

The Role of Vector AutoRegressive Modeling in Subspace Identification

Alessandro Chiuso

Abstract—Some recent subspace procedures make use, directly or indirectly, of Vector AutoRegressive with eXogenous inputs (VARX) models in a preliminary step. This was first noticed for the CCA method; more recently it has also been proved that the first oblique projection step of a subspace algorithm based on predictor identification (PBSID) is asymptotically equivalent to the SSARX algorithm by Jansson which performs a preliminary VARX modeling step.

For the purpose of comparison with more classical methods like CCA a recent work have introduced also an “optimized” version of PBSID.

In this paper we shall show that indeed also this latter “optimized” PBSID is equivalent to estimating a long VARX model followed by the “classical” steps of subspace identification. This latter step can be seen as a sort of model reduction. Besides the theoretical interest, we shall argue that this may have also important implications regarding computational complexity.

Index Terms—Identification, Statistical Analysis, Subspace Methods

I. INTRODUCTION

Subspace identification has attracted a lot of attention in the last two decades. It is also fair to say that the last few years have witnessed a renewed interest in this topic for essentially two reasons: first the introduction of new methods which have allowed subspace identification to be applied with closed loop data [8], [26], [20], [12] and second a whole body of results on the asymptotic properties of subspace methods which have allowed, on the one side, to assess accuracy of subspace estimators [2], [9], [6], [19] and on the other to compare different methods [2], [10], [4], [5].

Some analysis regarding the role of Vector AutoRegressive (VAR) models in subspace identification was performed in [14] where it was shown that the CCA algorithm introduced in [21] is asymptotically equivalent (in the sense of having the same asymptotic distribution of the estimators) to a procedure which first estimates a long VAR model and then does balanced model reduction.

Extension of subspace algorithms to closed loop operating conditions have required, at a certain stage, the introduction of two step procedures (see [20], [28], [22]) which were needed to eliminate undesired terms due to feedback. This was due to the lack of stochastic realization procedures indicating how the state space could be constructed in the presence of feedback. A brief overview of these realization procedures can be found in [12] and references therein.

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

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

We shall see that the preliminary step based on Vector AutoRegressive with eXogenous inputs (VARX) models, originally taken from [24], [20] is actually present, in a way or another, in a whole class of algorithms.

In fact, it has been recently proved in [5] that the preliminary VARX modeling step in the SSARX algorithm by Jansson [20] is asymptotically equivalent to the first step of a subspace algorithm based on predictor identification¹ (PBSID) (see [12], [13], [4]). In this work we extend this comparison² to the “optimized” version of PBSID which has been introduced in the recent work [4] for the purpose of comparison with more classical methods like CCA [21], [30]. This “optimized” PBSID method delivers consistent estimators even when there is feedback and, at the same time, it compares favorably (in the sense of asymptotic variance) with the standard CCA for any choice of input signal; see the paper [4], Theorem 5.3, for details.

We recall that CCA is known to be asymptotically efficient when there are no inputs (time series identification) and also to be “optimal” when inputs are white (see [2] for a nice survey), making it a quite natural benchmark against which other subspace procedures should be compared.

We shall see that the “optimally-weighted” projection step involved in the “optimized” PBSID in [4] is actually equivalent to estimating a single VARX model. The results of this paper also imply that the “optimized” PBSID can be implemented with a much lower computational complexity since the “optimally weighted” regression step can be substituted with estimation of a suitable VARX model of given length.

After this VARX model has been estimated, the subsequent steps can be seen as a sort of model reduction. It is interesting to observe that this second step has a fundamental role. In fact, it was shown in [14] that, for time series identification (i.e. no inputs) VAR modeling followed by balanced model reduction is asymptotically equivalent to the CCA method, which is asymptotically efficient (see [2] and references therein). It has also been shown in [4] that PBSID (and therefore its “optimized” version) is asymptotically equivalent to CCA for time-series identification and when input signals are white.

Hence, at least for white inputs and time series identification PBSID “does model reduction right”. The situation is

¹This algorithm was introduced in [12] under the name “whitening filter algorithm”.

²Even though some preliminary results along these lines have already been presented in [5], the author would like to thank an anonymous reviewer of the paper [4] which have underlined the relevance of the comparison performed in this paper; part of the merit of this paper should also go to him.

different when there are inputs, and they are colored. The “optimized” PBSID performs better than CCA but it is not efficient in general (see also the simulation results in Section V). Whether this lack of efficiency³ should be attributed to the first step, which is equivalent to estimating a VARX model as mentioned above and proved in Theorem 4.1, or to the second “model reduction” step, is not clear at the moment. We believe this is worth investigating.

