On the Relation Between CCA and Predictor-Based Subspace Identification

Alessandro Chiuso, *Senior Member, IEEE*

***Abstract*—In this paper, we investigate the relation between a recently proposed subspace method based on predictor identification (PBSID), known also as "whitening filter algorithm," and the classical CCA algorithm. The comparison is motivated by i) the fact that CCA is known to be asymptotically efficient for time series identification and optimal for white measured inputs and ii) some recent results showing that a number of recently developed algorithms are very closely related to PBSID. We show that PBSID is asymptotically equivalent to CCA precisely in the situations in which CCA is optimal while an "optimized" version of PBSID behaves no worse than CCA also for nonwhite inputs. Even though PBSID (and its optimized version) are consistent regardless of the presence of feedback, in this paper we work under the assumption that there is no feedback to make the comparison with CCA meaningful. The results of this paper imply that the "optimized" PBSID, besides being able to handle feedback, is to be preferred to CCA also when there is no feedback; only in very specific cases (white or no inputs) are the two algorithms (asymptotically) equivalent.**

***Index Terms*—Multiple-input–multiple-output (MIMO) identification, relative efficiency, statistical analysis, subspace identification.**

I. Introduction

A CERTAIN number of subspace algorithms have been developed during the last two decades. For time series identification, i.e., when there are no observed inputs, the algorithm developed by Larimore [36], [37] and Van Overschee and De Moor [47] is known to provide asymptotically efficient estimators [5].[1] Sometimes this algorithm goes under the name of CCA (or CVA) to remind one that the state construction is performed using canonical correlation analysis [32], as pioneered by Akaike [1], [2] and Desai and Pal [23]. The same ideas can be applied also when there are measured inputs, provided the canonical correlation analysis between "past" and "future" is performed "conditionally" on the future inputs [37], [35], [49]. It has become standard in the area of subspace identification to use the acronym CCA (or sometimes CVA) for this class of algorithms (with inputs [37], [42], [49], [35], [7], [4] or without inputs [36], [5]); we shall henceforth use the same terminology (CCA) also in this paper.

Manuscript received June 2, 2005; revised February 8, 2006, September 29, 2006, and December 7, 2006. Recommended by Associate Editor W. X. Zheng. This work was supported by MIUR under national project "New Methods and Algorithms for Identification and Adaptive Control of Technological Systems."

The author is with the Dipartimento di Tecnica e Gestione dei Sistemi Industriali, Università di Padova, 36100 Vicenza, 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.2007.906159

[1]Asymptotically, both in the number of data and "past" and "future" horizons. The word "efficient" is always used assuming Gaussian innovations in this paper.

Besides CCA, the most widely known procedures go under the acronyms N4SID [48] and MOESP [50]. Recently, several researchers have studied the asymptotic statistical properties of these algorithms [3], [33], [6], [7], [4], [16], [18] and compared, to some extent, existing procedures [7], [4], [17]. Also optimality of the CCA method when measured inputs are white has been established in [7]. The situation is not clear when inputs are not white. The interested reader is referred to [7].

It is our opinion, as has already been stressed in [21], that some new ideas have been introduced into the field by the study of subspace algorithms in the presence of feedback.

It is well known, in fact, that standard procedures such as MOESP, N4SID, and CCA are not consistent when data are collected in closed loop. Very recently, two subspace procedures have been introduced by Qin and Ljung (innovation estimation algorithm [43]) and Jansson (SSARX [34]) that, to some extent, are able to deal with feedback. A related algorithm is discussed also in [38]. The recent work [20] studies the statistical consistency of these two algorithms. In [20], also a "geometrical" version of the SSARX algorithm proposed by Jansson [34] was introduced and called "whitening-filter" algorithm.

This procedure forms the basis of our analysis and will be referred to as the "predictor-based subspace identification" (PBSID) algorithm in this paper. We refer the reader to [21] for an explanation of this terminology; in [21] also the relation between classical prediction error methods (PEM) and PBSID has been investigated and similarities pointed out. For reasons of space, we shall not discuss this issue further here.

It has also recently been proved (see [12] and [9]) that PBSID and SSARX, which uses vector autoregressive with exogenous inputs (VARX) modeling [34], are asymptotically equivalent. This relation further motivates the analysis of this paper; furthermore, it is shown in [14] and [9] that the optimized PBSID presented here is exactly equivalent (i.e., gives the same numerical results) to fitting a long VARX model followed by a sort of "model reduction." This also has important implication as far as computational complexity is concerned.[2] Experimental evidence shows that the behavior of PBSID/SSARX algorithm cannot be distinguished to any practical purpose from PEM in a number of simple examples; see the simulations reported in [34] and [20].

Using some recently derived formulas (see [13]) for the asymptotic variance of PBSID, one can verify that it is efficient in a number of examples when measured inputs are white. This observation raises the question: is PBSID optimal and, if so, under which conditions?

[2]The papers [14] and [9] had been submitted after this paper and actually benefitted of some comments raised by the reviewers of this paper, which are gratefully acknowledged.

We believe, therefore, that the relation of this procedure with more classical approaches is worth studying; some preliminary results have been presented in [11].

Most of the literature on the analysis of subspace methods (see [4] for a recent survey) has concentrated on open-loop procedures. It is well known (see [5]) that the CCA algorithm developed in [36] and [47] is efficient for time series identification and optimal (see [5] and [7]) for the white input case. It is also conjectured, even though not yet formally proved, that indeed CCA is efficient also for white inputs. It is therefore quite natural to compare new subspace algorithms to CCA, which provides in the situations mentioned above a sort of lower bound on the achievable accuracy by subspace procedures. Of course the comparison makes sense only in the situations where the CCA algorithm is consistent, i.e., when there is no feedback. Therefore, even though PBSID works regardless of the presence of feedback [20], [21], [34], in this paper we shall work under the assumption that *no feedback is present*. See [28], [26], and [20] for a formal definition. The main contributions are as follows.

1) We show that PBSID is asymptotically[3] equivalent to CCA in the time series case and also when measured inputs are white (see Section IV and Theorem 4.1).
2) We introduce an "optimized version" ($\text{PBSID}_{\text{opt}}$) of PBSID that performs no worse (in the sense of asymptotic variance) than CCA regardless of the input spectrum (see Section V and Theorem 5.3); $\text{PBSID}_{\text{opt}}$ can handle closed-loop data as PBSID does.

The reason why equivalence does not hold with arbitrary input signals will be made clear later on. Suffice it to say that standard procedures use "unnecessary" future input data in the regression used to construct the basis for the state space, meaning that present outputs are regressed both on past joint input–output and on future inputs [48]; PBSID instead enforces causality of the predictors [see (15)]; state constructions advocating for causal predictors have already been proposed in [42], [35], [43], and [44]. In the white input case, these "unnecessary future input data" are uncorrelated with past input and output and present output, and therefore do not influence the statistical properties (as briefly discussed in [42, p. 168]).

These are, we believe, important steps in understanding "predictor based" subspace identification; the results imply that the PBSID algorithm is asymptotically optimal for time series identification and for identification with white exogenous inputs, and also that its "optimized version" is always to be preferred to CCA. In addition, recall that both PBSID and its optimized version have a much wider range of applicability than CCA, being able to deal with closed-loop data.

The question regarding optimality in more general cases as well as the influence of the length of the "future horizon" remains open (see Section VI). In the particular example of this paper, the best (in terms of asymptotic variance) performance is reached, for colored input, when the future horizon is chosen equal to the state dimension, departing sharply from the behavior of CCA with white inputs (see [7]). See [10] for some preliminary consideration regarding the choice of the future horizon.

[3]Both in the number of data and in the length of the past horizon; see Section II for a precise definition.

The structure of this paper is as follows. In Section II, we introduce some basic notation. The details of the two algorithms analyzed are reported in Section III, while Section IV contains the statement of the result regarding white (or absent) inputs. Section V contains the results for general input signals; first the modified PBSID algorithm is presented, and then its relation to CCA is established in Theorem 5.3. Section VI contains some simulation results, and in Section VII we report some conclusions and discussion on future work. Part of the proofs are deferred to the Appendix.

II. Basic Notation and Preliminaries

Let $\{\mathbf{z}(t)\}, t \in \mathbb{Z}, \mathbf{z}(t) := [\mathbf{y}^\top(t)\, \mathbf{u}^\top(t)]^\top$ be a (weakly) stationary second-order ergodic stochastic process where $\mathbf{y}(t)$ and $\mathbf{u}(t)$ are, respectively, the output (m_y dimensional) and input (p_u dimensional) signals of a linear stochastic system in innovation form

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

We assume that there is *no feedback* from $\{\mathbf{y}(t)\}$ to $\{\mathbf{u}(t)\}$ [28], [8], [26]. The state process $\mathbf{x}(t)$ has dimension n and, without loss of generality, the realization (1) is minimal [15]. For simplicity, we assume that $D = 0$, i.e., there is no direct feedthrough[4] from $\mathbf{u}$ to $\mathbf{y}$. For time series identification (i.e., no measured inputs), one just has to let $B = 0$, $D = 0$ in (1). The white noise process $\mathbf{e}$, the innovation of $\mathbf{y}$ given the past of $\mathbf{z}$, is defined (see (3)) as the one step ahead (linear) prediction error of $\mathbf{y}(t)$ given the joint (strict) past of $\mathbf{z}$ up to time t. We define $\Lambda := \text{Var}\{\mathbf{e}(t)\}$.

We also assume that the process $\mathbf{z}$ satisfies the following.

Assumption 2.1: The joint spectrum $\Phi_{\mathbf{z}}$ is bounded and bounded away from zero on the unit circle, i.e., $\exists 0 < c \leq M < \infty$ such that $cI \leq \Phi_{\mathbf{z}}(e^{j\omega}) \leq MI \forall\, \omega \in [0, 2\pi)$.

For future reference, we define $\bar{A} := A - KC$ and let $\rho := \lambda_{\max}(\bar{A})$ be an eigenvalue of maximum modulus of $\bar{A}$; from Assumption 2.1, both $|\rho| < 1$ and $|\lambda_{\max}(A)| < 1$ hold.

The symbol I shall denote the identity matrix (of suitable dimension), $A^\top$ shall denote the transpose of the matrix A, and $||A||_2$ shall be the 2-norm. For a symmetric positive semidefinite matrix $A = A^\top \geq 0$, the symbol $A^{1/2}$ shall denote any matrix such that $A = A^{1/2}(A^{1/2})^\top$.

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

Boldface (lowercase) letters denote random variables (or semi-infinite tails). Lowercase letters denote sample values of a certain random variable. For example, we shall denote with $\mathbf{z}(t)$ the random vector denoting the joint input–output or, equivalently, the semi-infinite tail $[z_t z_{t+1}, \ldots z_{t+k} \ldots]$, where z_t is the sample value of $\mathbf{z}(t)$. It can be shown (see [40] and [19]) that the Hilbert spaces of random variables of second-order stationary and ergodic process and the Hilbert space of semi-infinite tails

[4]This assumption can be removed in our situation but is useful when there is feedback; see [28], [8], [26], and [20]. Since the "predictor based" algorithm is designed to work without assumptions on the feedback structure, we prefer to keep $D = 0$ also here.

containing sample values of the same process are isometrically isomorphic, and therefore random variables and semi-infinite tails can be regarded as being the same object. For this reason, we shall use the same symbol without risk of confusion.

We shall instead use capitals to denote the tail of length N. For instance, $Z_t := [z_t z_{t+1}, \dots z_{t+N-1}]$. These are the block rows of the usual *data block Hankel matrices* that appear in subspace identification.

For $t_0 \leq t \leq T$, we define the Hilbert space $\mathcal{Z}_{[t_0,t)}$ of random (zero-mean finite variance) variables $\mathcal{Z}_{[t_0,t)} := \overline{\text{span}}\{\mathbf{z}_k(s); k = 1, \dots, p_u + m_y, t_0 \leq s < t\}$; the bar denotes closure in mean square, i.e., in the metric defined by the inner product $\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle := \mathbb{E}\{\boldsymbol{\xi}\boldsymbol{\eta}\}$, where $\mathbb{E}\{\cdot\}$ denotes mathematical expectation. This is the *past space* at time t of the processes $\mathbf{z}$. Similarly, let $\mathcal{Z}_{[t,T]}$ be the future input and output space up to time T. The length $T - t$ of the future will be denoted with $\nu := T - t$, while we define $\bar{\nu} := \nu - 1$. Similar definitions hold for $\mathcal{Y}_{[t_0,t)}, \mathcal{U}_{[t_0,t)}$, etc. By convention, the past spaces do not include the present. When $t_0 = -\infty$, we shall use the shorthands $\mathcal{Z}_t^-$ for $\mathcal{Z}_{(-\infty,t)}$. Subspaces spanned by random vectors at just one time instant (e.g., $\mathcal{Z}_{[t,t]}$, etc.) are simply denoted with just one time subscript (e.g., $\mathcal{Z}_t$), while $\mathcal{Z} := \mathcal{Z}_{(-\infty,\infty)}$.

Let $\mathcal{B}, \mathcal{C}$ be two subspaces of $\mathcal{Z}$ such that $\mathcal{B} \cap \mathcal{C} = \{0\}$. With a slight abuse of notation, we shall denote with $E[\cdot \mid \mathcal{C}]$ the orthogonal projection onto $\mathcal{C}$, which coincides with conditional expectation in the Gaussian case. The symbol $E_{\|\mathcal{B}}[\cdot \,|\, \mathcal{C}]$ shall denote the oblique projection onto $\mathcal{C}$ along $\mathcal{B}$ (see [27] and [19]). We adopt the notation $\Sigma_{\mathbf{ad}} := \mathbb{E}[\mathbf{a}\mathbf{d}^\top]$ to denote the covariance matrix between the random vectors $\mathbf{a} \in \mathcal{Z}$ and $\mathbf{d} \in \mathcal{Z}$. If $\mathbf{c}$ is a finite-dimensional basis ($\Sigma_{\mathbf{cc}}$ invertible) for $\mathcal{C}$, one has $E[\mathbf{a}|\mathcal{C}] = \Sigma_{\mathbf{ac}}\Sigma_{\mathbf{cc}}^{-1}\mathbf{c}$. The symbol $\Sigma_{\mathbf{ad|c}}$ will denote projection error covariance (conditional covariance in the Gaussian case) $\Sigma_{\mathbf{ad|c}} := \mathbb{E}[(\mathbf{a} - E[\mathbf{a}|\mathcal{C}])(\mathbf{d} - E[\mathbf{d}|\mathcal{C}])^\top] = \Sigma_{\mathbf{ad}} - \Sigma_{\mathbf{ac}}\Sigma_{\mathbf{cc}}^{-1}\Sigma_{\mathbf{cd}}$. If $\mathbf{b}$ is a basis for $\mathcal{B}$, the oblique projection $E_{\|\mathcal{B}}[\mathbf{a}|\mathcal{C}]$ can be computed using the formula $E_{\|\mathcal{B}}[\mathbf{a}|\mathcal{C}] = \Sigma_{\mathbf{ac|b}}\Sigma_{\mathbf{cc|b}}^{-1}\mathbf{c}$.

We shall also use the notation $\mathbf{z}_{[t,s]} := [\mathbf{z}^\top(t) \;\; \mathbf{z}^\top(t+1) \;\; \cdots \;\; \mathbf{z}^\top(s)]^\top$ and the shorthands $\mathbf{z}^+ := \mathbf{z}_{[t,T-1]}$ and $\bar{\mathbf{z}}^+ := \mathbf{z}_{[t,T]}$. Similarly, the (finite) block Hankel data matrices will be denoted as $Z_{[t,s]} := [Z_t^\top \;\; Z_{t+1}^\top \;\; \cdots \;\; Z_s^\top]^\top$.

Sample covariances of finite sequences will be denoted with the same symbol used for the corresponding random variables with a "hat" on top. For example, given finite sequences $A_t := [a_t, a_{t+1}, \dots, a_{t+N-1}]$ and $B_t := [b_t, b_{t+1}, \dots, b_{t+N-1}]$ containing sample values of the processes $\{\mathbf{a}(t)\}, \{\mathbf{b}(t)\}$, we shall define $\hat{\Sigma}_{\mathbf{ab}} = (A_t B_t^\top / N)$. Under our ergodic assumption, $\lim_{N\to\infty} \hat{\Sigma}_{\mathbf{ab}} \overset{\text{a.s}}{=} \Sigma_{\mathbf{ab}}$. Similarly, given a third sequence (say, C_t, $\hat{\Sigma}_{\mathbf{ab|c}}$ is defined as[5] $\hat{\Sigma}_{\mathbf{ab|c}} := \hat{\Sigma}_{\mathbf{ab}} - \hat{\Sigma}_{\mathbf{ac}}\hat{\Sigma}_{\mathbf{cc}}^{-1}\hat{\Sigma}_{\mathbf{cb}}$. Orthogonal and oblique projections on (row) spaces of finite tails will be denoted with the symbol $\hat{E}$; e.g., $\hat{E}[\cdot|Z_{[t_0,t)}]$ and $\hat{E}_{\|U_{[t,T]}}[\cdot|Z_{[t_0,t)}]$. As above, the oblique projection can be computed using the formula [27], [19] $\hat{E}_{\|B_t}[A_t|C_t] = \hat{\Sigma}_{\mathbf{ac|b}}\hat{\Sigma}_{\mathbf{cc|b}}^{-1}C_t$. When projecting onto the space generated by the rows of two (or more) matrices, say, B_t and C_t, we shall use the notation $\hat{E}[\cdot|B_t, C_t]$.

[5]Provided $\hat{\Sigma}_{\mathbf{cc}}$ is invertible.

