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

Alessandro Chiuso

***Abstract*— There is experimental evidence that a recently proposed subspace algorithm based on predictor identification 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 known that time series identification using the Canonical Correlation Analysis (CCA) approach is asymptotically efficient. Asymptotic optimality of CCA has also been proved when measured inputs are white. In this paper we study the relation between the standard CCA approach and the recently proposed subspace procedure based on predictor identification (PBSID[1] from now on).

In this paper we work under the assumption that there is no feedback; 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.

Using some recent results on the asymptotic variance we report counter-examples showing that PBSID is not efficiency in general when measured inputs are not white.

I. Introduction

A certain number of subspace algorithms have been developed during the last two decades. For time series identification the algorithm developed by Van Overschee and De Moor [27] is known to provide asymptotically[2] efficient estimators [2]. Sometimes this algorithm goes under the name of CCA (or CVA) to remind that the state construction is performed using Canonical Correlation Analysis. In the presence of measured inputs the situation is different. The most widely known procedures go under the acronyms N4SID [28], CCA [23], [24] and MOESP [30]. Recently several researchers have studied the asymptotic statistical properties of these algorithms [1], [21], [4], [5], [3], [9], [11] and compared, to some extent, existing procedures [5], [3], [10]. Also optimality of the CCA method when measured inputs are white has been established in [5]. The situation is not clear when inputs are not white. The interested reader is referred to the paper [5].

It is our opinion, as has already been stressed in [14], 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, do not work when data are collected in closed loop. Very recently two subspace procedures have been introduced by Qin and Ljung [26], and Jansson [22] which, to some extent, are able to deal with feedback. The recent work [13] studies the statistical consistency of these two algorithms. In [13] also a "geometrical" version of the algorithm proposed by Jansson [22] 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 [14] for an explanation of this terminology.

Experimental evidence shows that the behavior of this algorithm cannot be distinguished to any practical purpose from PEM in a number of simple examples, see the simulations reported in [22], [13]. Using some recently derived formulas (see [6], [7]) 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 classical approaches is worth studying. To the best of our knowledge this relation has not been investigated yet.

In this paper we work under the assumption that *no feedback is present*. The main contribution is to show that PBSID is asymptotically equivalent to CCA in the time series case and also when measured inputs are white. 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; in the white input case these "unnecessary input data" are (asymptotically) uncorrelated with past input and output and present output and therefore do not influence the statistical properties.

This is, we believe, an important step in understanding "predictor based" subspace identification; our result implies that also the PBSID algorithm is asymptotically optimal for time series identification and for identification with white exogenous inputs. We stress that our proof is not based on the asymptotic variance expressions but rather on the analysis of the state estimation step on which PBSID and CCA are based.

The question regarding optimality in more general cases remains open of course; simulation results and computations based on the asymptotic variance expressions (see Section V) suggest that PBSID is in general not efficient for colored input. This with a notable exception: we have observed in a number of examples with ARX systems that the asymptotic variance is indistinguishable from the Cramér Rao lower

This work has been supported by MIUR

Alessandro Chiuso is with the Dipartimento di Ingegneria dell'Informazione, Università di Padova Via Gradenigo 6/A, 35131 Padova, Italy chiuso@dei.unipd.it

[1]Short for "predictor-based subspace identification".

[2]Both in the number of data and "past" and "future" horizons.

bound also for colored inputs and also when feedback is present.[3] Whether or not this depends on the particular example or can be generalized will be subject of future research.