It should also be stressed that all the statistical results in this paper and, to the best of our knowledge, available in the literature as of today, are of the asymptotic type, i.e. holds for large samples. Experimental evidence shows that indeed there might be significant differences for finite samples. Investigating (and optimizing) the finite sample behavior is certainly one of the main (and unfortunately rather difficult) aspects open in subspace identification.

It is the author opinion that the theoretical background of stochastic realization complemented with the results found in this paper and in [14], [2], [4], [5], could eventually turn out to be very helpful in the future development of subspace algorithms and for their analysis.

The structure of the paper is as follows. In Section II we state the problem and set up notation; Section III briefly recalls the algorithmic steps while Section IV states the main result of this paper. We continue with some experimental results in Section V while Section VI contains some conclusions.

II. STATEMENT OF THE PROBLEM AND NOTATION

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

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

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

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

³To the best of the author’s knowledge there is no efficient subspace procedure for general input signal. Even though efficiency is claimed in [22], there is evidence (see [4]) that this claim is not correct.

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

The symbol \doteq shall denote equality in probability 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 [15]) terms which are $o_P(1/\sqrt{N})$ can be neglected when studying the asymptotic statistical properties. We shall use the notation $\underline{o}_P(\cdot)$ to denote a random vector whose components are $o_P(\cdot)$.

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

The analysis reported in this paper requires that both N , the length of the finite tails⁴ and the past horizon $t - t_0$ ⁵ go to infinity. We remind the reader that $t - t_0$ has to go to infinity at a certain rate depending on the number N of data available. Details can be found, for instance, in [2] where the following assumption is made:

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

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

Under this assumption the effect of terms due to mishandling of the initial condition at time t_0 are $o(1/\sqrt{N})$ and therefore can be neglected. Moreover, (II.2) ensures that, when regressing onto past data and taking the limit as N goes to infinity, the computation of sample covariance matrices of increasing size (with $t - t_0$) does not pose any complication in the sense that their limit is well defined and equal to the population counterpart (see the discussion after Lemma 4 in [3]).

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

We shall use capitals to denote the tail of length N . For instance $Y_t := [y_t \ y_{t+1} \ \dots \ y_{t+N-1}]$, and $Z_t := [Y_t^\top \ U_t^\top]^\top$. These are the block rows of the usual *block Hankel data matrices* which appear in subspace identification. When using tails of length (say M), different from N , we shall explicitly append a superscript, e.g. $Y_t^M := [y_t \ y_{t+1} \ \dots \ y_{t+M-1}]$.

For $-\infty \leq t_0 \leq t \leq T \leq +\infty$ we define the Hilbert space of scalar zero-mean random variables

$$\mathcal{U}_{[\tau, t]} := \overline{\text{span}}\{\mathbf{u}_k(s); k = 1, \dots, m, \tau \leq s < t\}$$

where the bar denotes closure in mean square, i.e. in the metric defined by the inner product $\langle \xi, \eta \rangle := \mathbb{E}\{\xi\eta\}$, the operator \mathbb{E} denoting mathematical expectation. Similar

⁴This is the parameter j in the notation of Van Overschee and De Moor [31] i.e. the number of columns in the Hankel data matrices used in subspace identification.

⁵The number of block rows in the Hankel data matrix containing the past data.

definitions hold for $\mathcal{Y}_{[\tau, t]}$ and $\mathcal{Z}_{[\tau, t]}$, the symbol \vee denoting closed vector sum.

When $\tau = -\infty$ we shall use the shorthands \mathcal{U}_t^- , \mathcal{Y}_t^- for $\mathcal{U}_{[-\infty, t]}$, $\mathcal{Y}_{[-\infty, t]}$, and $\mathcal{Z}_t^- := \mathcal{U}_t^- \vee \mathcal{Y}_t^-$. The spaces generated by $\mathbf{u}(s)$ and $\mathbf{y}(s)$, $-\infty < s < \infty$ shall be denoted with the symbols \mathcal{U} , \mathcal{Y} , respectively. For convenience of notation we denote with $\nu := T - t$ the future horizon.

Given a subspace $\mathcal{C} \subseteq \mathcal{U} \vee \mathcal{Y}$, we shall denote with $E[\mathbf{a} | \mathcal{C}]$ the orthogonal projection of the random variable \mathbf{a} onto \mathcal{C} .

