$ and $\alpha = -0.9$. For each pair of values $\nu_0 \in [1,9]$ and $N \in [500,1000,1500,2000]$ we have performed 10 Monte Carlo experiments, obtaining overall 360 estimated values $\hat{\nu}_{opt}$. In both cases, see Table III, $\alpha = -0.2$ (top) and $\alpha = -0.9$ (bottom), the average of the efficiency numbers $Eff(\hat{\nu}_{opt})$ is very close to the "optimal" efficiency $Eff(\nu_{opt})$ and also the standard deviation of $Eff(\hat{\nu}_{opt})$ is rather small (zero for $\alpha = -0.9$). Instead, while for $\alpha = -0.9$ $\nu_{opt} = 1$ is always estimated correctly, for $\alpha = -0.2$ the estimated optimal value shows a significant variability and $\nu_{opt} = 2$ is slightly overestimated (compare also with Table II).

For sake of illustration we also consider two closed loop systems, example 2 and example 3 in Table I. Again we have chosen somewhat arbitrarily $\nu = 10$ and compared with the results obtained using $\hat{\nu}_{opt}$. Note that for example 3, $\nu = 10$ (see Fig. 5) gives essentially efficient estimators (indistinguishable from PEM) and hence no improvement can be expected by optimizing ν. Instead when we consider example 2, $\nu = 10$ is certainly not optimal (see Fig. 4) and therefore, as expected, optimizing ν should (and does) improve performance (of about a factor 2.5), see Fig. 5.

IX. CONCLUSION

In this paper it has been shown that the ADAPTx algorithm presented in [31] and the SSARX method in [26] are asymptotically equivalent. Using this fact and the results in [10], [11] 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). In particular Theorem 5.1 extends, under Condition 1, the results of [7]. We regard this extension as non trivial since, even though this paper uses in part methodologies from [7], the setup we consider includes closed loop identification; moreover we hope that the qualitative analysis of the role of ν on the state estimation step will give inspiration for future research.

As a byproduct we have also extended the results of [7] showing that, under Condition 1 the CCA weighting is optimal (see Proposition 7.1).

As also suggested in [7] we have discussed, and demonstrated on a few examples, a computational procedure based on some recent results on the asymptotic distribution of subspace estimators [9], [18] which allows in principle to chose the "optimal" value of the future horizon for a given set of data.

APPENDIX

Lemma A.1: With the choice of Π and Π_c in (V.8)

$$\hat{\Sigma}_{\mathbf{x}\tilde{\mathbf{x}}} = \underline{o}_P\left(\frac{1}{\sqrt{N}}\right) \tag{A1}$$

where $\tilde{X}_t$ is defined in (V.11).

Proof: For a (fixed) $\bar{\mathcal{C}}_p$ (of size $n \times (m_u + m_y)p$), analogously to (V.11), define $\tilde{\mathbf{x}}(t) := \bar{\mathcal{C}}_p \Pi_c \mathbf{z}_{[t_0,t)}$. Then, using $\mathbf{x}(t) \doteq \mathcal{C}_p \mathbf{z}_{[t_0,t)}$

$$\mathbb{E}\mathbf{x}(t)\tilde{\mathbf{x}}^\top(t) \doteq \mathcal{C}_p \Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}} \Pi_c^\top \bar{\mathcal{C}}_p^\top = 0 \tag{A2}$$

holds. This essentially states that the matrix Π_c allows to decouple the state estimate in uncorrelated components. The implication of this fact is that the sample covariance $\hat{\Sigma}_{\mathbf{x}\tilde{\mathbf{x}}}$ will be $\underline{o}_P(1/\sqrt{N})$. In fact, for any estimator $\hat{\mathcal{C}}_p$ the state $\hat{X}_t := \hat{\mathcal{C}}_p Z_{[t_0,t)}$ can be decoupled as above in the form $\hat{X}_t := \hat{T} X_t + \tilde{X}_t$ where $\hat{T} X_t \doteq \hat{T} \mathcal{C}_p Z_{[t_0,t)} := \hat{\mathcal{C}}_p \Pi Z_{[t_0,t)}$ and $\tilde{X}_t := \hat{\mathcal{C}}_p \Pi_c Z_{[t_0,t)}$. Equation (A2) implies that

$$\frac{X_t \tilde{X}_t^\top}{N} = \hat{\Sigma}_{\mathbf{x}\tilde{\mathbf{x}}} \doteq \mathcal{C}_p \hat{\Sigma}_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}} \Pi_c^\top \hat{\mathcal{C}}_p^\top = \underline{o}_P\left(\frac{1}{\sqrt{N}}\right) \tag{A3}$$

the last equality follows from the fact that both $\mathcal{C}_p \hat{\Sigma}_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}} \Pi_c^\top$ and $\hat{\mathcal{C}}_p \Pi_c$ are certainly $\underline{o}_P(N^{-1/2+\epsilon})$, for any choice of $\epsilon > 0$. Since these matrices have p columns, the product in (A3) is of order $\underline{o}_P(pN^{-1+2\epsilon}) = \underline{o}_P(1/\sqrt{N})$ where the last equality follows from Assumption 1 that $p = o((logN)^\alpha)$. This concludes the proof. ∎

