On the relation between CCA and predictor-based subspace identification.

Alessandro Chiuso, *Senior Member, IEEE*

Abstract

There is experimental evidence that a recently proposed subspace algorithm based on predictor identification, known also as "whitening filter algorithm", has a behavior which is very close to prediction error methods in certain simple examples; this observation raises a question concerning its optimality. It is also known that time series identification using the Canonical Correlation Analysis (CCA) approach is asymptotically efficient. Asymptotic optimality of CCA has also been proved when the measured inputs are white. For these reasons CCA provides a natural benchmark against which other subspace procedures should be compared. In this paper we study the relation between the standard CCA approach and the recently proposed subspace procedure based on predictor identification (PBSID from now on).

Even though PBSID is 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; it is shown that CCA and PBSID are asymptotically equivalent precisely in the situations when CCA is optimal. The equivalence holds only asymptotically in the number of data and in the limit as the past horizon goes to infinity. We also show that a slightly modified version of PBSID behaves no worse than CCA also for non-white input signals.

The results of this paper imply that the "optimized" PBSID, besides being able to handle feedback, is to be preferred to CCA when there is no feedback; only in very specific cases (white or no inputs) the two algorithms are (asymptotically) equivalent.

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

This work was supported by MIUR under national project *New methods and algorithms for identification and adaptive control of technological systems*.

Alessandro Chiuso is with the Dipartimento di Tecnica e Gestione dei Sistemi Industriali, Università di Padova, Stradella San Nicola, 3 36100 Vicenza, Italy, ph. +39-049-8277709, fax. +39-049-8277699, e-mail: chiuso@dei.unipd.it

by Larimore [33], [34], Van Overschee and De Moor [44] is known to provide asymptotically efficient estimators[1] [5]. Sometimes this algorithm goes under the name of CCA (or CVA) to remind that the state construction is performed using Canonical Correlation Analysis [29], as pioneered by Akaike [1], [2] and Desai and Pal [20]. 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 [34], [32], [46]. 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 [34], [39], [46], [32], [7], [4] or without inputs [33], [5]); we shall henceforth use the same terminology (CCA) also in this paper.

Besides CCA, the most widely known procedures go under the acronyms N4SID [45], and MOESP [47]. Recently several researchers have studied the asymptotic statistical properties of these algorithms [3], [30], [6], [7], [4], [13], [15] and compared, to some extent, existing procedures [7], [4], [14]. 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 the paper [7].

It is our opinion, as has already been stressed in [18], 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, 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", [40]), and Jansson (SSARX, [31]) which, to some extent, are able to deal with feedback. A related algorithm is discussed also in [35]. The recent work [17] studies the statistical consistency of these two algorithms. In [17] also a "geometrical" version of the SSARX algorithm proposed by Jansson [31] 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" for short) algorithm in this paper. We refer the reader to the paper [18] for an explanation of this terminology; in [18] also the relation between classical PEM and PBSID has been investigated and similarities pointed out. For reasons of space we

[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.

shall not discuss further this issue here.

It has also recently been proved (see [10]) that PBSID [17] and SSARX [31] are asymptotically equivalent. This relation further motivates the analysis of this paper; the results contained in [10] also suggest that there is a close connection between PBSID (and possibly its optimized version presented here) and VARX (Vector AutoRegressive with eXogenous inputs) modeling.[2] For reasons of space we shall have to postpone a more detailed analysis of this relation to future work and refer the reader to the paper [10] for some preliminary results.

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 [31], [17].

Using some recently derived formulas (see [11]) 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?*

We believe therefore that the relation of this procedure with more classical approaches is worth studying; some preliminary results have been presented in the paper [9].

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 [33], [44] is efficient for time series identification and optimal (see [5], [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 a sort of lower bound on the achievable accuracy in the situations mentioned above 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 [17], [18], [31], in this paper we shall work under the assumption that *no feedback is present*. See [25], [23], [17] 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]As an anonymous reviewer was suggesting.

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

2) we introduce an "optimized version" of PBSID which performs no worse (in the sense of asymptotic variance) than CCA regardless of the input spectrum (see Section V and Theorem 5.3); the "optimized" PBSID 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. Suffices 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 future inputs [45]; PBSID instead enforces causality of the predictors (see formula (19)); state constructions advocating for causal predictors have already been proposed in [39], [32], [40], [41]. 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 [39], page 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 remains open of course; simulation results and computations based on the asymptotic variance expressions (see Section VI) suggest that PBSID is a candidate for being efficient also for colored input. 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]).

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

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

We assume that there is *no feedback* from $\{\mathbf{y}(t)\}$ to $\{\mathbf{u}(t)\}$ [25], [8], [23]. Without loss of generality we shall assume that the dimension n of the state vector $\mathbf{x}(t)$ is as small as possible, i.e. the representation (1) is minimal. For simplicity we assume that $D = 0$, i.e. there is no direct feedthrough[4] from $\mathbf{u}$ to $\mathbf{y}$. This setup also encompasses time series identification (i.e. no measured inputs) provided one lets $B = 0$, $D = 0$ in (1).

For future reference we define $\bar{A} := A - KC$ and let $\rho := \lambda_{max}(\bar{A})$ be an eigenvalue of maximum modulus of $\bar{A}$; we shall assume that both $|\rho|$ and $|\lambda_{max}(A)|$ are strictly less than 1.

The white noise process $\mathbf{e}$, the innovation of $\mathbf{y}$ given the joint past of $\mathbf{y}, \mathbf{u}$, is defined (see formula (6)) as the one step ahead (linear) prediction error of $\mathbf{y}(t)$ given the joint (strict) past of $\mathbf{u}$ and $\mathbf{y}$ up to time t.

The symbol I shall denote the identity matrix (of suitable dimension), $A^\top$ shall denote the transpose of the matrix A, $\|A\|_2$ shall be the 2-norm. For a symmetric positive semidefinite matrix $A = A^\top \geq 0$ the symbol $A^{1/2}$ shall denote (slightly different from the standard notation) 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 $G(z) = C(zI - A)^{-1}K + I$, starting from input-output data $\{y_s, u_s\}$, $s \in [t_0, T + N]$, generated by the system (1).

In this paper we are interested in assessing the quality, i.e. variance, of the subspace estimators; therefore we shall have to deal with random fluctuations due to finite sample length (e.g. approximating expectations with finite time averages, etc.). The analysis of this paper will be concerned with asymptotic distribution and variance, i.e. quantification of these random

[4]This assumption can be removed in our situation but is useful when there is feedback, see [25], [8], [23], [17]. Since the "predictor based" algorithm is designed to work without assumptions on the feedback structure we prefer to keep $D = 0$ also here.

fluctuations for "large samples". Our concern is to show the link between CCA and "predictor based" algorithm asymptotically as the number of data N goes to infinity.

We shall use the standard notation of boldface (lowercase) letters to denote random variables (or semi-infinite tails). Lowercase letters denote sample values of a certain random variable. For example we shall denote with $\mathbf{y}(t)$ the random vector denoting the output or equivalently the semi-infinite tail $[y_t\ y_{t+1},\ \dots\ y_{t+k}\ \dots]$ where y_t is the sample value of $\mathbf{y}(t)$. It can be shown (see [37], [16]) that the Hilbert spaces of random variables of second order stationary and ergodic process and the Hilbert space of semi-infinite tails 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 $Y_t := [y_t\ y_{t+1},\ \dots\ y_{t+N-1}]$, $U_t := [u_t\ u_{t+1},\ \dots\ u_{t+N-1}]$ and $Z_t := [Y_t^\top\ U_t^\top]^\top$. These are the block rows of the usual *data block Hankel matrices* which appear in subspace identification.

Recall that, in order to deal with realistic algorithms which can only regress on a finite amount of data, in subspace identification one usually keeps *finite past and future horizons*. This setting we describe as using data from a *finite observation interval*. The analysis reported in this paper requires that both N, the length of the finite tails[5] and the past horizon $t - t_0$ [6] go to infinity. We remind the reader that $t - t_0$ has to go to infinity at a certain rate depending on the number N of data available. Details can be found, for instance, in [7] where the following assumption is made:

Assumption 2.1: The past horizon $t - t_0$ goes to infinity with N while satisfying:

$$\begin{aligned} t - t_0 &\geq \tfrac{log N^{-d/2}}{log|\rho|} \qquad 1 < d < \infty \\ t - t_0 &= o\left(log(N)^\alpha\right) \qquad \alpha < \infty \end{aligned} \tag{2}$$

The first condition states that $t-t_0$ goes to infinity "fast enough" as compared to the predictor dynamics (eigenvalues of $\bar{A}$), while the second ensures that it grows slower than logarithmically (to some power α) in the number of data points. This assumption shall be made throughout. The first condition, together with $|\rho| < 1$, implies that $(A - KC)^{t-t_0} = o(1/\sqrt{N})$; therefore the

[5]This is the parameter j in the notation of Van Overschee and De Moor [45] i.e. the number of columns in the block Hankel data matrices used in subspace identification.

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

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 $o(1/\sqrt{N})$ and therefore can be neglected for the purpose of asymptotic analysis. Moreover, (2) ensures that, when regressing onto past data and taking the limit as N goes to infinity, the computation of sample covariance matrices of increasing size (with $t - t_0$) does not pose any complication in the sense that their limit is well defined and equal to the population counterpart (see the discussion after Lemma 4 in [7]).

For $t_0 \leq t \leq T$ we define the Hilbert space $\mathcal{U}_{[t_0,\,t)}$ of random (zero mean finite variance) variables

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

the bar denotes closure in mean square, i.e. in the metric defined by the inner product $\langle\, \xi, \eta \,\rangle := \mathbb{E}\{\xi\eta\}$ where $\mathbb{E}\{\cdot\}$ denotes mathematical expectation. Closure is necessary since t_0 might go to $-\infty$. These are the *past spaces* at time t of the processes $\mathbf{u}$ and $\mathbf{y}$. Similarly, let $\mathcal{U}_{[t,\,T]},\ \mathcal{Y}_{[t,\,T]}$ be the future input and output spaces up to time T. The length $T - t$ of the future will be denoted with $\nu := T - t$ while we define $\bar{\nu} := \nu - 1$.

We define the *joint* future, $\mathcal{Z}_{[t,\,T]} := \mathcal{U}_{[t,\,T]} \vee \mathcal{Y}_{[t,\,T]}$ and joint past $\mathcal{Z}_{[t_0,\,t)} := \mathcal{U}_{[t_0,\,t)} \vee \mathcal{Y}_{[t_0,\,t)}$ the $\vee$ denoting closed vector sum. By convention the past spaces do not include the present. When $t_0 = -\infty$ we shall use the shorthands $\mathcal{U}_t^-,\ \mathcal{Y}_t^-$ for $\mathcal{U}_{(-\infty,\,t)},\ \mathcal{Y}_{(-\infty,\,t)}$, and $\mathcal{Z}_t^- := \mathcal{U}_t^- \vee \mathcal{Y}_t^-$. Subspaces spanned by random vectors at just one time instant (e.g. $\mathcal{U}_{[t,\,t\,]}$, etc) are simply denoted $\mathcal{U}_t$, etc. while for the spaces generated by $\mathbf{u}$ and $\mathbf{y}$ when t goes from $-\infty$ to $+\infty$ we shall use the symbols $\mathcal{U}$, $\mathcal{Y}$, respectively.

With a slight abuse of notation, given a subspace $\mathcal{A} \subseteq \mathcal{U} \vee \mathcal{Y}$, we shall denote with $E[\cdot \mid \mathcal{A}]$ the orthogonal projection onto $\mathcal{A}$, which coincides with conditional expectation in the Gaussian case. Given two non-intersecting subspaces $\mathcal{A} \subseteq \mathcal{U} \vee \mathcal{Y}$, $\mathcal{B} \subseteq \mathcal{U} \vee \mathcal{Y}$, $\mathcal{A} \cap \mathcal{B} = \{0\}$, $E_{\|\mathcal{B}}[\cdot \mid \mathcal{A}]$ shall denote the oblique projection onto $\mathcal{A}$ along $\mathcal{B}$ (see [24], [16]).

We adopt the notation $\Sigma_{\mathbf{ab}} := \mathbb{E}\left[\mathbf{a}\mathbf{b}^\top\right]$ to denote the covariance matrix between the zero mean random vectors $\mathbf{a}$ and $\mathbf{b}$. In the finite dimensional case the orthogonal projection of the random vector $\mathbf{a}$ onto the space $\mathcal{C} := \operatorname{span}\{\mathbf{c}\}$ spanned by the vector $\mathbf{c}$ will be given, provided

$\Sigma_{\mathbf{cc}}$ is invertible, by the usual formula

$$E[\mathbf{a}|\mathcal{C}] = \Sigma_{\mathbf{ac}}\Sigma_{\mathbf{cc}}^{-1}\mathbf{c}.$$

Defining the projection errors $\tilde{\mathbf{a}} := \mathbf{a} - E[\mathbf{a}|\mathcal{C}]$ and $\tilde{\mathbf{b}} := \mathbf{b} - E[\mathbf{b}|\mathcal{C}]$, the symbol $\Sigma_{\mathbf{ab|c}}$ will denote projection error covariance (conditional covariance in the Gaussian case) $\Sigma_{\mathbf{ab|c}} := \Sigma_{\tilde{\mathbf{a}}\tilde{\mathbf{b}}} = \Sigma_{\mathbf{ab}} - \Sigma_{\mathbf{ac}}\Sigma_{\mathbf{cc}}^{-1}\Sigma_{\mathbf{cb}}$. If we denote $\mathcal{B} := \text{span}\{\mathbf{b}\}$, $\mathcal{C} := \text{span}\{\mathbf{c}\}$, and assume that $\mathcal{B} \cap \mathcal{C} = \{0\}$, the oblique projection $E_{\|\mathcal{B}}[\mathbf{a}|\mathcal{C}]$ can be computed using the formula:

$$E_{\|\mathcal{B}}[\mathbf{a}|\mathcal{C}] = \Sigma_{\mathbf{ac|b}}\Sigma_{\mathbf{cc|b}}^{-1}\mathbf{c}. \tag{3}$$

We shall also use the notation: $\mathbf{y}_{[t,s]} := \begin{bmatrix} \mathbf{y}^\top(t) & \mathbf{y}^\top(t+1) & \dots & \mathbf{y}^\top(s) \end{bmatrix}^\top$ and the shorthands $\mathbf{y}^+ := \mathbf{y}_{[t,T-1]}$, and $\bar{\mathbf{u}}^+ := \mathbf{u}_{[t,T]}$.