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 main result. Section V contains some simulation results and in Section VI we report some conclusions and discussion on future work. Part of the proof is 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 p and m respectively, which are representable as the output and input signals of a linear stochastic system in innovation form

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

we assume that there is *no feedback* from $\{\mathbf{y}(t)\}$ to $\{\mathbf{u}(t)\}$ [18], [16]. Without loss of generality we shall assume that the dimension n of the state vector $\mathbf{x}(t)$ is as small as possible, i.e. the representation (II.1) is minimal. For simplicity we assume that $D = 0$, i.e. there is no direct feedthrough[4] from $\mathbf{u}$ to $\mathbf{y}$. 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 $|\rho|$ is 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 as the one step ahead 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. The equality sign $=$ will have to be understood, when random variables are involved, as *almost sure* equality while $\doteq$ shall denote equality in probability up to $o(1/\sqrt{N})$ terms, which we shall call *asymptotic equivalence*. In fact, from standard results in asymptotic analysis (see for instance [15]) terms which are $o(1/\sqrt{N})$ can be neglected when studying the asymptotic statistical properties. We shall also use the same symbol when the difference in the equated terms produces nonsingular change of basis $\hat{T}_N$ (up to $o(1/\sqrt{N})$ 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 $\mathbf{x}_1 - \hat{T}_N \mathbf{x}_2 = o(1/\sqrt{N})$ in probability.

Our aim is to identify the system parameters (A, B, C, K), or equivalently the transfer functions $F(z) = C(zI - A)^{-1}B$ and $G(z) = C(sI - A)^{-1}K + I$, starting from input-output data $\{y_s, u_s\}$, $s \in [t_0, T+N]$, generated by the system (II.1). This setup also encompasses time series identification (i.e. no measured inputs) provided one lets $B = 0$ in (II.1).

In this paper we shall have to deal with random fluctuations due to finite sample length (e.g. approximating expectations with finite time averages, etc.). 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 [25], [12]) that the Hilbert spaces of second order stationary random variables and the Hilbert space of semi-infinite tails containing sample values of a (second order) stationary stochastic 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 $Y_t := [y_t \;\; y_{t+1}, \;\; \dots \;\; y_{t+N-1}]$, $U_t := [y_t \; y_{t+1}, \; \dots \; y_{t+N-1}]$ and $Z_t := [Y_t^\top \; U_t^\top]^\top$. These are the block rows of the usual *data Hankel matrices* which appear in subspace identification.

Remind 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 [5] where the following assumption is made:

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

$$\begin{array}{ll} t - t_0 \geq \frac{log N^{-d/2}}{log|\rho|} & 1 < d < \infty \\ t - t_0 = o\left(log(N)^\alpha\right) & \alpha < \infty \end{array} \tag{II.2}$$

Under this assumption the effect of terms due to mishandling of the initial condition at time t_0 are $o(1/\sqrt{N})$ and therefore can be neglected. Moreover, (II.2) ensures that, when regressing onto past data and taking the limit as N goes to infinity, the computation of sample covariance matrices of increasing size (with $t - t_0$) does not pose any complication in the sense that their limit is well defined and equal to the population counterpart (see the discussion after Lemma 4 in [5]).

[3] We remind that something similar *does not* happen with CCA.

[4] This assumption can be removed in our situation but is useful when there is feedback, see [18], [16], [13]. Since the "predictor based" algorithm is designed to work without assumptions on the feedback structure we prefer to keep $D = 0$ also here.

[5] This is the parameter j in the notation of Van Overschee and De Moor [28] i.e. the number of columns in the Hankel data matrices used in subspace identification.

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

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

$$\mathcal{U}_{[t_0,\,t)} \quad := \quad \overline{\operatorname{span}}\,\{\mathbf{u}_k(s);\; k=1,\dots,p,\; t_0 \leq s < t\,\}$$

the bar denotes closure in mean square, i.e. in the metric defined by the inner product $\langle\,\xi,\eta\,\rangle := E\{\xi\eta\}$ where $E\{\cdot\}$ denote mathematical expectation. Similarly we define $\mathcal{Y}_{[t_0,\,t)}$. 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. We shall use $\nu := T - t$.

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 [17], [12]).

