Available online at www.sciencedirect.com
ScienceDirect
Automatica 43 (2007) 1034–1048
automatica
www.elsevier.com/locate/automatica

The role of vector autoregressive modeling in predictor-based subspace identification☆

Alessandro Chiuso

Dipartimento di Tecnica e Gestione dei Sistemi Industriali Università di Padova (sede di Vicenza), stradella San Nicola, 3-36100 Vicenza, Italy

Received 21 April 2006; received in revised form 27 September 2006; accepted 6 December 2006

Abstract

Subspace identification for closed loop systems has been recently studied by several authors. A class of new and consistent closed-loop subspace algorithms is based on identification of a predictor model, in a way similar as prediction error methods (PEM) do. Experimental evidence suggests that these methods have a behavior which is very close to PEM in certain examples. The asymptotical statistical properties of one of these methods have been studied recently allowing to show (i) its relation with CCA and (ii) that Cramér–Rao lower bound is not reached in general. Very little, however, is known concerning their relative performance.

In this paper we shall discuss the link between these "predictor-based" methods; to this purpose we exploit the role which Vector Auto Regressive with eXogenous input models play in all these algorithms. The results of this paper provide a unifying framework under which all these algorithms can be viewed; also the link with VARX modeling have important implications as to computational complexity is concerned, leading to very computationally attractive implementations.

We also hope that this framework, and in particular the relation with VARX modeling followed by model reduction will turn out to be useful in future developments of subspace identification, such as the quest for efficient procedures and the statistical analysis with finite-data.

Keywords: Closed-loop identification; Subspace methods; Asymptotic properties; Variance; Relative efficiency

1. 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 (Chiuso & Picci, 2003a, 2003b, 2005a; Jansson, 2003; Qin & Ljung, 2003a) and second a whole body of results on the asymptotic statistical properties of subspace methods which have allowed, on the one side, to asses accuracy of subspace estimators (Bauer, 2005a; Bauer & Jansson, 2000; Chiuso, 2006b; Chiuso & Picci, 2004a; Jansson, 2000) and on the other to compare different methods (Bauer, 2005a, 2005b; Bauer & Ljung, 2002; Chiuso, 2006a, 2007b; Chiuso & Picci, 2004b).

Extension of subspace algorithms to closed loop operating conditions have required, at a certain stage, the introduction of two step procedures (see Jansson, 2003; Larimore, 2004; Shi & MacGregor, 2001) 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. An overview of these realization procedures can be found in Chiuso and Picci (2003a, 2003b, 2005a) and references therein. The reader is also referred to Peternell, Scherrer, and Deistler (1996) and Peternell (1995) for early contributions advocating for two-step procedures even for "open loop" identification. There, an "iterative"

☆ This work has been supported in part by the national project *New techniques and applications of identification and adaptive control* funded by MIUR. Part of this work has been presented at the 2006 IFAC SYSID Symposium held in Newcastle, Australia and at the 2006 IEEE CDC conference held in San Diego, USA. This paper was recommended for publication in revised form by Associate Editor Michel Verhaegen under the direction of Editor Torsten Söderström.

E-mail address: chiuso@dei.unipd.it.

0005-1098/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2006.12.009

CCA which makes use of the Markov parameters estimated in a preliminary stage was proposed. In Peternell et al. (1996) it was also conjectured that the two step CCA could possibly lead to asymptotic efficiency.[1]

Often the preliminary estimation has been performed using Vector AutoRegressive with eXogenous inputs (VARX) models. Some analysis regarding the role of VAR models in subspace identification was performed in Dahlén and Scherrer (2004) where it was shown that the CCA algorithm introduced in Larimore (1983) 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.

Also some preliminary recent work relating VARX models with subspace procedures can be found in Onodera, Emoto, and Qin (2006), Qin and Ljung (2006).

In this paper we shall be concerned with a class of algorithms which we group under the name predictor-based subspace identification. This terminology stems from the fact that these algorithms aim at identifying a "predictor model" in a way that reminds of prediction error methods (PEM). We refer the reader to the papers Chiuso and Picci, 2005a, 2005b for a thorough discussion of the basic issues. The algorithms that shall be discussed here are the SSARX algorithm by Jansson (2003), the PBSID[2] algorithm (introduced in Chiuso & Picci (2005a) under the name "whitening filter"), its "optimized" version[3] ($\text{PBSID}_{\text{opt}}$ hereafter) introduced in Chiuso (2007b) and the algorithm presented by Ljung and McKelvey (1996).

We shall expand on two recent contributions (see Chiuso, 2006a, 2006c) and discuss the role of VARX models in subspace algorithms based on predictor identification; we shall show that the preliminary step based on VARX models, explicitly used in Jansson (2003) is actually present, in a way or another, in all these algorithms. We also observe that the bank of predictors used in Ljung and McKelvey (1996) to overcome problems due to feedback were constructed using VARX models. It turns out, as we shall see later in this paper, that the algorithm proposed in Ljung and McKelvey (1996) is very much related to the $\text{PBSID}_{\text{opt}}$ introduced in Chiuso (2007b). For this reason, even though $\text{PBSID}_{\text{opt}}$ has been developed independently from Ljung and McKelvey (1996) and actually derives from a theoretically sound optimization, we regard the paper (Ljung & McKelvey, 1996) as a fundamental early contribution to closed-loop subspace identification.

In particular the main results of this paper can be enumerated as follows:

(a) SSARX by Jansson (2003), which requires a preliminary VARX modeling step, is asymptotically equivalent, in the sense of yielding the same asymptotic distribution of the estimators, to PBSID. Some preliminary results have appeared in Chiuso (2006a).
(b) The "optimally weighted" projection step involved in $\text{PBSID}_{\text{opt}}$ in Chiuso (2007b) is actually equivalent (here in the sense of giving the *same numerical results*) to estimating a VARX model followed by the usual steps of subspace identification.[4] Some preliminary results can be found in Chiuso (2006c).
(c) The algorithm presented by Ljung and McKelvey (1996) is equivalent to a weighted version of $\text{PBSID}_{\text{opt}}$.

In our opinion the significance of these results with respect to the current state of the art in subspace identification can be described as follows:

(a) One contribution of this paper, which can be seen as a natural continuation of previous works (Bauer, 2005b; Bauer & Ljung, 2002; Chiuso, 2007b; Chiuso & Picci, 2004b), is to provide a comparison between recently proposed methods, trying to obtain a more unified picture of subspace algorithms; we believe this is useful since subspace algorithms have grown rapidly in number in the last few years (Chiuso & Picci, 2005a; Jansson, 2003; Ljung & McKelvey, 1996; Qin & Ljung, 2003b) with very little insight, if any, concerning their relative efficiency.
(b) Second, by showing that $\text{PBSID}_{\text{opt}}$ is numerically equivalent to estimating a VARX model followed by the usual steps of subspace identification, we provide a way to implement the $\text{PBSID}_{\text{opt}}$ algorithm with a much lower computational complexity then originally discussed in Chiuso (2007b). We would like to remind that $\text{PBSID}_{\text{opt}}$ share the advantages of PBSID in that it delivers consistent estimators with closed loop data while comparing favorably, in the sense of asymptotic variance, to CCA for the open loop case. In fact it was shown in Chiuso (2007b, Theorem 5.3), that the asymptotic variance of $\text{PBSID}_{\text{opt}}$ is less or equal than that of CCA for *any* choice of the input signal. The fact that CCA is known to be asymptotically efficient[5] for time series identification (= no inputs) (Bauer, 2005b) and optimal for white inputs (Bauer & Ljung, 2002) strengthens the significance of our result.
(c) Last but not least, the relation with VARX modeling followed by model reduction, together with the results in Bauer (2005a), Bauer and Ljung (2002), Chiuso (2007b) and Dahlén and Scherrer (2004), might be very helpful, in the author's opinion, in future developments of subspace identification. We also refer the reader to the paper (Wahlberg, 1989) for a discussion and early references on the use of high order AR models for identification of

[1] In this paper we shall always use the word efficient assuming Gaussian distributions of the data.

[2] Short for "Predictor-Based Subspace IDentification".

[3] The word "optimized" refers to the fact that a projection step is replaced by an "optimally weighted" (Markov) estimator.

[4] Even though some preliminary results along these lines have already been presented in Chiuso (2006a), the author would like to thank an anonymous reviewer of the paper (Chiuso, 2007b) which have underlined the relevance of the comparison performed in this paper; part of the merit of this paper should also go to him.

[5] When both the past and future horizon go to infinity with the number of data.

AutoRegressive Moving Average (ARMA) models. In particular all these results could provide:

(i) suggestions on how subspace procedures could be modified such as to reach asymptotic efficiency; recall for instance that in the case of no observed inputs (Wahlberg, 1989) derives an asymptotically efficient ARMA parameter estimation method based on AR modeling followed by model reduction.