Similarly the (finite) block Hankel data matrices will be denoted as $Y_{[t,s]} := \begin{bmatrix} Y_t^\top & Y_{t+1}^\top & \dots & Y_s^\top \end{bmatrix}^\top$

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 $A_t := [a_t, a_{t+1}.., a_{t+N-1}]$ and $B_t := [b_t, b_{t+1}.., b_{t+N-1}]$ containing sample values of the processes $\{\mathbf{a}(t)\}$, $\{\mathbf{b}(t)\}$, we shall define

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

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

Similarly, given a third sequence (say $C_t := [c_t, c_{t+1}, .., c_{t+N-1}]$), $\hat{\Sigma}_{\mathbf{ab|c}}$ is defined as $\hat{\Sigma}_{\mathbf{ab|c}} := \hat{\Sigma}_{\mathbf{ab}} - \hat{\Sigma}_{\mathbf{ac}}\hat{\Sigma}_{\mathbf{cc}}^{-1}\hat{\Sigma}_{\mathbf{cb}}$. Orthogonal and oblique projections on spaces of finite tails will be denoted with the symbol $\hat{E}$; e.g. $\hat{E}[\cdot|U_{[t_0,t)}]$ will be the orthogonal projection on the space generated by the rows of $U_{[t_0,t)}$ and $\hat{E}_{\|U_{[t,T]}}[\cdot|Z_{[t_0,t)}]$ will be the oblique projection along the space generated by the rows of future inputs $U_{[t,T]}$ onto the space generated by the rows of the joint past $Z_{[t_0,t)}$ [24]. As above, the oblique projection can be computed using the formula:

$$\hat{E}_{\|B_t}[A_t|C_t] = \hat{\Sigma}_{\mathbf{ac|b}}\hat{\Sigma}_{\mathbf{cc|b}}^{-1}C_t. \tag{4}$$

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 $\hat{E}[\cdot|B_t, C_t]$

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

$$\mathcal{Z}_{[t_0,T]} = \mathcal{Z}_{[t_0,t)} + \mathcal{Z}_{[t,T]}, \qquad t_0 \le t \le T \tag{5}$$

the $+$ sign denoting direct sum of subspaces. The symbol $\oplus$ will be reserved for *orthogonal* direct sum. Various conditions ensuring sufficient richness are known. For example, it is well-known that for a full-rank purely non deterministic (p.n.d.) process $\mathbf{z}$ to be sufficiently rich it is necessary and sufficient that the determinant of the spectral density matrix $\Phi_{\mathbf{z}}$ should have no zeros on the unit circle [28]. Therefore, whenever needed, we shall make the following assumption:

Assumption 2.2: The joint spectrum $\Phi_{\mathbf{z}}$ is bounded and bounded away from zero on the unit circle, i.e. $\exists\, 0 < c \leq M < \infty$ s.t.

$$cI \leq \Phi_{\mathbf{z}}(e^{j\omega}) \leq MI \quad \forall\, \omega \in [0, 2\pi)$$

Whenever necessary we shall assume that (5) holds also for finite sequences, i.e. that $Z_{[t_0,T]}$ is of full row rank; note that this will hold almost surely for ergodic sequences once (5) is satisfied.

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

$$\mathbf{e}(t) := \mathbf{y}(t) - E\left[\mathbf{y}(t) \mid \mathcal{Z}_t^-\right]. \tag{6}$$

We shall use the symbol $\mathcal{E}_t$ to denote the space generated by the components of $\mathbf{e}(t)$; we also define $\Lambda := \text{Var}\{\mathbf{e}(t)\}$.

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 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 [36]).

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 2.3 below:

Assumption 2.3: (i) $\exists\, M < \infty\, :\, \forall\, N\ \eta_N^\top \eta_N < M$; (ii) $\exists\, \eta \in \ell_2\ :\ \lim_{N\to\infty} \left\| [\eta_N^\top\ 0] - \eta^\top \right\|_2 = 0$ and (iii) $\lim_{N\to\infty} \text{Var}\left(\sqrt{N}\eta_N^\top \mathbf{v}_N\right) = c < \infty$.

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 $\|\Sigma_\infty\|_2 < \infty$)[7] 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 to deal with the expressions for $N = \infty$ rather than with the limit, as done also in [7].

Given two sequences of random variables $\mathbf{x}_N$ and $\mathbf{g}_N$, we shall say that $\mathbf{x}_N = o_P(\mathbf{g}_N)$ if, $\forall \delta > 0$,

$$\lim_{N\to\infty} P[\mid \mathbf{x}_N/\mathbf{g}_N \mid > \delta] = 0$$

The notation $\mathbf{x}_N = O(\mathbf{g}_N)$ means that the sequence $\mathbf{x}_N/\mathbf{g}_N$ is bounded almost surely, i.e.

$$\limsup_{N\to\infty} \mathbf{x}_N/\mathbf{g}_N \leq M \qquad a.s.$$

for some $0 \leq M < \infty$. The symbol $O_P(\cdot)$ has the same meaning in probability i.e. $\mathbf{x}_N = O_P(\mathbf{g}_N)$, if, $\forall \epsilon$, $\exists M$ s.t.

$$\sup_N P[\mid \mathbf{x}_N/\mathbf{g}_N \mid > M] < \epsilon$$

Similarly $o(\cdot)$ shall denote a.s. convergence to zero, i.e. $\mathbf{x}_N = o(\mathbf{g}_N)$ means that

$$\lim_{N\to\infty} \mathbf{x}_N/\mathbf{g}_N = 0 \qquad a.s.$$

If both $\mathbf{x}_N$ and $\mathbf{g}_N$ are deterministic sequences, say x_N and g_N, then $x_N = o(g_N)$ has the usual meaning

$$\lim_{N\to\infty} x_N/g_N = 0.$$

The symbol $\dot{=}$ 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 [21]) terms which are $o_P(1/\sqrt{N})$ can be neglected when studying the asymptotic statistical properties. 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) which elements are respectively $o_P(\cdot)$, $O_P(\cdot)$, $o(\cdot)$ or $O(\cdot)$; (e.g., given random matrices V_N^1 and V_N^2, the notation $V_N^1 = V_N^2 + \underline{o}_P(1/\sqrt{N})$ means that the elements of $V_N^1 - V_N^2$ are $o_P(1/\sqrt{N})$).

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

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 are 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 non singular $\hat{T}_N$, with $\lim_{N\to\infty}\hat{T}_N = I$, so that $\mathbf{x}_1 - \hat{T}_N\mathbf{x}_2 = \underline{o}_P(1/\sqrt{N})$.

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 $\mathbf{x}_N = o_P(1/\sqrt{N})$.

When dealing with tail matrices, e.g. A_t and B_t, containing the sample values a_{t+i}, b_{t+i}, $i = 0, .., N-1$ of the random vectors $\mathbf{a}(t)$ and $\mathbf{b}(t)$, the notation $A_t \doteq B_t$ really means that $\mathbf{a}(t) \doteq \mathbf{b}(t)$.

For future reference we also define the extended observability matrices

$$\Gamma_k := \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^k \end{bmatrix} \quad , \quad \bar{\Gamma}_k := \begin{bmatrix} C \\ C\bar{A} \\ C\bar{A}^2 \\ \vdots \\ C\bar{A}^k \end{bmatrix} \tag{7}$$

and the Toeplitz matrices containing the Markov parameters of the "stochastic" part:

$$H_k = \begin{bmatrix} I & 0 & \dots & 0 \\ CK & I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{k-1}K & CA^{k-2}K & \dots & I \end{bmatrix}. \tag{8}$$

Whenever needed we shall also make the following assumption on the innovation process $\mathbf{e}(t)$

Assumption 2.4: 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.

$$\begin{aligned} \mathbb{E}[\mathbf{e}(t)|\mathcal{F}_{t-1}^-] &= 0 \\ \mathbb{E}[\mathbf{e}(t)\mathbf{e}^\top(t)|\mathcal{F}_{t-1}^-] &= \Lambda. \end{aligned} \tag{9}$$

III. State Space Construction

It is well known [44], [45], [37], [16] 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

$$\begin{cases} \hat{X}_{t+1} & \simeq & A\hat{X}_t + BU_t + KE_t \\ Y_t & \simeq & C\hat{X}_t + E_t \end{cases} \tag{10}$$

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 which follow the so called "state" or "Larimore" approach [4]; in this paper we shall not be concerned with algorithms based on the so-called "shift invariance" (or MOESP-type) methods [4]. For this reason we compare algorithms on the basis of step 1). We shall identify procedures which 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], [13], [30]:

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

$$\hat{X}_t^1 \doteq \hat{X}_t^2 \tag{11}$$

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 (10) 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 [44], [45], [37]. The situation is different here since t_0 is let going to $-\infty$ according to Assumption 2.1 ♢

In this Section we shall review the state construction step for the CCA algorithm [33], [7] and for the PBSID algorithm [31], [17].

A. CCA Algorithm

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

$$\begin{aligned} \hat{Y}_{[t,T-1]} &= \hat{E}_{\|U_{[t,T]}}\left[Y_{[t,T-1]} \mid Z_{[t_0,t)}\right] = \\ &= \Gamma_{\bar{\nu}}\hat{E}_{\|U_{[t,T]}}\left[X_t \mid Z_{[t_0,t)}\right] + H_{\bar{\nu}}\hat{E}_{\|U_{[t,T]}}\left[E_{[t,T-1]} \mid Z_{[t_0,t)}\right] \\ &\simeq \Gamma_{\bar{\nu}}X_t. \end{aligned} \tag{12}$$

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

$$\hat{\mathbf{y}}_{[t,T-1]} = E_{\|\mathcal{U}_{[t,T]}}\left[\mathbf{y}_{[t,T-1]} \mid \mathcal{Z}_t^-\right] = \Gamma_{\bar{\nu}}\mathbf{x}(t) \tag{13}$$

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

$$\begin{aligned} \hat{W}_{CCA}^{-1}\hat{Y}_{[t,T-1]} &= USV^\top \\ &= [U_n\ \tilde{U}_n]\begin{bmatrix} S_n & 0 \\ 0 & \tilde{S}_n \end{bmatrix}\begin{bmatrix} V_n^\top \\ \tilde{V}_n^\top \end{bmatrix} \end{aligned} \tag{14}$$

The CCA algorithm corresponds to the choice [8]

$$\hat{W}_{CCA} := \hat{\Sigma}^{1/2}_{\mathbf{y}^+\mathbf{y}^+|\bar{\mathbf{u}}^+}. \tag{15}$$

An estimate of the observability matrix is obtained discarding the "less significant" singular values (i.e. pretending $\tilde{S}_n \simeq 0$) from[9]

$$\hat{\Gamma}_{\bar{\nu}} = \hat{W}_{CCA}U_n\hat{T} \tag{16}$$

where $\hat{T}$ can be any non-singular matrix providing a choice of basis. We acknowledge that the presence of the matrix $\hat{T}$ may appear non-standard. 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 $\hat{T}$ which shall guarantee that the estimated system matrices converge regardless of the

[8]The reader may argue that 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 so called "state approach", i.e. implementing step 2) above. See for instance [7], [13].

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

possible ambiguities, due to orientation of singular vectors and multiple singular values (if any), in the SVD (14). To this purpose we propose to use

$$\hat{T} := U_n^\top \hat{W}_{CCA}^{-1} \Gamma_{\bar{\nu}} \tag{17}$$

where $\Gamma_{\bar{\nu}}$ is the "true" (but unknown) observability matrix. The reader may argue that such choice is infeasible in practice; we stress however that 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; the choices $\hat{T} = I$ or $\hat{T} = S_n^{1/2}$ are usually made as mentioned above.

Lemma 3.2: The estimator $\hat{\Gamma}_{\bar{\nu}}$ in (16), with the choice of $\hat{T}$ in (17) converges a.s. to $\Gamma_{\bar{\nu}}$ as $N \to \infty$.