We adopt the notation $\Sigma_{\mathbf{ab}} := E\left[\mathbf{a}\mathbf{b}^\top\right]$ 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 spanned by the vector $\mathcal{C} := \operatorname{span}\{\mathbf{c}\}$ will be given by the usual formula (with $\Sigma_{\mathbf{cc}}$)

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

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

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

For column vectors formed by stacking past and/or future random variables (or semi-infinite Hankel matrices) we shall use the following notation: $\mathbf{y}_{[t,s]} := \left[\begin{array}{cccc} \mathbf{y}^\top(t) & \mathbf{y}^\top(t+1) & \dots & \mathbf{y}^\top(s) \end{array}\right]^\top$. We shall also use the shorthand $\mathbf{u}^+ := \mathbf{u}_{[t,T]}$.

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

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 $B_t := [b_t, b_{t+1}.., b_{t+N-1}]$ we shall define

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

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

Similarly, given a third sequence (say $C_t := [c_t, c_{t+1}, .., c_{N-1}]$), $\hat{\Sigma}_{\mathbf{ab}|\mathbf{c}}$ is defined as $\hat{\Sigma}_{\mathbf{ab}|\mathbf{c}} := \hat{\Sigma}_{\mathbf{ab}} - \hat{\Sigma}_{\mathbf{ac}}\hat{\Sigma}_{\mathbf{cc}}^{-1}\hat{\Sigma}_{\mathbf{cb}}$. 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 $Z_{[t_0,t)}$ [17]. As above, the oblique projection can be computed using the formula:

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

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

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

$$\mathcal{Z}_{[t_0,\,T]} = \mathcal{Z}_{[t_0,\,t)} + \mathcal{Z}_{[t,\,T]}, \qquad t_0 \leq t < T \tag{II.5}$$

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 Φ_z should have no zeros on the unit circle [20].

Whenever necessary we shall assume that (II.5) holds also for finite sequences, i.e. that $Z_{[t_0,T]}$ is of full row rank.

For future reference we also define the extended observability matrices

$$\Gamma_k := \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{k-1} \end{bmatrix}, \quad \bar{\Gamma}_k := \begin{bmatrix} C \\ C\bar{A} \\ C\bar{A}^2 \\ \vdots \\ C\bar{A}^{k-1} \end{bmatrix} \tag{II.6}$$

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

$$H_k = \begin{bmatrix} I & 0 & \dots & 0 \\ CK & I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{k-2}K & CA^{k-3}K & \dots & I \end{bmatrix}. \tag{II.7}$$

III. State Space Construction

It is well known [27], [28], [25], [12] 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 (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

$$\left\{ \begin{array}{rcl} \hat{X}_{t+1} & \simeq & A\hat{X}_t + B\hat{U}_t + KE_t \\ Y_t & \simeq & C\hat{X}_t + E_t \end{array} \right. \tag{III.1}$$

in the least squares sense

Different subspace algorithms have different implementations of the first step while the second remains the same for virtually al algorithms[7]. 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.

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 [27], [28], [25]. The situation is different here since t_0 is let going to $-\infty$ according to Assumption 2.1 $\diamondsuit$

In this Section we shall review the state construction step for the CCA algorithm [23], [29], [5] and for the PBSID algorithm [22], [13].

A. CCA Algorithm

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

$$\begin{aligned} \hat{Y}_{[t,T]} &= \hat{E}_{\|U_{[t,T]}}\left[Y_{[t,T]} \mid Z_{[t_0,t)}\right] \\ &\simeq \Gamma_\nu X_t. \end{aligned} \tag{III.2}$$

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

$$\hat{\mathbf{y}}_{[t,T]} = E_{\|\mathcal{U}_{[t,T]}}\left[\mathbf{y}_{[t,T]} \mid \mathcal{Z}_t^-\right] = \Gamma_\nu \mathbf{x}(t) \tag{III.3}$$

holds. The matrix $\hat{Y}_{[t,T]}$ has in general full row rank.