(ii) a tool to introduce structure in the identification problem (delays, inputs which do not affect certain outputs, etc.) which might turn out to be very useful when handling systems with large numbers of input–output channels (see e.g. the plenary lecture given by Zhu (2006) at the recent SYSID in Newcastle and also Hong, Harmse, Guiver, & Canney, 2006).

Concerning the relation of subspace methods with VARX modeling, recall that it was shown in Dahlén and Scherrer (2004) 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 as shown in Bauer (2005b). It has also been shown in Chiuso (2007b) 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 different when there are inputs, and they are colored. The $\text{PBSID}_{\text{opt}}$ performs better than CCA but it is not clear whether it is efficient in general; note that in Peternell et al. (1996) it was conjectured that pre-estimation of certain Markov parameters might be a way to obtain efficient subspace procedures.

In Larimore (2004) it is claimed that a procedure which is very much related to the SSARX algorithm might be efficient; however, in Larimore (2004) it is claimed that efficiency is reached for both past and future horizon which go to infinity. This claim appears to be wrong since, on the contrary, depending upon the input characteristics, the asymptotic variance might increase or decrease as a function of the future horizon (see Chiuso, 2007a, 2007b). Note also that in Chiuso, (2007a) the algorithm discussed in Larimore (2004) is shown to be asymptotically equivalent to SSARX and hence, from the results of this paper, also to PBSID.

We believe existence of an efficient subspace procedure is worth investigating.

We warn the reader that this paper does not mean to provide an exhaustive coverage of the state of the art in subspace identification but rather an analysis of a specific class of algorithms as mentioned earlier in the introduction. Many algorithms are not discussed (Jansson, 2005; Katayama, Kawauchi, & Picci, 2005; Katayama, Tanaka, & Enomoto, 2005) or just mentioned in passing (Onodera et al., 2006; Qin & Ljung, 2003b, 2006; Shi & MacGregor, 2001).

The structure of the paper is as follows. In Section 2 we state the problem and set up notation; Section 3 briefly recalls the algorithmic steps while Section 4 states the main results of this paper contained in Theorems 4.1, 4.2 and Proposition 4.6 together with some simulation results. Section 5 contains some conclusions. The most technical parts of the proofs are postponed to the Appendix.

2. Statement of the problem and notation

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

$$\begin{cases} \mathbf{x}(t+1) = A\mathbf{x}(t) + B\mathbf{u}(t) + K\mathbf{e}(t), \\ \mathbf{y}(t) = C\mathbf{x}(t) + D\mathbf{u}(t) + \mathbf{e}(t), \end{cases} \quad t \geqslant t_0. \tag{2.1}$$

We allow for *feedback* from $\{\mathbf{y}(t)\}$ to $\{\mathbf{u}(t)\}$ (Granger, 1963), 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 (2.1) 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 assume that spectral density matrix of $\mathbf{z}$, $\Phi(z)$ 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 which are inside the closed unit disc. We define $\rho := \max(|\mu_i|)$. From the assumption $\Phi(e^{j\omega}) > cI > 0$ it follows that $\rho < 1$. Note in particular that $1 > \rho \geqslant \max(|\lambda_i(\bar{A})|)$ where $\lambda_i(\bar{A})$ is the ith eigenvalue of $\bar{A}$.

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

Given two sequences of (scalar) random variables $\mathbf{x}_N$ and $\mathbf{g}_N$, we shall say that $\mathbf{x}_N$ is $O_P(\mathbf{g}_N)$, which we shall write $\mathbf{x}_N = O_P(\mathbf{g}_N)$, if, $\forall \varepsilon$, $\exists M$ s.t.

$$\sup_N P[|\mathbf{x}_N/\mathbf{g}_N| > M] < \varepsilon.$$

In particular if $\mathbf{x}_N = O_P(1)$ we say that $\mathbf{x}_N$ is bounded in probability,[6] note that $\mathbf{x}_N = O_P(\mathbf{g}_N)$ means that $\mathbf{x}_N/\mathbf{g}_N$ is bounded in probability.

Similarly, $\mathbf{x}_N = o_P(\mathbf{g}_N)$ means that, $\forall \delta > 0$,

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

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 $\doteq$ 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 Ferguson, 1996) terms which are $o_P(1/\sqrt{N})$ can be neglected when studying the asymptotic distribution.

We shall use the notations $\underline{o}_P(\cdot)$, $\underline{O}_P(\cdot)$, $\underline{o}(\cdot)$ or $\underline{O}(\cdot)$ to denote random matrices (of suitable dimensions possibly depending on N) whose elements are respectively $o_P(\cdot)$, $O_P(\cdot)$, $o(\cdot)$ or

[6] Sometimes this is stated saying that $\mathbf{x}_N$ is "uniformly tight". For instance every sequence of random variables converging in distribution is uniformly tight.

$O(\cdot)$ uniformly. Uniformity is needed when the matrices' sizes increase with N. In this paper uniformity shall be guaranteed by stationarity of the processes involved.

Note also, for future reference, that if $\mathbf{x}_N = O_P(1/\sqrt{N})$ and $\mathbf{y}_N = O_P(1/\sqrt{N})$, then $\mathbf{x}_N\mathbf{y}_N = o_P(1/\sqrt{N})$.

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 (2.1).

Throughout the paper the symbol t shall denote "present", t_0 shall be the initial time from which data are collected, so that $t - t_0$ is the "past horizon"; T shall be a design parameter so that $T - t$ is the number of future lags used to form predictors, commonly known as the "future horizon",[7] while N shall be the length of the finite tails.[8]

The analysis reported in this paper requires that both N, and $t - t_0$ 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 Bauer and Ljung (2002) where the following assumption is made:

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

$$\begin{aligned} t - t_0 &\geqslant \frac{\log N^{-d/2}}{\log|\rho|}, \quad 1 < d < \infty, \\ t - t_0 &= o(\log(N)^{\alpha}), \qquad \alpha < \infty. \end{aligned} \tag{2.2}$$

Under this assumption the effect of terms due to mishandling of the initial condition at time t_0 are $o_P(1/\sqrt{N})$ and therefore can be neglected. Moreover, (2.2) 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 Bauer & Ljung (2002, Lemma 4)).

In order to simplify the analysis in this paper we shall keep the future horizon ν fixed (and finite). We warn the reader that, for instance, asymptotic efficiency of the CCA method for time series identification requires ν to grow with the sample size (see Bauer, 2005a). We believe, however, that our results are significant even under this assumption since the comparison holds for fixed yet arbitrarily large ν.

According to Assumption 1 the total number of data is $T + N - t_0 + 1 = T - t + N + o(\log(N)^{\alpha})$; therefore, $\lim_{N\to\infty}(T + N - t_0 + 1)/N = 1$ (which implies, in particular, $O(N) = O(T + N - t_0 + 1)$). For this reason (and for convenience of notation) when dealing with asymptotic results, we shall refer to the length of the finite tails N rather to the total number of data $T + N - t_0 + 1$ (e.g. we shall use $o_P(1/\sqrt{N})$ and not $o_P(1/\sqrt{T + N - t_0 + 1})$).

[7] Respectively, the number of block rows in the block Hankel data matrix containing the past and future data.

[8] This is the parameter j in the notation of Van Overschee and De Moor (1994) i.e. the number of columns in the block Hankel data matrices used in subspace identification.

Remark 2.1 (*On the necessity of Assumption 1*). One may argue that having to deal with an "infinite" past horizon might not be very attractive. This condition, as discussed in detail in Chiuso and Picci (2005a), is needed to ensure consistency in closed loop. There is, however, another important reason to keep this assumption. Essentially subspace algorithms are "covariance-based" methods[9]; therefore, as discussed in Walker (1961) and Porat and Friedlander (1985) for the MA/ARMA case, it is necessary to estimate an infinite number of covariances to obtain asymptotically efficient estimators (see also Wahlberg, 1989, Remark 4, p. 289). This requires that $t - t_0$ goes to infinity. Therefore, a necessary condition for a subspace algorithm to reach the Cramér–Rao lower bound is that $t - t_0$ goes to infinity, as required by Assumption 1. Of course this does not mean that an algorithm allowing $t - t_0 \to \infty$ will automatically be efficient.

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 $\mathbf{z}(t)$ the random vector denoting the joint process and with z_t the sample value of $\mathbf{z}(t)$. We shall use capitals to denote the tail of length N. For instance $Z_t := [z_t \;\; z_{t+1}, \;\; \dots \;\; z_{t+N-1}]$. These are the block rows of the usual *block Hankel data matrices* which appear in subspace identification.

When dealing with tails of length different from N we shall add the number of columns as a superscript; for instance $Z_t^M := [z_t \;\; z_{t+1}, \;\; \dots \;\; z_{t+M-1}]$.