Proof: See Appendix A. ■

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

$$\begin{aligned} \hat{X}_t^{CCA} &:= \hat{\Gamma}_{\bar{\nu}}^{-L} \hat{Y}_{[t,T-1]} \\ \hat{X}_{t+1}^{CCA} &:= \hat{\Gamma}_{\bar{\nu}}^{-L} \hat{E}_{\|U_{[t+1,T]}} \left[Y_{[t+1,T]} \mid Z_{[t_0,t+1)} \right] \end{aligned} \tag{18}$$

These formulas for constructing the state space at time t and $t+1$ are discussed and motivated, for instance, in [13], 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 $\hat{W}_{CCA}$) can be substituted with their (a.s.) limit (say W_{CCA}), as discussed for instance in [6], [7], [4], [13].

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

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

Lemma 3.3: (Bauer-Ljung [7]) Assume the parameters are estimated following steps 1) and 2) above and the state is constructed according to (18). Then any choice $W_{CCA} = [H_{\bar{\nu}}(I \otimes \Lambda) H_{\bar{\nu}}^\top + \Gamma_{\bar{\nu}} \Sigma \Gamma_{\bar{\nu}}^\top]^{1/2}$ with $\Sigma = \Sigma^\top \geq 0$, provides the same asymptotic accuracy of the estimators of any system invariant.

Proof: See Appendix A. ■

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 $\hat{X}_t^{CCA}$ any state sequence resulting from a choice of W_{CCA} of the form $W_{CCA} = [H_{\bar{\nu}}(I \otimes \Lambda)H_{\bar{\nu}}^\top + \Gamma_{\bar{\nu}}\Sigma\Gamma_{\bar{\nu}}^\top]^{1/2}$ 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, but may differ for a nonsingular change of basis of course. From now on we shall always use $W_{CCA} := H_{\bar{\nu}}(I \otimes \Lambda^{1/2})$. ◊

B. PBSID algorithm

This algorithm inherits its name from the similarity with PE methods and from the fact that it is based on identification of the predictor model. As mentioned in the introduction, this algorithm was introduced in [17], inspired by [31], under the name "whitening filter algorithm". For reasons of space we shall refer the interested reader to the paper [18] 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 of all one computes the oblique projections[10]

$$\begin{aligned} \hat{Y}_{t+h}^p &:= \hat{E}_{\|Z_{[t,t+h)}}\left[Y_{t+h} \mid Z_{[t_0,t)}\right] \\ &\simeq C\bar{A}^{h-1}X_t \\ &\qquad h = 0, 1, .., \bar{\nu}. \end{aligned} \tag{19}$$

Also here the last approximate equality has to be understood in the sense that, asymptotically in N,

$$\begin{aligned} \hat{\mathbf{y}}^p(t+h) := E_{\|\mathcal{Z}_{[t,t+h)}}\left[\mathbf{y}(t+h) \mid \mathcal{Z}_t^-\right] = C\bar{A}^{h-1}\mathbf{x}(t) \\ h = 0, 1, .., \bar{\nu} \end{aligned} \tag{20}$$

holds.

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

Then one stacks all the predictors

$$\hat{Y}^p_{[t,T-1]} := \begin{bmatrix} \hat{Y}^p_t \\ \hat{Y}^p_{t+1} \\ \vdots \\ \hat{Y}^p_{T-1} \end{bmatrix} \simeq \bar{\Gamma}_{\bar{\nu}} X_t.$$

From the weighted Singular Value Decomposition[11]

$$W^{-1}\hat{Y}^p_{[t,T-1]} = PDQ^\top = [P_n \tilde{P}_n] \begin{bmatrix} D_n & 0 \\ 0 & \tilde{D}_n \end{bmatrix} \begin{bmatrix} Q_n^\top & \tilde{Q}_n^\top \end{bmatrix} \tag{21}$$

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

$$\hat{\bar{\Gamma}}_{\bar{\nu}} = W P_n \hat{\bar{T}} \tag{22}$$

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

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

$$\hat{\bar{T}} := P_n^\top W^{-1} \bar{\Gamma}_{\bar{\nu}} \tag{23}$$

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

Lemma 3.4: The estimator $\hat{\bar{\Gamma}}_{\bar{\nu}}$ in (22), with the choice of $\hat{\bar{T}}$ in (23) 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 $\hat{\bar{\Gamma}}^{-L}_{\bar{\nu}} := \left(\hat{\bar{\Gamma}}^\top_{\bar{\nu}} W^{-\top} W^{-1} \hat{\bar{\Gamma}}_{\bar{\nu}}\right)^{-1} \hat{\bar{\Gamma}}^\top_{\bar{\nu}} W^{-\top} W^{-1}$ a basis for the state space is given by

$$\begin{aligned} \hat{X}^{PBSID}_t &:= \hat{\bar{\Gamma}}^{-L}_{\bar{\nu}} \hat{Y}^p_{[t,T-1]} \\ \hat{X}^{PBSID}_{t+1} &:= \hat{\bar{\Gamma}}^{-L}_{\bar{\nu}} \begin{bmatrix} \hat{E}\left[Y_{t+1} \mid Z_{[t_0,t+1)}\right] \\ \hat{E}_{\| Z_{t+1}}\left[Y_{t+2} \mid Z_{[t_0,t+1)}\right] \\ \vdots \\ \hat{E}_{\| Z_{[t,T)}}\left[Y_T \mid Z_{[t_0,t+1)}\right] \end{bmatrix} \end{aligned} \tag{24}$$

[11] We introduce a weighting matrix W which will be chosen appropriately.

IV. WHITE INPUTS

In this section we shall study the link between the state constructions (18) and (24) under the assumption that the input signal is a white noise process (or it is absent). We leave the analysis of the general case to the next Section; since we shall need to introduce a modified PBSID algorithm to perform the comparison in that case we prefer to keep well separated the two situations and deal first with the standard algorithm.

We now state the main result of this section:

Theorem 4.1: Let Λ denote the innovation noise covariance. Under the conditions stated in Assumption 2.1, assuming that inputs are white or absent and provided W is chosen according to $W = I \otimes \Lambda^{1/2}$ and $W_{CCA} = H_{\bar{\nu}}(I \otimes \Lambda^{1/2})$, the state constructions in (18) and (24) satisfy

$$\hat{X}_t^{PBSID} \dot{=} \hat{X}_t^{CCA}$$

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 $\hat{Y}_{t+h} := \hat{E}_{\|U_{[t,T]}}\left[Y_{t+h} \mid Z_{[t_0,t)}\right]$ satisfies:

$$\hat{Y}_{t+h} \dot{=} \hat{Y}^p_{t+h} + \sum_{i=1}^{h} \hat{\Phi}_{hi} \hat{Y}_{t+h-i} \tag{25}$$

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

$$\hat{Y}^p_{[t,T-1]} \dot{=} \begin{bmatrix} I & 0 & \dots & 0 \\ -\hat{\Phi}_{11} & I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\Phi}_{\bar{\nu}-1,\bar{\nu}-1} & -\hat{\Phi}_{\bar{\nu}-1,\bar{\nu}-2} & \dots & I \end{bmatrix} \hat{Y}_{[t,T-1]} \tag{26}$$

The lower triangular matrix in formula (26) converges to a block Toeplitz matrix which is the inverse of $H_{\bar{\nu}}$ defined in (8), i.e.

$$\lim_{N\to\infty} \hat{H}_{\bar{\nu}}^{-1} = H_{\bar{\nu}}^{-1}.$$

Then we shall write (26) as

$$\hat{Y}^p_{[t,T-1]} \doteq H_{\bar{\nu}}^{-1} \hat{Y}_{[t,T-1]} \tag{27}$$

Proof: See Appendix A. ■

Remind that (27) should be understood in the sense that the left and right hand side give rise to state sequences which differ, up to $o_P(1/\sqrt{N})$ terms, only for a non-singular 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 (21) that in the predictor-based algorithm one takes the SVD of $W^{-1}\hat{Y}^p_{[t,T-1]}$ while, from (14), $W_{CCA}^{-1}\hat{Y}_{[t,T-1]}$ is used in the CCA algorithm.

Note that, the CCA algorithm corresponds to the choice

$$W_{CCA} = \Sigma^{1/2}_{\mathbf{y}^+\mathbf{y}^+|\bar{\mathbf{u}}^+} = (\Gamma_{\bar{\nu}} \Sigma_{\mathbf{xx}|\bar{\mathbf{u}}^+} \Gamma_{\bar{\nu}}^\top + H_{\bar{\nu}}(I \otimes \Lambda) H_{\bar{\nu}}^\top)^{1/2}$$

where Λ is the variance of the innovation.

However, by letting $\Sigma = 0$ in Lemma 3.3, $W_{CCA} = (H_{\bar{\nu}}(I \otimes \Lambda) H_{\bar{\nu}}^\top)^{1/2}$ provides the same asymptotic behavior.

If we now pre-multiply both sides of (27) after Lemma 4.2 by $W^{-1/2} = (I \otimes \Lambda)^{-1/2}$ we obtain that

$$W^{-1/2}\hat{Y}^p_{[t,T-1]} \doteq (I \otimes \Lambda)^{-1/2} H_{\bar{\nu}}^{-1} \hat{Y}_{[t,T-1]} = W_{CCA}^{-1} \hat{Y}_{[t,T-1]}. \tag{28}$$

As described in Section III the right hand side is used in CCA while 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 $\hat{X}_t^{CCA}$ and $\hat{X}_t^{PBSID}$ are asymptotically equivalent, which using Proposition 3.1 concludes the proof. ■

V. NON-WHITE 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 PBSID$_{opt}$. We shall explain the details later in this section.

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 PBSID$_{opt}$.

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

Defining $\mathcal{K} := \left[\bar{A}^{t-t_0-1}[K\ B]\ \bar{A}^{t-t_0-2}[K\ B]\ \dots [K\ B]\right]$ the output tail Y_{t+h} can be written as:

$$\begin{aligned} Y_{t+h} &= C\bar{A}^h X_t + \textstyle\sum_{i=1}^{h} C\bar{A}^{i-1}\left(KY_{t+h-i} + BU_{t+h-i}\right) + E_{t+h} \\ &= C\bar{A}^h \mathcal{K} Z_{[t_0,t)} + \textstyle\sum_{i=1}^{h} C\bar{A}^{i-1}\left(KY_{t+h-i} + BU_{t+h-i}\right) + E_{t+h} + \underline{o}_P(1/\sqrt{N}) \\ &:= \Xi_h Z_{[t_0,t)} + \textstyle\sum_{i=1}^{h} \Phi_{hi} Y_{t+h-i} + \Psi_{hi} U_{t+h-i} + E_{t+h} + \underline{o}_P(1/\sqrt{N}) \end{aligned} \tag{29}$$

where, thanks to Assumption 2.1, 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 construction for CCA and PBSID are based on the oblique projections (12) and (19); we shall make use of the relations

$$\hat{Y}_{t+h} = \hat{E}_{\|U_{[t,T]}}\left[Y_{t+h} \mid Z_{[t_0,t)}\right] = \hat{E}_{\|U_{[t,T]}}\left[\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}, U_{[t+h,T]}\right] \mid Z_{[t_0,t)}\right] \tag{30}$$

and

$$\hat{Y}^p_{t+h} = \hat{E}_{\|Z_{[t,t+h)}}\left[Y_{t+h} \mid Z_{[t_0,t)}\right] = \hat{E}_{\|Z_{[t,t+h)}}\left[\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}\right] \mid Z_{[t_0,t)}\right]. \tag{31}$$

A. Optimized PBSID Algorithm

Stacking the data and using (29) (discarding[12] $o_P(1/\sqrt{N})$ terms) we obtain:

$$\begin{bmatrix} Y_t \\ Y_{t+1} \\ \vdots \\ Y_T \end{bmatrix} \doteq \begin{bmatrix} \Xi_0 \\ \Xi_1 \\ \vdots \\ \Xi_\nu \end{bmatrix} Z_{[t_0,t)} + \begin{bmatrix} 0 & 0 & \dots & 0 \\ \Phi_{11} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{\nu\nu} & \dots & \Phi_{\nu 1} & 0 \end{bmatrix} Y_{[t,T]} + \begin{bmatrix} 0 & 0 & \dots & 0 \\ \Psi_{11} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{\nu\nu} & \dots & \Psi_{\nu 1} & 0 \end{bmatrix} U_{[t,T]} + \begin{bmatrix} E_t \\ E_{t+1} \\ \vdots \\ E_T \end{bmatrix} \tag{32}$$

Observe that the lower triangular matrices in (32) 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 (31) is equivalent to solving (32) "row by row"; hence

[12]See Appendix B for a formal justification.

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 $i \neq i'$.