Lemma A.2: Under Condition 1, the errors $\mathrm{vec}(\hat{A} - A)$, $\mathrm{vec}(\hat{B} - B)$, $\mathrm{vec}(\hat{C} - C)$, $\mathrm{vec}(\hat{K} - K)$ are asymptotically equivalent to terms of the form

$$M_1 \mathrm{vec}\left(\hat{\Sigma}_{\mathbf{e}\mathbf{z}_{[t_0,t)}} \mathcal{C}_p^\top\right) + M_2 \mathrm{vec}\left(\tilde{\mathcal{C}}_p^\nu \Pi_c \Upsilon_p\right) = M_1 \tilde{\theta}_1(N) + M_2 \tilde{\theta}_2(N,\nu) \tag{A4}$$

for suitable[25] M_1, M_2 and where

$$\Upsilon_p = \left[\Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}} \mathbb{E}\mathbf{z}_{[t_0,t)}\mathbf{z}^\top(t)\right] \begin{bmatrix} 0_{m_z \times n} \\ \mathcal{C}_p^\top \end{bmatrix} N_1 + \begin{bmatrix} 0_{(p-1)m_z \times m_z} \\ I_{m_z \times m_z} \end{bmatrix} N_2 \tag{A5}$$

[25]The matrices M_1 and M_2 have as many rows as the elements of $\mathrm{vec}\left(\hat{A} - A\right)$, $\mathrm{vec}\left(\hat{B} - B\right)$, $\mathrm{vec}\left(\hat{C} - C\right)$, $\mathrm{vec}\left(\hat{K} - K\right)$, i.e., $n^2 + n(m+p) + np$.

for matrices N_1 and N_2 which expressions are irrelevant to our purposes. The columns of Υ, the limit of Υ_p as $N \to \infty$, are in ℓ^2 thanks to the assumptions on the spectrum $\Phi_{\mathbf{z}}$.

Proof: First of all note that, from (V.11), $\hat{X}_t = \tilde{\mathcal{C}}_p^{\nu} Z_{[t_0,t)}$ and $\hat{X}_{t+1} = \tilde{\mathcal{C}}_p^{\nu} Z_{[t_0+1,t+1)}$. Let us first consider $\tilde{C} := \hat{C} - C$; from $Y_t = CX_t + E_t = C\hat{X}_t - C\tilde{X}_t + E_t$ and using (III.2), (A1) it follows that:

$$\tilde{C} = \left[\hat{\Sigma}_{\mathbf{ex}} - C\hat{\Sigma}_{x\mathbf{x}}\right]\hat{\Sigma}_{\mathbf{xx}}^{-1} \doteq \hat{\Sigma}_{\mathbf{ex}}\Sigma_{\mathbf{xx}}^{-1} \tag{A6}$$

here Lemma A.1 has been used in the last equation. Let us now consider $\tilde{A} := \hat{A} - A$, $\tilde{B}$ and $\tilde{K}$. From $X_{t+1} = AX_t + BU_t + KE_t$ it follows that:

$$\hat{X}_{t+1} = A\hat{X}_t + BU_t + KE_t + \tilde{X}_{t+1} - A\tilde{X}_t.$$

Using now $E_t = \hat{E}_t + \hat{C}\hat{X}_t - CX_t \doteq \hat{E}_t + \tilde{C}\hat{X}_t + C\tilde{X}_t$ it follows that:

$$\hat{X}_{t+1} = (A + K\tilde{C})\hat{X}_t + BU_t + K\hat{E}_t + \tilde{X}_{t+1} - \bar{A}\tilde{X}_t.$$