In order for oblique projections to be well defined, we shall need that $\mathcal{Z}_{[t_0,T]}$ admits the direct sum decomposition

$$\mathcal{Z}_{[t_0,T]} = \mathcal{Z}_{[t_0,t)} + \mathcal{Z}_{[t,T]}, \qquad t_0 \leq t \leq T \tag{2}$$

with the $+$ sign denoting direct sum of subspaces. Condition (2) is implied by Assumption 2.1. In fact it is well known [31] that (2) holds for a full-rank purely nondeterministic process $\mathbf{z}$ if and only if the determinant of the spectral density matrix $\Phi_{\mathbf{z}}$ has no zeros on the unit circle. Note that (2) will imply also that $Z_{[t_0,T]}$ is of full row rank almost surely for ergodic sequences.

With the notation introduced above, the innovation process $\mathbf{e}(t)$, i.e., the one step ahead (linear) prediction error of $\mathbf{y}(t)$ based on the joint past $\mathcal{Z}_t^-$ is written in the form

$$\mathbf{e}(t) := \mathbf{y}(t) - E[\mathbf{y}(t) \,|\, \mathcal{Z}_t^-]. \tag{3}$$

We shall assume that the innovation process satisfies the following.

Assumption 2.2: Let $\mathcal{F}_t^-$ be the σ-algebra generated by the random variables $\{\mathbf{y}(s), -\infty < s \leq t\}$ and $\{\mathbf{u}(s), -\infty < s < \infty\}$ (past outputs and past plus future inputs). The innovation process $\mathbf{e}(t)$ is an $\mathcal{F}_{t-1}^-$-martingale difference sequence with constant conditional variance, i.e.,

$$\begin{aligned} &\mathbb{E}[\mathbf{e}(t)|\mathcal{F}_{t-1}^-] = 0 \\ &\mathbb{E}[\mathbf{e}(t)\mathbf{e}^\top(t)|\mathcal{F}_{t-1}^-] = \Lambda. \end{aligned} \tag{4}$$

The analysis reported in this paper requires that both the length of the finite tails[6] N and the past horizon $p := t - t_0$[7] go to infinity. We remind the reader 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 [7], where the following assumption is made.

Assumption 2.3: The past horizon $p = t - t_0$ goes to infinity with N while satisfying

$$\begin{aligned} &p \geq \frac{\log N^{-d/2}}{\log|\rho|}, \qquad 1 < d < \infty \\ &p = o(\log(N)^\alpha), \qquad \alpha < \infty. \end{aligned} \tag{5}$$

The first condition shall be needed to ensure that the difference between the stationary predictor (i.e., the predictor based on past data in $(-\infty, t)$) and its truncated version (i.e., using past data in a finite window $[t_0, t)$) is $o(1/\sqrt{N})$, while the second ensures that the computation of sample covariance matrices of increasing size (with $p = t - t_0$) does not pose any complication in the sense that their limit is well defined and equal to the population counterpart (see the discussion after Lemma 4 in [7]).

We now need to recall some basic definitions in asymptotic statistics. Given a sequence of random vectors $\mathbf{v}_N$, we say that $\sqrt{N}\mathbf{v}_N$ is asymptotically normal if $\sqrt{N}\mathbf{v}_N$ converges in law to

[6]This is the parameter j in the notation of Van Overschee and De Moor [48], i.e., the number of columns in the block Hankel data matrices used in subspace identification.

[7]The number of block rows in the block Hankel data matrix containing the past data.

a Gaussian random vector. The variance of the limiting distribution is called *asymptotic variance* of $\sqrt{N}\mathbf{v}_N$. If the number of elements in the random vector $\mathbf{v}_N$ increases with N, we shall need a slight extension of the definition of asymptotic normality (see [39]). We shall say that $\sqrt{N}\mathbf{v}_N$ is asymptotically normal if the random variable $\sqrt{N}\eta_N^\top \mathbf{v}_N$ is asymptotically normal for any column vector η_N (of suitable dimensions) satisfying Assumption 2.4.

Assumption 2.4: i) $\exists M < \infty : \forall N \eta_N^\top \eta_N < M$; ii) $\exists \eta \in \ell_2 : \lim_{N\to\infty} \|[\eta_N^\top 0] - \eta^\top\|_2 = 0$; and iii) $\lim_{N\to\infty} \mathrm{Var}(\sqrt{N}\eta_N^\top \mathbf{v}_N) = c < \infty$.

With some abuse of terminology, we shall talk about asymptotic variance matrix also for vectors of increasing size. For instance, when we shall say that $\sqrt{N}\mathbf{v}_N$ has asymptotic variance Σ_∞ (with $\|\Sigma_\infty\|_2 < \infty$),[8] we shall really mean that the asymptotic variance of $\sqrt{N}\eta_N^\top \mathbf{v}_N$ is $\eta^\top \Sigma_\infty \eta$.

Similarly, given two asymptotically normal random vectors $\mathbf{v}_N$ and $\mathbf{w}_N$, we shall say that $\mathrm{AsVar}\{\sqrt{N}\mathbf{v}_N\} \geq \mathrm{AsVar}\{\sqrt{N}\mathbf{w}_N\}$ if, $\forall \eta_N$, $\mathrm{AsVar}\{\sqrt{N}\eta_N^\top \mathbf{v}_N\} \geq \mathrm{AsVar}\{\sqrt{N}\eta_N^\top \mathbf{w}_N\}$. Convergence allows one to deal with the expressions for $N = \infty$ rather than with the limit, as done also in [7].

Given two deterministic sequences x_N, g_N, the symbol $x_N = o(g_N)$ has the usual meaning $\lim_{N\to\infty} x_N/g_N = 0$ while $x_N = O(g_N)$ means that $\forall \epsilon, \exists M$ such that $\limsup_{N\to\infty} x_N/g_N \leq M$. If $\mathbf{x}_N$ is a sequence of random variables, we shall say that $\mathbf{x}_N = o_P(g_N)$ if, $\forall \delta > 0$,, $\lim_{N\to\infty} P[|\mathbf{x}_N/g_N| > \delta] = 0$, while $\mathbf{x}_N = O_P(g_N)$ indicates that $\mathbf{x}_N/g_N$ is bounded in probability, i.e., $\forall \epsilon$, $\exists M$ such that $\sup_N P[|\mathbf{x}_N/g_N| > M] < \epsilon$. Similarly, $o(\cdot)$ and $O(\cdot)$ have the same meaning almost surely, i.e., $\mathbf{x}_N = o(g_N)$ means that $\lim_{N\to\infty} \mathbf{x}_N/g_N = 0$ a.s. while $\mathbf{x}_N = O(g_N)$ means that $\limsup_{N\to\infty} \mathbf{x}_N/g_N \leq M$ a.s. for some $0 \leq M < \infty$.

The symbol $\doteq$ shall denote equality up to $o_P(1/\sqrt{N})$ terms, which we shall call *asymptotic equivalence*. In fact, from standard results in asymptotic analysis (see, for instance, [24]), terms that are $o_P(1/\sqrt{N})$ can be neglected when studying the asymptotic distribution. We shall use the notations $\underline{o}_P(\cdot)$, $\underline{O}_P(\cdot)$, $\underline{o}(\cdot)$, or $\underline{O}(\cdot)$ to denote random matrices (of suitable dimensions possibly depending on N) whose elements are (uniformly), respectively, $o_P(\cdot)$, $O_P(\cdot)$, $o(\cdot)$, or $O(\cdot)$; Uniformity is needed when the matrices' sizes increase with N. In this paper, uniformity shall be guaranteed by stationarity of the processes involved.

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 $o_P(1/\sqrt{N})$ 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 estimation of system invariants is concerned. For instance, if $\mathbf{x}_1$ and $\mathbf{x}_2$ are two candidate state variables, we shall write $\mathbf{x}_1 \doteq \mathbf{x}_2$ if there exists a nonsingular $\hat{T}_N$, with $\lim_{N\to\infty} \hat{T}_N = I$, so that $\mathbf{x}_1 - \hat{T}_N \mathbf{x}_2 = \underline{o}_P(1/\sqrt{N})$. When dealing with tails, e.g., A_t and B_t, containing the sample values a_{t+i}, b_{t+i}, $i = 0, \ldots, N-1$, of the random vectors $\mathbf{a}(t)$ and $\mathbf{b}(t)$, $A_t \doteq B_t$ really means that $\mathbf{a}(t) \doteq \mathbf{b}(t)$. Note also, for future reference, that if $\mathbf{x}_N = O(f_N)$ and $\mathbf{y}_N = O(g_N)$, then $\mathbf{x}_N \mathbf{y}_N = o(h_N)$ provided $f_N g_N = o(h_N)$. Recall also that almost sure convergence implies convergence in probability, which, in particular, means that $\mathbf{x}_N = o(1/\sqrt{N})$ implies $\mathbf{x}_N = o_P(1/\sqrt{N})$.

For future reference, we also define the extended observability matrix of the pair (A, C) as $\Gamma_k^\top := [C^\top, A^\top C^\top, \ldots, (CA^k)^\top]$. Similarly, $\bar{\Gamma}_k$ shall be the observability matrix of the pair $(\bar{A}, C)$. We also introduce the block Toeplitz matrices containing the Markov parameters of the "stochastic" part

$$H_k = \begin{bmatrix} I & 0 & \cdots & 0 \\ CK & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{k-1}K & CA^{k-2}K & \cdots & I \end{bmatrix}. \tag{6}$$

Let us also define $Q_N := \log(\log(N))/N$.

III. State Space Construction

It is well known [47], [48], [40], [19] that identification using subspace methods can be seen as a two-step procedure as follows.

1) Construct a basis $\hat{X}_t$ for the state space via suitable projection operations on data sequences (block Hankel data matrices).
2) Given (coherent) bases for the state space at time $t(\hat{X}_t)$ and $t+1(\hat{X}_{t+1})$, solve

$$\begin{cases} \hat{X}_{t+1} \simeq A\hat{X}_t + BU_t + KE_t \\ Y_t \simeq C\hat{X}_t + E_t \end{cases} \tag{7}$$

in the least squares sense.

Different subspace algorithms have different implementations of the first step while the second remains the same for virtually all algorithms that follow the "state" or "Larimore" approach [4]; in this paper, we shall not be concerned with algorithms based on the "shift invariance" (or MOESP-type) methods [4]. For this reason, we compare algorithms on the basis of step 1). We shall identify procedures that are (asymptotically) equivalent, modulo change of basis, as the first step is concerned.

To make this statement precise, we report the following result, which has been extensively used in the literature on asymptotic analysis of subspace procedures [4], [16], [33].

Proposition 3.1: Assume $\hat{X}_t^1$ and $\hat{X}_t^2$ are two candidate state sequences where

$$\hat{X}_t^1 \doteq \hat{X}_t^2 \tag{8}$$

and assume a similar property holds also for the state at time $t+1$. Then the least squares estimators $\hat{A}^1, \hat{B}^1, \hat{C}^1, \hat{K}^1$ and $\hat{A}^2, \hat{B}^2, \hat{C}^2, \hat{K}^2$ of A, B, C, K obtained from (7) using, respectively, $\hat{X}_t^1(\hat{X}_{t+1}^1)$ and $\hat{X}_t^2(\hat{X}_{t+1}^2)$ are asymptotically equivalent (modulo change of basis).

Proof: See Appendix A. ■

Remark III.1: We remind the reader that for t_0 finite, the estimation of the Kalman gain K involves the solution of a Riccati equation. See, for instance, [47], [48], and [40]. The situa-

[8] This shall be guaranteed by the fact that the generic elements of Σ_∞ shall go to zero exponentially as a function of the difference between row and column indexes.

tion is different here since t_0 is let going to $-\infty$ according to Assumption 2.3. ♢

In this section, we shall review the state construction step for the CCA algorithm [36], [7] and for the PBSID algorithm [34], [20].

A. CCA Algorithm

The basic object that allows one to construct a basis for the state space is the "oblique predictor"

$$\begin{aligned}\hat{Y}_{[t,T-1]} &= \hat{E}_{\|U_{[t,T]}}\left[Y_{[t,T-1]} \,|\, Z_{[t_0,t)}\right] \\ &= \Gamma_{\bar{\nu}}\hat{E}_{\|U_{[t,T]}}\left[X_t \,|\, Z_{[t_0,t)}\right] \\ &\quad + H_{\bar{\nu}}\hat{E}_{\|U_{[t,T]}}\left[E_{[t,T-1]} \,|\, Z_{[t_0,t)}\right] \\ &\simeq \Gamma_{\bar{\nu}}X_t. \end{aligned} \tag{9}$$

The approximate equality has to be understood in the sense that, asymptotically in N

$$\hat{\mathbf{y}}_{[t,T-1]} = E_{\|\mathcal{U}_{[t,T]}}[\mathbf{y}_{[t,T-1]} \,|\, \mathcal{Z}_t^-] = \Gamma_{\bar{\nu}}\mathbf{x}(t) \tag{10}$$

holds. The matrix $\hat{Y}_{[t,T-1]}$ has full row rank for finite N.

The reduction to rank n, the system order, is implemented via the weighted singular value decomposition

$$\hat{W}_{\text{cca}}^{-1}\hat{Y}_{[t,T-1]} = \begin{bmatrix}U_n & \tilde{U}_n\end{bmatrix}\begin{bmatrix}S_n & 0\\ 0 & \tilde{S}_n\end{bmatrix}\begin{bmatrix}V_n^\top \\ \tilde{V}_n^\top\end{bmatrix}. \tag{11}$$

In the CCA algorithm to the choice[9] $\hat{W}_{\text{cca}} := \hat{\Sigma}^{1/2}_{\mathbf{y}^+\mathbf{y}^+\,|\,\bar{\mathbf{u}}^+}$ is made. An estimate of the observability matrix is obtained discarding the "less significant" singular values (i.e., pretending $\tilde{S}_n \simeq 0$) from[10]

$$\hat{\Gamma}_{\bar{\nu}} = \hat{W}_{\text{cca}}U_n\hat{T} \tag{12}$$

where $\hat{T}$ can be any nonsingular matrix providing a choice of basis. We acknowledge that the presence of the matrix $\hat{T}$ may appear nonstandard. Standard procedures correspond to the choices $\hat{T} = I$ or $\hat{T} = S_n^{1/2}$. It is useful in the analysis reported in this paper to make a specific choice of $\hat{T}$ that shall guarantee that the estimated system matrices converge regardless of the possible ambiguities, due to orientation of singular vectors and multiple singular values (if any) in the singular value decomposition (SVD) (11). To this purpose, we propose to use

$$\hat{T} := U_n^\top\hat{W}_{\text{cca}}^{-1}\Gamma_{\bar{\nu}} \tag{13}$$

where $\Gamma_{\bar{\nu}}$ is the "true" (but unknown) observability matrix.[11]

[9]This procedure differs from the original CCA by the choice of a "right" weight. We remind that this "right weight" has no influence on the asymptotic accuracy of the estimates using the "state approach", i.e., implementing step 2) above. See, for instance, [7] and [16].

[10]We do not discuss order estimation in this paper. We shall always assume that the correct order is selected.

[11]Note that such choice is infeasible in practice; however, such $\hat{T}$ serves only for the purpose of asymptotic analysis and need not be known when implementing the algorithm. In fact, for a fixed data size N, the choice of $\hat{T}$ influences only the basis in which the system matrices are estimated and hence does not affect any system invariant.

Lemma 3.2: The estimator $\hat{\Gamma}_{\bar{\nu}}$ in (12), with the choice of $\hat{T}$ in (13), converges almost surely (a.s.) to $\Gamma_{\bar{\nu}}$ as $N \to \infty$.

Proof: See Appendix A. ■

Defining the left inverse $\hat{\Gamma}_{\bar{\nu}}^{-L} := (\hat{\Gamma}_{\bar{\nu}}^\top\hat{W}_{\text{cca}}^{-\top}\hat{W}_{\text{cca}}^{-1}\hat{\Gamma}_{\bar{\nu}})^{-1}\hat{\Gamma}_{\bar{\nu}}^\top\hat{W}_{\text{cca}}^{-\top}\hat{W}_{\text{cca}}^{-1}$, a basis for the state space is constructed from

$$\begin{aligned}\hat{X}_t^{\text{cca}} &:= \hat{\Gamma}_{\bar{\nu}}^{-L}\hat{Y}_{[t,T-1]} \\ \hat{X}_{t+1}^{\text{cca}} &:= \hat{\Gamma}_{\bar{\nu}}^{-L}\hat{E}_{\|U_{[t+1,T]}}\left[Y_{[t+1,T]} \,|\, Z_{[t_0,t+1)}\right]. \end{aligned} \tag{14}$$

These formulas for constructing the state space are discussed and motivated, for instance, in [16, pp. 276–277].

Remark III.2: We remind that all weighting matrices are, in practice, data dependent. However, for the purpose of asymptotic analysis, data-dependent weights (say, $\hat{W}_{\text{cca}}$) can be substituted with their (a.s.) limit (say, W_{cca}), as discussed, for instance, in [6], [7], [4], and [16].

Therefore, to streamline notation, we prefer to work directly with the population version of all weights. ♢

We quote now a result that first appeared in [7], which shows that the CCA weight $W_{\text{cca}} = \Sigma^{1/2}_{\mathbf{y}^+\mathbf{y}^+\,|\,\bar{\mathbf{u}}^+}$ can be substituted with $[H_{\bar{\nu}}(I\otimes\Lambda)H_{\bar{\nu}}^\top]^{1/2}$ without changing the asymptotic properties.

Lemma 3.3 [7]: Assume the parameters are estimated following steps 1) and 2) above and the state is constructed according to (14). Then any choice $W_{\text{cca}} = [H_{\bar{\nu}}(I\otimes\Lambda)H_{\bar{\nu}}^\top + \Gamma_{\bar{\nu}}\Sigma\Gamma_{\bar{\nu}}^\top]^{1/2}$ with $\Sigma = \Sigma^\top \geq 0$ provides the same asymptotic accuracy of the estimators of any system invariant.