This is equivalent to solving the least squares problem obtained vectorizing (32):

$$Y := \begin{bmatrix} \mathrm{vec}\,(Y_t) \\ \mathrm{vec}\,(Y_{t+1}) \\ \vdots \\ \mathrm{vec}\,(Y_T) \end{bmatrix} \doteq S^P \Omega^P + \begin{bmatrix} \mathrm{vec}\,(E_t) \\ \mathrm{vec}\,(E_{t+1}) \\ \vdots \\ \mathrm{vec}\,(E_T) \end{bmatrix} = S^P \Omega^P + E \tag{33}$$

where the matrix S^P has the form

$$S^P = \begin{bmatrix} (Z^\top_{[t_0,t)} \otimes I) & 0 & \dots & 0 \\ 0 & (Z^\top_{[t_0,t+1)} \otimes I) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (Z^\top_{[t_0,T)} \otimes I) \end{bmatrix} \tag{34}$$

and Ω^P is given by

$$\Omega^P = \begin{bmatrix} \mathrm{vec}^\top(\Xi_0) & \mathrm{vec}^\top(\Xi_1) & \mathrm{vec}^\top(\Phi_{11}) & \mathrm{vec}^\top(\Psi_{11}) & \dots & \mathrm{vec}^\top(\Xi_\nu) & \dots & \mathrm{vec}^\top(\Psi_{\nu 1}) \end{bmatrix}^\top; \tag{35}$$

note that the "noise term" E can be written in the form

$$E = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ 0 & 0 & 1 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} e_t \\ e_{t+1} \\ \vdots \\ e_{t+N-1} \\ e_{t+N} \\ \vdots \\ e_{T+N-1} \end{bmatrix} =: L E_I \tag{36}$$

where the last equality defines the $Nm\bar{\nu} \times (N+\bar{\nu})m$ matrix L and the vector E_I. Equation (36) 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 $\hat{\Omega}^{P_{opt}}$ of Ω^P which has the smallest asymptotic variance among all linear (asymptotically unbiased) estimators based on (33). Being the noise covariance R singular and the regression matrix S^P of full rank, $\hat{\Omega}^{P_{opt}}$ can be obtained as described in [43][Complement C.4.3]. However, it is possible to reduce (33) to a smaller least squares problem with full rank noise and equality constraints (see [42], [48], [43], [24] just to cite a few references). We refer the reader to Appendix B for further details. Note however that this 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 the present implementation the "optimized" algorithm has a computational complexity which is $O(N^2(log(N))^\beta)$ while the original PBSID algorithm as well as CCA have a computational complexity which is $O(N(log(N))^{2\beta})$, where β is the rate at which the past horizon $t - t_0$ grows with N (i.e. $t - t_0 = O((log(N))^\beta)$, see Assumption 2.1). 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 play an important role.

As an anonymous reviewer has suggested, both PBSID and its optimized version have strong similarities with VARX identification. We acknowledge that this is worth investigating; exploiting these similarities might also be advantageous for computational reasons. Indeed some work along these lines has already been done in [10], where the relation between PBSID and SSARX (which uses VARX modeling) has been studied and in [12], where the relation between PBSID$_{opt}$ and VARX modeling have been elucidated. However entering into the fine structure of the constrained least squares problem (33) would require far more space than available here and therefore we refer the reader to the references above.

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

Using the estimator $\hat{\Omega}^{P_{opt}}$, the oblique projections $\hat{Y}_{t+h}^p$ can be substituted with $\hat{Y}_{t+h}^{P_{opt}} = \hat{\Xi}_h^{P_{opt}} Z_{[t_0,t)}$ in the SVD step (21) and an estimator for the state be given by

$$\hat{X}_t^{P_{opt}} := \left(\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{opt}}\right)^{-L} \begin{bmatrix} \hat{Y}_t^{P_{opt}} \\ \hat{Y}_{t+1}^{P_{opt}} \\ \vdots \\ \hat{Y}_{T-1}^{P_{opt}} \end{bmatrix}. \tag{37}$$

Also the "shifted" oblique projections used for the computation of the state at time $t+1$ (see (24)) can be substituted by

$$\hat{X}_{t+1}^{P_{opt}} := \left(\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{opt}}\right) \begin{bmatrix} \hat{\Xi}_1^{P_{opt}} & \hat{\Phi}_{11}^{P_{opt}} & \hat{\Psi}_{11}^{P_{opt}} \\ \hat{\Xi}_2^{P_{opt}} & \hat{\Phi}_{22}^{P_{opt}} & \hat{\Psi}_{22}^{P_{opt}} \\ & \vdots & \\ \hat{\Xi}_\nu^{P_{opt}} & \hat{\Phi}_{\nu\nu}^{P_{opt}} & \hat{\Psi}_{\nu\nu}^{P_{opt}} \end{bmatrix} Z_{[t_0,t+1)}.$$

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

$$\hat{E}_t^{P_{opt}} := Y_t - E\left[\hat{Y}_t^{P_{opt}} | \hat{X}_t^{P_{opt}}\right] = Y_t - \hat{C}_N^{P_{opt}} \hat{X}_t^{P_{opt}} \tag{38}$$

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

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

in the least squares sense.

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

$$\bar{T}_N^{P_{opt}} := \left(\hat{\bar{\Gamma}}_{\bar{\nu}}^{P_{opt}}\right)^{-L} \bar{\Gamma}_{\bar{\nu}}$$

i.e. $A_N^{P_{opt}} := \bar{T}_N^{P_{opt}} A \left(\bar{T}_N^{P_{opt}}\right)^{-1}$ etc.

Then, with the choice of $\hat{\bar{T}}$ in (23), $\bar{T}_N^{P_{opt}}$ converges to the identity matrix and the errors $\tilde{A}_N^{P_{opt}} = \hat{A}_N^{P_{opt}} - A_N^{P_{opt}}, \tilde{B}_N^{P_{opt}} = \hat{B}_N^{P_{opt}} - B_N^{P_{opt}}, \tilde{C}_N^{P_{opt}} = \hat{C}_N^{P_{opt}} - C_N^{P_{opt}}, \tilde{K}_N^{P_{opt}} = \hat{K}_N^{P_{opt}} - K_N^{P_{opt}}$ satisfy:

$$\begin{bmatrix} \mathrm{vec}\left(\tilde{A}_N^{P_{opt}}\right) \\ \mathrm{vec}\left(\tilde{B}_N^{P_{opt}}\right) \\ \mathrm{vec}\left(\tilde{C}_N^{P_{opt}}\right) \\ \mathrm{vec}\left(\tilde{K}_N^{P_{opt}}\right) \end{bmatrix} \doteq M_P \left(\hat{\Omega}^{P_{opt}} - \Omega^P\right) \tag{39}$$

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 (32) and (33).

Using (29) and discarding $o_P(1/\sqrt{N})$ terms[13], the inner projection in (30) can be computed solving in the least squares sense:

$$\begin{bmatrix} Y_t \\ Y_{t+1} \\ \vdots \\ Y_T \end{bmatrix} \doteq \begin{bmatrix} \Xi_0 \\ \Xi_1 \\ \vdots \\ \Xi_\nu \end{bmatrix} Z_{[t_0,t)} + \begin{bmatrix} 0 & 0 & \dots & 0 \\ \Phi_{11} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{\nu\nu} & \dots & \Phi_{\nu 1} & 0 \end{bmatrix} Y_{[t,T]} + \begin{bmatrix} * & * & \dots & * \\ \Psi_{11} & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{\nu\nu} & \dots & \Psi_{\nu 1} & * \end{bmatrix} U_{[t,T]} + \begin{bmatrix} E_t \\ E_{t+1} \\ \vdots \\ E_T \end{bmatrix} \quad (40)$$

where the $*$'s denote parameters which are estimated but which value has no interest (and actually should be equal to zero). Since each row is parameterized independently the orthogonal projections $\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}, U_{[t+h,T]}\right]$, for $h = 0, .., \nu$, are equivalent to solving the least squares problem obtained by vectorizing (40)

$$\begin{bmatrix} \mathrm{vec}\,(Y_t) \\ \mathrm{vec}\,(Y_{t+1}) \\ \vdots \\ \mathrm{vec}\,(Y_T) \end{bmatrix} \doteq S^{CCA}\Omega^{CCA} + \begin{bmatrix} \mathrm{vec}\,(E_t) \\ \mathrm{vec}\,(E_{t+1}) \\ \vdots \\ \mathrm{vec}\,(E_T) \end{bmatrix} = S^{CCA}\Omega^{CCA} + E \quad (41)$$

where the matrices Ω^{CCA} and S^{CCA} have the form

$$\Omega^{CCA} = \begin{bmatrix} \Omega^P \\ * \end{bmatrix} \qquad S^{CCA} = \begin{bmatrix} S^P & * \end{bmatrix}; \quad (42)$$

the terms denoted with $*$ contain the vectorization of all the $*$'s in (40).

As we have just stressed the inner projection in (30) is equivalent to solving (41) in the least squares sense (with no weighting). We shall denote with $\hat{\Omega}^{CCA}$ the corresponding estimator of Ω^{CCA} and with $\hat{\Omega}^{P_{CCA}}$, $\hat{\Xi}_k^{CCA}$, $\hat{\Phi}_{ij}^{CCA}$, $\hat{\Psi}_{ij}^{CCA}$ the estimators of Ω^P, Ξ_k, Φ_{ij}, Ψ_{ij} extracted from its components.

[13]See Appendix B for a formal justification.

With this observation the oblique projections (30) are written in the from:

$$\hat{Y}_{t+h} = \hat{\Xi}_h^{CCA} Z_{[t_0,t)} + \sum_{i=1}^{h} \hat{\Phi}_{hi}^{CCA} \hat{Y}_{t+h-i} \tag{43}$$

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

$$W_{CCA}^{-1} \hat{Y}_{[t,T-1]} \doteq (I \otimes \Lambda^{-1/2}) \begin{bmatrix} \hat{\Xi}_0^{CCA} \\ \vdots \\ \hat{\Xi}_{\bar{\nu}}^{CCA} \end{bmatrix} Z_{[t_0,t)} \tag{44}$$

and

$$W_{CCA}^{-1} \hat{E}_{\| U_{[t+1,T]}} \left[Y_{[t+1,T]} \mid Z_{[t_0,t+1)} \right] \doteq (I \otimes \Lambda^{-1/2}) \begin{bmatrix} \hat{\Xi}_1^{CCA} & \hat{\Phi}_{11}^{CCA} & \hat{\Psi}_{11}^{CCA} \\ \hat{\Xi}_2^{CCA} & \hat{\Phi}_{22}^{CCA} & \hat{\Psi}_{22}^{CCA} \\ & \vdots & \\ \hat{\Xi}_{\nu}^{CCA} & \hat{\Phi}_{\nu\nu}^{CCA} & \hat{\Psi}_{\nu\nu}^{CCA} \end{bmatrix} Z_{[t_0,t+1)} \tag{45}$$

The state sequences $\hat{X}_t^{CCA}$, $\hat{X}_{t+1}^{CCA}$ are constructed as described in Section III. Substituting the right hand sides of (44) and (45) 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:

$$\hat{E}_t^{CCA} := Y_t - \hat{E}\left[Y_t \mid \hat{X}_t^{CCA} \right] = Y_t - \hat{C}_N^{CCA} \hat{X}_t^{CCA} \tag{46}$$

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

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

in the least squares sense.

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

$$T_N^{CCA} := \hat{\Gamma}_{\bar{\nu}}^{-L} \Gamma_{\bar{\nu}} \tag{47}$$

i.e. $A_N^{CCA} := T_N^{CCA} A \left(T_N^{CCA}\right)^{-1}$ etc.

Then, with the choice of $\hat{T}$ in (17), T_N^{CCA} converges to the identity matrix and the errors $\tilde{A}_N^{CCA} = \hat{A}_N^{CCA} - A_N^{CCA}, \tilde{B}_N^{CCA} = \hat{B}_N^{CCA} - B_N^{CCA}, \tilde{C}_N^{CCA} = \hat{C}_N^{CCA} - C_N^{CCA}, \tilde{K}_N^{CCA} = \hat{K}_N^{CCA} - K_N^{CCA}$ satisfy:

$$\begin{bmatrix} \mathrm{vec}\left(\tilde{A}_N^{CCA}\right) \\ \mathrm{vec}\left(\tilde{B}_N^{CCA}\right) \\ \mathrm{vec}\left(\tilde{C}_N^{CCA}\right) \\ \mathrm{vec}\left(\tilde{K}_N^{CCA}\right) \end{bmatrix} \doteq M_P\left(\hat{\Omega}^{P_{opt}} - \Omega^P\right) + M_1^{CCA}\left(\hat{\Omega}^{P_{CCA}} - \hat{\Omega}^{P_{opt}}\right) + M_2^{CCA}\mathrm{vec}\left(\hat{\Xi}_0 - \hat{\Xi}_0^{P_{opt}}\right) \quad (48)$$

where $\hat{\Xi}_0$ is defined in (A.67). The matrix M_P is the same which appears in equation (39) while M_1^{CCA} and M_2^{CCA} are suitably defined matrices with rows in ℓ_2.

Proof: See Appendix A. ■

C. Comparison between CCA and PBSID$_{opt}$

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 which depends differentiably on the system matrices (A, B, C, D). Denote with $\hat{\Theta}^{CCA}$ and $\hat{\Theta}^{P_{opt}}$ the estimators of any such Θ using respectively CCA and PBSID$_{opt}$; then,

$$\mathrm{AsVar}\left\{\sqrt{N}\hat{\Theta}^{CCA}\right\} \geq \mathrm{AsVar}\left\{\sqrt{N}\hat{\Theta}^{P_{opt}}\right\}. \quad (49)$$

Proof:

As seen above in formulas (41) and (33), the first step of the algorithms can be seen as a linear least squares problem; there are, however, three "complications" which 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^{CCA} are data dependent;
3) the dimension of the parameter vector Ω^P grows with the sample size.

We shall deal with problems, respectively, as follows:

1) We shall see in Appendix B, (see Lemma 7.1 and the discussion following the Lemma) that the $o_P(1/\sqrt{N})$ terms which have been omitted in (33) have a special form which allows, indeed, to neglect them.
2) Under Assumption 2.4, it is tedious but easy to adapt standard properties of Markov estimators (i.e. unbiasedness and minimum variance; see for instance [43][Lemma 4.3, Proof A]) to the case in which the regression matrix (S^P, S^{CCA}) is "data dependent". Of course unbiasedness and minimum variance hold only asymptotically.
3) We shall follow the approach of [36], studying the (scalar) estimators $\eta_N^\top \hat{\Omega}^P$ with η_N satisfying Assumption 2.3.