Using the notation $\Sigma_{\mathbf{ab}} := \mathbb{E}[\mathbf{ab}^\top]$ for the covariance matrix between the random vectors⁶ \mathbf{a} and \mathbf{b} , in the finite dimensional case $E[\mathbf{a} | \mathcal{C}]$ will be given by the usual formula ($\Sigma_{\mathbf{cc}}$ invertible) $E[\mathbf{a} | \mathcal{C}] = \Sigma_{\mathbf{ac}} \Sigma_{\mathbf{cc}}^{-1} \mathbf{c}$.

Defining also 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 $\Sigma_{\mathbf{ab} | \mathcal{C}}$ will denote projection error covariance (conditional covariance in the Gaussian case) $\Sigma_{\mathbf{ab} | \mathcal{C}} := \Sigma_{\tilde{\mathbf{a}}\tilde{\mathbf{b}}} = \Sigma_{\mathbf{ab}} - \Sigma_{\mathbf{ac}} \Sigma_{\mathbf{cc}}^{-1} \Sigma_{\mathbf{cb}}$. Given two non-intersecting subspaces $\mathcal{A} \subseteq \mathcal{U} \vee \mathcal{Y}$, $\mathcal{B} \subseteq \mathcal{U} \vee \mathcal{Y}$, $\mathcal{A} \cap \mathcal{B} = \{0\}$, $E_{\|\mathcal{B}}[\cdot | \mathcal{A}]$ shall denote the oblique projection onto \mathcal{A} along \mathcal{B} (see [16]) and can be computed by the formula: $E_{\|\mathcal{B}}[\mathbf{a} | \mathcal{C}] = \Sigma_{\mathbf{ac} | \mathcal{B}} \Sigma_{\mathbf{cc} | \mathcal{B}}^{-1} \mathbf{c}$.

For column vectors formed by stacking past and/or future random variables we shall use the notation: $\mathbf{y}_{[t, s]} := [\mathbf{y}^\top(t) \ \mathbf{y}^\top(t+1) \ \dots \ \mathbf{y}^\top(s)]^\top$. Finite block Hankel data matrices will be denoted using capitals, i.e. $Y_{[t, s]} := [Y_t^\top \ Y_{t+1}^\top \ \dots \ Y_s^\top]^\top$

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

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

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

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

As above, given a matrix C_t , we define the projection errors $\tilde{A}_t := A_t - \hat{E}[A_t | C_t]$ and $\tilde{B}_t := B_t - \hat{E}[B_t | C_t]$. The sample covariance (conditional sample covariance) of the projection errors is denoted with the symbol $\hat{\Sigma}_{\mathbf{ab} | \mathcal{C}} := \hat{\Sigma}_{\tilde{\mathbf{a}}\tilde{\mathbf{b}}}$ and computed by the formula $\hat{\Sigma}_{\mathbf{ab} | \mathcal{C}} := \hat{\Sigma}_{\mathbf{ab}} - \hat{\Sigma}_{\mathbf{ac}} \hat{\Sigma}_{\mathbf{cc}}^{-1} \hat{\Sigma}_{\mathbf{cb}}$.

We shall denote with $\hat{E}_{\|\mathcal{U}_{[t, T]}}[\cdot | Z_{[t_0, t]}]$ the oblique projection along the space generated by the rows of future inputs $\mathcal{U}_{[t, T]}$ onto the space generated by the rows of the joint past

$Z_{[t_0, t]}$. As above, the oblique projection can be computed as $\hat{E}_{\|\mathcal{B}_t}[A_t | C_t] = \hat{\Sigma}_{\mathbf{ac} | \mathcal{B}} \hat{\Sigma}_{\mathbf{cc} | \mathcal{B}}^{-1} C_t$.

For future reference we also define the extended observability matrix

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

III. ALGORITHMS

It is well known [31], [11] that identification using subspace methods can be seen as a two step procedure as follows:

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

$$\begin{cases} \hat{X}_{t+1} \simeq A \hat{X}_t + B \hat{U}_t + K E_t \\ Y_t \simeq C \hat{X}_t + E_t \end{cases} \quad (\text{III.1})$$

in the least squares sense

Different subspace algorithms have different implementations of the first step while the second remains the same for virtually all algorithms⁷. 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.