Let us define $\mathbf{s}(t) := \left[\mathbf{x}^\top(t) \quad \mathbf{u}^\top(t) \quad \mathbf{e}^\top(t)\right]^\top$ and similarly S_t. For convenience of notation let us use $\mathbf{x}'$ as a subscript when referring to $\mathbf{x}(t+1)$, e.g., $\hat{\Sigma}_{\mathbf{x}'\mathbf{s}} = \hat{X}_{t+1}S_t^\top/N$. Since A, B and K are estimated solving (III.2) it follows that the errors are of the form:

$$\left[\tilde{A} \quad \tilde{B} \quad \tilde{K}\right] \doteq \left[K\tilde{C} \quad 0 \quad 0\right] + \hat{\Sigma}_{x'\mathbf{s}}\Sigma_{\mathbf{ss}}^{-1} - \bar{A}\hat{\Sigma}_{x'\mathbf{s}}\Sigma_{\mathbf{ss}}^{-1}. \tag{A7}$$

Using Lemma A.1 we have $\hat{\Sigma}_{x\mathbf{x}} \doteq 0$; from $\mathbb{E}\mathbf{z}_{[t_0,t)}\mathbf{e}^\top(t) = 0$ also $\hat{\Sigma}_{x\mathbf{e}} \doteq 0$ follows. It is now crucial to observe that under Condition 1, and using (V.2), the input process $\mathbf{u}$ can be written, in terms of the state $\mathbf{x}(t)$ as

$$\mathbf{u}(t) = \bar{C}_u\mathbf{x}(t) + D_u\mathbf{e}(t) + \mathbf{n}(t) \tag{A8}$$

where $\bar{C}_u := C_uT_{xu} + D_uC$. This, again using Lemma A.1, implies $\hat{\Sigma}_{x\mathbf{u}} \doteq 0$. Therefore $\hat{\Sigma}_{x\mathbf{s}} \doteq 0$. It remains to compute the quantities related to $\hat{\Sigma}_{x'\mathbf{s}} = \left[\hat{\Sigma}_{x'\mathbf{x}} \quad \hat{\Sigma}_{x'\mathbf{u}} \quad \hat{\Sigma}_{x'\mathbf{e}}\right]$

$$\begin{aligned}
\hat{\Sigma}_{x'\mathbf{x}} &\doteq \tilde{\mathcal{C}}_p^{\nu}\Pi_c\mathbb{E}\mathbf{z}_{[t_0+1,t+1)}\mathbf{z}_{[t_0,t)}^\top\mathcal{C}_p^\top \\
&= \tilde{\mathcal{C}}_p^{\nu}\Pi_c\left[\Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}} \quad \mathbb{E}\mathbf{z}_{[t_0,t)}\mathbf{z}^\top(t)\right] \cdot \begin{bmatrix} 0_{m_z\times n} \\ \mathcal{C}_p^\top \end{bmatrix} \\
\hat{\Sigma}_{x'\mathbf{u}} &\doteq \hat{\Sigma}_{x'\mathbf{x}}\bar{C}_u^\top + \tilde{\mathcal{C}}_p^{\nu}\Pi_c\mathbb{E}\mathbf{z}_{[t_0+1,t+1)}\mathbf{e}^\top(t)D_u^\top \\
&\quad + \tilde{\mathcal{C}}_p^{\nu}\Pi_c\mathbb{E}\mathbf{z}_{[t_0+1,t+1)}\mathbf{n}^\top(t) \\
\hat{\Sigma}_{x'\mathbf{e}} &\doteq \tilde{\mathcal{C}}_p^{\nu}\Pi_c\mathbb{E}\mathbf{z}_{[t_0+1,t+1)}\mathbf{e}^\top(t)
\end{aligned}$$

where

$$\mathbb{E}\mathbf{e}(t)\mathbf{z}_{[t_0+1,t+1)}^\top = \Lambda\left[0_{m_y\times(p-1)m_z} \quad D_u^\top \quad I_{m_y}\right]$$

and

$$\mathbb{E}\mathbf{n}(t)\mathbf{z}_{[t_0+1,t+1)}^\top = \Lambda_{\mathbf{n}}\left[0_{m_u\times(p-1)m_z} \quad I_{m_u} \quad 0_{m_u\times m_y}\right].$$

Since, as shown above, $\hat{\Sigma}_{x\mathbf{s}} \doteq 0$, using (A7) and (A6) if follows that the errors $\tilde{A}$, $\tilde{B}$, $\tilde{K}$, $\tilde{C}$ are (asymptotically equivalent to) linear combinations of the columns of the matrices $\hat{\Sigma}_{x'\mathbf{x}}$, $\hat{\Sigma}_{x'\mathbf{u}}$, $\hat{\Sigma}_{x'\mathbf{e}}$ and $\hat{\Sigma}_{\mathbf{ex}} \doteq \hat{\Sigma}_{\mathbf{e}\mathbf{z}_{[t_0,t)}}\mathcal{C}_p^\top$. This implies (A4) with Υ_p as defined in (A5), concluding the proof. ∎

Lemma A.3: Under Condition 1 the errors $\tilde{\theta}_1(N)$ and $\tilde{\theta}_2(N,\nu)$ are asymptotically uncorrelated.