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

$$\begin{aligned} W^{-1}\hat{Y}_{[t,T]} &= USV^\top \\ &= [U_n \tilde{U}_n]\begin{bmatrix} S_n & 0 \\ 0 & \tilde{S}_n \end{bmatrix}\begin{bmatrix} V_n^\top & \tilde{V}_n^\top \end{bmatrix} \end{aligned} \tag{III.4}$$

The CCA algorithm corresponds to the choice[8] $W := \Sigma^{1/2}_{\mathbf{y}^+\mathbf{y}^+|\mathbf{u}^+}$. An estimate of the observability matrix is obtained discarding the "less significant" singular values (i.e. pretending $\tilde{S}_n \simeq 0$) from

$$\hat{\Gamma}_\nu = WU_nS_n^{1/2}$$

and consequently a basis for the state space given by:

$$\hat{X}_t^{CCA} := S_n^{-1/2}U_n^\top W^{-1}\hat{Y}_{[t,T]} \tag{III.5}$$

Remark III.2 We remind that all weighting matrices are in practice data dependent. However, for the purpose of asymptotic analysis, data dependent weights can be substituted with their (a.s.) limit. Therefore, to streamline notation, we prefer to work directly with the population version of all weights. $\diamondsuit$

[7]In this paper we shall not be concerned with algorithms based on the so-called "shift invariance" method.

[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 [5], [9].

We quote now a result first appeared in [5] which shows that the CCA weight $W = \Sigma^{1/2}_{\mathbf{y}^+\mathbf{y}^+|\mathbf{u}^+}$ can be substituted with $[H_\nu(I\otimes\Lambda)H_\nu^\top]^{1/2}$ without changing the asymptotic properties:

Lemma 3.1: (*Bauer Ljung* [5]) Assume the parameters are estimated following steps 1) and 2) above and the state is constructed according to (III.5). Then any choice $W = [H_\nu(I\otimes\Lambda)H_\nu^\top + \Gamma_\nu\Sigma\Gamma_\nu^\top]^{1/2}$ with $\Sigma = \Sigma^\top \geq 0$, provides the same asymptotic accuracy of the estimates of any system invariant.

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 shall denote with $\hat{X}_t^{CCA}$ any state sequence resulting from a choice of W of the form $W = [H_\nu(I\otimes\Lambda)H_\nu^\top + \Gamma_\nu\Sigma\Gamma_\nu^\top]^{1/2}$. Lemma 3.1 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. $\diamondsuit$

B. PBSID algorithm

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[9]

$$\begin{aligned} \hat{Y}^p_{t+h} &:= \hat{E}_{\|Z_{[t,t+h)}}\left[Y_{t+h} \mid Z_{[t_0,t)}\right] \\ &\simeq C\bar{A}^{h-1}X_t \\ &\quad h = 0,1,..,\nu. \end{aligned} \tag{III.6}$$

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

$$\begin{gathered} \hat{\mathbf{y}}^p(t+h) := E_{\|\mathcal{Z}_{[t,t+h)}}\left[\mathbf{y}(t+h) \mid \mathcal{Z}_t^-\right] = C\bar{A}^{h-1}\mathbf{x}(t) \\ h = 0,1,..,\nu \end{gathered} \tag{III.7}$$

holds. Then one stacks all the predictors

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

From the Singular Value Decomposition

$$W_p^{-1}\hat{Y}^p_{[t,T]} = PDQ^\top = [P_n\tilde{P}_n]\begin{bmatrix} D_n & 0 \\ 0 & \tilde{D}_n \end{bmatrix}\begin{bmatrix} Q_n^\top & \tilde{Q}_n^\top \end{bmatrix} \tag{III.8}$$

where W_p is a weighting matrix which will be chosen appropriately, an estimate of the observability matrix $\bar{\Gamma}_\nu$ is obtained discarding the "less significant" singular values (i.e. pretending $\tilde{D}_n \simeq 0$) from

$$\hat{\bar{\Gamma}}_\nu = W_pP_nD_n^{1/2}$$

and consequently a basis for the state space

$$\hat{X}_t^{PBSID} := D_n^{-1/2}P_n^\top W_p^{-1}\hat{Y}^p_{[t,T]} \tag{III.9}$$

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

IV. Main Result

Our purpose in this section is to study the link between the state constructions (III.5) and (III.9). We first state the main result of the paper and then proceed with a derivation of the result.