A. PBSID algorithm

The construction of the state space using the PBSID algorithm, introduced in [12] under the name “whitening filter”, involves several oblique projections. The projection of each (block) row Y_{t+h} , $0 = 1, \dots, \nu$, can be seen as a long ARX model as follows

$$\begin{aligned} \hat{Y}_{t+h} &:= \hat{E}[Y_{t+h} | Z_{[t_0, t+h]}] = \\ &= \hat{\Psi}_{h,1} Z_{t+h-1} + \dots + \hat{\Psi}_{h, t+h-t_0} Z_{t_0} \end{aligned} \quad (\text{III.2})$$

from which the oblique projections⁸

$$\begin{aligned} \hat{Y}_{t+h}^P &:= \hat{E}_{\|\mathcal{Z}_{[t, t+h]}}[Y_{t+h} | Z_{[t_0, t]}] = \\ &= \sum_{i=h+1}^{t-t_0+h} \hat{\Psi}_{hi} Z_{t+h-i} \simeq C \bar{A}^{h-1} X_t \end{aligned} \quad (\text{III.3})$$

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

$$\hat{Y}^P(t+h) := E_{\|\mathcal{Z}_{[t, t+h]}}[\mathbf{y}(t+h) | \mathcal{Z}_t^-] = C \bar{A}^{h-1} \mathbf{x}(t) \quad (\text{III.4})$$

holds. Then one stacks all the predictors

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

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

⁸The superscript P reminds that the quantity has to do with the “predictor-based” algorithm.

⁶Zero mean.

From the Singular Value Decomposition

$$W^{-1}\hat{Y}_{[t,T-1]}^P = PDQ^\top = [P_n \tilde{P}_n] \begin{bmatrix} D_n & 0 \\ 0 & \tilde{D}_n \end{bmatrix} \begin{bmatrix} Q_n^\top \\ \tilde{Q}_n^\top \end{bmatrix}, \quad (\text{III.5})$$

where W is a weighting matrix which can 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 $\hat{\Gamma}_\nu = WP_n D_n^{1/2}$ and consequently a basis for the state space

$$\hat{X}_t^{PBSID} := D_n^{-1/2} P_n^\top W^{-1} \hat{Y}_{[t,T-1]}^P. \quad (\text{III.6})$$

Similarly one constructs a basis \hat{X}_{t+1}^{PBSID} for the state at time $t+1$ and an estimate of the innovation sequence $\hat{E}_t := Y_t - E[\hat{Y}_t^P | \hat{X}_t^{PBSID}]$.

B. “Optimized” PBSID Algorithm

The optimized version of PBSID introduced in [4] differs from the original PBSID algorithm in the computation of the predictors (III.2); in fact in the optimized algorithm the estimation of the predictors \hat{Y}_{t+h} is formulated as a weighted least squares problem as described in this Section.

Let us first recall that

$$\begin{aligned} Y_{t+h} &= C\bar{A}^h X_t + E_{t+h} + \\ &\quad + \sum_{i=1}^h C\bar{A}^{i-1} (KY_{t+h-i} + BU_{t+h-i}) \\ &= C\bar{A}^h \mathcal{K} Z_{[t_0,t]} + \\ &\quad + \sum_{i=1}^h C\bar{A}^{i-1} (KY_{t+h-i} + BU_{t+h-i}) + \\ &\quad + E_{t+h} + o_P(1/\sqrt{N}) \\ &:= \Xi_h Z_{[t_0,t]} + \sum_{i=1}^h \Psi_{hi} Z_{t+h-i} + \varrho_P(1/\sqrt{N}) \end{aligned} \quad (\text{III.7})$$

where the last equality defines the matrices Ξ_h and Ψ_{hi} . Stacking the data and using (III.7) (discarding $o_P(1/\sqrt{N})$ terms) we obtain:

$$\begin{bmatrix} Y_t \\ Y_{t+1} \\ \vdots \\ Y_T \end{bmatrix} = \begin{bmatrix} \Xi_0 \\ \Xi_1 \\ \vdots \\ \Xi_\nu \end{bmatrix} Z_{[t_0,t]} + \begin{bmatrix} 0 & 0 & \dots & 0 \\ \Psi_{11} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{\nu\nu} & \dots & \Psi_{\nu 1} & 0 \end{bmatrix} Z_{[t,T]} + \begin{bmatrix} E_t \\ E_{t+1} \\ \vdots \\ E_T \end{bmatrix} \quad (\text{III.8})$$

Observe that the lower triangular matrices in (III.8) are Toeplitz, since $\Psi_{ij} = C\bar{A}^{j-1}[K \ B]$, $\forall i, j$. The projection in (III.2) is equivalent to solving (III.8) “row by row”; hence the Toeplitz structure is not preserved after estimation, i.e. $\hat{\Psi}_{ij} \neq \hat{\Psi}_{i'j}$, almost surely when $i \neq i'$.