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

$$\mathscr{Z}_{[\tau,t)} := \overline{\text{span}}\,\{\mathbf{z}_k(s);\, k = 1, \dots, m + p,\, \tau \leqslant s < t\},$$

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 definitions hold for $\mathscr{Y}_{[\tau,t)}$ and $\mathscr{U}_{[\tau,t)}$.

When $\tau = -\infty$ we shall use the shorthands $\mathscr{Z}_t^-$ for $\mathscr{Z}_{[-\infty,t)}$. The space generated by $\mathbf{z}(s)$, $-\infty < s < \infty$ shall be denoted with the symbol $\mathscr{Z}$. For convenience of notation we denote with $\nu := T - t$ the future horizon.

Given a subspace $\mathscr{C} \subseteq \mathscr{Z}$, we shall denote with $E[\mathbf{a}|\mathscr{C}]$ the orthogonal projection of the random variable $\mathbf{a}$ onto $\mathscr{C}$; in the Gaussian case the linear projection coincides with conditional expectation, i.e. $\mathbb{E}[\cdot|\mathscr{C}] = E[\cdot|\mathscr{C}]$. Let $\mathbf{c}$ be a (finite) basis for $\mathscr{C}$. Using the notation $\Sigma_{\mathbf{ab}} := \mathbb{E}\left[\mathbf{a}\mathbf{b}^{\top}\right]$ for the covariance matrix between the zero mean random vectors $\mathbf{a}$ and $\mathbf{b}$, in the finite dimensional case $E[\mathbf{a}|\mathscr{C}]$ will be given by the usual formula

$$E[\mathbf{a}|\mathscr{C}] = \Sigma_{\mathbf{ac}}\Sigma_{\mathbf{cc}}^{-1}\mathbf{c}. \tag{2.3}$$

Defining also the projection errors $\tilde{\mathbf{a}} := \mathbf{a} - E[\mathbf{a}|\mathscr{C}]$ and $\tilde{\mathbf{b}} := \mathbf{b} - E[\mathbf{b}|\mathscr{C}]$, the symbol $\Sigma_{\mathbf{ab|c}}$ will denote projection error covariance (conditional covariance in the Gaussian case) $\Sigma_{\mathbf{ab|c}} :=$

[9] One may argue that there are "data-based" methods, but this is just a computational aspect.

$\Sigma_{\tilde{\mathbf{a}}\tilde{\mathbf{b}}} = \Sigma_{\mathbf{ab}} - \Sigma_{\mathbf{ac}}\Sigma_{\mathbf{cc}}^{-1}\Sigma_{\mathbf{cb}}$. Given two trivially intersecting subspaces $\mathscr{C} \subseteq \mathscr{L}$, $\mathscr{B} \subseteq \mathscr{L}$, $\mathscr{C} \cap \mathscr{B} = \{0\}$, $E_{\|\mathscr{B}}[\cdot|\mathscr{C}]$ shall denote the oblique projection onto $\mathscr{C}$ along $\mathscr{B}$ (see Golub & Van Loan, 1989) and can be computed by the formula:

$$E_{\|\mathscr{B}}[\mathbf{a}|\mathscr{C}] = \Sigma_{\mathbf{ac}|\mathbf{b}}\Sigma_{\mathbf{cc}|\mathbf{b}}^{-1}\mathbf{c}. \tag{2.4}$$

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

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 $\lim_{N\to\infty}\hat{\Sigma}_{\mathbf{ab}} \overset{\text{a.s}}{=} \Sigma_{\mathbf{ab}}$.

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

$$\hat{E}[A_t|C_t] = \hat{\Sigma}_{\mathbf{ac}}\hat{\Sigma}_{\mathbf{cc}}^{-1}C_t. \tag{2.5}$$

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 $\hat{\Sigma}_{\mathbf{ab}|\mathbf{c}} := \hat{\Sigma}_{\tilde{\mathbf{a}}\tilde{\mathbf{b}}}$ and computed by the formula

$$\hat{\Sigma}_{\mathbf{ab}|\mathbf{c}} := \hat{\Sigma}_{\mathbf{ab}} - \hat{\Sigma}_{\mathbf{ac}}\hat{\Sigma}_{\mathbf{cc}}^{-1}\hat{\Sigma}_{\mathbf{cb}}.$$

We shall denote with $\hat{E}_{\|B_t}[\cdot|C_t]$ the oblique projection along the space generated by the rows B_t onto the space generated by the rows of C_t (provided the intersect only at zero). As above, the oblique projection can be computed using the formula:

$$\hat{E}_{\|B_t}[A_t|C_t] = \hat{\Sigma}_{\mathbf{ac}|\mathbf{b}}\hat{\Sigma}_{\mathbf{cc}|\mathbf{b}}^{-1}C_t. \tag{2.6}$$

For future reference we also define the extended observability matrix

$$\bar{\Gamma}_v^\top := [\,C^\top \quad \bar{A}^\top C^\top \quad (\bar{A}^\top)^2 C^\top \quad \dots \quad (\bar{A}^\top)^{v-1}C^\top\,]. \tag{2.7}$$

3. State space construction

It is well known (Chiuso & Picci, 2004c; Lindquist & Picci, 1996; Van Overschee & De Moor, 1994) that identification using subspace methods can be seen as a two step procedure as follows:

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

$$\begin{cases} \hat{X}_{t+1} \simeq A\hat{X}_t + B\hat{U}_t + KE_t, \\ Y_t \simeq C\hat{X}_t + E_t \end{cases} \tag{3.1}$$

in the least squares sense.

Different subspace algorithms have different implementations of the first step while the second remains the same for virtually all algorithms.[10] For this reason we compare algorithms on the basis of step (a). We shall identify procedures which are (asymptotically) equivalent, modulo change of basis, as the first step is concerned.

3.1. PBSID algorithm

The construction of the state space using this algorithm involves several oblique projections. The projection of each (block) row Y_{t+h}, $h = 0, \dots, v$, can be seen as a long VARX model as follows:

$$\begin{aligned}\hat{Y}_{t+h} &:= \hat{E}[Y_{t+h} \mid Z_{[t_0,t+h)}] \\ &= \hat{\Psi}_{1,h}Z_{t+h-1} + \cdots + \hat{\Psi}_{t+h-t_0,h}Z_{t_0}\end{aligned} \tag{3.2}$$

from which the oblique projections[11]

$$\begin{aligned}\hat{Y}_{t+h}^{\mathrm{P}} &:= \hat{E}_{\|Z_{[t,t+h)}}[Y_{t+h} \mid Z_{[t_0,t)}] \\ &= \sum_{i=h+1}^{t-t_0+h} \hat{\Psi}_{i,h}Z_{t+h-i} \simeq C\bar{A}^{h-1}X_t.\end{aligned} \tag{3.3}$$

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

$$\hat{\mathbf{y}}^{\mathrm{P}}(t+h) := E_{\|\mathscr{L}_{[t,t+h)}}[\mathbf{y}(t+h) \mid \mathscr{L}_t^-] = C\bar{A}^{h-1}\mathbf{x}(t) \tag{3.4}$$

holds. Then one stacks all the predictors

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

[10] In this paper we shall not be concerned with algorithms based on the so-called "shift invariance" method (Bauer, 2005a).

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

From the singular value decomposition

$$W_p^{-1}\hat{Y}^{\mathrm{P}}_{[t,T)} = PDQ^\top = [P_n \;\; \tilde{P}_n]\begin{bmatrix} D_n & 0 \\ 0 & \tilde{D}_n \end{bmatrix}\begin{bmatrix} Q_n^\top \\ \tilde{Q}_n^\top \end{bmatrix}, \tag{3.5}$$

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

$$\hat{\bar{\Gamma}}_v = W_p P_n D_n^{1/2}$$

and consequently a basis for the state space

$$\hat{X}^{\mathrm{PBSID}}_t := \hat{\bar{\Gamma}}_v^{-L}\hat{Y}^{\mathrm{P}}_{[t,T)},$$

$$\hat{X}^{\mathrm{PBSID}}_{t+1} := \hat{\bar{\Gamma}}_v^{-L}\hat{Y}^{\mathrm{P}}_{[t+1,T]}, \tag{3.6}$$

where $\hat{\bar{\Gamma}}_v^{-L}$ is the left inverse defined by

$$\hat{\bar{\Gamma}}_v^{-L} := (\hat{\bar{\Gamma}}_v^\top W_p^{-\top} W_p^{-1}\hat{\bar{\Gamma}}_v)^{-1}\hat{\bar{\Gamma}}_v^\top W_p^{-\top} W_p^{-1}. \tag{3.7}$$

3.2. SSARX algorithm

The algorithm described in the previous section can be seen as a "geometric" version of the SSARX algorithm by Jansson (2003). Instead of computing the oblique projections (3.3), or, equivalently, instead of estimating $v+1$ long VARX models, Jansson estimates just one (long) VARX model

$$Y_T \simeq \hat{\Phi}_1 Z_{T-1} + \hat{\Phi}_2 Z_{T-2} + \cdots + \hat{\Phi}_{T-t_0} Z_{t_0}, \tag{3.8}$$

where without loss of generality we have taken the length of the VARX model equal to $T - t_0$; then the effect of the future inputs/outputs is removed using the estimated parameters $\hat{\Phi}_k$ as[12] :

$$\hat{Y}^{\mathrm{S}}_{[t,T)} := \hat{E}[Y_{[t,T)} - \hat{\bar{H}}^{\mathrm{S}}_v Z_{[t,T)} \mid Z_{[t_0,t)}], \tag{3.9}$$

where

$$\hat{\bar{H}}^{\mathrm{S}}_v := \begin{bmatrix} 0 & 0 & \dots & 0 \\ \hat{\Phi}_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \hat{\Phi}_v & \dots & \hat{\Phi}_1 & 0 \end{bmatrix}.$$

The remaining part is essentially the same as in the previous section provided[13] $\hat{Y}^{\mathrm{S}}_{[t,T)}$ is substituted to $\hat{Y}^{\mathrm{P}}_{[t,T)}$.