Proof: For simplicity let us define[26] $\bar{\mathbf{v}}(t,\nu) := T_\nu S_n^{-1}(\nu)U_n^\top(\nu)\bar{\mathbf{e}}_{[t,T]}$. For future use note that $\mathbb{E}[\bar{\mathbf{v}}(t,\nu)\bar{\mathbf{v}}^\top(t,\nu) \mid \mathcal{Z}_{t-1}^-] = \mathbb{E}\bar{\mathbf{v}}(t,\nu)\bar{\mathbf{v}}^\top(t,\nu) = T_\nu S_n^{-2}(\nu)T_\nu^\top$. After some simple but tedious matrix manipulations it can be seen that $\mathrm{AsCov}\left\{\sqrt{N}\tilde{\theta}_1(N), \sqrt{N}\tilde{\theta}_2(N,\nu)\right\}$ has the form

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=0}^{N-1}\sum_{j=\max(0,i-\nu)}^{i}\Big[\left(\mathcal{C}_p\mathbb{E}\mathbf{z}_{[t_0+i,t+i)}\mathbf{z}_{[t_0+j,t+j)}^\top\bar{\Upsilon}_p\right) \otimes \left(\mathbb{E}\mathbf{e}_{t+i}\bar{\mathbf{v}}^\top(t+j,\nu)\right)\Big]$$

where $\bar{\Upsilon}_p := \Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}^{-1}\Pi_c\Upsilon_p$; also the property that

$$\mathrm{vec}(ab^\top)\mathrm{vec}^\top(cd^\top) = (bd^\top)\otimes(ac^\top)$$

for column vectors a, b, c, d has been used. The summation is limited to $\max(0,i-\nu) \le j \le i$ from the martingale difference property of $\mathbf{e}(t)$, i.e., $\mathbb{E}\left[\mathbf{e}(t) \mid \mathcal{Z}_{t-1}^-\right] = 0$, which also implies $\mathbb{E}\left[\bar{\mathbf{v}}(t,\nu) \mid \mathcal{Z}_{t-1}^-\right] = 0$.

For $i = j$ we have that $\mathcal{C}_p\mathbb{E}\mathbf{z}_{[t_0+i,t+i)}\mathbf{z}_{[t_0+i,t+i)}^\top\Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}^{-1}\Pi_c\Upsilon_p = \mathcal{C}_p\Pi_c\Upsilon_p = 0$. It remains to consider the case $i > j$. Note that, under Condition 1, there exists a matrix Γ_{i-j} such that

$$\begin{aligned}
&\mathcal{C}_p\mathbb{E}\mathbf{z}_{[t_0+i,t+i)}\mathbf{z}_{[t_0+j,t+j)}^\top\bar{\Upsilon}_p \\
&\quad = [0_{n\times(i-j)m_z}\mathcal{C}_p]\mathbb{E}\mathbf{z}_{[t_0+j,t+i)}\mathbf{z}_{[t_0+j,t+j)}^\top\bar{\Upsilon}_p \\
&\quad \doteq \mathcal{C}_{p+i-j}\begin{bmatrix} \mathbb{E}\mathbf{z}_{[t_0+j,t+j)}\mathbf{z}_{[t_0+j,t+j)}^\top\bar{\Upsilon}_p \\ \Gamma_{i-j}\mathcal{C}_p\mathbb{E}\mathbf{z}_{[t_0+j,t+j)}\mathbf{z}_{[t_0+j,t+j)}^\top\bar{\Upsilon}_p \end{bmatrix} \\
&\quad \doteq \mathcal{C}_{p+i-j}\begin{bmatrix} I_{pm_z} \\ 0_{(i-j)m_z\times pm_z} \end{bmatrix}\Pi_c\Upsilon_p
\end{aligned}$$

where $\mathcal{C}_{p+i-j} \doteq \left[0_{n\times(i-j)m_z} \quad \mathcal{C}_p\right]$ has been used. From $\mathcal{C}_p\Pi_c = 0$, it follows that (see, for more details, [7]):

$$\lim_{N\to\infty}\mathcal{C}_p\mathbb{E}\mathbf{z}_{[t_0+i,t+i)}\mathbf{z}_{[t_0+j,t+j)}^\top\Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}^{-1}\Pi_c\Upsilon_p = 0$$

which proves that also for $i > j$ the terms in the sum are zero, concluding the proof. ∎