Proof: For a proof of this result, see an extended version of this paper.[12] ■

This fact will be useful later on to study the relation between CCA and predictor-based subspace identification.

Remark III.3: With some abuse of notation, we can denote with $\hat{X}_t^{\text{cca}}$ any state sequence resulting from a choice of W_{cca} of the form $W_{\text{cca}} = [H_{\bar{\nu}}(I\otimes\Lambda)H_{\bar{\nu}}^\top + \Gamma_{\bar{\nu}}\Sigma\Gamma_{\bar{\nu}}^\top]^{1/2}$ for some $\Sigma = \Sigma^\top \geq 0$. Lemma 3.3 ensures that these state sequences are asymptotically equivalent as far as estimation of system invariants is concerned. From now on, we shall always use $W_{\text{cca}} := H_{\bar{\nu}}(I\otimes\Lambda^{1/2})$. ♢

B. PBSID Algorithm

This algorithm inherits its name from the similarity with PE methods, i.e., that it is based on identification of the predictor model. As mentioned in the Introduction, this algorithm was introduced in [20], inspired by [34], under the name "whitening filter algorithm." For reasons of space, we shall refer the interested reader to [21] for further comments regarding the relation between PBSID and PEM.

The construction of the state space using this algorithm is slightly more complicated and involves several oblique projections. First, one computes the oblique projections[13]

$$\begin{aligned}\hat{Y}_{t+h}^P &:= \hat{E}_{\|Z_{[t,t+h)}}\left[Y_{t+h} \,|\, Z_{[t_0,t)}\right] \\ &\simeq C\bar{A}^{h-1}X_t, \qquad h = 0,1,\ldots,\bar{\nu}. \end{aligned} \tag{15}$$

[12]www.dei.unipd.it/~chiuso.

[13]The superscript ^{P} reminds that the quantity has to do with the "predictor-based" algorithm.

Also here the approximate equality has to be understood in the sense that, asymptotically in $N (h = 0, 1, \ldots, \bar{\nu})$

$$\hat{\mathbf{y}}^P(t+h) := E_{\|\mathcal{Z}_{[t,t+h)}}[\mathbf{y}(t+h) \,|\, \mathcal{Z}_t^-] = C\bar{A}^{h-1}\mathbf{x}(t) \quad (16)$$

holds. Then one stacks all the predictors $\hat{Y}^P_{[t,T-1]} := [(\hat{Y}^P_t)^\top, \ldots, (\hat{Y}^P_{T-1})^\top]^\top$ so that $\hat{Y}^P_{[t,T-1]} \simeq \bar{\Gamma}_{\bar{\nu}} X_t$. From the weighted SVD[14]

$$W^{-1}\hat{Y}^P_{[t,T-1]} = [P_n \tilde{P}_n] \begin{bmatrix} D_n & 0 \\ 0 & \tilde{D}_n \end{bmatrix} [Q_n^\top \quad \tilde{Q}_n^\top] \quad (17)$$

an estimate of the observability matrix $\bar{\Gamma}_{\bar{\nu}}$ is obtained discarding the "less significant" singular values (i.e., pretending $\tilde{D}_n \simeq 0$) from

$$\hat{\bar{\Gamma}}_{\bar{\nu}} = W P_n \hat{\bar{T}} \quad (18)$$

where $\hat{\bar{T}}$ can be any nonsingular matrix providing a choice of basis.

As done in the previous section, for the purpose of analysis, we shall make the specific choice

$$\hat{\bar{T}} := P_n^\top W^{-1} \bar{\Gamma}_{\bar{\nu}} \quad (19)$$

where $\bar{\Gamma}_{\bar{\nu}}$ is the "true" (but unknown) observability matrix.

Lemma 3.4: The estimator $\hat{\bar{\Gamma}}_{\bar{\nu}}$ in (18), with the choice of $\hat{\bar{T}}$ in (19), converges a.s. to $\bar{\Gamma}_{\bar{\nu}}$ as $N \to \infty$.

Proof: It is analogous to the proof of Lemma 3.2 and shall be omitted. ■

Defining the left inverse $\hat{\bar{\Gamma}}_{\bar{\nu}}^{-L} := (\hat{\bar{\Gamma}}_{\bar{\nu}}^\top W^{-\top} W^{-1} \hat{\bar{\Gamma}}_{\bar{\nu}})^{-1} \hat{\bar{\Gamma}}_{\bar{\nu}}^\top W^{-\top} W^{-1}$, a basis for the state space is given by

$$\begin{aligned} \hat{X}^P_t &:= \hat{\bar{\Gamma}}_{\bar{\nu}}^{-L} \hat{Y}^P_{[t,T-1]} \\ \hat{X}^P_{t+1} &:= \hat{\bar{\Gamma}}_{\bar{\nu}}^{-L} \begin{bmatrix} \hat{E}[Y_{t+1} \,|\, Z_{[t_0,t+1)}] \\ \hat{E}_{\|Z_{t+1}}[Y_{t+2} \,|\, Z_{[t_0,t+1)}] \\ \vdots \\ \hat{E}_{\|Z_{[t,T)}}[Y_T \,|\, Z_{[t_0,t+1)}] \end{bmatrix}. \end{aligned} \quad (20)$$

IV. White Inputs

In this section, we shall study the link between the state constructions (14) and (20) under the assumption that the input signal is a white noise process (or it is absent). Since we shall need to introduce a modified PBSID algorithm to perform the comparison in the general case, we prefer to keep well separated the two situations and deal first with the standard algorithm, and hence we postpone the analysis of the general case to the next section.

We now state the main result of this section.

Theorem 4.1: Under the conditions stated in Assumption 2.3, assuming that inputs are white or absent, provided W is chosen according to $W = I \otimes \Lambda^{1/2}$ and $W_{\mathrm{cca}} = H_{\bar{\nu}}(I \otimes \Lambda^{1/2})$, the state constructions in (14) and (20) satisfy $\hat{X}^P_t \doteq \hat{X}^{\mathrm{cca}}_t$ and therefore yield asymptotically the same accuracy as far as estimation of any system invariant is concerned.

The proof of this theorem relies on an intermediate result, which we state in the form of a lemma.

Lemma 4.2: If $\mathbf{u}(t)$ is absent or white, the oblique predictor $\hat{Y}_{t+h} := \hat{E}_{\|U_{[t,T]}}[Y_{t+h} \,|\, Z_{[t_0,t)}]$ satisfies

$$\hat{Y}_{t+h} \doteq \hat{Y}^P_{t+h} + \sum_{i=1}^{h} \hat{\Phi}_{hi} \hat{Y}_{t+h-i} \quad (21)$$

for suitable matrix coefficients $\hat{\Phi}_{ij}$ satisfying $\lim_{N\to\infty} \hat{\Phi}_{ij} = \Phi_{i,j} = C\bar{A}^{i-1}K$. This relation can be written in compact form as

$$\hat{Y}^P_{[t,T-1]} \doteq H_{\bar{\nu}}^{-1} \hat{Y}_{[t,T-1]}. \quad (22)$$

Proof: See Appendix A. ■

Recall that (22) should be understood in the sense that the left- and right-hand side give rise to state sequences that differ, up to $o_P(1/\sqrt{N})$ terms, only for a nonsingular change of basis $\hat{T}_N$ converging to the identity matrix as N grows to infinity. We shall use the same notation without further notice in the rest of the paper. This is well known in the literature of subspace identification and corresponds to the fact that sample dependent weights can be substituted with their a.s. limit without changing the asymptotic properties of any system invariant (see, e.g., [4, Theorem 7]).

Proof of Theorem 4.1: We recall from (17) that, in the predictor-based algorithm, one takes the SVD of $W^{-1}\hat{Y}^P_{[t,T-1]}$, while, from (11), $W_{\mathrm{cca}}^{-1}\hat{Y}_{[t,T-1]}$ is used in the CCA algorithm.

Note that the CCA algorithm corresponds to the choice $W_{\mathrm{cca}} = \Sigma^{1/2}_{\mathbf{y}^+\mathbf{y}^+ \,|\, \bar{\mathbf{u}}^+} = (\Gamma_{\bar{\nu}} \Sigma_{\mathbf{x}\mathbf{x}|\bar{\mathbf{u}}^+} \Gamma_{\bar{\nu}}^\top + H_{\bar{\nu}}(I \otimes \Lambda) H_{\bar{\nu}}^\top)^{1/2}$. However, by letting $\Sigma = 0$ in Lemma 3.3, $W_{\mathrm{cca}} = (H_{\bar{\nu}}(I \otimes \Lambda) H_{\bar{\nu}}^\top)^{1/2}$ provides the same asymptotic behavior.

If we now premultiply both sides of (22) after Lemma 4.2 by $W^{-1/2} = (I \otimes \Lambda)^{-1/2}$, we obtain that

$$\begin{aligned} W^{-1/2}\hat{Y}^P_{[t,T-1]} &\doteq (I \otimes \Lambda)^{-1/2} H_{\bar{\nu}}^{-1} \hat{Y}_{[t,T-1]} \\ &= W_{\mathrm{cca}}^{-1} \hat{Y}_{[t,T-1]}. \end{aligned} \quad (23)$$

As described in Section III, the right-hand side is used in CCA and the left-hand side in PBSID. This means that the matrices of which one computes SVD are asymptotically equivalent for the two algorithms. As a consequence, also the estimated state sequences $\hat{X}^{\mathrm{cca}}_t$ and $\hat{X}^P_t$ are asymptotically equivalent, which using Proposition 3.1 concludes the proof. ■

V. Nonwhite Inputs

In this section, we shall address the case when measured inputs are not white. Unfortunately, it seems not possible to compare CCA with PBSID in the form presented in Section III. We shall need to consider an "optimized" version of PBSID, which we shall call $\mathrm{PBSID}_{\mathrm{opt}}$. We shall show that, in this case, the variance of the estimators obtained using CCA is greater than or equal to the variance of the estimators obtained using $\mathrm{PBSID}_{\mathrm{opt}}$.

First we shall explain the reasons for modifying the PBSID algorithm and then use these arguments to show that indeed $\mathrm{PBSID}_{\mathrm{opt}}$ performs no worse than CCA.

[14]We introduce a weighting matrix W that will be chosen appropriately.

Defining $\mathcal{K} := [\bar{A}^{p-1}[K\ B]\ \bar{A}^{p-2}[K\ B]\ \dots [K\ B]]$, the output tail Y_{t+h} can be written as

$$\begin{aligned} Y_{t+h} &= C\bar{A}^h X_t + E_{t+h} \\ &\quad + \sum_{i=1}^{h} C\bar{A}^{i-1}(KY_{t+h-i} + BU_{t+h-i}) \\ &= C\bar{A}^h \mathcal{K} Z_{[t_0,t)} + E_{t+h} + \underline{o}_P(1/\sqrt{N}) \\ &\quad + \sum_{i=1}^{h} C\bar{A}^{i-1}\left(KY_{t+h-i} + BU_{t+h-i}\right) \\ &:= \Xi_h Z_{[t_0,t)} + E_{t+h} + \underline{o}_P(1/\sqrt{N}) \\ &\quad + \sum_{i=1}^{h} \Phi_{hi} Y_{t+h-i} + \Psi_{hi} U_{t+h-i} \end{aligned} \tag{24}$$

where, thanks to Assumption 2.3, the $\underline{o}_P(1/\sqrt{N})$ term accounts for mishandling of the initial condition and $\Xi_h := C\bar{A}^h\mathcal{K}$, $\Phi_{hi} := C\bar{A}^{i-1}K$, $\Psi_{hi} := C\bar{A}^{i-1}B$. The parameters Φ_{hi} and Ψ_{hi} do not depend on h, but this notation shall be useful in the sequel.

The state constructions for CCA and PBSID are based on the oblique projections (9) and (15); we shall make use of the relations

$$\begin{aligned} \hat{Y}_{t+h} &= \hat{E}_{/\!/U_{[t,T]}}\left[Y_{t+h} \,|\, Z_{[t_0,t)}\right] \\ &= \hat{E}_{/\!/U_{[t,T]}}\left[\hat{E}\left[Y_{t+h} \,|\, Z_{[t_0,t+h)}, U_{[t+h,T]}\right] \,|\, Z_{[t_0,t)}\right] \end{aligned} \tag{25}$$

and

$$\begin{aligned} \hat{Y}^P_{t+h} &= \hat{E}_{/\!/Z_{[t,t+h)}}\left[Y_{t+h} \,|\, Z_{[t_0,t)}\right] \\ &= \hat{E}_{/\!/Z_{[t,t+h)}}\left[\hat{E}\left[Y_{t+h} \,|\, Z_{[t_0,t+h)}\right] \,|\, Z_{[t_0,t)}\right]. \end{aligned} \tag{26}$$

A. Optimized PBSID Algorithm

Stacking the data and using (24) (discarding[15] $o_P(1/\sqrt{N})$ terms), we obtain

$$\begin{aligned} \begin{bmatrix} Y_t \\ Y_{t+1} \\ \vdots \\ Y_T \end{bmatrix} &\doteq \begin{bmatrix} \Xi_0 \\ \Xi_1 \\ \vdots \\ \Xi_\nu \end{bmatrix} Z_{[t_0,t)} + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \Phi_{11} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{\nu\nu} & \cdots & \Phi_{\nu 1} & 0 \end{bmatrix} Y_{[t,T]} \\ &\quad + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \Psi_{11} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{\nu\nu} & \cdots & \Psi_{\nu 1} & 0 \end{bmatrix} U_{[t,T]} + \begin{bmatrix} E_t \\ E_{t+1} \\ \vdots \\ E_T \end{bmatrix}. \end{aligned} \tag{27}$$

Observe that the lower triangular matrices in (27) are Toeplitz, since $\Phi_{ij} = C\bar{A}^{j-1}K$, $\Psi_{ij} = C\bar{A}^{j-1}B$, $\forall i,j$. The inner projection in (26) is equivalent to solving (27) "row by row"; hence the Toeplitz structure is not preserved after estimation, i.e., $\hat{\Phi}_{ij} \neq \hat{\Phi}_{i'j}$, $\hat{\Psi}_{ij} \neq \hat{\Psi}_{i'j}$, almost surely when $i \neq i'$.

[15]See Appendix B for a formal justification.

This is equivalent to solving the least squares problem obtained vectorizing (27)

$$\begin{aligned} Y &:= \begin{bmatrix} \mathrm{vec}(Y_t) \\ \mathrm{vec}(Y_{t+1}) \\ \vdots \\ \mathrm{vec}(Y_T) \end{bmatrix} \doteq S^P \Omega^P + \begin{bmatrix} \mathrm{vec}(E_t) \\ \mathrm{vec}(E_{t+1}) \\ \vdots \\ \mathrm{vec}(E_T) \end{bmatrix} \\ &= S^P \Omega^P + E \end{aligned} \tag{28}$$

where the matrix S^P has the form

$$S^P = \mathrm{block\,diag}\left\{(Z^\top_{[t_0,t)} \otimes I), \dots, (Z^\top_{[t_0,T)} \otimes I)\right\} \tag{29}$$

and Ω^P is given by

$$\begin{aligned} \Omega^P = [&\mathrm{vec}^\top(\Xi_0) \quad \mathrm{vec}^\top(\Xi_1) \quad \mathrm{vec}^\top(\Phi_{11}) \quad \mathrm{vec}^\top(\Psi_{11}) \\ &\cdots \quad \mathrm{vec}^\top(\Xi_\nu) \quad \cdots \quad \mathrm{vec}^\top(\Psi_{\nu 1})]^\top. \end{aligned} \tag{30}$$

Note that the "noise term" E can be written as a linear function of $E_I := [e_t^\top\ e_{t+1}^\top\ \cdots\ e_{T+N-1}^\top]^\top$

$$E =: LE_I \tag{31}$$

where L is an $Nm\bar{\nu} \times (N+\bar{\nu})m$ "selection" matrix (which elements are either zero or one). Equation (31) shows that indeed E has a singular covariance matrix $R = \mathrm{Var}\{E\} = L(I \otimes \Lambda)L^\top$.

This structure can be used to form an estimator $\hat{\Omega}^{P_{\mathrm{opt}}}$ of Ω^P, which has the smallest asymptotic variance among all linear (asymptotically unbiased) estimators based on (28). With the noise covariance R being singular and the regression matrix S^P of full rank, $\hat{\Omega}^{P_{\mathrm{opt}}}$ can be obtained as described in [46, Complement C.4.3]. However, it is possible to reduce (28) to a smaller least squares problem with full rank noise and equality constraints (see [45], [51], [46], and [27] just to cite a few references). We refer the reader to Appendix B for further details. Note, however, that this is just a matter of computational cost, which is of course fundamental when it comes to implementing algorithms, but does not influence the results of this paper. In this implementation, the "optimized" algorithm has a computational complexity that is $O(N^2(\log(N))^\beta)$, while the original PBSID algorithm as well as CCA has a computational complexity that is $O(N(\log(N))^{2\beta})$, where β is the rate at which the past horizon p grows with N (i.e., $p = O((\log(N))^\beta)$; see Assumption 2.3). Of course, this rough evaluation does not take into account constants, which may strongly influence the computation time. For instance, the dimension of input and output signals as well as the length of the future horizon plays an important role.

It should be noted that both PBSID and its optimized version have strong similarities with VARX identification. Results along these lines have been presented in [12], where the relation between PBSID and SSARX (which uses VARX modeling) has been studied; and in [14] and [9], where the relation between $\mathrm{PBSID}_{\mathrm{opt}}$ and VARX modeling has been elucidated. This relation has allowed proposing an implementation of $\mathrm{PBSID}_{\mathrm{opt}}$ of much lower computational complexity than discussed here. However, entering into the fine structure of the constrained least

squares problem (28) would require far more space than available here, and therefore we refer the reader to [9] for details.