3.3. "Optimized" PBSID algorithm

The optimized version of PBSID introduced in Chiuso (2007b) ($\mathrm{PBSID}_{\mathrm{opt}}$) differs from the original PBSID algorithm in the computation of the predictors (3.2); 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 define $\mathscr{K} := [\bar{A}^{t-t_0-1}[K\ B]\bar{A}^{t-t_0-2}[K\ B]\cdots[K\ B]]$. Recall that

$$\begin{aligned} Y_{t+h} &= C\bar{A}^h X_t \\ &\quad + \sum_{i=1}^{h} C\bar{A}^{i-1}(KY_{t+h-i} + BU_{t+h-i}) + E_{t+h} \\ &= C\bar{A}^h \mathscr{K} Z_{[t_0,t)} \\ &\quad + \sum_{i=1}^{h} C\bar{A}^{i-1}(KY_{t+h-i} + BU_{t+h-i}) \\ &\quad + E_{t+h} + \underline{o}_P(1/\sqrt{N}) \\ &:= \Xi_h Z_{[t_0,t)} \\ &\quad + \sum_{i=1}^{h} \Psi_{hi} Z_{t+h-i} + E_{t+h} + \underline{o}_P(1/\sqrt{N}), \end{aligned} \tag{3.10}$$

where the last equality defines the matrices Ξ_h and Ψ_{hi}. Stacking the data and using (3.10) (discarding $o_P(1/\sqrt{N})$ terms; this is a delicate matter see Appendix B in Chiuso (2007b) for details) we obtain:

$$\begin{aligned} \begin{bmatrix} Y_t \\ Y_{t+1} \\ \vdots \\ Y_T \end{bmatrix} &\doteq \begin{bmatrix} \Xi_0 \\ \Xi_1 \\ \vdots \\ \Xi_v \end{bmatrix} Z_{[t_0,t)} \\ &\quad + \begin{bmatrix} 0 & 0 & \dots & 0 \\ \Psi_{11} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{vv} & \dots & \Psi_{v1} & 0 \end{bmatrix} Z_{[t,T]} + \begin{bmatrix} E_t \\ E_{t+1} \\ \vdots \\ E_T \end{bmatrix}. \end{aligned} \tag{3.11}$$

Observe that the lower triangular matrices in (3.11) are Toeplitz, since $\Psi_{ij} = C\bar{A}^{j-1}[K\ B], \forall i, j$. The projection in (3.2) is equivalent to solving (3.11) "row by row"; hence the Toeplitz structure is not preserved after estimation, i.e. $\hat{\Psi}_{ij} \neq \hat{\Psi}_{i'j}$, $i \neq i'$ almost surely.

This is equivalent to solving the least squares problem obtained vectorizing (3.11):

$$\begin{aligned} Y := \begin{bmatrix} \mathrm{vec}(Y_t) \\ \mathrm{vec}(Y_{t+1}) \\ \vdots \\ \mathrm{vec}(Y_T) \end{bmatrix} &\doteq S^{\mathrm{P}}\Omega^{\mathrm{P}} + \begin{bmatrix} \mathrm{vec}(E_t) \\ \mathrm{vec}(E_{t+1}) \\ \vdots \\ \mathrm{vec}(E_T) \end{bmatrix} \\ &= S^{\mathrm{P}}\Omega^{\mathrm{P}} + E, \end{aligned} \tag{3.12}$$

[12] The superscript $^{\mathrm{S}}$ stands for "SSARX".

[13] Also a specific choice of W_p is done by Jansson. We leave this choice unspecified here since equivalence holds for every choice of W_p.

where the matrix S^{P} has the form

$$S^{\mathrm{P}} = \text{block diag}\{(Z^{\top}_{[t_0,t)} \otimes I), \ldots, (Z^{\top}_{[t_0,T)} \otimes I)\} \tag{3.13}$$

and Ω^{P} is given by

$$\Omega^{\mathrm{P}} = [\mathrm{vec}^{\top}(\Xi_0)\ \ \mathrm{vec}^{\top}(\Xi_1)\ \ \mathrm{vec}^{\top}(\Psi_{11}) \quad \cdots \mathrm{vec}^{\top}(\Xi_v)\ \ \cdots\ \ \mathrm{vec}^{\top}(\Psi_{v1})]^{\top}, \tag{3.14}$$

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

$$Y \doteq S^{\mathrm{P}}\Omega^{\mathrm{P}} + E, \tag{3.15}$$

where $o_P(1/\sqrt{N})$ terms have been neglected[14], gives an estimator $\hat{\Omega}^{\mathrm{P_{opt}}}$ of Ω^{P} which has the smallest asymptotic variance among all linear (asymptotically unbiased) estimators based on (3.12). Incidentally, this has allowed to show in Chiuso (2007b) 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 (Larimore, 1983; Van Overschee & De Moor, 1993).

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

$$E = L\mathrm{vec}(E_t^{N+v}), \tag{3.16}$$

where L is a "selection[15] matrix" of size $pNv \times p(v+N)$. We refer the reader to the paper (Chiuso, 2007b) 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. Eq. (3.16) shows that indeed E has a singular covariance matrix $R = \mathrm{Var}\{E\} = L(I \otimes \Lambda)L^{\top}$.

In the paper (Chiuso, 2007b) it is shown how (3.15) can be converted into a least squares problem with full rank noise covariance and equality constraints (see also Golub & Van Loan, 1989; Rao, 1973; Söderström & Stoica, 1989; Werner & Yapar, 1996). 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 $[t_0, T + N - 1]$.

Using the estimator $\hat{\Omega}^{\mathrm{P_{opt}}}$, the oblique projections $\hat{Y}^{\mathrm{P}}_{t+h}$ (3.3) can be substituted with $\hat{Y}^{\mathrm{P_{opt}}}_{t+h} = \hat{\Xi}^{\mathrm{P_{opt}}}_h Z_{[t_0,t)}$ in the SVD step (3.5); hence, defining $\hat{Y}^{\mathrm{P_{opt}}}_{[t,T)} := [(\hat{Y}^{\mathrm{P_{opt}}}_t)^{\top}, (\hat{Y}^{\mathrm{P_{opt}}}_{t+1})^{\top}, \ldots, (\hat{Y}^{\mathrm{P_{opt}}}_{T-1})^{\top}]^{\top}$, an estimator for the state shall be given by

$$\hat{X}^{\mathrm{P_{opt}}}_t := (\hat{\bar{\Gamma}}^{\mathrm{P_{opt}}}_v)^{-L} \hat{Y}^{\mathrm{P_{opt}}}_{[t,T)}, \tag{3.17}$$

where $\hat{\bar{\Gamma}}^{\mathrm{P_{opt}}}_v$ is the estimate obtained substituting $\hat{Y}^{\mathrm{P_{opt}}}_{t+k}$ to $\hat{Y}^{\mathrm{P}}_{t+h}$ in (3.5).

[14] See Chiuso (2007b, Appendix B) for a rigorous discussion.

[15] We call "selection matrix" a matrix formed with zeros and ones in which each row all entries are zero except for one.