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_p is chosen according to $W_p = I \otimes \Lambda^{1/2}$, the state constructions in (III.5) and (III.9) yield asymptotically the same accuracy as far as estimation of any system invariant is concerned i.e.:

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

Remark IV.4 Using state sequences which cannot be distinguished up to $o(1/\sqrt{N})$ terms guarantees that the estimated system matrices obtained from step 2) above and consequently any system invariant share the same asymptotic properties. $\diamondsuit$

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:

$$\hat{Y}_{t+h} \dot{=} \hat{Y}_{t+h}^p + \sum_{i=1}^{h} C\bar{A}^{i-1} K \hat{Y}_{t+h-i} \qquad \text{(IV.1)}$$

which can be written in compact form as

$$\hat{Y}_{[t,T]}^p \dot{=} \begin{bmatrix} I & 0 & \dots & 0 \\ -CK & I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -C\bar{A}^{\nu-2}K & -C\bar{A}^{\nu-3}K & \dots & I \end{bmatrix} \hat{Y}_{[t,T]} \qquad \text{(IV.2)}$$

The block Toeplitz matrix which appears in formula (IV.2) is the inverse of H_ν defined above, i.e.

$$\hat{Y}_{[t,T]}^p \dot{=} H_\nu^{-1} \hat{Y}_{[t,T]}. \qquad \text{(IV.3)}$$

This lemma shows that $\hat{Y}_{[t,T]}^p$ is asymptotically equivalent to a weighted version of $\hat{Y}_{[t,T]}$

Proof: The proof can be found in the Appendix ■

Proof of Theorem 4.1.

We recall from (III.8) that in the predictor-based algorithm one takes SVD of $W_p^{-1}\hat{Y}_{[t,T]}^p$ while, from (III.4), $W^{-1}\hat{Y}_{[t,T]}$ is used in the CCA algorithm.

Note that, the CCA algorithm corresponds to the choice

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

where Λ is the variance of the innovation.

However, by letting $\Sigma = 0$ in Lemma 3.1, $W = (H_\nu (I \otimes \Lambda) H_\nu^\top)^{1/2}$ provides the same asymptotic behavior.

If we now pre-multiply both sides of (IV.3) in Lemma 4.2 by $W_p^{-1/2} = (I \otimes \Lambda)^{-1/2}$ we obtain that

$$W_p^{-1/2} \hat{Y}_{[t,T]}^p \dot{=} (I \otimes \Lambda)^{-1/2} H_\nu^{-1} \hat{Y}_{[t,T]}. \qquad \text{(IV.4)}$$

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 $\hat{X}^{CCA}(t)$ and $\hat{X}^{PBSID}(t)$ are asymptotically equivalent, which concludes the proof. $\square$

V. 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

$$A_1 = 0.5 \quad C_1 = 1 \quad B_1 = 1 \quad C_1 = 1 \quad K_1 = 0.5 \quad D_1 = 0$$

while the second is a first order ARMAX model

$$A_2 = 0.5 \quad C_2 = 1 \quad B_2 = 1 \quad C_2 = 1 \quad K_2 = 1 \quad D_2 = 0$$

The input is either unit variance white noise or unit variance white noise passed through the filter with state space realization:

$$A_u = \begin{bmatrix} 0 & 1 \\ -0.9 & 0.5 \end{bmatrix} \quad B_u = \begin{bmatrix} 1.3 \\ .3 \end{bmatrix}$$
$$C_u = \begin{bmatrix} 1 & 0 \end{bmatrix} \quad D_u = 1$$

We report results concerning the asymptotic variance (sample variance estimated over 500 Monte Carlo runs multiplied by the number $N = 1000$ of data point used in each experiment) of the deterministic transfer functions $F_i(z) = C_i(zI - A_i)^{-1}B_i,\ i = 1, 2$. The past horizon has been chosen to be 5 and 10 respectively for Example 1 and Example 2. As a reviewer mentioned, this might be a critical point. Let us just mention that the length of the past horizon for the equivalence to hold in practice depends on ρ. The closer ρ to 1, the larger $t - t_0$ needs to be. Investigations concerning such choice are postponed to future publications for reasons of space.