We shall use the notation $\hat{\Xi}_h^{P_{\rm opt}}$, $\hat{\Phi}_{ij}^{P_{\rm opt}}$ for the estimators of Ξ_h, Φ_{ij} extracted from the components of $\hat{\Omega}^{P_{\rm opt}}$.

Using the estimator $\hat{\Omega}^{P_{\rm opt}}$, the oblique projections $\hat{Y}_{t+h}^P$ can be substituted with $\hat{Y}_{t+h}^{P_{\rm opt}} = \hat{\Xi}_h^{P_{\rm opt}} Z_{[t_0,t)}$ in the SVD step (17) and an estimator for the state be given by

$$\hat{X}_t^{P_{\rm opt}} := \left(\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{\rm opt}}\right)^{-L} \hat{Y}_{[t,T-1]}^{P_{\rm opt}}. \tag{32}$$

Also the "shifted" oblique projections used for the computation of the state at time $t+1$ [see (20)] can be substituted by

$$\hat{X}_{t+1}^{P_{\rm opt}} := \left(\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{\rm opt}}\right) \begin{bmatrix} \hat{\Xi}_1^{P_{\rm opt}} & \hat{\Phi}_{11}^{P_{\rm opt}} & \hat{\Psi}_{11}^{P_{\rm opt}} \\ \vdots & \vdots & \vdots \\ \hat{\Xi}_\nu^{P_{\rm opt}} & \hat{\Phi}_{\nu\nu}^{P_{\rm opt}} & \hat{\Psi}_{\nu\nu}^{P_{\rm opt}} \end{bmatrix} Z_{[t_0,t+1)}.$$

Similarly, an estimator of the innovation sequence E_t can be found by

$$\hat{E}_t^{P_{\rm opt}} := Y_t - E\left[\hat{Y}_t^{P_{\rm opt}} | \hat{X}_t^{P_{\rm opt}}\right] = Y_t - \hat{C}_N^{P_{\rm opt}} \hat{X}_t^{P_{\rm opt}}. \tag{33}$$

Proposition 5.1: Let the estimators $\hat{A}_N^{P_{\rm opt}}, \hat{B}_N^{P_{\rm opt}}, \hat{C}_N^{P_{\rm opt}}, \hat{K}_N^{P_{\rm opt}}$ be obtained solving[16]

$$\begin{cases} \hat{X}_{t+1}^{P_{\rm opt}} \simeq A\hat{X}_t^{P_{\rm opt}} + BU_t + K\hat{E}_t^{P_{\rm opt}} \\ \hat{Y}_t^{P_{\rm opt}} \simeq C\hat{X}_t^{P_{\rm opt}} \end{cases}$$

in the least squares sense.

Let also $(A_N^{P_{\rm opt}}, B_N^{P_{\rm opt}}, C_N^{P_{\rm opt}}, K_N^{P_{\rm opt}})$ denote the "true" system matrices expressed in the basis corresponding to $\bar{T}_N^{P_{\rm opt}} := (\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{\rm opt}})^{-L}\bar{\Gamma}_{\bar{\nu}}$, i.e., $A_N^{P_{\rm opt}} := \bar{T}_N^{P_{\rm opt}} A (\bar{T}_N^{P_{\rm opt}})^{-1}$, etc.

Then, with the choice of $\hat{\bar{T}}$ in (19), $\bar{T}_N^{P_{\rm opt}}$ converges to the identity matrix and the errors $\tilde{A}_N^{P_{\rm opt}} = \hat{A}_N^{P_{\rm opt}} - A_N^{P_{\rm opt}}$, etc., satisfy

$$\begin{bmatrix} \operatorname{vec}\left(\tilde{A}_N^{P_{\rm opt}}\right) \\ \operatorname{vec}\left(\tilde{B}_N^{P_{\rm opt}}\right) \\ \operatorname{vec}\left(\tilde{C}_N^{P_{\rm opt}}\right) \\ \operatorname{vec}\left(\tilde{K}_N^{P_{\rm opt}}\right) \end{bmatrix} \doteq M_P \left(\hat{\Omega}^{P_{\rm opt}} - \Omega^P\right) \tag{34}$$

for a suitably defined matrix M_P with rows in ℓ_2.

Proof: See Appendix A. ∎

B. CCA Algorithm Revisited

We now move to the CCA algorithm and study its relation to the procedure just described. To make the comparison easier, we rephrase the CCA algorithm using the least square formulation analogous to (27) and (28).

[16]Note that the equation used to estimate C is nonstandard since $\hat{Y}_t^{P_{\rm opt}}$ is used instead of Y_t.

Using (24) and discarding $o_P(1/\sqrt{N})$ terms,[17] the inner projection in (25) can be computed solving in the least squares sense

$$\begin{bmatrix} Y_t \\ Y_{t+1} \\ \vdots \\ Y_T \end{bmatrix} \doteq \begin{bmatrix} \Xi_0 \\ \Xi_1 \\ \vdots \\ \Xi_\nu \end{bmatrix} Z_{[t_0,t)} + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \Phi_{11} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{\nu\nu} & \cdots & \Phi_{\nu 1} & 0 \end{bmatrix} Y_{[t,T]} + \begin{bmatrix} * & * & \cdots & * \\ \Psi_{11} & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{\nu\nu} & \cdots & \Psi_{\nu 1} & * \end{bmatrix} U_{[t,T]} + \begin{bmatrix} E_t \\ E_{t+1} \\ \vdots \\ E_T \end{bmatrix} \tag{35}$$

where the asterisks denote parameters that are estimated but whose value has no interest (and actually should be equal to zero). Since each row is parameterized independently, the orthogonal projections $\hat{E}[Y_{t+h} \,|\, Z_{[t_0,t+h)}, U_{[t+h,T]}]$, for $h = 0, \ldots, \nu$, are equivalent to solving the least squares problem obtained by vectorizing (35) [see also (28)]

$$Y \doteq S^{\rm cca}\Omega^{\rm cca} + E \tag{36}$$

where the matrices $\Omega^{\rm cca}$ and $S^{\rm cca}$ have the form

$$\Omega^{\rm cca} = \begin{bmatrix} \Omega^P \\ * \end{bmatrix} \quad S^{\rm cca} = [S^P \quad *]. \tag{37}$$

The terms denoted with $*$ contain the vectorization of all the $*$s in (35).

As we have just stressed, the inner projection in (25) is equivalent to solving (36) in the least squares sense (with no weighting). We shall denote with $\hat{\Omega}^{\rm cca}$ the corresponding estimator of $\Omega^{\rm cca}$ and with $\hat{\Omega}^{P_{\rm cca}}$, $\hat{\Xi}_k^{\rm cca}$, $\hat{\Phi}_{ij}^{\rm cca}$, and $\hat{\Psi}_{ij}^{\rm cca}$ the estimators of Ω^P, Ξ_k, Φ_{ij}, and Ψ_{ij} extracted from its components.

With this observation, the oblique projections (25) are written in the form

$$\hat{Y}_{t+h} = \hat{\Xi}_h^{\rm cca} Z_{[t_0,t)} + \sum_{i=1}^{h} \hat{\Phi}_{hi}^{\rm cca} \hat{Y}_{t+h-i}. \tag{38}$$

With the same argument used in the proof of Lemma 4.2, we obtain that

$$W_{\rm cca}^{-1} \hat{Y}_{[t,T-1]} \doteq (I \otimes \Lambda^{-1/2}) \begin{bmatrix} \hat{\Xi}_0^{\rm cca} \\ \vdots \\ \hat{\Xi}_{\bar{\nu}}^{\rm cca} \end{bmatrix} Z_{[t_0,t)} \tag{39}$$

and

$$W_{\rm cca}^{-1} \hat{E}_{\|U_{[t+1,T]}} \left[Y_{[t+1,T]} \,|\, Z_{[t_0,t+1)}\right] \doteq (I \otimes \Lambda^{-1/2}) \begin{bmatrix} \hat{\Xi}_1^{\rm cca} & \hat{\Phi}_{11}^{\rm cca} & \hat{\Psi}_{11}^{\rm cca} \\ \vdots & \vdots & \vdots \\ \hat{\Xi}_\nu^{\rm cca} & \hat{\Phi}_{\nu\nu}^{\rm cca} & \hat{\Psi}_{\nu\nu}^{\rm cca} \end{bmatrix} Z_{[t_0,t+1)}. \tag{40}$$

[17]See Appendix B for a formal justification.

The state sequences $\hat{X}_t^{\rm cca}$, $\hat{X}_{t+1}^{\rm cca}$ are constructed as described in Section III. Substituting the right-hand sides of (39) and (40) in the CCA algorithm does not change its asymptotic properties as stated in Lemma 3.1.

The estimator for the innovation sequence is taken here of the form

$$\hat{E}_t^{\rm cca} := Y_t - \hat{E}\left[Y_t \,|\, \hat{X}_t^{\rm cca}\right] = Y_t - \hat{C}_N^{\rm cca}\hat{X}_t^{\rm cca}. \tag{41}$$

Proposition 5.2: Let the estimators $\hat{A}_N^{\rm cca}, \hat{B}_N^{\rm cca}, \hat{C}_N^{\rm cca}, \hat{K}_N^{\rm cca}$ be obtained solving

$$\begin{cases} \hat{X}_{t+1}^{\rm cca} \simeq A\hat{X}_t^{\rm cca} + BU_t + K\hat{E}_t^{\rm cca} \\ \hat{Y}_t \simeq C\hat{X}_t^{\rm cca} \end{cases}$$

in the least squares sense.

Let also $(A_N^{\rm cca}, B_N^{\rm cca}, C_N^{\rm cca}, K_N^{\rm cca})$ denote the "true" system matrices expressed in the basis corresponding to

$$T_N^{\rm cca} := \hat{\Gamma}_{\bar{\nu}}^{-L}\Gamma_{\bar{\nu}} \tag{42}$$

i.e., $A_N^{\rm cca} := T_N^{\rm cca}A(T_N^{\rm cca})^{-1}$, etc. Then, with the choice of $\hat{T}$ in (13), $T_N^{\rm cca}$ converges to the identity matrix and the errors $\tilde{A}_N^{\rm cca} = \hat{A}_N^{\rm cca} - A_N^{\rm cca}$, etc., satisfy

$$\begin{aligned} &\begin{bmatrix} \operatorname{vec}\left(\tilde{A}_N^{\rm cca}\right) \\ \operatorname{vec}\left(\tilde{B}_N^{\rm cca}\right) \\ \operatorname{vec}\left(\tilde{C}_N^{\rm cca}\right) \\ \operatorname{vec}\left(\tilde{K}_N^{\rm cca}\right) \end{bmatrix} \\ &\quad \doteq M_P(\hat{\Omega}^{P_{\rm opt}} - \Omega^P) + M_1^{\rm cca}(\hat{\Omega}^{P_{\rm cca}} - \hat{\Omega}^{P_{\rm opt}}) \\ &\qquad + M_2^{\rm cca}\operatorname{vec}(\hat{\Xi}_0 - \hat{\Xi}_0^{P_{\rm opt}}) \end{aligned} \tag{43}$$

where $\hat{\Xi}_0$ is defined in (A.61). The matrix M_P is the same that appears in (34), while $M_1^{\rm cca}$ and $M_2^{\rm cca}$ are suitably defined matrices with rows in ℓ_2.

Proof: See Appendix A. ∎

C. Comparison Between CCA and PBSID$_{\rm opt}$

The optimized version of PBSID introduced above can now be easily compared to the standard CCA algorithm without making assumptions on the input spectrum (besides, of course, persistency of excitation conditions and absence of feedback). We first state the main result as a theorem; the remaining part of this section shall be devoted to the proof.

Theorem 5.3: Let Θ be any system invariant that depends differentiably on the system matrices (A, B, C, K). Denote with $\hat{\Theta}^{\rm cca}$ and $\hat{\Theta}^{P_{\rm opt}}$ the estimators of any such Θ using, respectively, CCA and PBSID$_{\rm opt}$; then

$$\operatorname{AsVar}\{\sqrt{N}\hat{\Theta}^{\rm cca}\} \geq \operatorname{AsVar}\{\sqrt{N}\hat{\Theta}^{P_{\rm opt}}\}. \tag{44}$$

Proof: As seen in (36) and (28), the first step of the algorithms can be seen as a linear least squares problem; there are, however, three "complications" that make the analysis more difficult.

1) The effect of the initial condition (the $o_P(1/\sqrt{N})$ terms).
2) The regression matrices S^P and $S^{\rm cca}$ are data dependent.
3) The dimension of the parameter vector Ω^P grows with the sample size.

We shall deal with these problems, respectively, as follows.

1) We shall see in Appendix B (see Lemma B.1 and the discussion following the lemma) that the $o_P(1/\sqrt{N})$ terms that have been omitted in (28) have a special form that allows, indeed, to neglect them.
2) Under Assumption 2.2, it is tedious but easy to adapt standard properties of Markov estimators (i.e., unbiasedness and minimum variance; see, for instance, [46, Lemma 4.3, Proof A]) to the case in which the regression matrix $(S^P, S^{\rm cca})$ is "data dependent." Of course, unbiasedness and minimum variance hold only asymptotically.
3) We shall follow the approach of [39], studying the (scalar) estimators $\eta_N^\top\hat{\Omega}^P$ with η_N satisfying Assumption 2.4.

As discussed in Appendix B, the estimators $\hat{\Omega}^{P_{\rm cca}}$ and $\hat{\Omega}^{P_{\rm opt}}$ satisfy

$$\operatorname{AsVar}\left\{\sqrt{N}\eta_N^\top\hat{\Omega}^{P_{\rm cca}}\right\} \geq \operatorname{AsVar}\left\{\sqrt{N}\eta_N^\top\hat{\Omega}^{P_{\rm opt}}\right\} \tag{45}$$

$\forall\, \eta_N$ satisfying Assumption 2.4. In particular, with $\tilde{\Omega}^{P_{\rm opt}} := \hat{\Omega}^{P_{\rm opt}} - \Omega^P$ being the "optimal" estimation error, $\tilde{\Omega}^{P_{\rm cca}} := \hat{\Omega}^{P_{\rm cca}} - \Omega^P$ can also be written as

$$\sqrt{N}\tilde{\Omega}^{P_{\rm cca}} = \sqrt{N}\tilde{\Omega}^{P_{\rm opt}} + \sqrt{N}(\hat{\Omega}^{P_{\rm cca}} - \hat{\Omega}^{P_{\rm opt}}) \tag{46}$$

where the two terms on the right-hand side are asymptotically uncorrelated, i.e.,

$$\operatorname{AsCov}\{\sqrt{N}\eta_N^\top\tilde{\Omega}^{P_{\rm opt}}, \sqrt{N}\gamma_N^\top(\hat{\Omega}^{P_{\rm cca}} - \hat{\Omega}^{P_{\rm opt}})\} = 0$$

$\forall \eta_N, \gamma_N$ satisfying Assumption 2.4. The reason for this inequality is twofold.

1) In (36), the parameters denoted with asterisks are estimated even though they are known to be zero (observe that instead in (27) and (28) the lower triangular structure of the matrix describing the link from future input to future output is enforced).
2) The "optimized" PBSID algorithm solves (28) in an optimal fashion.

Furthermore, a similar decomposition holds also for $\hat{\Xi}_0$, i.e., $\tilde{\Xi}_0 := \hat{\Xi}_0 - \Xi_0$ can be written as

$$\sqrt{N}\tilde{\Xi}_0 = \sqrt{N}\left(\hat{\Xi}_0^{P_{\rm opt}} - \Xi_0\right) + \sqrt{N}(\hat{\Xi}_0 - \hat{\Xi}_0^{P_{\rm opt}}) \tag{47}$$

where $\sqrt{N}\eta_N^\top\operatorname{vec}(\hat{\Xi}_0 - \hat{\Xi}_0^{P_{\rm opt}})$ is (asymptotically) uncorrelated from $\sqrt{N}\gamma_N^\top\hat{\Omega}^{P_{\rm opt}} = \sqrt{N}\gamma_N^\top(\hat{\Omega}^{P_{\rm opt}} - \Omega^P)\,\forall\,\eta_N, \gamma_N$ satisfying Assumption 2.4.

These latter observations imply that ($\sqrt{N}$ times) the last two terms on the right-hand side of (43) are asymptotically uncorrelated with ($\sqrt{N}$ times) the first. Therefore, from (34) and (43)

$$\begin{aligned}\mathrm{AsVar}&\left\{\begin{array}{l}\sqrt{N}\mathrm{vec}\left(\tilde{A}_N^{\mathrm{cca}}\right)\\ \sqrt{N}\mathrm{vec}\left(\tilde{B}_N^{\mathrm{cca}}\right)\\ \sqrt{N}\mathrm{vec}\left(\tilde{C}_N^{\mathrm{cca}}\right)\\ \sqrt{N}\mathrm{vec}\left(\tilde{K}_N^{\mathrm{cca}}\right)\end{array}\right\}\\ &=\mathrm{AsVar}\left\{\begin{array}{l}\sqrt{N}\mathrm{vec}\left(\tilde{A}_N^{P_{\mathrm{opt}}}\right)\\ \sqrt{N}\mathrm{vec}\left(\tilde{B}_N^{P_{\mathrm{opt}}}\right)\\ \sqrt{N}\mathrm{vec}\left(\tilde{C}_N^{P_{\mathrm{opt}}}\right)\\ \sqrt{N}\mathrm{vec}\left(\tilde{K}_N^{P_{\mathrm{opt}}}\right)\end{array}\right\}\\ &\quad+\mathrm{AsVar}\left\{\sqrt{N}\left[M_1^{\mathrm{cca}}\quad M_2^{\mathrm{cca}}\right]\begin{bmatrix}(\hat{\Omega}^{P_{\mathrm{cca}}}-\hat{\Omega}^{P_{\mathrm{opt}}})\\ \mathrm{vec}(\hat{\Xi}_0-\hat{\Xi}_0^{P_{\mathrm{opt}}})\end{bmatrix}\right\}\end{aligned}$$

where the fact that the rows of M_P, M_1^{cca}, and M_2^{cca} are in ℓ_2 has been used. This concludes the proof. ■

Remark V.4: It is our experience from the simulation experiments that the "optimized" PBSID algorithm does not introduce significant improvements with respect to PBSID (see Fig. 4), while it does increase (see Section III.B) the computational complexity due to the solution of the constrained least squares problem. See Appendix B for details.