Also the "shifted" oblique projections used for the computation of the state at time $t + 1$ (see (3.6)) can be substituted by

$$\hat{X}^{\mathrm{P_{opt}}}_{t+1} := (\hat{\bar{\Gamma}}^{\mathrm{P_{opt}}}_v)^{-L} \begin{bmatrix} \hat{\Xi}^{\mathrm{P_{opt}}}_1 & \hat{\Psi}^{\mathrm{P_{opt}}}_{11} \\ \hat{\Xi}^{\mathrm{P_{opt}}}_2 & \hat{\Psi}^{\mathrm{P_{opt}}}_{22} \\ \vdots & \vdots \\ \hat{\Xi}^{\mathrm{P_{opt}}}_v & \hat{\Psi}^{\mathrm{P_{opt}}}_{vv} \end{bmatrix} Z_{[t_0,t+1)}. \tag{3.18}$$

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

$$\hat{E}^{\mathrm{P_{opt}}}_t := Y_t - \hat{E}[\hat{Y}^{\mathrm{P_{opt}}}_t | \hat{X}^{\mathrm{P_{opt}}}_t]. \tag{3.19}$$

4. Main results

This section contains the main results of this paper; first we shall discuss the (asymptotic) equivalence between PBSID and SSARX and later we shall discuss how $\mathrm{PBSID_{opt}}$ can be implemented using VARX models. Last $\mathrm{PBSID_{opt}}$ is related to the algorithm of Ljung and McKelvey (1996) which, in some sense, might be seen as a predecessor of all these methods.

For clarity of exposition we divide this task in three separate and self-contained subsections, each complemented with some simulation results.

4.1. Asymptotic equivalence of PBSID and SSARX

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

Theorem 4.1. *Assume the past horizon $t - t_0$ grows with N according to Assumption* 1. *Denote with $\hat{\Theta}^{\mathrm{P}}$ and $\hat{\Theta}^{\mathrm{S}}$ the estimators of any system invariant Θ using respectively the PBSID algorithm and the SSARX algorithm. Then, under standard assumptions* (*see, e.g.* Bauer, 2005a; Chiuso, 2006b) *on the innovation process* **e**

$$\hat{\Theta}^{\mathrm{P}} \doteq \hat{\Theta}^{\mathrm{S}} \tag{4.1}$$

holds.

We first state the following technical lemma which shall be useful in the proof of this result:

Lemma 4.2. *Let the pair* $(\mathbf{y}, \mathbf{u})$ *satisfy the assumptions of Section* 2. *Assume also the coefficients of the following two VARX models*

$$\mathbf{y}(t) = \sum_{i=1}^{K_1} \alpha_i \mathbf{z}(t-i) + \mathbf{e}_{K_1}(t) \tag{4.2}$$

and

$$\mathbf{y}(t) = \sum_{i=1}^{K_2} \beta_i \mathbf{z}(t-i) + \mathbf{e}_{K_2}(t) \tag{4.3}$$

are estimated (in the least square sense) from data $\{y_s, u_s\}$, *respectively, in the intervals* $s \in [t - K_1, t + N - 1]$ *and* $s \in [t - K_2, t + N - 1]$. *Assume also that* $K_1 \geqslant K_2 \geqslant K_{\min}$ *go to infinity with* N *while* K_1, $K_{\min}$ *satisfy*[16] *Assumption* 1. *Then for any fixed and finite* f

$$\hat{\alpha}_j \doteq \hat{\beta}_j, \quad j = 1, \ldots, f. \tag{4.4}$$

The same holds if the parameters in (4.2) *and* (4.3) *are estimated using data in the intervals* $[t_1 - K_1, t_1 + N - 1]$ *and* $[t_2 - K_2, t_2 + N - 1]$, *respectively, as long as* $t_1 - t_2$ *is fixed and finite.*

Proof. The proof follows from Eqs. (4.4) and (4.5) in Kuersteiner (2005), by letting $K_1 = h_{\max}$, $K_2 = \hat{h}_n$, $K_{\min} = h_{\min}$ and $l(h) = [0, \ldots, 0, 1, 0, \ldots, 0, \ldots]^\top$. Of course, in our case $P[K_2 \in [K_{\min}, K_1]] = 1$. Note that the result in Kuersteiner (2005) is much stronger and holds for more general linear combinations $l(h)$ and for data dependent order selection rules K_2 s.t. $P[K_2 \in [K_{\min}, K_1]] \to 1$ as $N \to \infty$. This is useful here since it makes it easy to extend our results also to the case in which the length of past and future horizons are estimated from data provided the conditions in Kuersteiner (2005) are still verified. However, we shall not discuss this extension here. □

Proof of Theorem 4.1. Our goal is essentially to show that $\hat{Y}^{\mathrm{S}}_{[t,T)}$ and $\hat{Y}^{\mathrm{P}}_{[t,T)}$ can be used interchangeably as far as asymptotic properties are concerned.

To this purpose, note that defining

$$\hat{\bar{H}}^{\mathrm{P}}_v := \begin{bmatrix} 0 & 0 & \ldots & 0 \\ \hat{\Psi}_{1,1} & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \hat{\Psi}_{v,v} & \ldots & \hat{\Psi}_{1,v} & 0 \end{bmatrix}.$$

$\hat{Y}^{\mathrm{P}}_{[t,T)}$ can be rewritten as

$$\hat{Y}^{\mathrm{P}}_{[t,T)} = \hat{E}[Y_{[t,T)} - \hat{\bar{H}}^{\mathrm{P}}_v Z_{[t,T)} \mid Z_{[t_0,t)}] \tag{4.5}$$

which has the same form as (3.9) provided $\hat{\bar{H}}^{\mathrm{P}}_v$ is substituted with $\hat{\bar{H}}^{\mathrm{S}}_v$. Using this observation we can write

$$\hat{Y}^{\mathrm{S}}_{[t,T)} - \hat{Y}^{\mathrm{P}}_{[t,T)} = (\hat{\bar{H}}^{\mathrm{P}}_v - \hat{\bar{H}}^{\mathrm{S}}_v)\hat{E}[Z_{[t,T)} \mid Z_{[t_0,t)}]. \tag{4.6}$$

It is obvious that, provided we can show that

$$\hat{\bar{H}}^{\mathrm{P}}_v \doteq \hat{\bar{H}}^{\mathrm{S}}_v, \tag{4.7}$$

using $\hat{Y}^{\mathrm{P}}_{[t,T)}$ in lieu of $\hat{Y}^{\mathrm{S}}_{[t,T)}$ does not change the asymptotic properties; in fact, under (4.7), also the difference $\hat{Y}^{\mathrm{S}}_{[t,T)} - \hat{Y}^{\mathrm{P}}_{[t,T)}$ will be $\underline{o}_P(1/\sqrt{N})$.

[16] Where the role of t_0 is played respectively by $t_0^1 := t - K_1$ and $t_0^{\min} := t - K_{\min}$.

Inspecting the structure of the matrices $\hat{\bar{H}}^{\mathrm{S}}_v$ and $\hat{\bar{H}}^{\mathrm{P}}_v$, it is rather simple to see that showing (4.7) is equivalent to prove that

$$\hat{\Phi}_i \doteq \hat{\Psi}_{i,h}, \quad i = 1, \ldots, h, \quad h = 1, \ldots, v. \tag{4.8}$$

Hence the last part of the proof shall be concerned with (4.8).

Let us fix for a moment $h = \bar{h}$. Showing that $\hat{\Phi}_i \doteq \hat{\Psi}_{i,\bar{h}}$ for $i = 1, \ldots, \bar{h}$ amounts to prove that the estimators ($\hat{\Phi}_i$ and $\hat{\Psi}_{i,\bar{h}}$, $i = 1, \ldots, \bar{h}$) of the first $\bar{h}$ coefficients of two long VARX models satisfying:

(a) the orders $T - t_0$ and $t - t_0 + \bar{h}$ differ of exactly $v - \bar{h}$ and both go to infinity at a rate specified by Assumption 1;
(b) the parameters are estimated essentially using the same data (essentially here means that there might be a finite number of data points which are used in one of the two and are not used in the other and vice versa) are asymptotically equivalent. This result has been formalized in Lemma 4.2 above. Repeated application of Lemma 4.2 to the VARX regressions (3.2) and (3.8) allows indeed to prove (4.8) and hence (4.7), from which the statement of Theorem 4.1 follows. □

We now report some simulation results concerning the equivalence of PBSID and SSARX. We consider two systems in innovation from (2.1) where the input given in closed loop

$$\mathbf{u}(t) = \mathbf{r}(t) - H_i(z)\mathbf{y}(t).$$

Example 1 is a first order ARMAX system with $A_1 = 0.7$, $B_1 = 1$, $K_1 = 1$, $C_1 = 1$, $D_1 = 0$, $\mathrm{Var}\{\mathbf{e}_1\} = 1$, with a proportional controller $H_1(z) = 1.5$ and white reference signal $\mathbf{r}(t) = 5\mathbf{n}(t)$ where $\mathbf{n}(t)$ is zero mean unit variance white noise uncorrelated from $\mathbf{e}(t)$.