Lemma A.4: Under Condition 1 the asymptotic variance $V_{\theta_2}(\nu) = \mathrm{AsVar}\left\{\sqrt{N}\tilde{\theta}_2(N,\nu)\right\}$ can be written in the form

$$V_{\theta_2}(\nu) = \lim_{N\to\infty}\left(\Upsilon_p^\top\Pi_c^\top\Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}^{-1}\Pi_c\Upsilon_p\right) \otimes \left(T_\nu S_n^{-2}(\nu)T_\nu^\top\right). \tag{A9}$$

Proof: The proof of this lemma follows very closely [7Section 4.1]. However, thanks to the different choice of the projection Π_c, it is not necessary to separate the cases $n \le m_y + m_u = m_z$ and $n > m_y + m_u = m_z$.

After some simple but tedious matrix manipulations, defining $\bar{\Upsilon}_p := \Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}^{-1}\Pi_c\Upsilon_p$, it can be seen that $V_{\theta_2}(\nu)$ has the form

$$\lim_{N\to\infty}\frac{1}{N}\sum_{\substack{i,j=0 \\ |i-j|\le\nu}}^{N-1}\left(\bar{\Upsilon}_p^\top\mathbb{E}\mathbf{z}_{[t_0+i,t+i)}\mathbf{z}_{[t_0+j,t+j)}^\top\bar{\Upsilon}_p\right)$$

[26]This proof follows very closely the [7, proof of Lemma 4].

$$\otimes \mathbb{E}\bar{\mathbf{v}}(t+i,\nu)\bar{\mathbf{v}}(t+j,\nu). \quad \text{(A10)}$$

The summation is limited to $|i-j| \leq \nu$ since the other terms vanish due the martingale difference property of $\mathbf{e}(t)$ as in the previous Lemma.

Let us now consider the case $i > j$, being $i < j$ completely symmetric. As in the previous Lemma, under Condition 1

$$\begin{aligned}
&\bar{\Upsilon}_p^\top \mathbb{E}\mathbf{z}_{[t_0+i,t+i)}\mathbf{z}_{[t_0+j,t+j)}^\top \bar{\Upsilon}_p^\top \\
&\quad \doteq \bar{\Upsilon}_{p+i-j}^\top \begin{bmatrix} \mathbb{E}\mathbf{z}_{[t_0+j,t+j)}\mathbf{z}_{[t_0+j,t+j)}^\top \bar{\Upsilon}_p \\ \Gamma_{i-j}\mathcal{C}_p\Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}\bar{\Upsilon}_p + \underline{o}\left(\frac{1}{\sqrt{N}}\right) \end{bmatrix} \\
&\quad \doteq \bar{\Upsilon}_{p+i-j}^\top \begin{bmatrix} I_{pm_z} \\ 0_{(i-j)m_z \times pm_z} \end{bmatrix} \Pi_c \Upsilon_p.
\end{aligned}$$

where $\mathbb{E}\mathbf{z}(t)\mathbf{z}_{[t_0-(i-j),t)}^\top \doteq \begin{bmatrix} 0_{n\times(i-j)m_z} & \mathbb{E}\mathbf{z}(t)\mathbf{z}_{[t_0,t)}^\top \end{bmatrix}$ and $\mathcal{C}_{p+(i-j)}^\top \doteq \begin{bmatrix} 0_{n\times(i-j)m_z} & \mathcal{C}_p^\top \end{bmatrix}$ have been used. Define $\bar{\Upsilon}^\top := \Upsilon^\top \Pi_c^\top \Sigma_{\mathbf{z}^-\mathbf{z}^-}^{-1}$, where $\Sigma_{\mathbf{z}^-\mathbf{z}^-} := \lim_{N\to\infty}\Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}$. When $N \to \infty$ (and hence $p \to \infty$), the last matrix on the right hand side converges to

$$\begin{aligned}
&\Upsilon^\top \Pi_c^\top \Sigma_{\mathbf{z}^-\mathbf{z}^-}^{-1} \begin{bmatrix} I_\infty \\ 0_{(i-j)m_z\times\infty} \end{bmatrix} \Pi_c \Upsilon \\
&\quad = \Upsilon^\top \Sigma_{\mathbf{z}^-\mathbf{z}^-}^{-1} \begin{bmatrix} I_\infty \\ 0_{(i-j)m_z\times\infty} \end{bmatrix} \Pi_c \Upsilon \\
&\qquad \pm \Upsilon^\top \mathcal{C}_\infty^\top \left(\mathcal{C}_\infty \Sigma_{\mathbf{z}^-\mathbf{z}^-} \mathcal{C}_\infty^\top\right)^{-1} \mathcal{C}_\infty \begin{bmatrix} I_\infty \\ 0_{(i-j)m_z\times\infty} \end{bmatrix} \Pi_c \Upsilon \\
&\quad = \Upsilon^\top \begin{bmatrix} \Sigma_{\mathbf{z}^-\mathbf{z}^-}^{-1} \\ 0_{(i-j)m_z\times\infty} \end{bmatrix} \Pi_c \Upsilon \qquad \text{(A11)}
\end{aligned}$$