As discussed in Appendix B, the estimators $\hat{\Omega}^{P_{CCA}}$ and $\hat{\Omega}^{P_{opt}}$, satisfy the inequality:

$$\mathrm{AsVar}\left\{\sqrt{N}\eta_N^\top \hat{\Omega}^{P_{CCA}}\right\} \geq \mathrm{AsVar}\left\{\sqrt{N}\eta_N^\top \hat{\Omega}^{P_{opt}}\right\} \tag{50}$$

$\forall\ \eta_N$ satisfying Assumption 2.3. In particular, being $\tilde{\Omega}^{P_{opt}} := \hat{\Omega}^{P_{opt}} - \Omega^P$ the "optimal" estimation error, $\tilde{\Omega}^{P_{CCA}} := \hat{\Omega}^{P_{CCA}} - \Omega^P$ can also be written as

$$\sqrt{N}\tilde{\Omega}^{P_{CCA}} = \sqrt{N}\tilde{\Omega}^{P_{opt}} + \sqrt{N}(\hat{\Omega}^{P_{CCA}} - \hat{\Omega}^{P_{opt}}) \tag{51}$$

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

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

$\forall\ \eta_N$, γ_N satisfying Assumption 2.3.

The reason for this inequality is twofold:

1) in (41) the parameters denoted with $*$'s are estimated even though they are known to be zero (observe that instead in (32), (33) the lower triangular structure of the matrix describing the link from future input to future output is enforced)
2) the modified PBSID algorithm solves (33) 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

$$\sqrt{N}\left(\tilde{\Xi}_0\right) = \sqrt{N}\left(\hat{\Xi}_0^{P_{opt}} - \Xi_0\right) + \sqrt{N}(\hat{\Xi}_0 - \hat{\Xi}_0^{P_{opt}}) = \sqrt{N}\tilde{\Xi}_0^{P_{opt}} + \sqrt{N}(\hat{\Xi}_0 - \hat{\Xi}_0^{P_{opt}}) \tag{52}$$

where $\sqrt{N}\eta_N^\top \mathrm{vec}(\hat{\Xi}_0 - \hat{\Xi}_0^{P_{opt}})$ is (asymptotically) uncorrelated from $\sqrt{N}\gamma_N^\top \tilde{\Omega}^{P_{opt}} = \sqrt{N}\gamma_N^\top\left(\hat{\Omega}^{P_{opt}} - \Omega^P\right)$ $\forall\ \eta_N, \gamma_N$ satisfying Assumption 2.3.

These last observations imply that ($\sqrt{N}$ times) the last two terms on the right hand side of (48) are asymptotically uncorrelated with ($\sqrt{N}$ times) the first. Therefore, from equations (39) and (48),

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

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 (see Figure 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 [12]. The weighting step needed to find $\hat{\Omega}^{P_{opt}}$ (see Appendix B) is however necessary to obtain the inequality (50). ◊

VI. Simulation Results

The simulation setup is as follows: we consider two systems to be identified (in innovation form); the first is a first order ARX system

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

while the second is a first order ARMAX model

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

The input is either unit variance white noise or unit variance white noise passed through the filter $H_u(z)$

$$H_u(z) = \frac{z^2 + 0.8z + 0.55}{z^2 - 0.5z + 0.9};$$

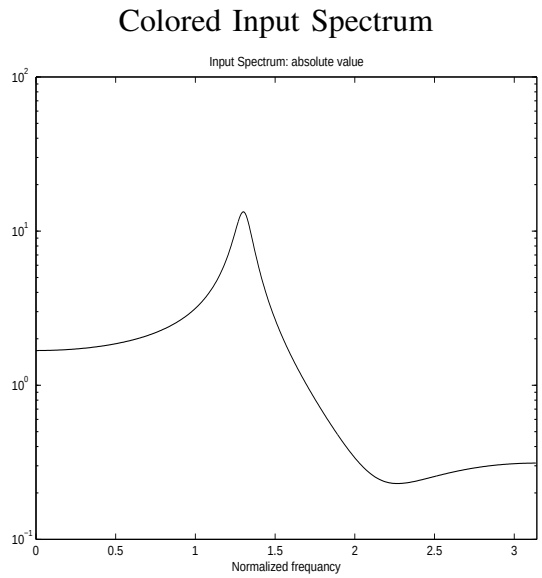

Fig. 1. Colored input spectrum: absolute value.

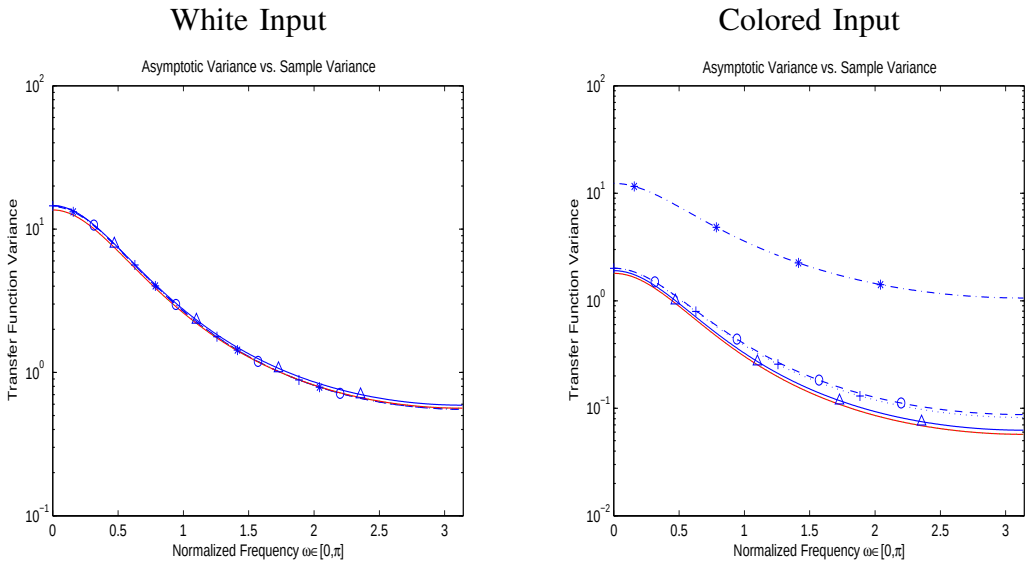

Fig. 2. EXAMPLE 1 (*ARX of order 1*): Asymptotic Variance (and its Monte Carlo estimate) vs. normalized frequency ($\omega \in [0, \pi]$) Solid with triangles ($\triangle$) PEM, dashed with stars ($*$): CCA, dotted with crosses ($+$):"predictor-based" algorithm (PBSID), dotted with circles (o): Jansson's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

the input spectrum is plotted in Figure 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 $F(z) = \frac{1}{z-0.5}$ (equal for the two examples); the future and past horizons are chosen to be $t - t_0 = \nu = 10$. In Figure 4 (left plot) we also report the results for $\nu = 1$, $t - t_0 = 10$. In Figure 4 we show the dependence of the asymptotic variance as a function of the future horizon ν measured by the efficiency index:

$$Eff(\nu) := \frac{\int_0^{2\pi} \mathrm{AsVar}\left\{\hat{F}_\nu(j\omega)\right\} d\omega}{\int_0^{2\pi} CRLB(j\omega)\, d\omega} \tag{53}$$

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

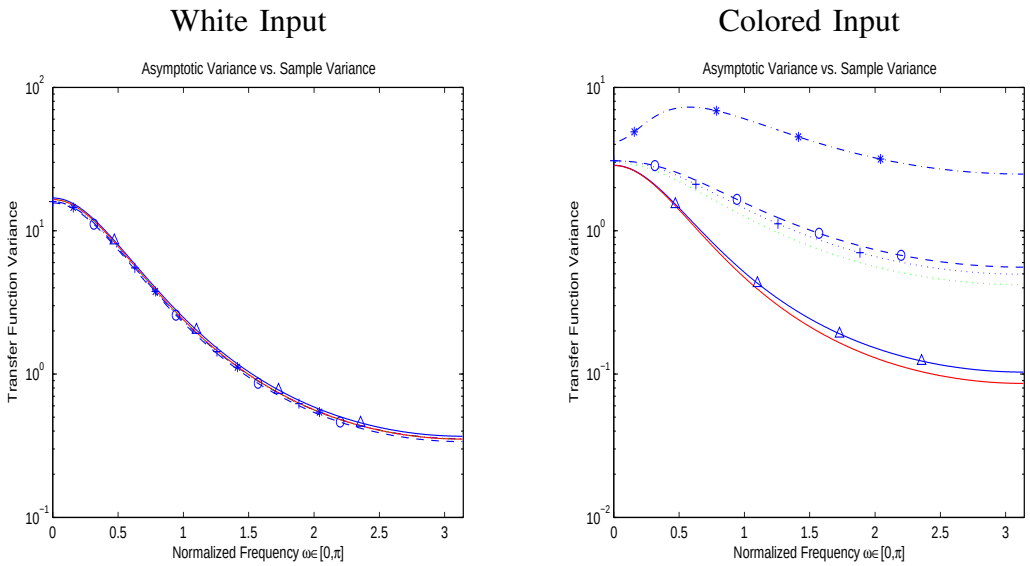

Fig. 3. EXAMPLE 2 (*ARMAX of order 1*): Asymptotic Variance (and its Monte Carlo estimate) vs. normalized frequency ($\omega \in [0, \pi]$) Solid with triangles ($\triangle$) PEM, dashed with stars ($*$): CCA, dotted with crosses ($+$):"predictor-based" algorithm (PBSID), dotted with circles (o): Jansson's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

Note that for the white input case (see Figures 2 and 3, left plots) 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 [11]) reaches the Cramér-Rao lower bound also

for colored inputs when the system is ARX (see Figure 2, right plot).

In the colored input case (see Figures 2 and 3, right plots and Figure 4) the results are fundamentally different: CCA behaves significantly worse than PEM and PBSID. We also report the asymptotic variance computed using the formulas which can be found in [11] (*dotted line*) and the Cramér-Rao Lower Bound (CRLB, *solid line*).

The algorithm by Jansson [31] is always indistinguishable from PBSID, as predicted by the results in [10].

It is interesting to observe that in this particular example and with colored inputs, the asymptotic variance of PBSID/PBSID$_{opt}$ is close to the Cramér-Rao lower bound (even though it does not reach it) for $\nu = n = 1$ (see Figure 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 Figure 4) that the modified version PBSID$_{opt}$ behaves as PBSID in the example considered.

VII. Conclusions

In this paper we have shown that the PBSID algorithm, introduced in [17] 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 [11].

The significance of this result is strengthened by the fact that, as shown in [10], PBSID and the SSARX algorithm in [31] are asymptotically equivalent.

We have also proposed a slightly modified version of PBSID which provably behaves always better than CCA. We remind the reader that both PBSID and its "optimized" version PBSID$_{opt}$ are able to deal with feedback.

An important question which remains open concerns efficiency.

In [35] it is claimed that a procedure which is essentially equivalent to the SSARX algorithm is efficient for general input signal. However in [35] the past and future horizons (see formula (2) and the definitions before formula (9) in [35]) are assumed to be equal and, for consistency reason, let to go to infinity. Computations based on the asymptotic variance (see Figure 4 right plot for an example with $t - t_0 = \nu = 10$, but essentially unchanged results are obtained

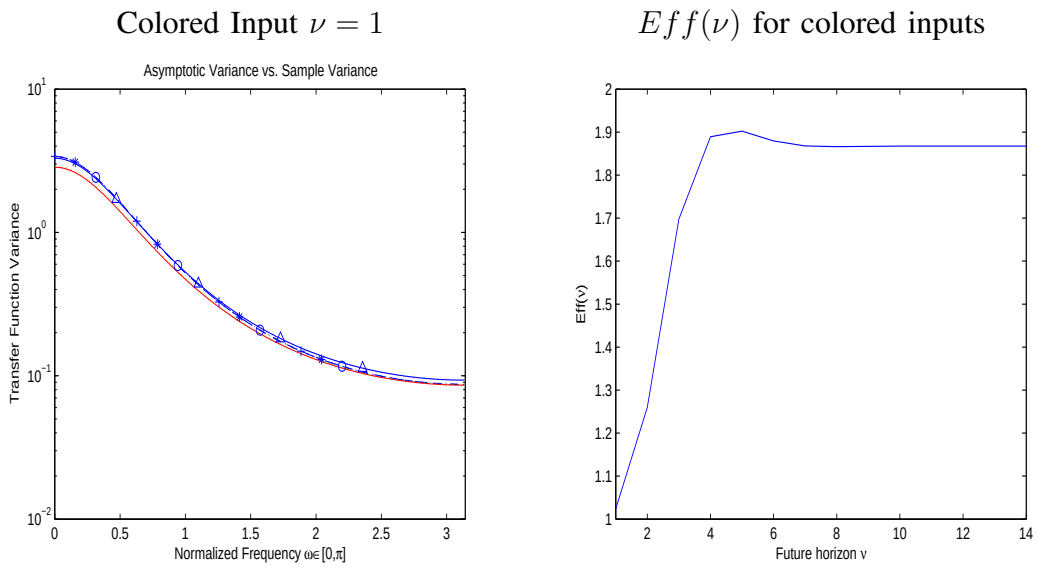

Fig. 4. EXAMPLE 3 (*ARMAX of order 1*). Left plot: asymptotic Variance (and its Monte Carlo estimate) vs. normalized frequency ($\omega \in [0, \pi]$) Solid with triangles ($\triangle$) PEM, dashed with stars ($*$): CCA, dotted with crosses ($+$):"predictor-based" algorithm (PBSID), dotted with circles (o): PBSID$_{opt}$ algorithm, dotted: asymptotic variance for PBSID, solid: Cramér-Rao lower bound. Right plot: Index $Eff(\nu)$ (see (53)) as a function of ν

increasing $t - t_0 = \nu$) 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 Figure 4 right plot).