Future horizon has been chosen equal to 5 in both examples.

Note that for the white input case both CCA and the predictor based algorithm are indistinguishable from PEM as predicted by the theory. The algorithm by Jansson [22] is indistinguishable from PBSID. The asymptotic variance of PBSID (computed using the formulas of [6] and estimated from the simulation) is indistinguishable from the Cramér Rao lower bound also for colored inputs when the system is ARX.

In the colored input case 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 [6], [7] and the Cramér Rao Lower Bound (CRLB).

VI. Conclusion

In this paper we have shown that the PBSID algorithm introduced in [13], which may be seen as a “geometrical” version of the algorithm in [22], 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 [6], [7].

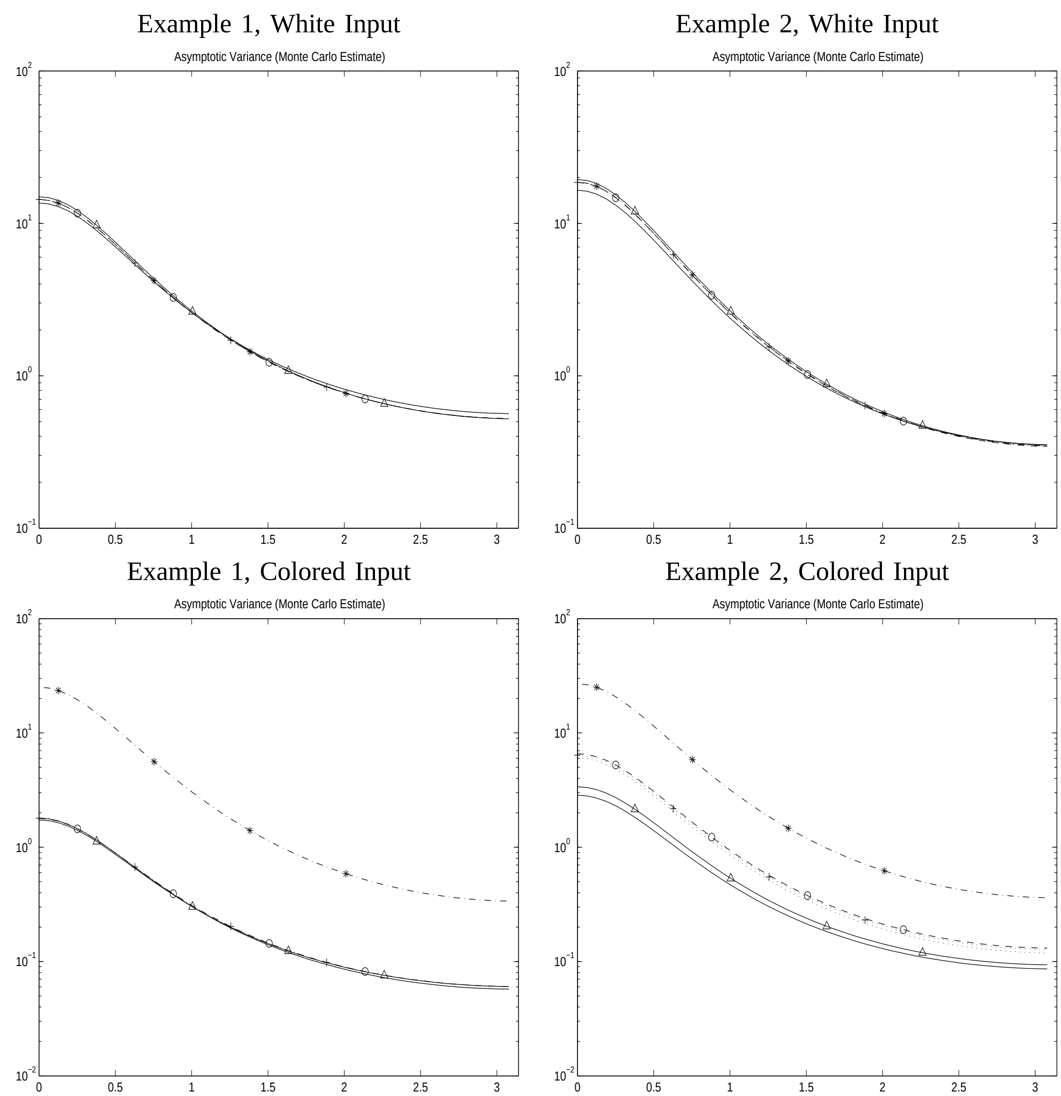

Fig. 1. Asymptotic Variance (Monte Carlo estimate) vs. normalized frequency ($\omega \in [0, \pi]$) Solid with triangles ($\triangle$) PEM, dashed-dotted with stars ($*$): CCA, dotted with crosses "predictor-based" algorithm (PBSID), dashed with circles (o): Jansson's algortihm, dotted: asymptotic variance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

APPENDIX: PROOFS

Proof of Lemma 4.2. First let us note that

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

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

$$\begin{aligned}\|\hat{\Sigma}_{\mathbf{fp}}\| &= O\left((t-t_0)\sqrt{\frac{log(log\, N)}{N}}\right)\\ \|\hat{\Sigma}_{\mathbf{yf}}\| &= O\left((t-t_0)\sqrt{\frac{log(log\, N)}{N}}\right)\end{aligned}$$

which implies