where the first equality follows from substituting the definition of Π_c while the second using the matrix inversion lemma; the second equality sign represents also the main difference with [7, eq. (8)]; in fact, thanks to the different choice of the projection (here Π_c, $\mathcal{P}_K$ in [7]) the last equality follows for any value of n and not only for $n \leq m_z$.

Using again the same argument used in [7] it is now easy to see that the last term in (A11) (and hence the first) is equal to zero for $\Upsilon = \lim_{N\to\infty}\Upsilon_p$ as defined in equation (A5); this holds for any choice of N_1, N_2.

Here, rather than the white noise assumption used in [7], Condition 1 is used to show $\mathbb{E}\mathbf{z}_{[t_0,t)}\mathbf{z}^\top(t) \doteq \Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}\mathcal{C}_p^\top \Gamma_1^\top$ for a suitable matrix Γ_1.

This shows that the terms in the summation (A10) vanish asymptotically for $i > j$. Symmetrically the same holds for $i < j$, so that only the term for $i = j$ matters, concluding the proof.

Even though this does not play any role in the proof of Theorem 5.1, it should also be noted that also the term for $i = j$ vanishes if

$$\Upsilon = \left[\Sigma_{\mathbf{z}_{[t_0,t)}\mathbf{z}_{[t_0,t)}}\mathbb{E}\mathbf{z}_{[t_0,t)}\mathbf{z}^\top(t)\right] \begin{bmatrix} 0_{m_z\times n} \\ \mathcal{C}_p^\top \end{bmatrix}.$$

Therefore, to the purpose of variance computation, only

$$\Upsilon = \begin{bmatrix} 0_{(p-1)m_z\times m_z} \\ I_{m_z} \end{bmatrix} N_2$$

can be considered, simplifying the variance expressions (under Condition 1). ■

Lemma A.5: Let us denote with $\hat{H}_\nu^s$ the estimate of the matrix H_ν^s obtained from the VARX estimators $\hat{\Phi}_k^y$, $\hat{\Phi}_k^u$. Then

$$\hat{H}_\nu^s = \left(I - \hat{\bar{H}}_\nu^y\right)^{-1} \quad \hat{H}_\nu^d = \hat{H}_\nu^s \hat{\bar{H}}_\nu^u$$

hold.

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

Proof of Proposition 6.1: It is sufficient to show that the terms $T_\nu S_n^{-1}(\nu)\hat{X}_t^e$ and $T_\nu S_n^{-1}(\nu)\hat{X}_t^u$ in (V.1) do not change when increasing $\nu > k$.

First of all note under the assumption $\bar{A}^k = 0$ the observability matrix $\bar{\mathcal{O}}_\nu$ has the structure

$$\bar{\mathcal{O}}_\nu^\top = [C^\top \quad \bar{A}^\top C^\top \quad \ldots \quad (\bar{A}^\top)^{k-1} \quad 0 \quad \ldots \quad 0].$$

It is also easy to verify that $(H_\nu^s)^{-1}\mathcal{O}_\nu = \bar{\mathcal{O}}_\nu$. Therefore, from (IV.4) $(H_\nu^s)^{-1}\mathcal{O}_\nu = (I\otimes\Lambda)^{1/2}U_n(\nu)S_n(\nu) = \bar{\mathcal{O}}_\nu$ and therefore also $(I\otimes\Lambda)^{1/2}U_n(\nu)S_n(\nu)$ has the structure[27]

$$\begin{aligned}
&\left((I\otimes\Lambda)^{1/2}U_n(\nu)S_n(\nu)\right)^\top \\
&\qquad = \Big[\underbrace{* \quad * \quad \ldots \quad *}_{k} \quad \underbrace{0 \quad \ldots \quad 0}_{\nu-k}\Big]
\end{aligned}$$

and the elements denoted with stars do not change as a function of $\nu \geq k$. This also implies that T_ν in (IV.4) can be taken, w.l.o.g., equal to the identity matrix for $\nu \geq k$. 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

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

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

$$S_n(\nu)U_n^\top(\nu)W_A^{-1}(\nu) = \Big[\underbrace{* \quad * \quad \ldots \quad *}_{k} \quad \underbrace{0 \quad \ldots \quad 0}_{\nu-k}\Big]$$

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 ν larger than k, i.e., $S_n(\nu) = S_n(k)$, $\nu \geq k$.