However, this algorithm can be implemented with a much lower computational complexity, as described in [14] and [9]. The weighting step needed to find $\hat{\Omega}^{P_{\mathrm{opt}}}$ (see Appendix B) is, however, necessary to obtain (45). ♢

VI. Simulation Results

Consider the first-order system

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

In example 1, we set $\gamma=0$ (ARX), while in example 2, $\gamma=0.5$ (ARMAX).

The input is either unit variance white noise or unit variance white noise passed through the filter $H_u(z)=(z^2+0.8z+0.55/z^2-0.5z+0.9)$; the input spectrum is plotted in Fig. 1.

We report results concerning the asymptotic variance and the sample variance estimated over 100 Monte Carlo runs (both multiplied by the number $N=1000$ of data points used in each experiment) of the deterministic transfer function $F(z)=(1/z-0.5)$ (equal for the two examples); the future and past horizons are chosen to be $p=\nu=10$. In Fig. 4 (top), we also report the results for $\nu=1$, $p=10$. In Fig. 4, we show the dependence of the asymptotic variance as a function of the future horizon ν measured by the efficiency index

$$\mathrm{Eff}(\nu):=\frac{\int_0^{2\pi}\mathrm{AsVar}\{\hat{F}_\nu(e^{j\omega})\}\,d\omega}{\int_0^{2\pi}\mathrm{CRLB}(e^{j\omega})\,d\omega}\qquad(48)$$

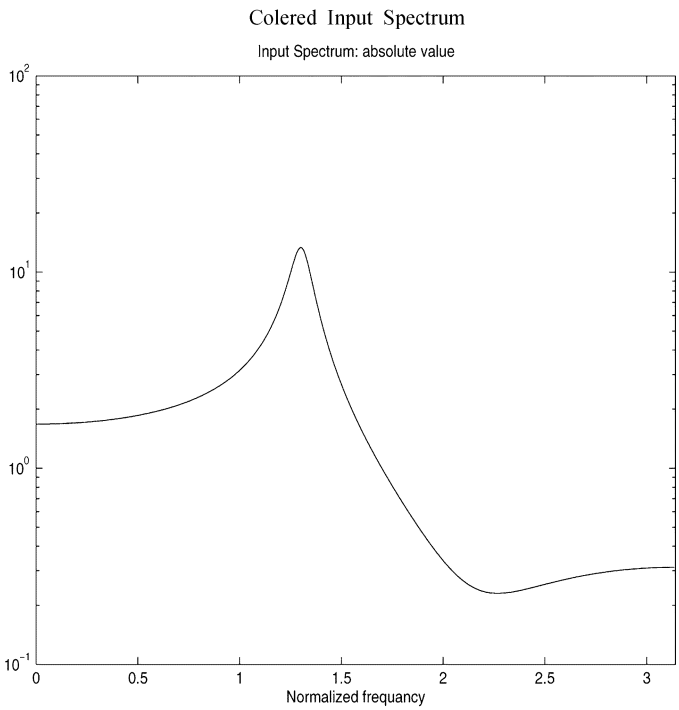

Fig. 1. Colored input spectrum: absolute value.

where $\hat{F}_\nu(e^{j\omega})$ is the PBSID-estimator of the frequency response $F(e^{j\omega})=C(e^{j\omega}I-A)^{-1}B$ when the future horizon is ν and $\mathrm{CRLB}(e^{j\omega})$ is the Cramér–Rao lower bound as a function of ω.

Note that for the white input case [see Figs. 2 and 3 (top)], both CCA and the predictor-based algorithm are indistinguishable from PEM; in fact all algorithms reach the Cramér–Rao lower bound in the examples considered.

It is also remarkable that the sample variance estimated from the simulations (and its "theoretical" value computed using the formulas of [13]) reaches the Cramér–Rao lower bound also for colored inputs when the system is ARX [see Fig. 2 (bottom)].

In the colored input case [see Figs. 2 and 3 (bottom) and Fig. 4] the results are fundamentally different: CCA behaves significantly worse than PEM and PBSID. We also report the asymptotic variance computed using the formulas that can be found in [13] *(dotted line)* and the Cramér–Rao lower bound (Crlb, *solid line*).

The algorithm by Jansson [34] is always indistinguishable from PBSID, as predicted by the results in [12] and [9].

It is interesting to observe that in this particular example and with colored inputs, the asymptotic variance of PBSID/$\mathrm{PBSID}_{\mathrm{opt}}$ is close to the Cramér–Rao lower bound (even though it does not reach it) for $\nu=n=1$ (see Fig. 4). Note that this behavior departs sharply from what happens to CCA with white inputs (and hence also to PBSID by Theorem 4.1); in that case, in fact, the asymptotic variance decreases monotonically as a function of ν (see [7]).

Note also (see Fig. 4) that the modified version $\mathrm{PBSID}_{\mathrm{opt}}$ behaves as PBSID in the example considered.

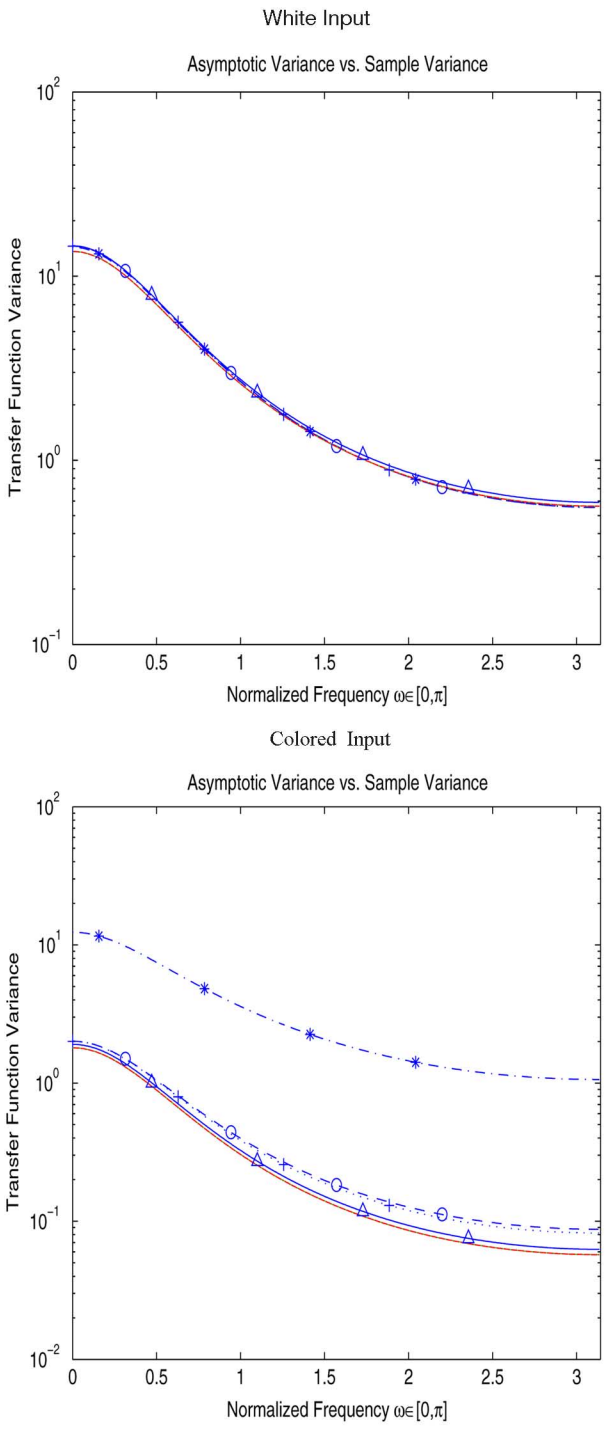

Fig. 2. Example 1 *(ARX of order 1):* Asymptotic variance (and its Monte Carlo estimate) versus normalized frequency ($\omega \in [0, \pi]$). Solid with triangles ($\triangle$): PEM; dashed with stars ($*$): CCA; dotted with crosses ($+$): "predictor-based" algorithm (PBSID); dotted with circles (o): Jansson's algorithm; dotted: asymptotic variance for PBSID; solid: Cramér–Rao lower bound.

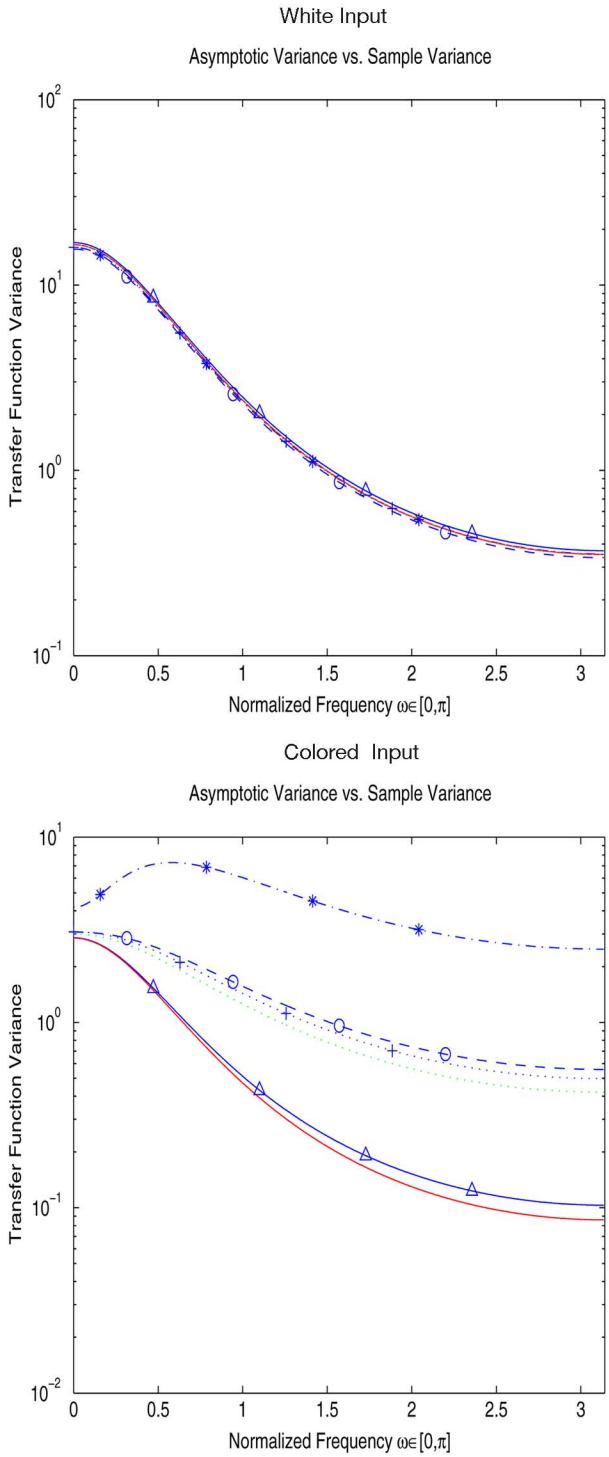

Fig. 3. Example 2 *(ARMAX of order 1):* Asymptotic variance (and its Monte Carlo estimate) versus normalized frequency ($\omega \in [0, \pi]$). Solid with triangles ($\triangle$): PEM; dashed with stars ($*$): CCA; dotted with crosses ($+$): "predictor-based" algorithm (PBSID); dotted with circles (o): Jansson's algorithm; dotted: asymptotic variance for PBSID; solid: Cramér–Rao lower bound.

VII. Conclusion

In this paper, we have shown that the PBSID algorithm, introduced in [20] under the name "whitening filter" algorithm, is asymptotically equivalent to CCA when measured inputs are white or absent. Our analysis is supported by both the simulation results and the asymptotic variance formulas computed in [13].

The significance of this result is strengthened by the fact that, as shown in [12] and [9], PBSID and the SSARX algorithm in [34] are asymptotically equivalent.

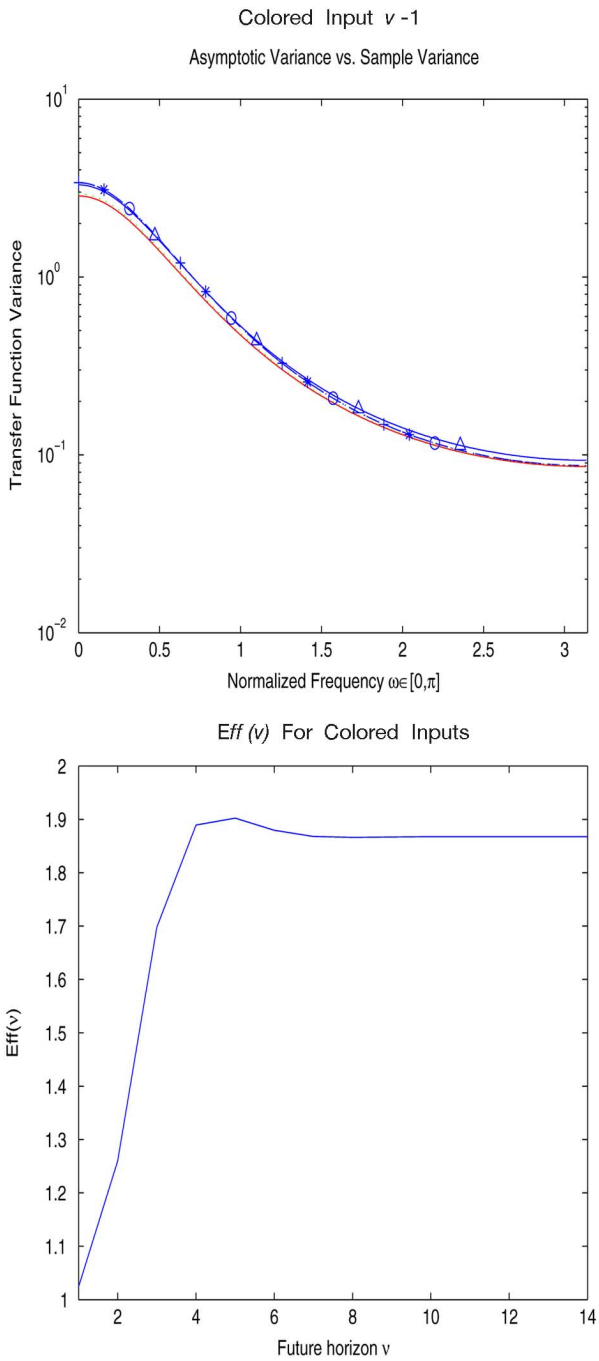

Fig. 4. Example 3 *(ARMAX of order 1).* (Top) asymptotic variance (and its Monte Carlo estimate) versus normalized frequency ($\omega \in [0, \pi]$). Solid with triangles ($\triangle$): PEM; dashed with stars ($*$): CCA; dotted with crosses ($+$): "predictor-based" algorithm (PBSID); dotted with circles (o): $\text{PBSID}_{\text{opt}}$ algorithm; dotted: asymptotic variance for PBSID, solid: Cramér–Rao lower bound. (Bottom) Index $\text{Eff}(\nu)$ [see (48)] as a function of ν.

We have also proposed a slightly modified version of PBSID, called $\text{PBSID}_{\text{opt}}$, which provides estimators with a lower asymptotic variance than CCA; see Theorem 5.3. We remind the reader that both PBSID and its "optimized" version $\text{PBSID}_{\text{opt}}$ are able to deal with feedback. We refer the reader to [14] and [9] for computationally efficient implementations of $\text{PBSID}_{\text{opt}}$.

An important question that remains open concerns efficiency. In [38], it is claimed that a procedure that is essentially equivalent to the SSARX algorithm is efficient for general input signal. However, in [38], the past and future horizons [see [38, (2) and the definitions before (9)]] are assumed to be equal and, for consistency, let to go to infinity. Computations based on the asymptotic variance [see Fig. 4 (bottom) for an example with $p = \nu = 10$, but essentially unchanged results are obtained increasing $p = \nu$] show that indeed that claim is not correct and instead efficiency is never attained; note also that, in this example, the "optimal" future horizon is $\nu = 1$ [see Fig. 4 (bottom) and [10]].

Therefore, we believe, the quest for an asymptotically efficient subspace procedure with inputs is still open; also a methodology to optimally chose ν is missing[18]; see [10] for some preliminary results. We hope the results of this paper have shed some light towards this direction.

Also the question of nonasymptotic relative performance (i.e., with "finite data"), which is of primary importance for practical purposes, remains open and, in our opinion, deserves further investigation. This is, we believe, one of the main open research directions in the area of subspace identification.