This is equivalent to solving the least squares problem obtained vectorizing (III.8):

$$Y := \begin{bmatrix} \text{vec}(Y_t) \\ \text{vec}(Y_{t+1}) \\ \vdots \\ \text{vec}(Y_T) \end{bmatrix} = S^P \Omega^P + \begin{bmatrix} \text{vec}(E_t) \\ \text{vec}(E_{t+1}) \\ \vdots \\ \text{vec}(E_T) \end{bmatrix} = S^P \Omega^P + E \quad (\text{III.9})$$

where the matrix S^P has the form

$$S^P = \begin{bmatrix} (Z_{[t_0,t]}^\top \otimes I) & 0 & \dots & 0 \\ 0 & (Z_{[t_0,t+1]}^\top \otimes I) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (Z_{[t_0,T]}^\top \otimes I) \end{bmatrix} \quad (\text{III.10})$$

and Ω^P is given by

$$\Omega^P = \begin{bmatrix} \text{vec}^\top(\Xi_0) & \text{vec}^\top(\Xi_1) & \text{vec}^\top(\Psi_{11}) & \dots \\ \dots & \text{vec}^\top(\Xi_\nu) & \dots & \text{vec}^\top(\Psi_{\nu 1}) \end{bmatrix}^\top. \quad (\text{III.11})$$

Finding an “optimal” solution $\hat{\Omega}^{P_{opt}}$ (Markov estimator) of

$$Y = S^P \Omega^P + E, \quad (\text{III.12})$$

where $o_P(1/\sqrt{N})$ terms have been neglected, gives an estimator $\hat{\Omega}^{P_{opt}}$ of Ω^P which has the smallest asymptotic variance among all linear (asymptotically unbiased) estimators based on (III.9). Incidentally, this has allowed to show in [4] that this “optimized” version yields, asymptotically, a lower variance of the estimators of any system invariant as compared to the standard PBSID and, more importantly, to the classical CCA algorithm [21], [30].

To this purpose it is very useful to observe that the “noise term” E can be written in the form

$$E^\top = [e_t^\top \quad e_{t+1}^\top \quad \dots \quad e_{t+N-1}^\top \quad \dots \quad e_{t+N-1}^\top] L^\top \quad (\text{III.13})$$

where L is a “selection”⁹ matrix of size $pN\nu \times p(\nu + N)$. We refer the reader to the paper [4] for an explicit expression of L ; suffices it to remind that L has full column rank. We shall later use the specific structure of the column space of L and of its left kernel.

Equation (III.13) shows that indeed E has a singular covariance matrix $R = \text{Var}\{E\} = L(I \otimes \Lambda)L^\top$.

In the paper [4] it is shown how (III.12) can be converted into a least squares problem with full rank noise covariance and equality constraints (see also [27], [29], [16]). Remarkably, as we shall see in the next Section, this is equivalent to estimating a long VARX model of length $t - t_0$, using data in the interval $[t_0, T + N]$.

Using the estimator $\hat{\Omega}^{P_{opt}}$, the oblique projections \hat{Y}_{t+h}^P (III.3) can be substituted with $\hat{Y}_{t+h}^{P_{opt}} = \hat{\Xi}_h^{P_{opt}} Z_{[t_0,t]}$ in the SVD step (III.5); hence an estimator for the state shall be given by

$$\hat{X}_t^{P_{opt}} := \left(\hat{\Gamma}_\nu^{P_{opt}} \right)^{-L} \begin{bmatrix} \hat{Y}_t^{P_{opt}} \\ \hat{Y}_{t+1}^{P_{opt}} \\ \vdots \\ \hat{Y}_{T-1}^{P_{opt}} \end{bmatrix}. \quad (\text{III.14})$$

Also the “shifted” oblique projections used for the computation of the state at time $t+1$ (see (III.6)) can be substituted

⁹We call “selection matrix” a matrix formed with zeros and ones in which each row all entries are zero except for one.

by

$$\hat{X}_{t+1}^{P_{opt}} := \left(\hat{\Gamma}_{\nu}^{P_{opt}} \right) \begin{bmatrix} \hat{\Xi}_0^{P_{opt}} & \hat{\Psi}_{11}^{P_{opt}} \\ \hat{\Xi}_1^{P_{opt}} & \hat{\Psi}_{22}^{P_{opt}} \\ \vdots & \vdots \\ \hat{\Xi}_{\nu}^{P_{opt}} & \hat{\Psi}_{\nu\nu}^{P_{opt}} \end{bmatrix} Z_{[t_0, t+1)}. \quad (\text{III.15})$$

Similarly an estimator of the innovation sequence E_t can be found by $\hat{E}_t^{P_{opt}} := Y_t - \hat{E} \left[Y_t^{P_{opt}} | \hat{X}_t^{P_{opt}} \right]$.