$$\begin{aligned}\|\hat{\Sigma}_{\mathbf{fp}}\|\|\hat{\Sigma}_{\mathbf{fp}}\| &= o(1/\sqrt{N})\\ \|\hat{\Sigma}_{\mathbf{yf}}\|\|\hat{\Sigma}_{\mathbf{fp}}\| &= o(1/\sqrt{N})\end{aligned} \tag{A.1}$$

Recall that

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

which using (A.1) becomes, for the purpose of asymptotic analysis

$$\hat{E}_{\|F}\left[Y_{t+h} \mid P\right] \doteq \hat{\Sigma}_{\mathbf{yp}}\hat{\Sigma}_{\mathbf{pp}}^{-1}P \tag{A.2}$$

Similarly one can show that

$$\hat{E}_{\|P}\left[Y_{t+h} \mid F\right] \doteq (\hat{\Sigma}_{\mathbf{yf}} - \hat{\Sigma}_{\mathbf{yp}}\hat{\Sigma}_{\mathbf{pp}}^{-1}\hat{\Sigma}_{\mathbf{pf}})\hat{\Sigma}_{\mathbf{ff}}^{-1}F \tag{A.3}$$

Using (A.2) and (A.3) we obtain

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

for a suitable matrix $\hat{\Theta}$ which follows from (A.3). Next, observe that $\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}\right]$ can be written in the form

$$\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}\right] = \hat{Y}^p_{t+h} + \sum_{i=1}^{h} \hat{\Phi}_i Y_{t+h-i} + \hat{\Psi}_i U_{t+h-i} \tag{A.4}$$

for suitable matrix coefficients $\hat{\Psi}_i$, $\hat{\Phi}_i$. Taking now the oblique projection $\hat{E}_{\|U_{[t,T]}}\left[\cdot \mid Z_{[t_0,t)}\right]$ of both sides of (A.4) we obtain:

$$\begin{aligned} \hat{Y}_{t+h} &= \hat{E}_{\|U_{[t,T]}}\left[Y_{t+h} \mid Z_{[t_0,t)}\right] \\ &\doteq \hat{E}_{\|U_{[t,T]}}\left[\hat{E}\left[Y_{t+h} \mid Z_{[t_0,t+h)}\right] \mid Z_{[t_0,t)}\right] \\ &= \hat{Y}^p_{t+h} + \textstyle\sum_{i=1}^{h} \hat{\Phi}_i \hat{Y}_{t+h-i} \end{aligned} \tag{A.5}$$