Using these considerations it follows that:

a) $T_\nu S_n^{-1}(\nu)\tilde{X}_t^e$ does not vary for $\nu \geq k$.
b) Using the lower triangular structure of $\tilde{H}_\nu^d$ it follows that $T_\nu S_n^{-1}(\nu)\tilde{X}_t^u$ 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. □

REFERENCES

[1] H. Akaike, "Markovian representation of stochastic processes by canonical variables," *SIAM J. Control*, vol. 13, pp. 162–173, 1975.

[27]The "stars" and zeros in these formulas have dimension $n \times m_y$.

[2] H. Akaike, "Canonical correlation analysis of time series and the use of an information criterion," in *System Identification: Advances and Case Studies*, R. Mehra and D. Lainiotis, Eds. New York: Academic Press, 1976, pp. 27–96.
[3] D. Bauer, "Order estimation for subspace methods," *Automatica*, vol. 37, pp. 1561–1573, 2001.
[4] D. Bauer, "Asymptotic properties of subspace estimators," *Automatica*, vol. 41, pp. 359–376, 2005.
[5] D. Bauer, "Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs," *J. Time Series Anal.*, vol. 26, pp. 631–668, 2005.
[6] D. Bauer, "Estimating linear dynamical systems using subspace methods," *Econometric Theory*, vol. 21, pp. 181–211, 2005.
[7] D. Bauer and L. Ljung, "Some facts about the choice of the weighting matrices in Larimore type of subspace algorithm," *Automatica*, vol. 38, pp. 763–773, 2002.
[8] P. Caines and C. Chan, "Estimation, identification and feedback," in *System Identification: Advances and Case Studies*, R. Mehra and D. Lainiotis, Eds. New York: Academic, 1976, pp. 349–405.
[9] A. Chiuso, "Asymptotic variance of closed-loop subspace identification algorithms," *IEEE Trans. Autom. Control*, vol. 51, no. 8, pp. 1299–1314, Aug. 2006.
[10] A. Chiuso, "On the relation between CCA and predictor-based subspace identification," *IEEE Trans. Autom. Control*, vol. 52, no. 10, pp. 1795–1812, Oct. 2007.
[11] A. Chiuso, "The role of vector autoregressive modeling in predictor based subspace identification," *Automatica*, vol. 43, no. 6, pp. 1034–1048, Jun. 2007.
[12] A. Chiuso, "Some insights on the choice of the future horizon in CCA-type subspace algortihms," in *Proc. ACC*, 2007, pp. 840–845.
[13] A. Chiuso and G. Picci, "Constructing the state of random processes with feedback," in *Proc. IFAC Int. Symp. Syst. Ident. (SYSID)*, Rotterdam, The Netherlands, Aug. 2003, pp. 855–860.
[14] A. Chiuso and G. Picci, *Geometry of Oblique Splitting, Minimality and Hankel Operators*, ser. Lect. Notes in Control and Information Sciences, A. Rantzer and C. Byrnes, Eds. New York: Springer, 2003, pp. 85–124.
[15] A. Chiuso and G. Picci, "On the ill-conditioning of subspace identification with inputs," *Automatica*, vol. 40, no. 4, pp. 575–589, 2004.
[16] A. Chiuso and G. Picci, "Consistency analysis of some closed-loop subspace identification methods," *Automatica*, vol. 41, no. 3, pp. 377–391, 2005.
[17] A. Chiuso and G. Picci, "Prediction error vs. subspace methods in closed-loop identification," in *Proc. 16th IFAC World Congress*, Prague, Czech Republic, Jul. 2005 [Online]. Available: http://www.ifac-papersonline.net
[18] A. Chiuso and G. Picci, "Estimating the asymptotic variance of closed loop subspace estimators," in *Proc. SYSID'06*, Newcastle, Australia, 2006 [Online]. Available: http://www.ifac-papersonline.net
[19] M. Deistler, K. Peternell, and W. Scherrer, "Consistency and relative efficiency of subspace methods," *Automatica*, vol. 31, no. 12, pp. 1865–1875, 1995.
[20] T. Ferguson, *A Course in Large Sample Theory*. London, U.K.: Chapman and Hall, 1996.
[21] G. Golub and C. Van Loan, *Matrix Computation*, 2nd ed. Baltimore, MA: The Johns Hopkins Univ. Press, 1989.
[22] C. Granger, "Economic processes involving feedback," *Inform. Control*, vol. 6, pp. 28–48, 1963.
[23] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. New York: Chelsea, 1958.
[24] E. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*. New York: Wiley, 1988.