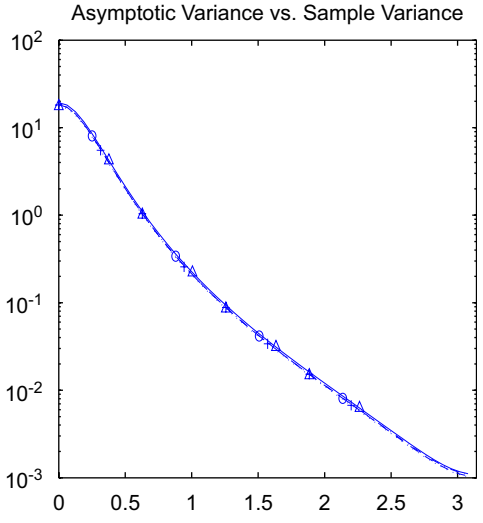

Fig. 1. Example 1. Sample variance (Monte Carlo estimate) vs. normalized frequency ($\omega \in [0, \pi]$). Solid with triangles (Δ): PEM. Dashed with crosses (+): PBSID. Dashed with circles (○): SSARX. Dotted with crosses (+): asymptotic variance for PBSID.

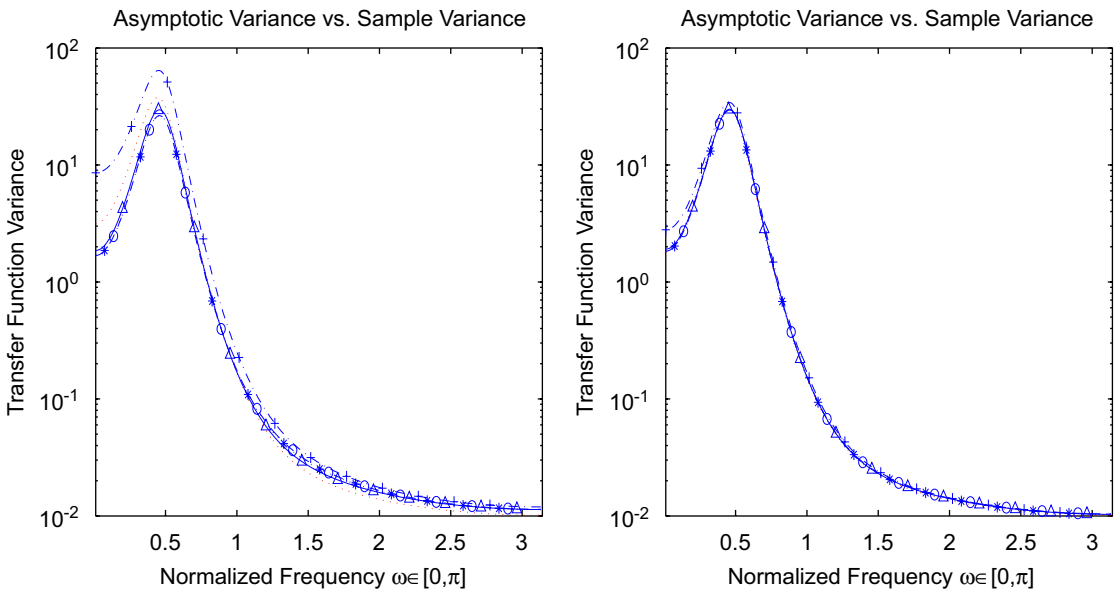

Fig. 2. Example 2. Left: "small" $t - t_0 = 10$. Right: "large" $t - t_0$. Variance (and its Monte Carlo estimate) vs. normalized frequency ($\omega \in [0, \pi]$). Solid with triangles (Δ) PEM. Dashed with crosses (+) PBSID. Dashed with circles ($\circ$) $\text{PBSID}_{\text{opt}}$. Dotted with stars ($*$): SSARX. Dotted: asymptotic variance for PBSID.

Table 1
Parameters chosen in the implementation

#	$t - t_0$	$T - t = v$	N
Fig. 1	10	10	1000
Fig. 2	10	10	1000
Fig. 2	30	10	3000

Example 2 is a second-order ARMAX system with

$$A_2 = \begin{bmatrix} 1.5 & -0.7 \\ 1 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad K_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$C_2 = [1 \;\; 0], \quad D_2 = 0, \quad \text{Var}\{\mathbf{e}_2\} = 1.$$

The reference signal is unit variance white noise uncorrelated with the innovation $\mathbf{e}(t)$ while the controller is a first order system of the from

$$H_2(z) = 0.2\frac{0.1z - 0.5}{z - 0.5}.$$

We compare the Monte Carlo estimate (500 Monte Carlo runs) of the transfer function estimator ($\hat{F}(z) := \hat{C}(zI - \hat{A})^{-1}\hat{B}$) variance (normalized by N) of SSARX and PBSID algorithms. The parameters chosen in the three simulation reported, respectively, in Figs. 1, 2 are summarized in Table 1. The results of Fig. 1 refer to Example 1 while those in Fig. 2 to Example 2. SSARX and PBSID are indistinguishable as predicted by the theory in this paper in Example 1. As far as Example 2 is concerned, while equivalence does not hold for small $t - t_0$ and N, it does indeed hold (see Fig. 2, right plot) when increasing N and $t - t_0$.

4.2. Vector AutoRegressive implementation of the $\text{PBSID}_{\text{opt}}$ method

The second main result of this paper shows that, indeed, the $\text{PBSID}_{\text{opt}}$ can be efficiently implemented via VARX estimation. Even though VARX models were introduced also in previous contributions, among which (Jansson, 2003; Ljung & McKelvey, 1996; Peternell, 1995), in our framework the VARX models pop up quite naturally from a theoretically sound "optimized" method. This consideration constitutes, in the author's opinion, a starting point for future investigations.

Theorem 4.3. *Consider the infinite VARX model*

$$y_t = \sum_{i=1}^{\infty} \Phi_i z_{t-i} + e_t \tag{4.9}$$

and denote with $\hat{\Phi}_i$, $i = 1, \ldots, t - t_0$, *the estimators of the first* $t - t_0$ *coefficients in* (4.9) *obtained solving*[17]

$$Y_t^{v+N} \simeq \sum_{i=1}^{t-t_0} \Phi_i Z_{t-i}^{v+N} \tag{4.10}$$

in the least squares sense.

The "optimally weighted" solution to (3.15), *i.e. the one that yields the least asymptotic variance of the estimators* $\hat{\Omega}^{\text{P}_{\text{opt}}}$ *among all linear, asymptotically unbiased estimators of* Ω^{P} *based on the regression* (3.15), *is equivalent to estimating the*

[17] Note that the estimators are function of N and $t - t_0$, which according to Assumption 1 grows with N. In order to streamline notation this dependence is not made explicit.

VARX model (4.10) *in the sense that*

$$\begin{bmatrix} \hat{\Xi}_0^{P_{opt}} \\ \hat{\Xi}_1^{P_{opt}} \\ \vdots \\ \hat{\Xi}_v^{P_{opt}} \end{bmatrix} = \begin{bmatrix} \hat{\Phi}_{t-t_0} & \cdots & \hat{\Phi}_{T-t_0} & \cdots & \hat{\Phi}_1 \\ 0 & \hat{\Phi}_{t-t_0} & \cdots & \cdots & \hat{\Phi}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \hat{\Phi}_{t-t_0} & \cdots & \hat{\Phi}_{v+1} \end{bmatrix} \tag{4.11}$$

and

$$\hat{\Psi}_{ji}^{P_{opt}} = \hat{\Phi}_i. \tag{4.12}$$

Proof. See the Appendix. □

Remark 4.4. 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 (3.11) is indeed Toeplitz (see Eq. (4.12)). 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 (3.2)) 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 Dahlén and Scherrer (2004), a more theoretically sound foundation to the (usually adopted) practice of determining $t - t_0$ using standard order selection criterions (Ljung (1997); Hannan & Deistler (1988); Söderström & Stoica (1989)) for vector autoregressive models (see, e.g. Bauer (2001)). The reader is also referred to the recent paper (Kuersteiner, 2005) which discusses automatic inference for infinite order autoregressions.

Using the result of Theorem 4.3 the PBSID$_{opt}$ algorithm can be implemented as follows:

(a) Estimate the VARX model (4.9) as described in (4.10); this may include estimation of the appropriate $t - t_0$ using standard criterions for VARX order estimation.
(b) Use the estimated coefficients as described in formulas (4.11) and (4.12) to form the predictors

$$\hat{Y}_{t+h}^{P_{opt}} = \sum_{i=h+1}^{t-t_0+h} \hat{\Psi}_{hi}^{P_{opt}} Z_{t+h-i} = \sum_{i=h+1}^{t-t_0} \hat{\Phi}_i Z_{t+h-i}; \tag{4.13}$$

the state sequences $\hat{X}_t^{P_{opt}}$ and $\hat{X}_{t+1}^{P_{opt}}$ are then obtained as described in formulas (3.17) and (3.18).