IV. MAIN RESULT

The main result of this paper can be summarized as follows:

Theorem 4.1: Consider the VARX model

$$y_t = \sum_{i=1}^{t-t_0} \Phi_i z_{t-i} + e_t \quad (\text{IV.1})$$

and denote with $\hat{\Phi}_i$ the estimators of the coefficients in (IV.1) obtained solving

$$Y_t^{\nu+N} \simeq \sum_{i=1}^{t-t_0} \Phi_i Z_{t-i}^{\nu+N} \quad (\text{IV.2})$$

in the least squares sense.

The ‘‘optimally-weighted’’ solution to (III.12), i.e. the one that yields the least asymptotic variance of the estimators $\hat{\Omega}^{P_{opt}}$ among all linear, asymptotically unbiased estimators of Ω^P based on the regression (III.12), is equivalent to estimating the VARX model (IV.1) in the sense that:

$$\begin{bmatrix} \hat{\Xi}_0^{P_{opt}} \\ \hat{\Xi}_1^{P_{opt}} \\ \vdots \\ \hat{\Xi}_{\nu}^{P_{opt}} \end{bmatrix} = \begin{bmatrix} \hat{\Phi}_{t-t_0} & \cdots & \hat{\Phi}_{T-t_0} & \cdots & \hat{\Phi}_1 \\ 0 & \hat{\Phi}_{t-t_0} & \cdots & \cdots & \hat{\Phi}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \hat{\Phi}_{t-t_0} & \cdots & \hat{\Phi}_{\nu+1} \end{bmatrix} \quad (\text{IV.3})$$

and

$$\hat{\Psi}_{ji}^{P_{opt}} = \hat{\Phi}_i \quad (\text{IV.4})$$

Proof: See [7]. ■

Remark IV.1 It is worth mentioning that, with the ‘‘optimally weighted’’ (Markov) estimator of the coefficients Ξ_i, Ψ_{ij} , the estimate of the lower triangular matrix in (III.8) is indeed Toeplitz (see eq. (IV.4)). As a side, note the idea of imposing the Toeplitz structure was first put forward in [25]. However, this structure is here the results of an ‘‘optimally weighted’’ procedure rather than an a priori constraint imposed when designing the estimation algorithm. It is also interesting to note that the estimate of the VARX coefficients weighting the ‘‘far’’ past (i.e. Ψ_{ji} for $i > t - t_0$ in (III.2)) are set to zero by the ‘‘optimal’’ estimator (i.e. $\hat{\Psi}_{ji}^{P_{opt}} = 0$ for $i > t - t_0$). This is reasonable since, according to Assumption 1, for $i > t - t_0$ the Ψ_{ji} 's go to zero faster than $1/\sqrt{N}$; on the contrary, estimating these coefficients would lead to errors which are of order $1/\sqrt{N}$ in probability. This also brings up the question of choosing the length of the past horizon $t - t_0$; the analysis of this paper gives, together with the results in [14], a more theoretically sound foundation to the (usually adopted) practice of determining $t - t_0$ using standard order

selection criterions [23], [18], [29] for vector autoregressive models (see, e.g. [1]). ◊

Using the result of Theorem 4.1 the ‘‘optimized’’ PBSID algorithm can be implemented as follows:

- Estimate the VARX model (IV.1) as described in (IV.2); this may include estimation of the appropriate $t - t_0$ using standard criterions for VARX order estimation.
- Use the estimated coefficients as described in formulas (IV.3) and (IV.4) to form the predictors

$$\hat{Y}_{t+h}^{P_{opt}} = \sum_{i=h+1}^{t-t_0+h} \hat{\Psi}_{hi}^{P_{opt}} Z_{t+h-i} = \sum_{i=h+1}^{t-t_0} \hat{\Phi}_i Z_{t+h-i};$$

the state sequences $\hat{X}_t^{P_{opt}}$ and $\hat{X}_{t+1}^{P_{opt}}$ are then obtained as described in formulas (III.14) and (III.15).