[25] H. Hotelling, "Relations between two set of variables," *Biometrika*, vol. 28, pp. 321–377, 1936.
[26] M. Jansson, "Subspace identification and ARX modeling," in *Proc. SYSID'03*, Rotterdam, The Netherlands, 2003, [CD ROM].
[27] G. Kuersteiner, "Automatic inference for infinite order vector autoregressions," *Econometric Theory*, vol. 21, pp. 85–115, 2005.
[28] W. Larimore, "System identification, reduced-order filtering and modeling via canonical variate analysis," in *Proc. Amer. Control Conf.*, 1983, pp. 445–451.
[29] W. Larimore, "Canonical variate analysis in identification, filtering, and adaptive control," in *Proc. 29th IEEE Conf. Decision Control*, Honolulu, HI, 1990, pp. 596–604.
[30] W. Larimore, *ADAPTx Automated System Identification Software Users Manual*. McLean, VA: Adaptic, Inc., 1992.
[31] W. Larimore, "Large sample efficiency for ADAPTx subspace identification with unknown feedback," in *Proc. IFAC DYCOPS'04*, Boston, MA, 2004, [CD ROM].
[32] R. Lewis and G. Reinsel, "Prediction of multivariate time series by autoregressive model fitting," *J. Multiv. Anal.*, vol. 16, pp. 393–411, 1985.
[33] A. Lindquist and G. Picci, "Canonical correlation analysis, approximate covariance extension and identification of stationary time series," *Automatica*, vol. 32, pp. 709–733, 1996.
[34] L. Ljung, *System Identification, Theory for the User*. Englewood Cliffs, NJ: Prentice Hall, 1997.
[35] L. Ljung and T. McKelvey, "Subspace identification from closed loop data," *Signal Processing*, vol. 52, no. 2, pp. 209–216, 1996.
[36] M. Moonen, B. De Moor, L. Vandeberghe, and J. Vandewalle, "On- and off-line identification of linear state-space models," *Int. J. Control*, vol. 49, no. 1, pp. 219–232, 1989.
[37] T. Ng, G. Goodwin, and B. Anderson, "Identifiability of MIMO linear dynamic systems operating in closed loop," *Automatica*, vol. 13, pp. 4776–485, 1977.
[38] K. Peternell, "Subspace Methods for Subspace Identification," Ph.D. dissertation, Tech. Univ. Vienna, Vienna, Austria, 1995.
[39] S. Qin and L. Ljung, "Closed-loop subspace identification with innovation estimation," in *Proc. SYSID'03*, Rotterdam, The Netherlands, 2003, pp. 887–892.
[40] S. Qin and L. Ljung, "On the role of future horizon in closed-loop subspace identification," in *Proc. SYSID'06*, Newcastle, Australia, 2006 [Online]. Available: www.ifac-papersonline.net
[41] F. Shi and J. MacGregor, "A framework for subspace identification," in *Proc. IEEE ACC*, Arlington, VA, 2001, pp. 3678–3683.
[42] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[43] P. Van Overschee and B. D. Moor, "Subspace algorithms for the stochastic identification problem," *Automatica*, vol. 29, pp. 649–660, 1993.
[44] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems*. Norwell, MA: Kluwer Academic Publications, 1996.
[45] M. Verhaegen, "Identification of the deterministic part of MIMO state space models given in innovations form from input-output data," *Automatica*, vol. 30, pp. 61–74, 1994.

Alessandro Chiuso (SM'06) received the "Laurea" degree (with highest honors) in telecommunication engineering from the University of Padova, Padova, Italy, in 1996 and the Ph.D. degree in systems engineering from the University of Bologna, Bologna, Italy, in 2000.

He is an Associate Professor with the Department of Management and Engineering, University of Padova. He has held visiting positions with Washington University, St. Louis MO, KTH (Sweden) UCLA, Los Angeles, CA. He is an Associate Editor of *Automatica* (2008-present) and a member of the editorial board of *IET Control Theory and Application* (2007-present). His research interest are mainly in estimation, identification theory and applications.

Dr. Chiuso is an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL (2010-present) and was an Associate Editor of the *IEEE CEB* (2004–2009). He serves or has served as member of several conference program committees and technical committees.