This implementation has a much lower computational complexity w.r.t. the implementation described in Chiuso (2007b) which involves solving the least squares problem (3.15) 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 (4.10) has complexity $O(N(\log N)^2)$ (see Golub & Van Loan, 1989, p. 248). The order and state estimation (step (b) above) can be performed on the "squared" version of the matrix $\hat{Y}_{[t,T)}^{P_{opt}}$. 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 v long VARX models, increasing the complexity of the first step roughly by a factor v.

Hence the implementation described above of the PBSID$_{opt}$ compares favorably to a variety of subspace procedures (among which PBSID or CCA) as far computational complexity is concerned while, according to Chiuso (2007b, Theorem 5.3), yielding lower asymptotic variance than CCA. We remind also that the PBSID$_{opt}$ 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).

Remark 4.5 (PBSID$_{opt}$ *vs. SSARX*). The main differences between the PBSID$_{opt}$ and SSARX algorithms are as follows: (i) the length of the ARX model estimated is (in general) different for the two methods; in particular SSARX uses order larger than v (but in Jansson, 2003 it is just required that the order be "high" to ensure consistency), possibly chosen according to infinite-order ARX models selection rules (Kuersteiner, 2005) only the first v coefficients are then used; instead PBSID$_{opt}$ the order is exactly $t - t_0$; this results in the PBSID$_{opt}$ filling with zeros the Toeplitz matrix used to construct the bank of predictors (see Eq. (4.11)); (ii) the SSARX methods projects the "corrected future" to form $\hat{Y}_{[t,T)}^{S}$ (see Eqs. (3.9)) while the PBSID$_{opt}$ uses directly the estimated coefficients from the VARX modeling step to form the bank of predictors $\hat{Y}_{t+k}^{P_{opt}}$ (see eq. (3.17) and (4.11)). This makes PBSID$_{opt}$ even more advantageous from the computational point of view, since it does not require computing the projection (3.9).

We consider the following examples, frequently used in the literature of subspace identification, to illustrate the result.

The first is an"open loop" experiment which contains all the essential features of the "optimized" method i.e.: (i) it is not efficient (it does not reach Cramér–Rao) 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.

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 $H_u(z)$

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

the input spectrum is plotted in Fig. 3.

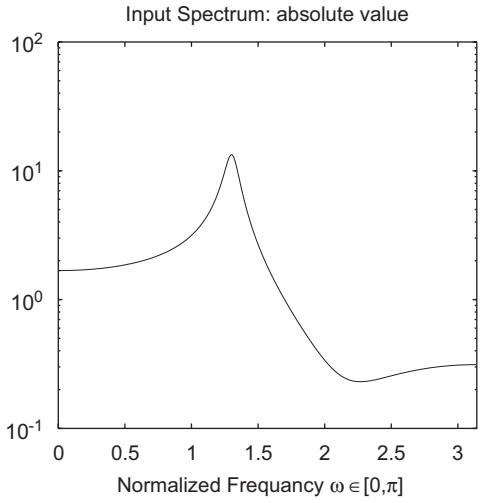

Fig. 3. Colored input spectrum: absolute value.

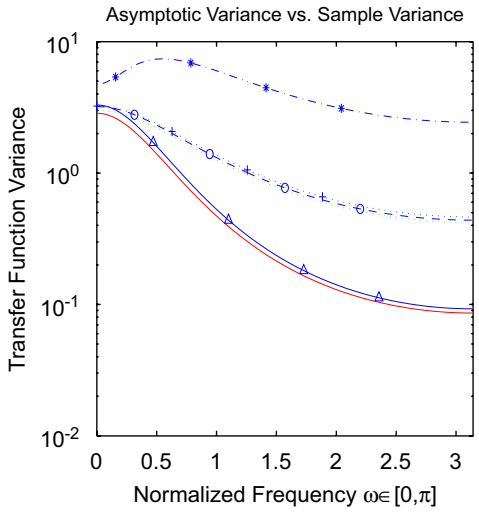

Fig. 4. Asymptotic variance (and its Monte Carlo estimate) vs. normalized frequency ($\omega \in [0, \pi]$) (*ARMAX of order 1*). Solid with triangles (Δ) PEM, dashed-dotted with stars ($*$): CCA, dotted with crosses ($+$): (PBSID), dashed with circles ($\circ$): $\text{PBSID}_{\text{opt}}$; dotted: asymptotic variance for PBSID, solid: Cramér–Rao lower bound.

We report in Fig. 4 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 in each experiment of the deterministic transfer function $F(z) = 1/(z - 0.5)$.

As a second example we consider a fifth-order (marginally stable) system in state space form (2.1) where (see Van Overschee & De Moor, 1997; Verhaegen, 1993)

$$A = \begin{bmatrix} 4.40 & 1 & 0 & 0 & 0 \\ -8.09 & 0 & 1 & 0 & 0 \\ 7.83 & 0 & 0 & 1 & 0 \\ -4 & 0 & 0 & 0 & 1 \\ 0.86 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad C = [1 \quad 0 \quad 0 \quad 0 \quad 0],$$

$$B = [0.00098 \quad 0.01299 \quad 0.01859 \quad 0.0033 \quad -0.00002]^{\top},$$

$$K = [2.3 \quad -6.64 \quad 7.515 \quad -4.0146 \quad 0.86336]^{\top}, \quad D = 0$$

and $\mathbf{e}(t)$ is unit variance white noise. The input $\mathbf{u}$ is generated in closed loop by $\mathbf{u}(t) = 5\mathbf{r}(t) - H(z)\mathbf{y}(t)$; the reference signal $\mathbf{r}(t)$ is unit variance white noise; $H(z)$ is given by

$$H(z) = \frac{0.63 - 2.083z^{-1} + 2.8222z^{-2} - 1.865z^{-3} + 0.4978z^{-4}}{1 - 2.65z^{-1} + 3.11z^{-2} - 1.75z^{-3} + 0.39z^{-4}}.$$

We have also used $t - t_0 = 30$, $v = 10$ and $N = 2000$.

In this example $\text{PBSID}_{\text{opt}}$ outperforms PBSID in the low-frequency band and performs slightly better than the innovation estimation method (IEM hereafter) (Qin & Ljung, 2003b) in the high-frequency band (see Fig. 4). The algorithm by Ljung and McKelvey, which as shown in the next section is a weighted version of $\text{PBSID}_{\text{opt}}$, performs worse than $\text{PBSID}_{\text{opt}}$ and IEM (Qin & Ljung, 2003b).

It has been checked that the original algorithm presented in Chiuso (2007b) and its alternative implementation presented in this paper give indeed the same result. In particular conditions (4.11) and (4.12) have been verified to hold for the estimated coefficients of the $\text{PBSID}_{\text{opt}}$ described in Chiuso (2007b).

4.3. Relation with the method by Ljung and McKelvey

In this section we shall briefly discuss the relation of $\text{PBSID}_{\text{opt}}$ with the algorithm presented in Ljung and McKelvey (1996). We shall not enter into a detailed description of the algorithm for which we refer the reader to the original paper; our description of the algorithm follows the Matlab code provided in Ljung and McKelvey (1995).

Suffices here to say that the first step is to construct a matrix $\hat{Y}^{LK}_{[t,T)}$ formed with predictors from which a basis of the state space is extracted; $\hat{Y}^{LK}_{[t,T)}$ shall play the same role as $\hat{Y}^{\text{P}_{\text{opt}}}_{[t,T)}$ and $\hat{Y}^{\text{S}}_{[t,T)}$ in $\text{PBSID}_{\text{opt}}$ and SSARX, respectively.

The main result can be stated as follows:

Proposition 4.6. *Assume the model orders n_a and n_b in* (4.14) *are chosen according to $n_a = n_b = t - t_0$ and the VARX coefficients $H_i := [H_{y,i} \;\; H_{u,i}]$ in* (4.14) *are estimated letting $\hat{H}_i := \hat{\Phi}_i$ where $\hat{\Phi}_i$ are the least squares solution of* (4.10). *Then the algorithm proposed in Ljung and McKelvey* (1996) *is a weighted version of* $\text{PBSID}_{\text{opt}}$, *in the sense that the two state construction steps differ only for the choice of a (row) weighting matrix W_{LK}, as made precise by formula* (4.20).

Proof. Consider the VARX model

$$\hat{Y}_{t|t-1} = \sum_{i=1}^{n_a} \hat{H}_{y,i} Y_{t-i} + \sum_{i=1}^{n_b} \hat{H}_{u,i} U_{t-i}. \tag{4.14}$$

Essentially the algorithm in Ljung and McKelvey (1996) construct the state space using a bank of predictors[18]

$$\hat{Y}^{LK}_{[t,T)} := [\hat{Y}^{\top}_{t|t-1}\ \hat{Y}^{\top}_{t+1|t-1}\ \dots\ \hat{Y}^{\top}_{T|t-1}]^{\top} \tag{4.15}$$

where $\hat{Y}_{t+k|t-1}$ is computed recursively as

$$\hat{Y}_{t+k|t-1} := \sum_{i=1}^{k} \hat{H}_{y,i}\hat{Y}_{t+k-i|t-1} + \sum_{i=k+1}^{n_a} \hat{H}_{y,i}Y_{t+k-i} + \sum_{i=k+1}^{n_b} \hat{H}_{u,i}U_{t+k-i}. \tag{4.16}$$

The remaining steps (i.e. state construction and estimation of A, B, C, K) follow the same lines as described in the previous Sections.