where $\hat{E}_{\|U_{[t,T]}}\left[\hat{\Theta}U_{[t+h,T]} \mid Z_{[t_0,t)}\right] = 0$ has been used. Once again, the sample values of $\hat{\Phi}_i$ can be substituted[10] with their (a.s.) limit Φ_i without changing the asymptotic properties, i.e.

$$\hat{Y}_{t+h} \doteq \hat{Y}^p_{t+h} + \sum_{i=1}^{h} \Phi_i \hat{Y}_{t+h-i} \tag{A.6}$$

The only step left is to prove that $\Phi_i = C\bar{A}^{i-1}K$. In order to do so, we look at the regression problem with infinite data and recall that convergence holds under Assumption 2.1. Writing the output in predictor form

$$\begin{aligned} \mathbf{y}(t+h) &= C\bar{A}^{h-1}\mathbf{x}(t) + \textstyle\sum_{i=1}^{h} C\bar{A}^{i-1}K\mathbf{y}(t+h-i) + \\ &\quad + \textstyle\sum_{i=1}^{h} C\bar{A}^{i-1}B\mathbf{u}(t+h-i) + \mathbf{e}(t+h) \end{aligned}$$

and projecting onto[11] $\mathcal{Z}_t^-$ we obtain

$$\begin{aligned} E\left[\mathbf{y}(t+h) \mid \mathcal{Z}_t^-\right] &= C\bar{A}^{h-1}\mathbf{x}(t) + \\ &\quad + \textstyle\sum_{i=1}^{h} C\bar{A}^{i-1}K\mathbf{y}(t+h-i) + \\ &\quad + \textstyle\sum_{i=1}^{h} C\bar{A}^{i-1}B\mathbf{u}(t+h-i) \\ &= \hat{\mathbf{y}}^p(t+h) + \textstyle\sum_{i=1}^{h} \Phi_i \mathbf{y}(t+h-i) + \\ &\quad + \textstyle\sum_{i=1}^{h} \Psi_i \mathbf{u}(t+h-i) \end{aligned}$$

Note that from (III.7) $C\bar{A}^{h-1}\mathbf{x}(t) = \hat{\mathbf{y}}^p(t+h)$; using Assumption II.5 the projection admits a unique representation as a function of $\mathbf{y}(s)$, $\mathbf{u}(s)$ and therefore, in particular, $\Phi_i = C\bar{A}^{i-1}K$. □

REFERENCES

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

[2] ——, "Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs," *Journal of Time Series Analysis (submitted)*, 2002.

[10] This substitution is a delicate matter. We refer the reader to the paper [8] for the details which unfortunately we cannot report here for reasons of space.

[11] Recall that also $t_0 \to -\infty$.

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

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

[5] D. Bauer and L. Ljung, "Some facts about the choice of the weighting matrices in Larimore type of subspace algorithm," *Automatica*, vol. 36, pp. 763–773, 2001.

[6] A. Chiuso, "Asymptotic variance of a certain closed-loop subspace identification method," in *Proc. of the 43rd IEEE Conf. on Dec. and Control*, Nassau, Bahamas, 2004.

[7] ——, "Asymptotic variance of closed-loop identification algorithms," *Submitted to IEEE Trans. on Aut. Control*, 2004, available at http://www.dei.unipd.it/∼chiuso.

[8] ——, "On the relation between CCA and predictor based subspace identification," *Submitted to IEEE Trans. on Aut. Control*, 2005, available at http://www.dei.unipd.it/∼chiuso.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[29] ——, "Choice of state-space basis in combied deterministic-stochastic subspace identification," *Automatica*, vol. 31, no. 12, pp. 1877–1883, 1995.

[30] 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.