This implementation has a much lower computational complexity w.r.t. the implementation described in [4] which involves solving the least squares problem (III.12) directly.

In fact, step a) above involves the estimation of a VARX model of length $t - t_0$ (which, according to Assumption 1, is $O(\log(N))$); solving (IV.2) has complexity $O(N(\log N)^2)$ (see [16] pag. 248). The order and state estimation (step b) above) can be performed on the ‘‘squared’’ version of the matrix $\hat{Y}_{[t, T-1]}^{P_{opt}}$. This second step is common to all subspace algorithms. Instead step a) has the same ‘‘order’’ of complexity than, e.g., CCA and PBSID; however both these algorithms essentially estimate ν long ARX models, increasing the complexity of the first step roughly by a factor ν .

Hence the implementation described above of the ‘‘optimized’’ PBSID compares favorably to a variety of subspace procedures (among which PBSID or CCA) as far computational complexity is concerned while, according to Theorem 5.3 in [4], yielding lower asymptotic variance than CCA. We remind also that the ‘‘optimized’’ PBSID algorithm works (i.e. is consistent) regardless of the presence of feedback.

These considerations make the algorithm described above a strong alternative to standard used methods for a variety of reasons, among which computational complexity and asymptotic statistical properties (it is consistent also in closed loop and gives lower variance than the original PBSID and CCA).

V. SIMULATION RESULTS

We consider the first order ARMAX model

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

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

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

We report results concerning the asymptotic variance and the sample variance estimated over 100 Monte Carlo runs multiplied by the number $N = 1000$ of data points used

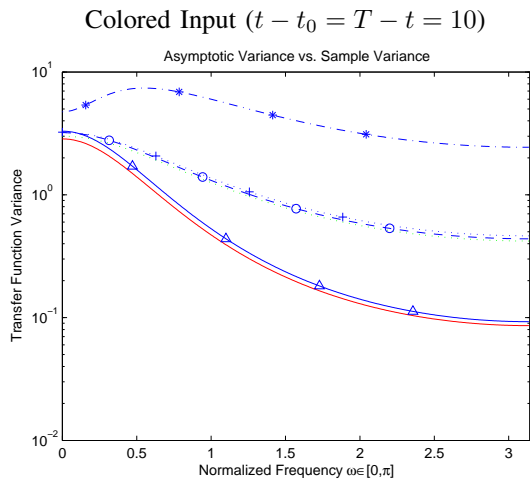


Fig. 1. *ARMAX* of order 1: Asymptotic Variance (and its Monte Carlo estimate) vs. normalized frequency ($\omega \in [0, \pi]$) Solid with triangles (Δ) PEM, dashed-dotted with stars (*): CCA, dotted with crosses (+): “predictor-based” algorithm (PBSID), dashed with circles (o): “optimized” PBSID algorithm described on page 5, second column; dotted: asymptotic variance for PBSID, solid (red): Cramér Rao lower bound.

in each experiment of the deterministic transfer function $F(z) = \frac{1}{z-0.5}$. It has been checked that the original algorithm presented in [4] and its alternative implementation presented in this paper give indeed the same result. In particular conditions (IV.3) and (IV.4) have been verified to hold for the estimated coefficients of the “optimized” PBSID described in [4].

We have chosen this simple example, which was used also in [4], since it contains all the essential features of the “optimized” method i.e.: (i) it is not efficient for increasing $T - t$ as instead claimed in [22] and (ii) it gives (strictly) lower asymptotic variance than CCA. Of course this example is performed in “open loop” to allow the comparison with CCA. In this example the original PBSID and the “optimized” version have the same asymptotic behavior. However, there is no proof, at the moment, that this can be generalized.

VI. CONCLUSIONS

In this paper we have shown that the “optimally weighted” projection in the optimized PBSID algorithm introduced in [4] is equivalent to estimating a single VARX model. Together with the results of [14], [20], [4], [5] this indeed shows the fundamental role played by VARX models in subspace identification.

This observation has important implications concerning computational complexity and, hopefully, also for the analysis of subspace methods.

VII. ACKNOWLEDGMENTS

The author would like to thank an anonymous reviewer of the paper [4] for some suggestions contained in his report which have stimulated the research documented in this paper.