Appendix A
Proofs

Proof of Proposition 3.1: Recall that, defining $\hat{E}_t^i = Y_t - \hat{E}[Y_t \mid \hat{X}_t^i]$, the least square estimators $\hat{A}^i, \hat{B}^i, \hat{C}^i$, $i = 1, 2$ are obtained from

$$[\hat{A}^i \quad \hat{B}^i \quad \hat{K}^i] := \hat{X}_{t+1}^i \begin{bmatrix} \left(\hat{X}_t^i\right)^\top & U_t^\top & \left(\hat{E}_t^i\right)^\top \end{bmatrix} \cdot \left\{ \begin{bmatrix} \hat{X}_t^i \\ U_t \\ \hat{E}_t^i \end{bmatrix} \begin{bmatrix} \left(\hat{X}_t^i\right)^\top & U_t^\top & \left(\hat{E}_t^i\right)^\top \end{bmatrix} \right\}^{-1} \tag{A.49}$$

$$\hat{C}^i := Y_t \left(\hat{X}_t^i\right)^\top \left\{ \hat{X}_t^i \left(\hat{X}_t^i\right)^\top \right\}^{-1}. \tag{A.50}$$

For simplicity of exposition we shall only deal with (A.50); using (8) and recalling that $\hat{X}_t^1 \doteq \hat{X}_t^2$ means that there exists $\hat{T}_N$, $\lim_{N\to\infty} \hat{T}_N = I$, so that $\hat{X}_t^1 = \hat{T}_N \hat{X}_t^2 + \underline{o}_P(1/\sqrt{N})$, we have

$$\begin{aligned} \hat{C}^1 &= \frac{Y_t \left(\hat{X}_t^1\right)^\top}{N} \left[\frac{\hat{X}_t^1 \left(\hat{X}_t^1\right)^\top}{N} \right]^{-1} \\ &= \frac{Y_t (T_N \hat{X}_t^2 + \underline{o}_P(1/\sqrt{N}))^\top}{N} \\ &\quad \cdot \left\{ \frac{(\hat{T}_N \hat{X}_t^2 + \underline{o}_P(1/\sqrt{N}))(\hat{T}_N \hat{X}_t^2 + \underline{o}_P(1/\sqrt{N}))^\top}{N} \right\}^{-1}. \end{aligned}$$

Using the fact that, for instance, $(1/N)\sum_{i=0}^{N-1} y_{t+i} o_P(1/\sqrt{N}) = o_P(1/\sqrt{N})$ and recalling that,

[18]Recall that here p is supposed to go to infinity according to Assumption 2.3.

given a positive definite matrix Σ, $(\Sigma + \underline{o}_P(1/\sqrt{N}))^{-1} = \Sigma^{-1} + \underline{o}_P(1/\sqrt{N})$, the last term can be simplified to yield

$$\hat{C}^1 = \frac{Y_t \left(\hat{X}_t^1\right)^\top}{N} \hat{T}_N^\top \left\{ \frac{\hat{T}_N \hat{X}_t^1 \left(\hat{X}_t^1\right)^\top \hat{T}_N^\top}{N} \right\}^{-1} + \underline{o}_P(1/\sqrt{N})$$
$$= \hat{C}^2 \hat{T}_N^{-1} + \underline{o}_P(1/\sqrt{N}).$$

Similarly

$$\hat{A}^1 = \hat{T}_N \hat{A}^2 \hat{T}_N^{-1} + \underline{o}_P(1/\sqrt{N})$$
$$\hat{B}^1 = \hat{T}_N \hat{B}^2 + \underline{o}_P(1/\sqrt{N})$$
$$\hat{K}^1 = \hat{T}_N \hat{K}^2 + \underline{o}_P(1/\sqrt{N})$$

can be proven to hold. ∎

Proof of Lemma 3.2: Denoting with Π_{U_n} the orthogonal projector operator onto the column space of U_n, (12) can be rewritten as

$$\hat{\Gamma}_{\bar{\nu}} := \hat{W}_{\text{cca}} U_n \hat{T} = \hat{W}_{\text{cca}} U_n \left(U_n^\top U_n\right)^{-1} U_n^\top \hat{W}_{\text{cca}}^{-1} \Gamma_{\bar{\nu}}$$
$$= \hat{W}_{\text{cca}} \Pi_{U_n} \hat{W}_{\text{cca}}^{-1} \Gamma_{\bar{\nu}}.$$

Let W_{cca} be the a.s. limit of $\hat{W}_{\text{cca}}$. It is well known that under mild conditions, e.g., $\bar{\nu}$ larger than the system order, Assumption 2.3, and (2), the column space of $\hat{W}_{\text{cca}} U_n = \hat{\Gamma}_{\bar{\nu}} \hat{T}^{-1}$ converges to the column space of $\Gamma_{\bar{\nu}}$ (see, for instance, [22, Theorems 9 and 10]). Therefore also the column space of U_n converges to the column space of $W_{\text{cca}}^{-1} \Gamma_{\bar{\nu}}$ and hence $\Pi_{U_n} \hat{W}_{\text{cca}}^{-1} \Gamma_{\bar{\nu}}$ converges to $W_{\text{cca}}^{-1} \Gamma_{\bar{\nu}}$, from which $\hat{\Gamma}_{\bar{\nu}} = \hat{W}_{\text{cca}} \Pi_{U_n} \hat{W}_{\text{cca}}^{-1} \Gamma_{\bar{\nu}}$ converges to $\Gamma_{\bar{\nu}}$. ∎

Proof of Lemma 4.2: First let us note that

$$\hat{E}_{\|U_{[t,T]}} \left[Y_{t+h} \,|\, Z_{[t_0,t)} \right]$$
$$= \hat{E}_{\|U_{[t,T]}} \left[\hat{E}\big[Y_{t+h} \,|\, Z_{[t_0,t+h)}, U_{[t+h,T]} \big] \,|\, Z_{[t_0,t)} \right]. \quad \text{(A.51)}$$

To simplify notation, let $P := Z_{[t_0,t+h)}$ (past) and $F := U_{[t+h,T]}$ (future). Under the assumption that $\mathbf{u}(t)$ is white (or absent of course), the rows of $U_{[t+h,T]}$ are asymptotically orthogonal to the rows of $Z_{[t_0,t+h)}$ and also to the rows of Y_{t+h}; therefore, from the uniform convergence of sample covariances (see, for instance, [30, Theorem 5.3.2]), it follows that $\hat{\Sigma}_{\mathbf{fp}} := (FP^\top/N)$ and $\hat{\Sigma}_{\mathbf{yf}} := (Y_{t+h}F^\top/N)$ satisfy

$$\|\hat{\Sigma}_{\mathbf{fp}}\|_2 = O(\sqrt{pQ_N}) \quad \|\hat{\Sigma}_{\mathbf{yf}}\|_2 = O(\sqrt{Q_N})$$

which implies

$$\|\hat{\Sigma}_{\mathbf{fp}}\|_2 \|\hat{\Sigma}_{\mathbf{fp}}\|_2 = O(pQ_N)$$
$$\|\hat{\Sigma}_{\mathbf{yf}}\|_2 \|\hat{\Sigma}_{\mathbf{fp}}\|_2 = O(\sqrt{p}Q_N). \quad \text{(A.52)}$$

Now we write the inner projection in (A.51) as follows:

$$\hat{E}\left[Y_{t+h} \,|\, Z_{[t_0,t+h)}, U_{[t+h,T]} \right]$$
$$= \hat{E}[Y_{t+h} \,|\, P, F]$$
$$= \hat{E}_{\|P}[Y_{t+h} \,|\, F] + \hat{E}_{\|F}[Y_{t+h} \,|\, P]. \quad \text{(A.53)}$$

Recall now that, given matrices $\Sigma_1(N)$, $\Delta\Sigma_1(N)$, $\Sigma_2(N)$, $\Delta\Sigma_2(N)$ of appropriate dimensions possibly dependent on N, with $\|\Sigma_1(N)\|_2 = O(1)$, $\Sigma_2(N)$ a.s. invertible with bounded inverse $\|\Sigma_2^{-1}(N)\|_2 = O(1)$, and $\|\Delta\Sigma_1(N)\|_2$, $\|\Delta\Sigma_2(N)\|_2$ infinitesimal (a.s.) as $N \to \infty$

$$\left(\Sigma_1(N) + \Delta\Sigma_1(N)\right)\left(\Sigma_2(N) + \Delta\Sigma_2(N)\right)^{-1}$$
$$= \Sigma_1(N)\Sigma_2^{-1}(N) + \Delta\Sigma_1(N)\Sigma_2^{-1}(N) +$$
$$- \Sigma_1(N)\Sigma_2^{-1}(N)\Delta\Sigma_2(N)\Sigma_2^{-1}(N)$$
$$+ \underline{o}(\|\Delta\Sigma_2(N)\|_2) \quad \text{(A.54)}$$

holds. Now we apply (A.54) to the oblique projection

$$\hat{E}_{\|F}[Y_{t+h} \,|\, P] := \hat{\Sigma}_{\mathbf{yp|f}} \hat{\Sigma}_{\mathbf{pp|f}}^{-1} P$$
$$= \left(\hat{\Sigma}_{\mathbf{yp}} - \hat{\Sigma}_{\mathbf{yf}} \hat{\Sigma}_{\mathbf{ff}}^{-1} \hat{\Sigma}_{\mathbf{fp}}\right)$$
$$\cdot \left(\hat{\Sigma}_{\mathbf{pp}} - \hat{\Sigma}_{\mathbf{pf}} \hat{\Sigma}_{\mathbf{ff}}^{-1} \hat{\Sigma}_{\mathbf{fp}}\right)^{-1} P$$

with $\Sigma_1(N) := \hat{\Sigma}_{\mathbf{yp}}$, $\Delta\Sigma_1(N) := -\hat{\Sigma}_{\mathbf{yf}} \hat{\Sigma}_{\mathbf{ff}}^{-1} \hat{\Sigma}_{\mathbf{fp}}$, $\Sigma_2(N) := \hat{\Sigma}_{\mathbf{pp}}$, and $\Delta\Sigma_2(N) := -\hat{\Sigma}_{\mathbf{pf}} \hat{\Sigma}_{\mathbf{ff}}^{-1} \hat{\Sigma}_{\mathbf{fp}}$.

Observe that the assumption $\Phi_{\mathbf{z}} \geq cI > 0$ implies (uniformly in p; see [29] and [25]) that $\Sigma_{\mathbf{pp}} \geq cI > 0$ and therefore $\|\Sigma_{\mathbf{pp}}^{-1}\|_2 \leq \sqrt{1/c} < \infty$. From the uniform convergence of sample covariances [30, Theorem 5.3.2], $\|\hat{\Sigma}_{\mathbf{pp}} - \Sigma_{\mathbf{pp}}\|_2 = O(p\sqrt{Q_N})$ holds and therefore $\hat{\Sigma}_{\mathbf{pp}}$ is a.s. invertible and $\|\hat{\Sigma}_{\mathbf{pp}}^{-1}\|_2 = O(1)$. With similar argument, which uses the fact that the covariance $E[\mathbf{y}(t)\mathbf{z}^\top(t-\tau)]$ goes to zero exponentially as a function of τ (since $|\lambda_{\max}(A)| < 1$ and $\mathbf{u}$ white), one can show that $\left\|\hat{\Sigma}_{\mathbf{yp}}\right\|_2 = O(1)$.

Therefore, we obtain

$$\hat{E}_{\|F}[Y_{t+h} \,|\, P] = \Big[\hat{\Sigma}_{\mathbf{yp}} \hat{\Sigma}_{\mathbf{pp}}^{-1} - \hat{\Sigma}_{\mathbf{yf}} \hat{\Sigma}_{\mathbf{ff}}^{-1} \hat{\Sigma}_{\mathbf{fp}} \hat{\Sigma}_{\mathbf{pp}}^{-1}$$
$$+ \hat{\Sigma}_{\mathbf{yp}} (\hat{\Sigma}_{\mathbf{pp}}^{-1} \hat{\Sigma}_{\mathbf{pf}} \hat{\Sigma}_{\mathbf{ff}}^{-1} \hat{\Sigma}_{\mathbf{fp}} \hat{\Sigma}_{\mathbf{pp}}^{-1})\Big] P$$
$$+ \underline{o}\left(\left\|\hat{\Sigma}_{\mathbf{pf}} \hat{\Sigma}_{\mathbf{ff}}^{-1} \hat{\Sigma}_{\mathbf{fp}}\right\|_2\right) P.$$

Using (A.52), $\|\hat{\Sigma}_{\mathbf{pp}}^{-1}\|_2 = O(1)$, $\|\hat{\Sigma}_{\mathbf{yp}}\|_2 = O(1)$, it is now easy to verify that $\tilde{\Psi} := -\hat{\Sigma}_{\mathbf{yf}} \hat{\Sigma}_{\mathbf{ff}}^{-1} \hat{\Sigma}_{\mathbf{fp}} \hat{\Sigma}_{\mathbf{pp}}^{-1} + \hat{\Sigma}_{\mathbf{yp}} (\hat{\Sigma}_{\mathbf{pp}}^{-1} \hat{\Sigma}_{\mathbf{pf}} \hat{\Sigma}_{\mathbf{ff}}^{-1} \hat{\Sigma}_{\mathbf{fp}} \hat{\Sigma}_{\mathbf{pp}}^{-1})$ satisfies $\|\tilde{\Psi}\|_2 = O(pQ_N)$. Since the elements of P are $O_P(1)$, it follows that the product $\tilde{\Psi}P = \underline{O}_P(p^2 Q_N) = \underline{o}_P(1/\sqrt{N})$.

Using the last equation the oblique projection $\hat{E}_{\|F}[Y_{t+h} \,|\, P]$ becomes, for the purpose of asymptotic analysis

$$\hat{E}_{\|F}[Y_{t+h} \,|\, P] \doteq \hat{\Sigma}_{\mathbf{yp}} \hat{\Sigma}_{\mathbf{pp}}^{-1} P = \hat{E}[Y_{t+h} \,|\, P]. \quad \text{(A.55)}$$

Similarly, one can show that

$$\hat{E}_{\|P}[Y_{t+h} \,|\, F] \doteq \left(\hat{\Sigma}_{\mathbf{yf}} - \hat{\Sigma}_{\mathbf{yp}} \hat{\Sigma}_{\mathbf{pp}}^{-1} \hat{\Sigma}_{\mathbf{pf}}\right) \hat{\Sigma}_{\mathbf{ff}}^{-1} F. \quad \text{(A.56)}$$

Using (A.55) and (A.56), we obtain

$$\hat{E}[Y_{t+h} \,|\, Z_{[t_0,t+h)}, U_{[t+h,T]}] \doteq \hat{E}\left[Y_{t+h} \,|\, Z_{[t_0,t+h)}\right] + \hat{\Theta} U_{[t+h,T]}$$

for a suitable matrix $\hat{\Theta}$, which follows from (A.56). Next, observe that $\hat{E}[Y_{t+h} \mid Z_{[t_0,t+h)}]$ can be written in the form

$$\hat{E}[Y_{t+h} \mid Z_{[t_0,t+h)}] = \hat{Y}^P_{t+h} + \sum_{i=1}^{h} \hat{\Phi}_{hi} Y_{t+h-i} + \hat{\Psi}_{hi} U_{t+h-i} \tag{A.57}$$

for suitable matrix coefficients $\hat{\Psi}_{hi}$, $\hat{\Phi}_{hi}$. Taking now the oblique projection $\hat{E}_{\|U_{[t,T]}}[\cdot \mid Z_{[t_0,t)}]$ of both sides of (A.57), we obtain

$$\begin{aligned}\hat{Y}_{t+h} &= \hat{E}_{\|U_{[t,T]}}[Y_{t+h} \mid Z_{[t_0,t)}] \\ &\doteq \hat{E}_{\|U_{[t,T]}}\left[\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}\right] \mid Z_{[t_0,t)}\right] \\ &= \hat{Y}^P_{t+h} + \sum_{i=1}^{h} \hat{\Phi}_{hi} \hat{Y}_{t+h-i}\end{aligned} \tag{A.58}$$

where $\hat{E}_{\|U_{[t,T]}}[\hat{\Theta} U_{[t+h,T]} \mid Z_{[t_0,t)}] = 0$ has been used.

In matrix form, this becomes

$$\hat{Y}^P_{[t,T]} \doteq \begin{bmatrix} I & 0 & \cdots & 0 \\ -\hat{\Phi}_{11} & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\Phi}_{\bar{\nu},\bar{\nu}} & -\hat{\Phi}_{\bar{\nu},\bar{\nu}-1} & \cdots & I \end{bmatrix} \hat{Y}_{[t,T]}.$$

Since $\hat{Y}^P_{[t,T]}$ represents a weighted version of $\hat{Y}_{[t,T]}$, only the asymptotic value of this weight matters as far as any system invariant is concerned. Therefore, we need to study the limit $\lim_{N\to\infty} \hat{\Phi}_{ij} := \Phi_{ij}$. Recall that, according to Assumption 2.3, also $t_0 - t_0 \to \infty$ when $N \to \infty$ and therefore Φ_{hi} are simply the coefficients of the (stationary) one step-ahead predictor. As already stated in (24), it is a standard fact (see, for instance [46], [30], and [41]) that $\Phi_{hi} = C\bar{A}^{i-1}K$.

The fact that

$$\begin{bmatrix} I & 0 & \cdots & 0 \\ -CK & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -C\bar{A}^{\bar{\nu}-2}K & -C\bar{A}^{\bar{\nu}-3}K & \cdots & I \end{bmatrix}$$

is the inverse of $H_{\bar{\nu}}$ is a simple exercise and is left to the reader. ∎

Proof of Lemma 5.1: We prove now only (34) regarding $\tilde{C}_N^{P_{\rm opt}}$, with those being related to $\tilde{A}_N^{P_{\rm opt}}$, $\tilde{B}_N^{P_{\rm opt}}$, $\tilde{K}_N^{P_{\rm opt}}$ completely analogous.