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.[14]

We hope the results of this paper have shed some light towards this direction.

Also the question of non-asymptotic 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.

[14]Remind that here $t - t_0$ is supposed to go to infinity according to Assumption 2.1.

APPENDIX A: PROOFS

Proof of Proposition 3.1. Recall that, defining

$$\hat{E}_t^i = Y_t - \hat{E}\left[Y_t \mid \hat{X}_t^i\right],$$

the least square estimators $\hat{A}^i, \hat{B}^i, \hat{C}^i$, $i = 1, 2$ are obtained from

$$\begin{bmatrix} \hat{A}^i & \hat{B}^i & \hat{K}^i \end{bmatrix} := \hat{X}_{t+1}^i \begin{bmatrix} \left(\hat{X}_t^i\right)^\top & U_t^\top & \left(\hat{E}_t^i\right)^\top \end{bmatrix} \left\{ \begin{bmatrix} \hat{X}_t^i \\ U_t \\ \hat{E}_t^i \end{bmatrix} \begin{bmatrix} \left(\hat{X}_t^i\right)^\top & U_t^\top & \left(\hat{E}_t^i\right)^\top \end{bmatrix} \right\}^{-1} \quad \text{(A.54)}$$

$$\hat{C}^i := Y_t \left(\hat{X}_t^i\right)^\top \left\{ \hat{X}_t^i \left(\hat{X}_t^i\right)^\top \right\}^{-1} \quad \text{(A.55)}$$

For simplicity of exposition we shall only deal with (A.55); using (11) and recalling that $\hat{X}_t^1 \dot{=} \hat{X}_t^2$ 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:

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

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

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

Similarly:

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

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, equation (16) can be rewritten as

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

Let W_{CCA} be the a.s. limit of $\hat{W}_{CCA}$. It is well known that under mild conditions, e.g. $\bar{\nu}$ larger than the system order, Assumption 2.1 and condition (5), the column space of $\hat{W}_{CCA}U_n = \hat{\Gamma}_{\bar{\nu}}\hat{T}^{-1}$ converges to the column space of $\Gamma_{\bar{\nu}}$ (see for instance [19], Theorems 9 and 10). Therefore also the column space of U_n converges to the column space of $W_{CCA}^{-1}\Gamma_{\bar{\nu}}$ and hence $\Pi_{U_n}\hat{W}_{CCA}^{-1}\Gamma_{\bar{\nu}}$ converges to $W_{CCA}^{-1}\Gamma_{\bar{\nu}}$ from which $\hat{\Gamma}_{\bar{\nu}} = \hat{W}_{CCA}\Pi_{U_n}\hat{W}_{CCA}^{-1}\Gamma_{\bar{\nu}}$ converges to $\Gamma_{\bar{\nu}}$. ■

Proof of Lemma 3.3. First note that, from the asymptotic variance analysis found for instance in [7], [13], [11] W_{CCA} appears only through the left inverse

$$\Gamma_{\bar{\nu}}^{-L} = \left(\Gamma_{\bar{\nu}}^\top W_{CCA}^{-\top}W_{CCA}^{-1}\Gamma_{\bar{\nu}}\right)^{-1}\Gamma_{\bar{\nu}}^\top W_{CCA}^{-\top}W_{CCA}^{-1}.$$

Let now $W_0 := (H_{\bar{\nu}}(I\otimes\Lambda)H_{\bar{\nu}}^\top)^{1/2}$ and define $R_0^{-1} := W_0^{-\top}W_0^{-1} = (H_{\bar{\nu}}(I\otimes\Lambda)H_{\bar{\nu}}^\top)^{-1}$; we shall show that if W_{CCA} (square invertible), is chosen such that $R^{-1} := W_{CCA}^{-\top}W_{CCA}^{-1}$ can be written in the form $R^{-1} = (R_0 + \Gamma_{\bar{\nu}}\Sigma\Gamma_{\bar{\nu}}^\top)^{-1}$ for some matrix $\Sigma = \Sigma^\top \geq 0$, then

$$\Gamma_{\bar{\nu}}^{-L} := \left(\Gamma_{\bar{\nu}}^\top W_{CCA}^{-\top}W_{CCA}^{-1}\Gamma_{\bar{\nu}}\right)^{-1}\Gamma_{\bar{\nu}}^\top W_{CCA}^{-\top}W_{CCA}^{-1} = \left(\Gamma_{\bar{\nu}}^\top W_0^{-\top}W_0^{-1}\Gamma_{\bar{\nu}}\right)^{-1}\Gamma_{\bar{\nu}}^\top W_0^{-\top}W_0^{-1} \quad \text{(A.56)}$$

holds, where the right hand side does not depend on Σ; therefore different choices of Σ do not change the asymptotic variance.

First, we use the matrix inversion lemma and rewrite $R^{-1} = (R_0 + \Gamma_{\bar{\nu}}\Sigma\Gamma_{\bar{\nu}}^\top)^{-1} = R_0^{-1} + R_0^{-1}\Gamma_{\bar{\nu}}S\Gamma_{\bar{\nu}}^\top R_0^{-1}$ for a suitable matrix S (depending upon Σ).

Define now $\Gamma_0 := W_0^{-1}\Gamma_{\bar{\nu}}$. Let us first rewrite $\Gamma_{\bar{\nu}}^{-L}$ as

$$\begin{aligned}\Gamma_{\bar{\nu}}^{-L} &= \left(\Gamma_0^\top(I+\Gamma_0 S\Gamma_0^\top)\Gamma_0\right)^{-1}\Gamma_0^\top(I+\Gamma_0 S\Gamma_0^\top)W_0^{-1} \\ &= \left[(I+\Gamma_0^\top\Gamma_0 S)\Gamma_0^\top\Gamma_0\right]^{-1}(I+\Gamma_0^\top\Gamma_0 S)\Gamma_0^\top W_0^{-1}.\end{aligned}$$

Since $\Gamma_0\Gamma_0^\top$ and $\left(\Gamma_0^\top(I+\Gamma_0 S\Gamma_0^\top)\Gamma_0\right)$ are nonsingular, also $(I+\Gamma_0^\top\Gamma_0)$ is invertible from which

$$\Gamma_{\bar{\nu}}^{-L} = \left(\Gamma_0^\top\Gamma_0\right)^{-1}\Gamma_0^\top W_0^{-1} = \left(\Gamma_{\bar{\nu}}^\top W_0^{-\top}W_0^{-1}\Gamma_{\bar{\nu}}\right)^{-1}\Gamma_{\bar{\nu}}^\top W_0^{-\top}W_0^{-1},$$

which yields (A.56), concluding the proof[15]. ■

Proof of Lemma 4.2.

[15]Thanks to an anonymous referee for a suggestion which have allowed to shorten the proof.

First let us note that

$$\begin{array}{l}\hat{E}_{\|U_{[t,T]}}\left[Y_{t+h} \mid Z_{[t_0,t)}\right] = \\ = \hat{E}_{\|U_{[t,T]}}\left[\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}, U_{[t+h,T]}\right] \mid Z_{[t_0,t)}\right]\end{array} \tag{A.57}$$

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 [27][Theorem 5.3.2]), it follows that $\hat{\Sigma}_{\mathbf{fp}} := \frac{FP^\top}{N}$ and $\hat{\Sigma}_{\mathbf{yf}} := \frac{Y_{t+h}F^\top}{N}$ satisfy:

$$\left\|\hat{\Sigma}_{\mathbf{fp}}\right\|_2 = O\left(\sqrt{\tfrac{(t-t_0)log(log\,N)}{N}}\right) \quad \left\|\hat{\Sigma}_{\mathbf{yf}}\right\|_2 = O\left(\sqrt{\tfrac{log(log\,N)}{N}}\right)$$

which implies

$$\begin{array}{c}\left\|\hat{\Sigma}_{\mathbf{fp}}\right\|_2 \left\|\hat{\Sigma}_{\mathbf{fp}}\right\|_2 = O\left((t-t_0)\tfrac{log(log(N)}{N}\right) \\ \left\|\hat{\Sigma}_{\mathbf{yf}}\right\|_2 \left\|\hat{\Sigma}_{\mathbf{fp}}\right\|_2 = O\left(\sqrt{(t-t_0)}\tfrac{log(log(N)}{N}\right)\end{array} \tag{A.58}$$

Now we write the inner projection in (A.57) as follows

$$\begin{array}{rcl}\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}, U_{[t+h,T]}\right] & = & \hat{E}\left[Y_{t+h} \mid P, F\right] \\ & = & \hat{E}_{\|P}\left[Y_{t+h} \mid F\right] + \hat{E}_{\|F}\left[Y_{t+h} \mid P\right]\end{array} \tag{A.59}$$

Recall now that, given matrices $\Sigma_1(N), \Delta\Sigma_1(N), \Sigma_2(N), \Delta\Sigma_2(N)$ of appropriate dimensions[16], with $\|\Sigma_1(N)\|_2 = O(1)$, $\Sigma_2(N)$ a.s. invertible with bounded inverse $\left\|\Sigma_2^{-1}(N)\right\|_2 = O(1)$, and $\|\Delta\Sigma_1(N)\|_2$, $\|\Delta\Sigma_2(N)\|_2$ infinitesimal (a.s.) as $N \to \infty$,

$$\begin{array}{rcl}\left(\Sigma_1(N) + \Delta\Sigma_1(N)\right)\left(\Sigma_2(N) + \Delta\Sigma_2(N)\right)^{-1} & = & \Sigma_1(N)\Sigma_2^{-1}(N) + \Delta\Sigma_1(N)\Sigma_2^{-1}(N) + \\ & & -\Sigma_1(N)\Sigma_2^{-1}(N)\Delta\Sigma_2(N)\Sigma_2^{-1}(N) + \\ & & +\underline{o}(\|\Delta\Sigma_2(N)\|_2)\end{array} \tag{A.60}$$

holds.

Now we apply (A.60) to the oblique projection:

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

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 $\Delta\Sigma_2(N) := -\hat{\Sigma}_{\mathbf{pf}}\hat{\Sigma}_{\mathbf{ff}}^{-1}\hat{\Sigma}_{\mathbf{fp}}$.

[16]The dimensions may also be a function of N.

Observe that the assumption $\Phi_{\mathbf{z}} \geq cI > 0$, implies (uniformly in $t - t_0$, see [26], [22]) that $\Sigma_{\mathbf{pp}} \geq cI > 0$ and therefore $\left\|\Sigma_{\mathbf{pp}}^{-1}\right\|_2 \leq \sqrt{1/c} < \infty$. From the uniform convergence of sample covariances (Theorem 5.3.2 in [27]), $\left\|\hat{\Sigma}_{\mathbf{pp}} - \Sigma_{\mathbf{pp}}\right\|_2 = O\left((t-t_0)\sqrt{\frac{log(log(N))}{N}}\right)$ holds and therefore $\hat{\Sigma}_{\mathbf{pp}}$ is a.s. invertible and $\left\|\hat{\Sigma}_{\mathbf{pp}}^{-1}\right\|_2 = O(1)$. With similar argument, which uses the fact that the covariance $E\left[\mathbf{y}(t)\mathbf{z}^\top(t-\tau)\right]$ goes to zero exponentially as a function of the 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

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

Using (A.58), $\left\|\hat{\Sigma}_{\mathbf{pp}}^{-1}\right\|_2 = O(1)$, $\left\|\hat{\Sigma}_{\mathbf{yp}}\right\|_2 = O(1)$ 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 $\left\|\tilde{\Psi}\right\|_2 = O\left((t-t_0)\frac{log(log(N))}{N}\right)$. Since the elements of P are $O_P(1)$, it follows that the product $\tilde{\Psi}P = \underline{O}_P\left((t-t_0)^2\frac{log(log(N))}{N}\right) = \underline{o}_P(1/\sqrt{N})$.