REFERENCES

- [1] D. Bauer. Order estimation for subspace methods. *Automatica*, 37:1561–1573, 2001.
- [2] D. Bauer. Asymptotic properties of subspace estimators. *Automatica*, 41:359–376, 2005.
- [3] D. Bauer and L. Ljung. Some facts about the choice of the weighting matrices in Larimore type of subspace algorithm. *Automatica*, 38:763–773, 2002.
- [4] A. Chiuso. On the relation between CCA and predictor based subspace identification. *Submitted to IEEE Trans. on Aut. Control*, 2005. available at <http://www.dei.unipd.it/~chiuso>.
- [5] A. Chiuso. Asymptotic equivalence of certain closed-loop subspace identification methods. In *Proc. of SYSID 2006*, Newcastle, Australia, 2006.
- [6] A. Chiuso. Asymptotic variance of closed-loop subspace identification algorithms. *IEEE Trans. on Aut. Control*, 51(8):1299–1314, 2006.
- [7] A. Chiuso. On the role of Vector AutoRegressive modeling in subspace identification. 2006. Submitted to *Automatica*, available at <http://www.dei.unipd.it/~chiuso>.
- [8] A. Chiuso and G. Picci. Constructing the state of random processes with feedback. In *Proc. of the IFAC Int. Symposium on System Identification (SYSID)*, Rotterdam, August 2003.
- [9] A. Chiuso and G. Picci. The asymptotic variance of subspace estimates. *Journal of Econometrics*, 118(1-2):257–291, 2004.
- [10] A. Chiuso and G. Picci. Asymptotic variance of subspace methods by data orthogonalization and model decoupling: A comparative analysis. *Automatica*, 40(10):1705–1717, 2004.
- [11] A. Chiuso and G. Picci. On the ill-conditioning of subspace identification with inputs. *Automatica*, 40(4):575–589, 2004.
- [12] A. Chiuso and G. Picci. Consistency analysis of some closed-loop subspace identification methods. *Automatica*, 41(3):377–391, 2005.
- [13] A. Chiuso and G. Picci. Prediction error vs. subspace methods in closed-loop identification. In *Proc. of the 16th IFAC World Congress*, Prague, July 2005.
- [14] A. Dahlén and W. Scherrer. The relation of CCA subspace method to a balanced reduction of an autoregressive model. *Journal of Econometrics*, 118(1-2):293–312, 2004.
- [15] T. Ferguson. *A Course in Large Sample Theory*. Chapman and Hall, 1996.
- [16] G.H. Golub and C.R. Van Loan. *Matrix Computation*. The Johns Hopkins Univ. Press., 2nd ed. edition, 1989.
- [17] C.W.J. Granger. Economic processes involving feedback. *Information and Control*, 6:28–48, 1963.
- [18] E.J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. Wiley, 1988.
- [19] M. Jansson. Asymptotic variance analysis of subspace identification methods. In *Proceedings of SYSID2000*, S. Barbara Ca., 2000.
- [20] M. Jansson. Subspace identification and ARX modeling. In *Proceedings of SYSID 2003*, Rotterdam, 2003.
- [21] W.E. Larimore. System identification, reduced-order filtering and modeling via canonical variate analysis. In *Proc. American Control Conference*, pages 445–451, 1983.
- [22] W.E. Larimore. Large sample efficiency for adaptix subspace identification with unknown feedback. In *Proc. of IFAC DYCOPS’04*, Boston, MA, USA, 2004.
- [23] L. Ljung. *System Identification; Theory for the User*. Prentice Hall, 1997.
- [24] K. Peternell. *Subspace methods for subspace identification*. PhD thesis, Technical University of Vienna, 1995.
- [25] K. Peternell, W. Scherrer, and M. Deistler. Statistical analysis of novel subspace identification methods. *Signal Processing*, 52:161–178, 1996.
- [26] S.J. Qin and L. Ljung. Closed-loop subspace identification with innovation estimation. In *Proceedings of SYSID 2003*, Rotterdam, 2003.
- [27] C.R. Rao. Representations of the best linear unbiased estimators in the Gauss-Markov model with a singular dispersion matrix. *J. Multivariate Anal.*, 3:276–292, 1973.
- [28] F. Shi and J.F. MacGregor. A framework for subspace identification. In *Proc. of IEEE ACC*, Arlington, VA, 2001.
- [29] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, 1989.
- [30] P. Van Overschee and B. De Moor. Subspace algorithms for the stochastic identification problem. *Automatica*, 29:649–660, 1993.
- [31] P. Van Overschee and B. De Moor. N4SID: Subspace algorithms for the identification of combined deterministic– stochastic systems. *Automatica*, 30:75–93, 1994.