First note that, from Lemma 3.4, $\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{\rm opt}}$ converges to $\bar{\Gamma}_{\bar{\nu}}$ and therefore $(\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{\rm opt}})^{-L}\bar{\Gamma}_{\bar{\nu}}$ converges to the identity matrix. Recall now that

$$\left(\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{\rm opt}}\right)^{-L} \begin{bmatrix} \Xi_0 \\ \vdots \\ \Xi_{\bar{\nu}} \end{bmatrix} Z_{[t_0,t)} \doteq T_N^{P_{\rm opt}} X_t$$

and that the state estimator $\hat{X}_t^{P_{\rm opt}}$, from (32), is of the form

$$\hat{X}_t^{P_{\rm opt}} = \left(\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{\rm opt}}\right)^{-L} \begin{bmatrix} \hat{\Xi}_0^{P_{\rm opt}} \\ \vdots \\ \hat{\Xi}_{\bar{\nu}}^{P_{\rm opt}} \end{bmatrix} Z_{[t_0,t)}.$$

Using also the definition of $\hat{E}_t^{P_{\rm opt}}$ in (33), and the fact that $CX_t \doteq \Xi_0 Z_{[t_0,t)}$, it follows that

$$\begin{aligned} Y_t^{P_{\rm opt}} &= \hat{\Xi}_0^{P_{\rm opt}} Z_{[t_0,t)} \\ &\doteq C_N^{P_{\rm opt}} \hat{X}_t^{P_{\rm opt}} + \left(\hat{\Xi}_0^{P_{\rm opt}} - \Xi_0\right) Z_{[t_0,t)} \\ &\quad + C_N^{P_{\rm opt}} \left(T_N^{P_{\rm opt}} X_t - \hat{X}_t^{P_{\rm opt}}\right) \\ &\doteq C_N^{P_{\rm opt}} \hat{X}_t^{P_{\rm opt}} + \tilde{\Xi}_0^{P_{\rm opt}} Z_{[t_0,t)} \\ &\quad - C_N^{P_{\rm opt}} \left(\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{\rm opt}}\right)^{-L} \begin{bmatrix} \tilde{\Xi}_0^{P_{\rm opt}} \\ \vdots \\ \tilde{\Xi}_{\bar{\nu}}^{P_{\rm opt}} \end{bmatrix} Z_{[t_0,t)}. \end{aligned}$$

Therefore (recall that $T_N^{P_{\rm opt}}$ converges to the identity matrix)

$$\begin{aligned} \tilde{C}_N^{P_{\rm opt}} \doteq \left(\tilde{\Xi}_0^{P_{\rm opt}} - C\bar{\Gamma}_{\bar{\nu}}^{-L} \begin{bmatrix} \tilde{\Xi}_0^{P_{\rm opt}} \\ \vdots \\ \tilde{\Xi}_{\bar{\nu}}^{P_{\rm opt}} \end{bmatrix} \right) \\ \times Z_{[t_0,t)} (\hat{X}_t)^\top [\hat{X}_t \hat{X}_t]^{-1}. \end{aligned} \tag{A.59}$$

Vectorizing (A.59) and substituting sample values with their a.s. limit, there exists a matrix M_C, so that

$$\operatorname{vec}\left\{\tilde{C}_N^{P_{\rm opt}}\right\} \doteq M_C \tilde{\Omega}^{P_{\rm opt}}. \tag{A.60}$$

The fact that the rows of M_C are in ℓ_2 follows from the exponential convergence to zero of the predictor impulse response. ∎

Proof of Proposition 5.2: First note that, from Lemma 3.2, $\hat{\Gamma}_{\bar{\nu}}$ converges to $\Gamma_{\bar{\nu}}$ and therefore $\hat{\Gamma}_{\bar{\nu}}^{-L}\Gamma_{\bar{\nu}}$ converges to the identity matrix.

We prove now only (43) regarding $\tilde{C}_N^{\rm cca}$, with those being related to $\tilde{A}_N^{\rm cca}$, $\tilde{B}_N^{\rm cca}$, $\tilde{K}_N^{\rm cca}$ completely analogous.

First, recall that the state estimator $\hat{X}_t^{\rm cca}$

$$\begin{aligned} \hat{X}_t^{\rm cca} &= \hat{\Gamma}_{\bar{\nu}}^{-L} \hat{Y}_{[t,T-1]} \\ &= \hat{\Gamma}_{\bar{\nu}}^{-L} W_{\rm cca} W_{\rm cca}^{-1} \hat{Y}_{[t,T-1]} \\ &\doteq \hat{\Gamma}_{\bar{\nu}}^{-L} W_{\rm cca} W^{-1} \begin{bmatrix} \hat{\Xi}_0^{\rm cca} \\ \vdots \\ \hat{\Xi}_{\bar{\nu}}^{\rm cca} \end{bmatrix} Z_{[t_0,t)} \end{aligned}$$

where the last (asymptotic) equality follows from (39).

We also recall that, by definition of $T_N^{\rm cca}$ in (42) and $\Gamma_{\bar{\nu}} = H_{\bar{\nu}}\bar{\Gamma}_{\bar{\nu}} = W_{\rm cca}W^{-1}\bar{\Gamma}_{\bar{\nu}}$,

$$\hat{\Gamma}_{\bar{\nu}}^{-L} W_{\rm cca} W^{-1} \begin{bmatrix} \Xi_0 \\ \vdots \\ \Xi_{\bar{\nu}} \end{bmatrix} Z_{[t_0,t)} \doteq T_N^{\rm cca} X_t.$$

Let us define

$$\hat{E}_t := Y_t - E[Y_t | Z_{[t_0,t)}] := Y_t - \hat{\Xi}_0 Z_{[t_0,t)}. \tag{A.61}$$

The equation above defines also $\hat{\Xi}_0$, which is, in general, different from $\hat{\Xi}_0^{\rm cca}$. Note also that $\hat{E}_t$ is different from $\hat{E}_t^{\rm cca}$ de-

fined in (41). It follows that $Y_t = CX_t + E_t \doteq \Xi_0 Z_{[t_0,t)} + E_t$ can also be written in the form

$$\begin{aligned} Y_t &\doteq C_N^{\rm cca}\hat{X}_t^{\rm cca} + \hat{E}_t \\ &\quad + C_N^{\rm cca}(T_N^{\rm cca}X_t - \hat{X}_t^{\rm cca}) + \left(\hat{\Xi}_0 - \Xi_0\right) Z_{[t_0,t)} \\ &\doteq C_N^{\rm cca}\hat{X}_t^{\rm cca} + \tilde{\Xi}_0 Z_{[t_0,t)} + \hat{E}_t + \\ &\quad - C_N^{\rm cca}\left(\hat{\Gamma}_{\bar{\nu}}^{-L} W_{\rm cca} W^{-1} \begin{bmatrix} \tilde{\Xi}_0^{\rm cca} \\ \vdots \\ \tilde{\Xi}_{\bar{\nu}}^{\rm cca} \end{bmatrix} Z_{[t_0,t)}\right). \end{aligned}$$

Using $\bar{\Gamma}_{\bar{\nu}} = H_{\bar{\nu}}^{-1}\Gamma_{\bar{\nu}}$, it is easy to prove that $\bar{\Gamma}_{\bar{\nu}}^{-L} W = \Gamma_{\bar{\nu}}^{-L} W_{\rm cca}$ and therefore $\hat{\Gamma}_{\bar{\nu}}^{-L} W_{\rm cca}$ converges to $\bar{\Gamma}_{\bar{\nu}}^{-L} W$.

Having this observation in mind, from orthogonality of the rows of $\hat{E}_t$ with those of $\hat{X}_t^{\rm cca}$, recalling that $T_N^{\rm cca}$ converges to the identity matrix, it follows that

$$\tilde{C}_N^{\rm cca} \doteq \left(\tilde{\Xi}_0 - C\bar{\Gamma}_{\bar{\nu}}^{-L} \begin{bmatrix} \tilde{\Xi}_0^{\rm cca} \\ \vdots \\ \tilde{\Xi}_{\bar{\nu}}^{\rm cca} \end{bmatrix}\right) Z_{[t_0,t)}\hat{X}_t^\top [\hat{X}_t\hat{X}_t]^{-1}. \tag{A.62}$$

Using $\tilde{\Omega}^{P_{\rm cca}} = \tilde{\Omega}^{P_{\rm opt}} + (\hat{\Omega}^{P_{\rm cca}} - \hat{\Omega}^{P_{\rm opt}})$ and $\tilde{\Xi}_0 = \tilde{\Xi}_0^{P_{\rm opt}} + (\hat{\Xi}_0 - \hat{\Xi}_0^{P_{\rm opt}})$, it is clear that, vectorizing (A.62) and substituting sample values with their a.s. limit, there exist matrices M_C, $M_{C1}^{\rm cca}$ and $M_{C2}^{\rm cca}$ so that

$$\begin{aligned} {\rm vec}\{\tilde{C}_N^{\rm cca}\} \doteq M_C\tilde{\Omega}^{P_{\rm opt}} + M_{C1}^{\rm cca}(\hat{\Omega}^{P_{\rm cca}} - \hat{\Omega}^{P_{\rm opt}}) \\ + M_{C2}^{\rm cca}{\rm vec}\left(\hat{\Xi}_0 - \hat{\Xi}_0^{P_{\rm opt}}\right) \end{aligned}$$

where clearly, from (A.62), M_C is the same as that in (A.60), which concludes the proof. As before, $M_{C1}^{\rm cca}$ and $M_{C2}^{\rm cca}$ have rows in ℓ_2 due to the exponential convergence to zero of the predictor impulse response. ∎

Appendix B
Least Squares With Equality Constraints

Consider the least squares problem (28), which we rewrite here for convenience

$$Y = S^P\Omega^P + E + \Delta_Y(N) = S^P\Omega^P + LE_I + \Delta_Y(N) \tag{A.63}$$

where

$$\Delta_Y(N) := [{\rm vec}(C\bar{A}^p X_{t_0})^\top, \dots, {\rm vec}(C\bar{A}^{p+\bar{\nu}} X_{t_0})^\top]^\top \tag{A.64}$$

satisfies $\Delta_Y(N) = \underline{o}_P(1/\sqrt{N})$. The noise covariance ${\rm Var}\{E\} = R = L(I\otimes\Lambda)L^\top$, where L has been defined in (31), is a singular matrix.

The following lemma shows that $\Delta_Y(N)$ has a very special structure that allows one to discard it when studying the asymptotic distribution of the estimated parameters $\eta_N^\top\hat{\Omega}^P := \eta_N^\top F Y$.

Lemma B.1: The term $\Delta_Y(N)$ in (B.63) and (B.64) can be written as

$$\Delta_Y(N) = S^P\Upsilon o(1/\sqrt{N}) + L\Delta_X(N) \tag{B.65}$$

with $\|\Upsilon\|_2 = O(1)$ and $\gamma_N^\top\Delta_X(N) = o_P(1/\sqrt{N})$, $\forall\gamma_N$, satisfying $\|\gamma_N\|_2 = O(1)$.

Proof: First, note that $\bar{A}^k X_{t_0} = X_{t_0+k} - \sum_{i=1}^k \bar{A}^{i-1} K Z_{t_0+k-i}$. Therefore, $C\bar{A}^{p+k}X_{t_0} = C\bar{A}^p X_{t_0+k} + C\bar{A}^p\Upsilon_k Z_{[t_0,t+k)}$ holds for some matrix of coefficients Υ_k, which, from the exponential decrease of the terms $\bar{A}^i$, satisfies $\|\Upsilon_k\|_2 = O(1)$.

Therefore, we decompose the vectorization in (B.64) in the two parts

$$[{\rm vec}(C\bar{A}^p X_{t_0})^\top, \dots, {\rm vec}(C\bar{A}^p X_{t_0+\bar{\nu}})^\top]^\top \tag{B.66}$$

$$\left[{\rm vec}\left(C\bar{A}^p\Upsilon_0 Z_{[t_0,t)}\right)^\top, \dots, {\rm vec}\left(C\bar{A}^p\Upsilon_{\bar{\nu}} Z_{[t_0,t+\bar{\nu})}\right)^\top\right]^\top. \tag{B.67}$$

It is rather straightforward then to see that (B.66) can be written as

$$L\begin{bmatrix} C\bar{A}^p x_{t_0} \\ \vdots \\ C\bar{A}^p x_{t_0+N+\bar{\nu}-1} \end{bmatrix} = L\Delta_X(N) \tag{B.68}$$

while (B.67) as

$$S^P\begin{bmatrix} {\rm vec}(C\bar{A}^p\Upsilon_0) \\ \vdots \\ {\rm vec}(C\bar{A}^p\Upsilon_{\bar{\nu}}) \end{bmatrix} = S^P\Upsilon o(1/\sqrt{N}) \tag{B.69}$$

with obvious meaning of the symbols Υ and $\Delta_X(N)$; also $\|\Upsilon\|_2 = O(1)$ holds.

Note also that, due to Assumption 2.1, the covariance matrix $\Lambda_{xx} := {\rm Var}\{[x_{t_0}^\top, \dots, x_{t_0+N+\bar{\nu}-1}^\top]^\top\}$ satisfies $\|\Lambda_{xx}\|_2 = O(1)$; therefore, from (B.68) and $\|\bar{A}^p\| = o(1/\sqrt{N})$

$$\|{\rm Var}\{\Delta_X(N)\}\|_2 = o(1/\sqrt{N})\|\Lambda_{xx}\|_2 = o(1/\sqrt{N})$$

holds true; it follows that also $\gamma_N^\top\Delta_X(N) = o_P(1/\sqrt{N})$ holds for all column vectors γ_N (of suitable dimensions) satisfying $\|\gamma_N\|_2 = O(1)$. ∎

Using (B.64), the original least squares problem (B.63) can be written as

$$Y = S^P(\Omega^P + \Upsilon o(1/\sqrt{N})) + L(E_I + \Delta_X(N)). \tag{B.70}$$

The noise term $L(E_I + \Delta_X(N))$ has a covariance that can be written in the form $R_o = L[(I\otimes\Lambda) + o(1/\sqrt{N})\tilde{\Sigma}]L$; using the structure of E_I and $\Delta_X(N)$, (B.68), and Assumption 2.1, it follows also that $\|\tilde{\Sigma}\|_2 = O(1)$.

We are now ready to derive the optimal [asymptotically best linear unbiased estimator (BLUE)] of $\eta_N^\top\Omega^P$. Let us define $W_{S^P,R} := R + S^P(S^P)^\top$ and $W_{S^P,R_o} := R_o + S^P(S^P)^\top$. It is easy to see (using for instance, [46, (C4.3.3)]) that the asymptotically BLUE is given by[19] $\eta_N^\top\hat{\Omega}^{P_{\rm opt}} := \eta_N^\top F_{\rm opt}^{R_o} Y$ with

$$F_{\rm opt}^{R_o} = [(S^P)^\top W_{S^P,R_o}^\dagger S^P]^{-1}(S^P)^\top W_{S^P,R_o}^\dagger. \tag{B.71}$$

For N large enough, both $(I\otimes\Lambda)$ and $(I\otimes\Lambda) + o(1/\sqrt{N})\tilde{\Sigma}$ are nonsingular and bounded away from zero. Under this assumption, it is possible to see that, asymptotically

$$F_{\rm opt} := [(S^P)^\top W_{S^P,R}^\dagger S^P]^{-1}(S^P)^\top W_{S^P,R}^\dagger \tag{B.72}$$

[19] † denotes the Moore–Penroose pseudoinverse [27].

which is the Markov estimator computed as if there were no $\underline{o}_P(1/\sqrt{N})$ terms in (B.63), gives the same distribution of the estimator. It is crucial here that the limit is computed with perturbations to R which do not alter its column space (nor its rank in the limit).

Note also that, given any linear estimator ${}_F Y$ such that ${}_F S^P = I$, the estimation error $\eta_N^\top \tilde{\Omega}^P := \eta_N^\top({}_F Y - \Omega^P)$ has the form

$$\eta_N^\top \tilde{\Omega}^P = \eta_N^\top \Upsilon o(1/\sqrt{N}) + \eta_N^\top {}_F E + \eta_N^\top {}_F L \Delta_X(N).$$

The first term can be thought as a "bias." However, since $\|\Upsilon\|_2 = O(1)$, also $\|\eta_N^\top \Upsilon\|_2 = O(1)$, and hence this bias term goes to zero faster than $1/\sqrt{N}$. The last term is instead the contribution due to the $\underline{o}_P(1/\sqrt{N})$ terms, which are in the columns space of L.

Note that, provided $\|{}_F\|_2 = O(1)$, the vector $\gamma_N^\top := \eta_N^\top {}_F L$ has uniformly bounded 2-norm $\|\gamma_N^\top\|_2 = \|\eta_N^\top {}_F L\|_2 \leq \|\eta_N^\top\|_2 \|{}_F\|_2 \|L\|_2 = O(1)$ and hence, according to Lemma B.1, $\eta_N^\top {}_F L \Delta_X(N) = \gamma_N^\top \Delta_X(N) = o_P(1/\sqrt{N})$, which can therefore be neglected. As we shall see at the end of this Appendix, the estimators we are interested in (namely, $\hat{\Omega}^{P_{\rm opt}} = {}_{F_{\rm opt}} Y$ and $\hat{\Omega}^{P_{\rm cca}} = {}_{F_{\rm cca}} Y$) satisfy the even stronger condition $\|{}_{F_{\rm opt}}\|_2 = O(1/\sqrt{N})$ and $\|{}_{F_{\rm cca}}\|_2 = O(1/\sqrt{N})$; therefore $\eta_N^\top \tilde{\Omega}^P \doteq \eta_N^\top {}_F E$, providing a proof that, indeed, $o_P(1/\sqrt{N})$ in (B.63) can be discarded both to the purpose of design and analysis of the estimator.