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

$$\hat{E}_{\|F}\left[Y_{t+h} \mid P\right] \doteq \hat{\Sigma}_{\mathbf{yp}}\hat{\Sigma}_{\mathbf{pp}}^{-1}P = \hat{E}\left[Y_{t+h} \mid P\right] \tag{A.61}$$

Similarly one can show that

$$\hat{E}_{\|P}\left[Y_{t+h} \mid F\right] \doteq (\hat{\Sigma}_{\mathbf{yf}} - \hat{\Sigma}_{\mathbf{yp}}\hat{\Sigma}_{\mathbf{pp}}^{-1}\hat{\Sigma}_{\mathbf{pf}})\hat{\Sigma}_{\mathbf{ff}}^{-1}F \tag{A.62}$$

Using (A.61) and (A.62) we obtain

$$\begin{aligned}\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}, U_{[t+h,T]}\right] &\doteq \hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}\right] + \\ &\quad + \hat{\Theta}U_{[t+h,T]}\end{aligned}$$

for a suitable matrix $\hat{\Theta}$ which follows from (A.62). Next, observe that $\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}\right]$ can be written in the form

$$\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}\right] = \hat{Y}_{t+h}^p + \sum_{i=1}^{h} \hat{\Phi}_{hi}Y_{t+h-i} + \hat{\Psi}_{hi}U_{t+h-i} \tag{A.63}$$

for suitable matrix coefficients $\hat{\Psi}_{hi}$, $\hat{\Phi}_{hi}$. Taking now the oblique projection $\hat{E}_{\|U_{[t,T]}}\left[\cdot \mid Z_{[t_0,t)}\right]$ of both sides of (A.63) we obtain:

$$\begin{aligned}\hat{Y}_{t+h} &= \hat{E}_{\|U_{[t,T]}}\left[Y_{t+h} \mid Z_{[t_0,t)}\right] \\ &\doteq \hat{E}_{\|U_{[t,T]}}\left[\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}\right] \mid Z_{[t_0,t)}\right] \\ &= \hat{Y}_{t+h}^p + \textstyle\sum_{i=1}^{h} \hat{\Phi}_{hi}\hat{Y}_{t+h-i}\end{aligned} \tag{A.64}$$

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

In matrix form this becomes:

$$\hat{Y}^p_{[t,T]} \dot{=} \begin{bmatrix} I & 0 & \dots & 0 \\ -\hat{\Phi}_{11} & I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\Phi}_{\bar{\nu},\bar{\nu}} & -\hat{\Phi}_{\bar{\nu},\bar{\nu}-1} & \dots & I \end{bmatrix} \hat{Y}_{[t,T]}$$

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 2.1, 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 (29), it is a standard fact (see for instance [43], [27], [38]) that $\Phi_{hi} = C\bar{A}^{i-1}K$.

The fact that

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

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

Proof of Proposition 5.1. We prove now only the relation (39) regarding $\tilde{C}_N^{P_{opt}}$, being those related to $\tilde{A}_N^{P_{opt}}$, $\tilde{B}_N^{P_{opt}}$, $\tilde{K}_N^{P_{opt}}$ completely analogous.

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

Recall now that

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

and that the state estimator $\hat{X}_t^{P_{opt}}$, from (37), is of the form

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

Using also the definition of $\hat{E}_t^{P_{opt}}$ in (38), and the fact that $CX_t \dot{=} \Xi_0 Z_{[t_0,t)}$, it follows that

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

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

$$\tilde{C}_N^{P_{opt}} \dot{=} \left(\tilde{\Xi}_0^{P_{opt}} - C \bar{\Gamma}_{\bar{\nu}}^{-L} \begin{bmatrix} \tilde{\Xi}_0^{P_{opt}} \\ \vdots \\ \tilde{\Xi}_{\bar{\nu}}^{P_{opt}} \end{bmatrix} \right) Z_{[t_0,t)} \left(\hat{X}_t \right)^\top \left[\hat{X}_t \hat{X}_t \right]^{-1} \tag{A.65}$$

Vectorizing equation (A.65) and substituting sample values with their $a.s.$ limit, there exists a matrix M_C, so that

$$\mathrm{vec}\{\tilde{C}_N^{P_{opt}}\} \dot{=} M_C \tilde{\Omega}^{P_{opt}}. \tag{A.66}$$

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 the relation (48) regarding $\tilde{C}_N^{CCA}$, being those related to $\tilde{A}_N^{CCA}$, $\tilde{B}_N^{CCA}$, $\tilde{K}_N^{CCA}$ completely analogous.

First of all recall that the state estimator $\hat{X}_t^{CCA}$

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

where the last (asymptotic) equality follows from (44).

We also recall that, by definition of T_N^{CCA} in (47) and the relation $\Gamma_{\bar{\nu}} = H_{\bar{\nu}}\bar{\Gamma}_{\bar{\nu}} = W_{CCA}W^{-1}\bar{\Gamma}_{\bar{\nu}}$,

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

Let us define

$$\hat{E}_t := Y_t - E\left[Y_t | Z_{[t_0,t)}\right] := Y_t - \hat{\Xi}_0 Z_{[t_0,t)}; \tag{A.67}$$

the equation above defines also $\hat{\Xi}_0$ which is, in general, different from $\hat{\Xi}_0^{CCA}$. Note also that $\hat{E}_t$ is different from $\hat{E}_t^{CCA}$ defined in (46). It follows that $Y_t = CX_t + E_t \dot{=} \Xi_0 Z_{[t_0,t)} + E_t$ can also be written in the form

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

Using the relation $\bar{\Gamma}_{\bar{\nu}} = H_{\bar{\nu}}^{-1}\Gamma_{\bar{\nu}}$ it is easy to prove that $\bar{\Gamma}_{\bar{\nu}}^{-L} W = \Gamma_{\bar{\nu}}^{-L} W_{CCA}$ and therefore $\hat{\Gamma}_{\bar{\nu}}^{-L} W_{CCA}$ 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 $\hat{X}_t^{CCA}$, recalling that T_N^{CCA} converges to the identity matrix, it follows that

$$\tilde{C}_N^{CCA} \dot{=} \left(\tilde{\Xi}_0 - C\bar{\Gamma}_{\bar{\nu}}^{-L} \begin{bmatrix} \tilde{\Xi}_0^{CCA} \\ \vdots \\ \tilde{\Xi}_{\bar{\nu}}^{CCA} \end{bmatrix}\right) Z_{[t_0,t)} \hat{X}_t^{\top} \left[\hat{X}_t \hat{X}_t\right]^{-1} \tag{A.68}$$

Using $\tilde{\Omega}^{P_{CCA}} = \tilde{\Omega}^{P_{opt}} + (\hat{\Omega}^{P_{CCA}} - \hat{\Omega}^{P_{opt}})$ and $\tilde{\Xi}_0 = \tilde{\Xi}_0^{P_{opt}} + (\hat{\Xi}_0 - \hat{\Xi}_0^{P_{opt}})$ it is clear that, vectorizing equation (A.68) and substituting sample values with their $a.s.$ limit, there exist matrices M_C, M_{C1}^{CCA} and M_{C2}^{CCA} so that

$$\mathrm{vec}\{\tilde{C}_N^{CCA}\} \dot{=} M_C \tilde{\Omega}^{P_{opt}} + M_{C1}^{CCA}(\hat{\Omega}^{P_{CCA}} - \hat{\Omega}^{P_{opt}}) + M_{C2}^{CCA}\mathrm{vec}(\hat{\Xi}_0 - \hat{\Xi}_0^{P_{opt}})$$

where clearly, from (A.68), M_C is the same as that in (A.66), which concludes the proof. As before M_{C1}^{CCA} and M_{C2}^{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 (33), which we rewrite here for convenience,

$$Y = S^P\Omega^P + E + \Delta_Y(N) = S^P\Omega^P + LE_I + \Delta_Y(N) \tag{B.69}$$

where $\Delta_Y(N) = \underline{o}_P(1/\sqrt{N})$; recall that the noise covariance $\mathrm{Var}\{E\} = R = L(I \otimes \Lambda)L^\top$, where L has been defined in (36), is a singular matrix. The $\underline{o}_P(1/\sqrt{N})$ has the form

$$\Delta_Y(N) := \left[\mathrm{vec}\left(C\bar{A}^{t-t_0}X_{t_0}\right)^\top, \mathrm{vec}\left(C\bar{A}^{t-t_0+1}X_{t_0}\right)^\top, ..., \mathrm{vec}\left(C\bar{A}^{t-t_0+\bar{\nu}}X_{t_0}\right)^\top\right]^\top. \tag{B.70}$$

The following Lemma shows that this term has a very special structure which allows to discard it when studying the asymptotic distribution of the estimated parameters $\eta_N^\top\hat{\Omega}^P := \eta_N^\top F Y$.

Lemma 7.1: The term $\Delta_Y(N) = \underline{o}_P(1/\sqrt{N})$ in (B.69), (B.70) can be written as

$$\Delta_Y(N) = S^P\Upsilon o(1/\sqrt{N}) + L\Delta_X(N) \tag{B.71}$$

with $\|\Upsilon\|_2 = O(1)$ and $\gamma_N^\top\Delta_X(N) = o_P(1/\sqrt{N})$ provided the column vector γ_N (of suitable dimensions) satisfies $\|\gamma_N\|_2 = O(1)$.

Proof: First of all note that $\bar{A}^k X_{t_0} = X_{t_0+k} - \sum_{i=1}^k \bar{A}^{i-1}KZ_{t_0+k-i}$. Therefore

$$C\bar{A}^{t-t_0+k}X_{t_0} = C\bar{A}^{t-t_0}X_{t_0+k} + C\bar{A}^{t-t_0}\Upsilon_k Z_{[t_0,t+k)}$$

for some matrix of coefficients Υ_k which satisfies $\|\Upsilon_k\|_2 = O(1)$ from the exponential decrease of the terms $\bar{A}^i$.

Therefore we decompose the vectorization in (B.70) in the two parts:

$$\left[\mathrm{vec}\left(C\bar{A}^{t-t_0}X_{t_0}\right)^\top, \mathrm{vec}\left(C\bar{A}^{t-t_0}X_{t_0+1}\right)^\top, ..., \mathrm{vec}\left(C\bar{A}^{t-t_0}X_{t_0+\bar{\nu}}\right)^\top\right]^\top \tag{B.72}$$

and

$$\left[\mathrm{vec}\left(C\bar{A}^{t-t_0}\Upsilon_0 Z_{[t_0,t)}\right)^\top, \mathrm{vec}\left(C\bar{A}^{t-t_0}\Upsilon_1 Z_{[t_0,t+1)}\right)^\top, ..., \mathrm{vec}\left(C\bar{A}^{t-t_0}\Upsilon_{\bar{\nu}} Z_{[t_0,t+\bar{\nu})}\right)^\top\right]^\top. \tag{B.73}$$

It is rather straightforward then to see that (B.72) can be written as

$$L\begin{bmatrix} C\bar{A}^{t-t_0}x_{t_0} \\ C\bar{A}^{t-t_0}x_{t_0+1} \\ \vdots \\ C\bar{A}^{t-t_0}x_{t_0+N+\bar{\nu}-1} \end{bmatrix} = L\Delta_X(N) \tag{B.74}$$

while (B.73) as

$$S^P \begin{bmatrix} \mathrm{vec}\left(C\bar{A}^{t-t_0}\Upsilon_0\right) \\ \mathrm{vec}\left(C\bar{A}^{t-t_0}\Upsilon_1\right) \\ \vdots \\ \mathrm{vec}\left(C\bar{A}^{t-t_0}\Upsilon_{\bar{\nu}}\right) \end{bmatrix} = S^P\Upsilon o(1/\sqrt{N}) \tag{B.75}$$

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

Note also that, due to Assumption 2.2, the covariance matrix $\Lambda_{xx} := \mathrm{Var}\left\{[x_{t_0}^\top, ..., x_{t_0+N+\bar{\nu}-1}^\top]^\top\right\}$ satisfies $\|\Lambda_{xx}\|_2 = O(1)$; therefore, since

$$\begin{aligned} \|\mathrm{Var}\left\{\Delta_X(N)\right\}\|_2 &= \left\|\mathrm{Var}\left\{\left[\ (C\bar{A}^{t-t_0}x_{t_0})^\top,\ \ \ldots,\ \ (C\bar{A}^{t-t_0}x_{t_0+N+\bar{\nu}-1})^\top\ \right]^\top\right\}\right\|_2 \\ &= o(1/\sqrt{N})\|\Lambda_{xx}\|_2 = o(1/\sqrt{N}) \end{aligned}$$

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

Using (B.70) the original least squares problem (B.69) can be written as

$$Y = S^P\left(\Omega^P + \Upsilon o(1/\sqrt{N})\right) + L\left(E_I + \Delta_X(N)\right). \tag{B.76}$$

The noise term $L\left(E_I + \Delta_X(N)\right)$ has a covariance which can be written in the form $R_o = L\left[(I\otimes\Lambda) + o(1/\sqrt{N})\tilde{\Sigma}\right]L$; using the structure of E_I and $\Delta_X(N)$, (B.74) and Assumption 2.2, it follows also that $\|\tilde{\Sigma}\|_2 = O(1)$.

We are now ready to derive the optimal (asymptotically BLUE) of $\eta_N^\top\Omega^P$. It is easy to see (using for instance, formula (C4.3.3) in [43]) that the asymptotically BLUE is given by[17] $\eta_N^\top\hat{\Omega}^{P_{opt}} := \eta_N^\top F_{opt}^{R_o}Y$ with

$$F_{opt}^{R_o} = \left[(S^P)^\top\left(R_o + S^P(S^P)^\top\right)^\dagger S^P\right]^{-1}(S^P)^\top\left(R_o + S^P(S^P)^\top\right)^\dagger. \tag{B.77}$$

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

$$F_{opt} := \left[(S^P)^\top\left(R + S^P(S^P)^\top\right)^\dagger S^P\right]^{-1}(S^P)^\top\left(R + S^P(S^P)^\top\right)^\dagger \tag{B.78}$$

which is the Markov estimator computed as if there was no $\underline{o}_P(1/\sqrt{N})$ terms in (B.69), 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).

[17] † denotes the Moore-Penroose pseudoinverse.