Remark B.5: As an indirect proof that ${}_{F_{\rm opt}^{R_o}}$ and ${}_{F_{\rm opt}}$ give asymptotically equivalent estimators, note the following. With an argument similar to that used at the end of this Appendix to show that $\|{}_{F_{\rm opt}}\|_2 = O(1/\sqrt{N})$, also $\|{}_{F_{\rm opt}^{R_o}}\|_2 = O(1/\sqrt{N})$ can be proved. Therefore, $\eta_N^\top({}_{F_{\rm opt}^{R_o}} Y - \Omega^P) \doteq \eta_N^\top {}_{F_{\rm opt}^{R_o}} E$ and $\eta_N^\top({}_{F_{\rm opt}} Y - \Omega^P) \doteq \eta_N^\top {}_{F_{\rm opt}^{R}} E$, which means that their asymptotic properties do not depend on the $\underline{o}_P(1/\sqrt{N})$ terms. Since the first is asymptotically optimal when the $o_P(1/\sqrt{N})$ terms are accounted for while the second would be optimal if there were no $o_P(1/\sqrt{N})$ terms, both $\mathrm{AsVar}\{\sqrt{N}\eta_N^\top({}_{F_{\rm opt}^{R_o}} Y - \Omega^P)\} \leq \mathrm{AsVar}\{\sqrt{N}\eta_N^\top({}_{F_{\rm opt}} Y - \Omega^P)\}$ and $\mathrm{AsVar}\{\sqrt{N}\eta_N^\top({}_{F_{\rm opt}} Y - \Omega^P)\} \leq \mathrm{AsVar}\{\sqrt{N}\eta_N^\top({}_{F_{\rm opt}^{R_o}} Y - \Omega^P)\}$ hold, proving that, indeed $\mathrm{AsVar}\{\sqrt{N}\eta_N^\top({}_{F_{\rm opt}^{R_o}} Y - \Omega^P)\} = \mathrm{AsVar}\{\sqrt{N}\eta_N^\top({}_{F_{\rm opt}} Y - \Omega^P)\}$. ♢

Consider now the SVD $L = U_L S_L V_L^\top$ and let $U_{L^\perp}$ be an orthonormal basis for the left kernel of L, i.e., $U_{L^\perp}^\top L = 0$, $U_{L^\perp}^\top U_{L^\perp} = I$. We recall that $L^\dagger = V_L S_L^{-1} U_L^\top$.

Then define

$$\begin{aligned} Y_L &:= (I \otimes \Lambda)^{-1/2} L^\dagger Y \\ &= (I \otimes \Lambda)^{-1/2} L^\dagger S^P \Omega^P + (I \otimes \Lambda)^{-1/2} L^\dagger E \\ &= S_L^P \Omega^P + E_L \end{aligned}$$

which now has a full rank noise term E_L, $\mathrm{Var}\{E_L\} = I$.

Finally, observe that $Y = L[y_t^\top, y_{t+1}^\top, \ldots, y_{T+N-1}^\top]^\top = LY_I$, which implies that also $U_{L^\perp}^\top Y = 0$. It follows that $U_{L^\perp}^\top Y = U_{L^\perp}^\top S^P \Omega^P + U_{L^\perp}^\top E = S_{L^\perp}^P \Omega^P = 0$, where $S_{L^\perp}^P := U_{L^\perp}^\top S^P$ and $U_{L^\perp}^\top L = 0$ have been used. Therefore, (B.63) can be converted into a least squares problem with equality constraints

$$\begin{cases} Y_L = S_L^P \Omega^P + E_L \\ \text{such that} \quad 0 = S_{L^\perp}^P \Omega^P \,. \end{cases} \tag{B.73}$$

Let d_Ω denote the number of parameters in Ω^P. The matrix $S_{L^\perp}^P$ has dimension $(N-1)\bar{\nu} m_y \times d_\Omega$. For N "large," $S_{L^\perp}^P$ has more rows than columns; let us denote with $r_S^\perp$ the rank of $S_{L^\perp}^P$. Of course[20] $r_S^\perp \leq d_\Omega$. Let $\bar{U}_{L^\perp}^\top$ be a selection of $r_S^\perp$ rows of $U_{L^\perp}^\top$ so that the rows of $\bar{S}_{L^\perp}^P := \bar{U}_{L^\perp}^\top S^P$ form a basis of the row space of $S_{L^\perp}^P$. Clearly the constraints $0 = S_{L^\perp}^P \Omega^P$ and $0 = \bar{S}_{L^\perp}^P \Omega^P$ are equivalent since the rows of $S_{L^\perp}^P$ are linear combinations of the rows of $\bar{S}_{L^\perp}^P$. Therefore, we rewrite the constrained least squares problem (B.73) as

$$\begin{cases} Y_L = S_L^P \Omega^P + E_L \\ \text{such that} \quad 0 = \bar{S}_{L^\perp}^P \Omega^P \,. \end{cases} \tag{B.74}$$

Consider now the QR-decomposition [27]

$$\bar{S}_{L^\perp}^P = \bar{R}^\top Q^\top = \begin{bmatrix} \bar{R}_1^\top & 0 \end{bmatrix} \begin{bmatrix} Q_1^\top \\ Q_2^\top \end{bmatrix}$$

with $\bar{R}_1$ square $(r_S^\perp \times r_S^\perp)$ invertible. It is easy to show that, $\forall \eta_N$ satisfying Assumption 2.4, the asymptotically BLUE of $\eta_N^\top \Omega^P$ is given by[21] $\eta_N^\top \hat{\Omega}^{P_{\rm opt}}$ where

$$\hat{\Omega}^{P_{\rm opt}} = Q_2 (S_L^P Q_2)^\dagger Y_L. \tag{B.75}$$

Observe that, with Q_2 being orthonormal, the conditioning of the problem depends on the matrix $S_L^P Q_2$, i.e., to S_L^P "restricted to" the orthogonal complement of the "constraint matrix" $\bar{S}_{L^\perp}^P$. Note that[22] the matrix S_L^P could be collinear (or almost collinear); however, from (where $\Lambda_I := I \otimes \Lambda$)

$$\begin{aligned} \begin{bmatrix} U_{L^\perp}^\top \\ \Lambda_I^{-1/2} L^\dagger \end{bmatrix} S^P [Q_1 \quad Q_2] &= \begin{bmatrix} S_{L^\perp}^P \\ S_L^P \end{bmatrix} [Q_1 \quad Q_2] \\ &= \begin{bmatrix} R_1^\top & 0 \\ S_L^P Q_1 & S_L^P Q_2 \end{bmatrix} \end{aligned} \tag{B.76}$$

in (B.76) we have used the fact that the rows of $S_{L^\perp}^P$ are linear combinations of the rows of $\bar{S}_{L^\perp}^P$, so that $S_{L^\perp}^P Q^\top = [R_1^\top 0]$. The condition number of $S_L^P Q_2$ is smaller than that of

$$\begin{bmatrix} U_{L^\perp}^\top \\ \Lambda_I^{-1/2} L^\dagger \end{bmatrix} S^P = \begin{bmatrix} I & 0 \\ 0 & \Lambda_I^{-1/2} V_L^\top S_L^{-1} \end{bmatrix} \begin{bmatrix} U_{L^\perp}^\top \\ U_L^\top \end{bmatrix} S^P \tag{B.77}$$

which guarantees that the constrained problem is well posed if the original problem is so.

Note that the "optimal" estimator $\hat{\Omega}^{P_{\rm opt}}$ is of the form $\hat{\Omega}^{P_{\rm opt}} = {}_{F_{\rm opt}} Y$ with

$${}_{F_{\rm opt}} = Q_2 \left(S_L^P Q_2\right)^\dagger (I \otimes \Lambda)^{-1/2} V_L S_L^{-1} U_L^\top. \tag{B.78}$$

[20]Using the structure of the matrices L and S^P, it is possible to check that, indeed, $d_\Omega - r_S^\perp = p m_y (m_y + p_u) > 0$. For reasons of space it is not possible to report the details here; see [14] and [9].

[21]See, e.g., [46, Remark 3, p. 70], for the case in which Ω^P has dimensions not depending on N.

[22]As an anonymous reviewer has pointed out.

We now need to verify that indeed $\|F_{\rm opt}\|_2 = O(1/\sqrt{N})$ holds true. From (B.78), it follows that

$$\|F_{\rm opt}\|_2 \leq \left\|\left(S_L^P Q_2\right)^\dagger\right\|_2 \|\Lambda^{-1/2}\|_2 \left\|S_L^{-1}\right\|_2. \qquad \text{(B.79)}$$

First note that $\|\Lambda^{-1/2}\|_2 = O(1)$ since Λ is the (estimated) noise covariance. The fact that $\|S_L^{-1}\|_2 = O(1)$ follows directly from the structure of the matrix L. Using now (B.76) and (B.77), $\sigma_{\min}(S_L^P Q_2) \geq \min(1, 1/\sqrt{\|\Lambda\|_2})\sigma_{\min}(S^P)$ since $\|\Lambda\|_2 = O(1)$.

Using now (29), it is immediate to see that $((S^P)^\top(S^P)/N)$ is a block diagonal matrix with block diagonal elements $(Z_{[t_0,t+k]}Z_{[t_0,t+k]}^\top/N) \otimes I$; from Assumption 2.1 and the uniform convergence of sample covariances (see [29] and [30]), $\sigma_{\min}^{-1}[((S^P)^\top S^P/N)]^{-1} = O(1)$, which implies $\sigma_{\min}^{-1}[S^P] = O(1/\sqrt{N})$.

It follows that $\|(S_L^P Q_2)^\dagger\|_2 = \sigma_{\min}^{-1}(S_L^P Q_2) = O(1/\sqrt{N})$, which, inserted in (B.79), gives the desired result $\|F_{\rm opt}\|_2 = O(1/\sqrt{N})$. With a similar calculation, we could also show that in $\hat{\Omega}^{P_{\rm cca}} := F_{\rm cca}Y$, $\|F_{\rm cca}\|_2 = O(1/\sqrt{N})$.

REFERENCES

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

[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, 1976, pp. 27–96.

[3] D. Bauer, "Some asymptotic theory for the estimation of linear systems using maximum likelihood methods or subspace algorithms," Ph.D. dissertation, TU Wien, Wien, Austria, 1998.

[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 and M. Jansson, "Analysis of the asymptotic properties of the MOESP type of subspace algorithms," *Automatica*, vol. 36, pp. 497–509, 2000.

[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, "The role of vector autoregressive modeling in predictor-based subspace identification," *Automatica*, vol. 43, no. 6, pp. 1034–1048, Jun. 2007.

[10] A. Chiuso, "Some insights on the choice of the future horizon in CCA-type subspace algortihms," in *Proc. Amer. Control Conf.*, New York, 2007.

[11] A. Chiuso, "On the relation between CCA and predictor-based subspace identification," in *Proc. 44rd IEEE Conf. Decision Contr.*, Sevilla, Spain, 2005.

[12] A. Chiuso, "Asymptotic equivalence of certain closed-loop subspace identification methods," in *Proc. SYSID 2006*, Newcastle, Australia, 2006.

[13] A. Chiuso, "Asymptotic variance of closed-loop subspace identification algorithms," *IEEE Trans. Autom. Control*, vol. 51, no. 8, pp. 1299–1314, 2006.

[14] A. Chiuso, "The role of vector auto regressive modeling in subspace identification," in *Proc. CDC 2006*, San Diego, CA, Dec. 2006.

[15] A. Chiuso and G. Picci, *Geometry of Oblique Splitting, Minimality and Hankel Operators*, ser. Lecture Notes in Control and Information Sciences. Berlin, Germany: Springer, 2003, pp. 85–124, no. 286.

[16] A. Chiuso and G. Picci, "The asymptotic variance of subspace estimates," *J. Econometrics*, vol. 118, no. 1-2, pp. 257–291, 2004.

[17] A. Chiuso and G. Picci, "Asymptotic variance of subspace methods by data orthogonalization and model decoupling: A comparative analysis," *Automatica*, vol. 40, no. 10, pp. 1705–1717, 2004.

[18] A. Chiuso and G. Picci, "Numerical conditioning and asymptotic variance of subspace estimates," *Automatica*, vol. 40, no. 4, pp. 677–683, 2004.

[19] A. Chiuso and G. Picci, "On the ill-conditioning of subspace identification with inputs," *Automatica*, vol. 40, no. 4, pp. 575–589, 2004.

[20] A. Chiuso and G. Picci, "Consistency analysis of some closed-loop subspace identification methods," *Automatica*, vol. 41, no. 3, pp. 377–391, 2005.

[21] A. Chiuso and G. Picci, "Prediction error vs. subspace methods in closed-loop identification," in *Proc. 16th IFAC World Congr.*, Prague, Czech Republic, Jul. 2005.

[22] N. Chui and J. Maciejowski, "Criteria for informative experiments for subspace identification," *Int. J. Contr.*, vol. 78, no. 5, pp. 326–344, 2005.

[23] U. Desai and D. Pal, "A realization approach to stochastic model reduction," *IEEE Trans. Autom. Control*, vol. AC-29, pp. 1097–1100, 1984.

[24] T. Ferguson, *A Course in Large Sample Theory*. New York: Chapman and Hall, 1996.

[25] H. Gazzah, P. Regalia, and J. Delmas, "Asymptotic eigenvalue distribution of block Toeplitz matrices and application to blind SIMO channel identification," *IEEE Trans. Inf. Theory*, vol. 47, pp. 1243–1251, Mar. 2001.

[26] M. Gevers and B. Anderson, "On jointly stationary feedback-free stochastic processes," *IEEE Trans. Autom. Control*, vol. AC-27, pp. 431–436, 1982.

[27] G. Golub and C. Van Loan, *Matrix Computation*, 2nd ed. Baltimore, MD: The Johns Hopkins Univ. Press., 1989.

[28] C. Granger, "Economic processes involving feedback," *Inf. Contr.*, vol. 6, pp. 28–48, 1963.

[29] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. New York: Chelsea, 1958.

[30] E. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*. New York: Wiley, 1988.

[31] E. Hannan and D. Poskitt, "Unit canonical correlations between future and past," *Ann. Statist.*, vol. 16, pp. 784–790, 1988.

[32] H. Hotelling, "Relations between two set of variables," *Biometrika*, vol. 28, pp. 321–377, 1936.

[33] M. Jansson, "Asymptotic variance analysis of subspace identification methods," in *Proc. SYSID2000*, Santa Barbara, CA, 2000.

[34] M. Jansson, "Subspace identification and ARX modeling," in *Proc. SYSID 2003*, Rotterdam, The Netherlands, 2003.

[35] T. Katayama and G. Picci, "Realization of stochastic systems with exogenous inputs and subspace system identification methods," *Automatica*, vol. 35, no. 10, pp. 1635–1652, 1999.

[36] W. Larimore, "System identification, reduced-order filtering and modeling via canonical variate analysis," in *Proc. Amer. Contr. Conf.*, 1983, pp. 445–451.

[37] W. Larimore, "Canonical variate analysis in identification, filtering, and adaptive control," in *Proc. 29th IEEE Conf. Decision Contr.*, Honolulu, HI, 1990, pp. 596–604.

[38] W. Larimore, "Large sample efficiency for ADAPTX subspace identification with unknown feedback," in *Proc. IFAC DYCOPS'04*, Boston, MA, 2004.

[39] R. Lewis and G. Reinsel, "Prediction of multivariate time series by autoregressive model fitting," *J. Multivariate Anal.*, vol. 16, pp. 393–411, 1985.

[40] A. Lindquist and G. Picci, "Canonical correlation analysis, approximate covariance extension and identification of stationary time series," *Automatica*, vol. 32, pp. 709–733, 1996.

[41] L. Ljung, *System Identification, Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1997.

[42] K. Peternell, W. Scherrer, and M. Deistler, "Statistical analysis of novel subspace identification methods," *Signal Process.*, vol. 52, pp. 161–178, 1996.

[43] S. Qin and L. Ljung, "Closed-loop subspace identification with innovation estimation," in *Proc. SYSID 2003*, Rotterdam, The Netherlands, 2003.

[44] S. Qin and L. Ljung, "Parallel QR implementation of subspace identification with parsimonious models," in *Proc. SYSID 2003*, Rotterdam, The Netherlands, 2003.

[45] C. Rao, "Representations of the best linear unbiased estimators in the Gauss-Markov model with a singular dispersion matrix," *J. Multivariate Anal.*, vol. 3, pp. 276–292, 1973.
[46] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[47] P. Van Overschee and B. De Moor, "Subspace algorithms for the stochastic identification problem," *Automatica*, vol. 29, pp. 649–660, 1993.
[48] P. Van Overschee and B. De Moor, "N4SID: Subspace algorithms for the identification of combined deterministic—Stochastic systems," *Automatica*, vol. 30, pp. 75–93, 1994.
[49] P. Van Overschee and B. De Moor, "A unifying theorem for three subspace system identification algorithms," *Automatica*, vol. 31, no. 12, pp. 1853–1864, 1995.
[50] 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.
[51] H. Werner and C. Yapar, "On inequality constrained generalized least squares selections in the general possibly singular Gauss-Markov model: A projector theoretical approach," *Linear Algebra Applicat. (Special Issue Honoring Calyampudi Radhakrishna Rao)*, vol. 237/238, no. 1–3, pp. 359–393, 1996.

Alessandro Chiuso (SM'06) received the "Laurea" degree *(summa cum laude)* in telecommunication engineering from the University of Padova, Italy, in 1996 and the Ph.D. degree (Dottorato di Ricerca) in system engineering from the University of Bologna, Italy, in 2000.

Currently, he is an Associate Professor with the Dipartimento di Tecnica e Gestione dei Sistemi Industriali, University of Padova. He has held visiting positions with the Department of Electrical Engineering, Washington University at St. Louis, the Department of Mathematics, Royal Institute of Technology, Sweden, and the Department of Computer Science, University of California, Los Angeles. His research interests are mainly in estimation, identification theory and applications (subspace methods, stochastic realization, nonlinear estimation, hybrid systems), and computer vision (structure from motion, texture and gait analysis).

Dr. Chiuso is a member of the Editorial Board of *IET Control Theory and Applications*, the IEEE Control Systems Society (CSS) Conference Editorial board, and the IEEE CSS Technical Committee on Identification and Adaptive Control, and a Reviewer for *Mathematical Reviews* of the AMS.