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 7.1, $\eta_N^\top F L \Delta_X(N) = \gamma_N^\top \Delta_X(N) = o_P(1/\sqrt{N})$, which can therefore be neglected. As we shall see at the end of this Appendix the estimators we are interested in (namely $\hat{\Omega}^{P_{opt}} = F_{opt} Y$ and $\hat{\Omega}^{P_{CCA}} = F_{CCA} Y$) satisfy the even stronger condition $\|F_{opt}\|_2 = O(1/\sqrt{N})$ and $\|F_{CCA}\|_2 = O(1/\sqrt{N})$; therefore $\eta_N^\top \tilde{\Omega}^P \doteq \eta_N^\top F E$, providing a proof that, indeed, the $o_P(1/\sqrt{N})$ in (B.69) can be discarded both to the purpose of design and analysis of the estimator.

Remark B.5 As an indirect proof that $F_{opt}^{R_o}$ and F_{opt} give asymptotically equivalent estimators note the following. With an argument similar to that used at the end of this Appendix to show that $\|F_{opt}\|_2 = O(1/\sqrt{N})$, also $\|F_{opt}^{R_o}\|_2 = O(1/\sqrt{N})$ can be proved. Therefore, $\eta_N^\top (F_{opt}^{R_o} Y - \Omega^P) \doteq \eta_N^\top F_{opt}^{R_o} E$ and $\eta_N^\top (F_{opt} Y - \Omega^P) \doteq \eta_N^\top F_{opt}^{R} E$, 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 $\mathrm{AsVar}\{\sqrt{N}\eta_N^\top (F_{opt}^{R_o} Y - \Omega^P)\} \leq \mathrm{AsVar}\{\sqrt{N}\eta_N^\top (F_{opt} Y - \Omega^P)\}$ and $\mathrm{AsVar}\{\sqrt{N}\eta_N^\top (F_{opt} Y - \Omega^P)\} \leq \mathrm{AsVar}\{\sqrt{N}\eta_N^\top (F_{opt}^{R_o} Y - \Omega^P)\}$ hold, proving that, indeed $\mathrm{AsVar}\{\sqrt{N}\eta_N^\top (F_{opt}^{R_o} Y - \Omega^P)\} = \mathrm{AsVar}\{\sqrt{N}\eta_N^\top (F_{opt} Y - \Omega^P)\}$. $\Diamond$

Consider now the SVD

$$L = \begin{bmatrix} U_L & U_{L^\perp} \end{bmatrix} \begin{bmatrix} S_L & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_L^\top \\ V_{L^\perp}^\top \end{bmatrix}$$

We recall that $L^\dagger = V_L S_L^{-1} U_L^\top$; note also that $U_{L^\perp}$ spans the left kernel of L, i.e. $U_{L^\perp}^\top L = 0$. Then define

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

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

Finally, observe that $Y = L[y_t^\top, y_{t+1}^\top, ..., y_{T+N-1}^\top]^\top = LY_I$, 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 $U_{L^\perp}^\top L = 0$ has been used. Therefore (B.69) can be converted into a least squares problem with equality constraints:

$$\begin{cases} Y_L &= S_L^P \Omega^P + E_L \\ s.t. & 0 = S_{L^\perp}^P \Omega^P \end{cases} \tag{B.79}$$

Let d_Ω denote the number of parameters in Ω^P. The matrix $S_{L^\perp}^P$ has dimension $(N-1)\bar{\nu}m \times d_\Omega$. 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[18] $r_S^\perp \leq d_\Omega$. 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 $\bar{S}_{L^\perp}^P$; note in fact that $0 = \bar{S}_{L^\perp}^P \Omega^P$ implies also that $0 = v^\top \Omega^P$ for any (row) vector $v^\top$ which is in the row span of $\bar{S}_{L^\perp}^P$. Therefore we rewrite the constrained least squares problem (B.79) as:

$$\begin{cases} Y_L &= S_L^P \Omega^P + E_L \\ s.t. & 0 = \bar{S}_{L^\perp}^P \Omega^P \end{cases} \tag{B.80}$$

Consider now the QR-decomposition [24]

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

with $\bar{R}_1$ square $(r_S^\perp \times r_S^\perp)$ invertible. It is easy to show that, $\forall \eta_N$ satisfying Assumption 2.3, the asymptotically best linear unbiased estimator (Asymptotically BLUE) of $\eta_N^\top \Omega^P$ is given by[19] $\eta_N^\top \hat{\Omega}^{P_{opt}}$ where

$$\hat{\Omega}^{P_{opt}} = Q_2 \left(S_L^P Q_2 \right)^\dagger Y_L \tag{B.81}$$

where $(\cdot)^\dagger$ denotes Moore-Penrose pseudoinverse [24]. Observe that, being Q_2 orthonormal, the conditioning of the problem depends on the matrix $S_L^P Q_2$, i.e. to S_L^P "restricted to" the

[18]Using the structure of the matrices L and S^P it is possible to check that, indeed, $d_\Omega - r_S^\perp = m(t - t_0)(m + p) > 0$. For reasons of space it is not possible to report the details here, see [12].

[19]See, e.g., [43] Remark 3, page 70, for the case in which Ω^P has dimensions not depending on N.

orthogonal complement of the "constraint matrix" $\bar{S}^P_{L^\perp}$. Note that[20] the matrix S^P_L could be collinear (or almost collinear); however from

$$\left[\begin{array}{c} U^\top_{L^\perp} \\ (I\otimes\Lambda)^{-1/2}L^\dagger \end{array}\right] S^P \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right] = \left[\begin{array}{c} S^P_{L^\perp} \\ S^P_L \end{array}\right] \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right] = \left[\begin{array}{cc} R_1^\top & 0 \\ S^P_L Q_1 & S^P_L Q_2 \end{array}\right], \quad \text{(B.82)}$$

in (B.82) we have used the fact that the rows of $S^P_{L^\perp}$ are linear combinations of the rows of $\bar{S}^P_{L^\perp}$, so that $S^P_{L^\perp} Q^\top = [R_1^\top \; 0]$.

It is immediate to see that the condition number of $S^P_L Q_2$ is smaller than that of

$$\left[\begin{array}{c} U^\top_{L^\perp} \\ (I\otimes\Lambda)^{-1/2}L^\dagger \end{array}\right] S^P = \left[\begin{array}{cc} I & 0 \\ 0 & (I\otimes\Lambda)^{-1/2}V_L^\top S_L^{-1} \end{array}\right] \left[\begin{array}{c} U^\top_{L^\perp} \\ U_L^\top \end{array}\right] S^P, \quad \text{(B.83)}$$

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

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

$$F_{opt} = Q_2 \left(S^P_L Q_2\right)^\dagger (I\otimes\Lambda)^{-1/2} V_L S_L^{-1} U_L^\top. \quad \text{(B.84)}$$

We now need to verify that indeed $\|F_{opt}\|_2 = O(1/\sqrt{N})$ holds true. From (B.84) it follows that

$$\|F_{opt}\|_2 \leq \left\|\left(S^P_L Q_2\right)^\dagger\right\|_2 \left\|\Lambda^{-1/2}\right\|_2 \left\|S_L^{-1}\right\|_2. \quad \text{(B.85)}$$

First note that $\left\|\Lambda^{-1/2}\right\|_2 = O(1)$ since Λ is the (estimated) noise covariance. The fact that $\left\|S_L^{-1}\right\|_2 = O(1)$ follows directly from the structure of the matrix L.

Using now equations (B.82) and (B.83), $\sigma_{min}\left(S^P_L Q_2\right) \geq \min\left(1, 1/\sqrt{\|\Lambda\|_2}\right) \sigma_{min}\left(S^P\right)$; since $\|\Lambda\|_2 = O(1)$.

Using now (34) it is immediate to see that $\frac{(S^P)^\top (S^P)}{N}$ is a block diagonal matrix with block diagonal elements $\frac{Z_{[t_0,t+k]} Z^\top_{[t_0,t+k]}}{N} \otimes I$; from Assumption 2.2 and the uniform convergence of sample covariances (see [26] and [27]), $\sigma^{-1}_{min}\left[\frac{(S^P)^\top S^P}{N}\right]^{-1} = O(1)$ which implies $\sigma^{-1}_{min}\left[S^P\right] = O(1/\sqrt{N})$.

It follows that $\left\|\left(S^P_L Q_2\right)^\dagger\right\|_2 = \sigma^{-1}_{min}\left(S^P_L Q_2\right) = O(1/\sqrt{N})$, which, inserted in (B.85) gives the desired result $\|F_{opt}\|_2 = O(1/\sqrt{N})$. With a similar calculation we could also show that in $\hat{\Omega}^{P_{CCA}} := F_{CCA} Y$, $\|F_{CCA}\|_2 = O(1/\sqrt{N})$.

[20] As an anonymous reviewer has pointed out.

REFERENCES

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

[2] ——, "Canonical correlation analysis of time series and the use of an information criterion," in *System Identification: Advances and Case Studies*, R. Mehra and D. Lainiotis, Eds. New York: Academic Press, 1976, pp. 27–96.

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

[4] ——, "Asymptotic properties of subspace estimators," *Automatica*, vol. 41, pp. 359–376, 2005.

[5] ——, "Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs," *Journal of Time Series Analysis*, 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. Academic, 1976, pp. 349–405.

[9] A. Chiuso, "On the relation between CCA and predictor-based subspace identification," in *Proc. of the 44rd IEEE Conf. on Dec. and Control*, Sevilla, Spain, 2005.

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

[11] ——, "Asymptotic variance of closed-loop subspace identification algorithms," *IEEE Trans. on Aut. Control*, vol. 51, no. 8, pp. 1299–1314, 2006.

[12] ——, "On the role of Vector AutoRegressive modeling in subspace identification," in *Proc. of CDC 2006 (to appear)*, San Diego (USA), Dec. 2006.

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

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

[15] ——, "Numerical conditioning and asymptotic variance of subspace estimates," *Automatica*, vol. 40, no. 4, pp. 677–683, 2004.

[16] ——, "On the ill-conditioning of subspace identification with inputs," *Automatica*, vol. 40, no. 4, pp. 575–589, 2004.

[17] ——, "Consistency analysis of some closed-loop subspace identification methods," *Automatica*, vol. 41, no. 3, pp. 377–391, 2005.

[18] ——, "Prediction error vs. subspace methods in closed-loop identification," in *Proc. of the 16th IFAC World Congress*, Prague, July 2005.

[19] N. Chui and J. Maciejowski, "Criteria for informative experiments for subspace identification," *Intl. Journal of Control*, vol. 78, no. 5, pp. 326–344, 2005.

[20] U. Desai and D. Pal, "A realization approach to stochastic model reduction," *IEEE Transactions Automatic Control*, vol. 29, pp. 1097–1100, 1984.

[21] T. Ferguson, *A Course in Large Sample Theory*. Chapman and Hall, 1996.

[22] H. Gazzah, P. Regalia, and J. Delmas, "Asymptotic eigenvalue distribution of block Toeplitz matrices and application to blind SIMO channel identification," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1243–1251, March 2001.

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

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

[25] C. Granger, "Economic processes involving feedback," *Information and Control*, vol. 6, pp. 28–48, 1963.

[26] U. Grenander and G. Szegö, *Toeplitz Forms and their Applications*. Chelsea, 1958.

[27] E. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*. Wiley, 1988.

[28] E. Hannan and D. Poskitt, "Unit canonical correlations between future and past," *The Annals of Statistics*, vol. 16, pp. 784–790, 1988.

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

[30] M. Jansson, "Asymptotic variance analysis of subspace identification methods," in *Proceedings of SYSID2000*, S. Barbara Ca., 2000.

[31] ——, "Subspace identification and ARX modeling," in *Proceedings of SYSID 2003*, Rotterdam, 2003.

[32] 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.

[33] W. Larimore, "System identification, reduced-order filtering and modeling via canonical variate analysis," in *Proc. American Control Conference*, 1983, pp. 445–451.

[34] ——, "Canonical variate analysis in identification, filtering, and adaptive control," in *Proc. 29th IEEE Conf. Decision & Control*, Honolulu, 1990, pp. 596–604.

[35] ——, "Large sample efficiency for ADAPTX subspace identification with unknown feedback," in *Proc. of IFAC DYCOPS'04*, Boston, MA, USA, 2004.

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

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

[38] L. Ljung, *System Identification; Theory for the User*. Prentice Hall, 1997.

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

[40] S. Qin and L. Ljung, "Closed-loop subspace identification with innovation estimation," in *Proceedings of SYSID 2003*, Rotterdam, 2003.

[41] ——, "Parallel QR implementation of subspace identification with parsimonious models," in *Proceedings of SYSID 2003*, Rotterdam, 2003.

[42] 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.

[43] T. Söderström and P. Stoica, *System Identification*. Prentice-Hall, 1989.

[44] P. Van Overschee and B. De Moor, "Subspace algorithms for the stochastic identification problem," *Automatica*, vol. 29, pp. 649–660, 1993.

[45] ——, "N4SID: Subspace algorithms for the identification of combined deterministic– stochastic systems," *Automatica*, vol. 30, pp. 75–93, 1994.

[46] ——, "A unifying theorem for three subspace system identification algorithms," *Automatica*, vol. 31, no. 12, pp. 1853–1864, 1995.

[47] 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.

[48] 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 and its Applications*, vol. 237/238, no. 1–3, pp. 359–393, 1996, special issue honoring Calyampudi Radhakrishna Rao.