In order to make clear the link between the "predictor" used in $\text{PBSID}_{\text{opt}}$ and $\hat{Y}^{LK}_{[t,T)}$, we rewrite (4.16) as follows:

$$\hat{Y}_{t+k|t-1} - \sum_{i=1}^{k} \hat{H}_{y,i}\hat{Y}_{t+k-i|t-1} = \sum_{i=k+1}^{n_a} \hat{H}_{y,i}Y_{t+k-i} + \sum_{i=k+1}^{n_b} \hat{H}_{u,i}U_{t+k-i}. \tag{4.17}$$

Using the assumption that $n_a = n_b = t - t_0$, letting $\hat{H}_i := [\hat{H}_{y,i}\ \hat{H}_{u,i}]$ and defining

$$W_{LK} := \begin{bmatrix} I & 0 & \dots & 0 \\ -\hat{H}_{y,1} & I & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ -\hat{H}_{y,v} & \dots & -\hat{H}_{y,1} & I \end{bmatrix}, \tag{4.18}$$

Eq. (4.17) can be rewritten in matrix form as follows:

$$W_{LK}\hat{Y}^{LK}_{[t,T)} = \begin{bmatrix} H_{t-t_0} & \dots & \dots & \dots & \dots & H_1 \\ 0 & H_{t-t_0} & \dots & \dots & \dots & H_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & H_{t-t_0} & \dots & H_{v+1} \end{bmatrix} Z_{[t_0,t)}, \tag{4.19}$$

from the assumption that the VARX model (4.14) has been estimated using the same data as (4.9) it also follows that $\hat{H}_i = \hat{\Phi}_i$. Therefore, using (4.19) the stacked predictors in (4.15) can be rewritten as

$$\hat{Y}^{LK}_{[t,T)} = W^{-1}_{LK} \begin{bmatrix} \hat{\Phi}_{t-t_0} & \dots & \dots & \dots & \dots & \hat{\Phi}_1 \\ 0 & \hat{\Phi}_{t-t_0} & \dots & \dots & \dots & \hat{\Phi}_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \hat{\Phi}_{t-t_0} & \dots & \hat{\Phi}_{v+1} \end{bmatrix} Z_{[t_0,t)} = W^{-1}_{LK}\hat{Y}^{\text{P}_{\text{opt}}}_{[t,T)} \tag{4.20}$$

where (4.13) has been used in the last equality. □

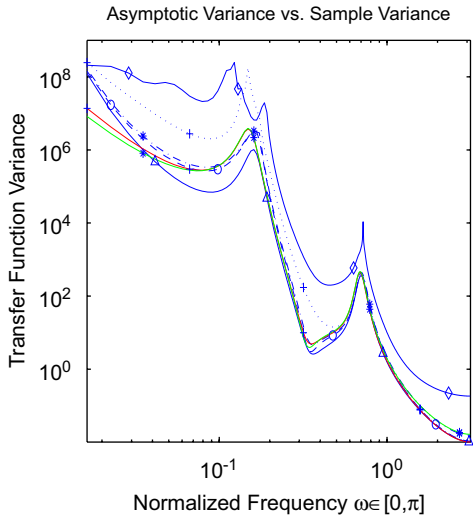

Fig. 5. Asymptotic variance (and its Monte Carlo estimate) Dashed-dotted with circles (○): $\text{PBSID}_{\text{opt}}$. Dashed-dotted with stars (∗): IEM (Qin & Ljung, 2003b). Dotted with crosses (+): PBSID. Solid with diamonds (◇): Ljung–McKelvey Ljung and McKelvey (1996). Dotted with triangles (△): PEM. Solid with stars (green) (∗): asymptotic variance for IEM. Dotted with crosses (red) (+): asymptotic variance for PBSID.

It is remarkable that the bank of predictors used in the original paper (Ljung & McKelvey, 1995) is indeed equivalent to a weighted version of the bank of predictors used in $\text{PBSID}_{\text{opt}}$. It should not be surprising that the original algorithm in Ljung and McKelvey (1995, 1996) does not perform as well as $\text{PBSID}_{\text{opt}}$ (see Fig. 5); in fact it is well known that the (row) weighting does affect the asymptotic statistical properties of the estimators using the "state sequence" approach (see Bauer, 2005a). Note that in Qin and Ljung (2006) it was conjectured (even though not proved) that an algorithm named HOARX is equivalent (asymptotically) to the algorithm by Ljung and McKelvey. As shown in this section (see, in particular, the right-hand side of (4.19) and formula (8) in Qin & Ljung (2006)) this is not exactly true; instead they differ up to a row weighting. To be precise also some Markov parameters (those weighting the "far past") are set to zero (see (4.19)). Using the techniques in Kuersteiner (2005) it would be possible to see that this does not make any difference asymptotically as long as v remains fixed (or bounded above).

[18] This implementation has been taken from the Matlab code reported in Ljung and McKelvey (1995).

5. Conclusions

In this paper we have discussed several subspace algorithms based on predictor model identification. It is shown that all these algorithms can be formulated as VARX estimation followed by model reduction.

In particular it has been shown that

(a) SSARX (Jansson, 2003) and PBSID (Chiuso & Picci, 2005a) are asymptotically equivalent;
(b) $\text{PBSID}_{\text{opt}}$ (Chiuso, 2007b) is exactly equivalent (i.e. give the same numerical results on finite data) to estimating a suitable VARX model followed by the usual steps of subspace identification (i.e. state estimation via SVD followed by estimation of the system matrices)
(c) The algorithm by Ljung and McKelvey (1995) is a weighted version of $\text{PBSID}_{\text{opt}}$.

Experimental results (on simulated data) are included which support the theoretical derivations. The results of this paper, together with the comparison performed in Chiuso (2007b), indicate that $\text{PBSID}_{\text{opt}}$ should be considered as one of the most appealing in this class of algorithms for the following reasons:

(a) it is consistent under closed-loop operating conditions;
(b) it performs no worse than CCA (regardless of the choice of input) for open loop data and better than SSARX/PBSID with feedback;
(c) it is prone to a very simple and computationally attractive implementation via VARX modeling.

The simulation results reported in this paper seem to support these statements. Even though $\text{PBSID}_{\text{opt}}$ can be verified to be asymptotically efficient in a number of examples,[19] it is not so in general.

It was conjectured in Larimore (2004) that an algorithm which is essentially equivalent to PBSID is (asymptotically) efficient for large v (actually for $t - t_0 = v \to \infty$). Instead, as clearly seen in Fig. 4 PBSID is not efficient for large v. We have verified that indeed the performance does not change increasing v. Instead PBSID is nearly efficient for this example with $v = 1$ (see Chiuso, 2007b, Fig. 4 and Chiuso, 2007a).

There is certainly much work to be done; in particular not completely clear is, at the moment, the relation of these methods with the new algorithms introduced in Onodera et al. (2006), Qin and Ljung (2006) and with the IEM of Qin and Ljung (2003b).

Also the question of finite-data behavior is certainly of interest and deserves, in our opinion, further investigation.

Acknowledgments

The author would like to thank Manfred Deistler, Lennart Ljung, Giorgio Picci and Joe Qin for fruitful discussions on the subject. Also an anonymous reviewer of the paper (Chiuso, 2007b) is gratefully acknowledged for comments and suggestions contained in his report.

[19] This might require that also v grows with N. However, the analysis in this paper deals only with the case of fixed v.

Appendix A. Proofs

Proof of Theorme 4.3. The proof makes use of the fine structure of the matrix L. Let us denote with L_I a matrix which columns span the image of L and with L_K a matrix spanning the left kernel of L, so that $[L_I\, L_K]$ is a full rank square $(pN(v+1) \times pN(v+1))$ matrix. The least squares problem (3.15) can be transformed into the equivalent form

$$\begin{bmatrix} L_I^\top \\ L_K^\top \end{bmatrix} Y = \begin{bmatrix} L_I^\top \\ L_K^\top \end{bmatrix} S^{\text{P}} \Omega^{\text{P}} + \begin{bmatrix} L_I^\top \\ L_K^\top \end{bmatrix} E. \tag{A.1}$$

Note that, by construction, $L_I^\top L$ has full rank and therefore $L_I^\top E$ has full rank covariance. Similarly $L_K^\top L = 0$ and hence $L_K^\top E = 0$ (in the mean square sense).

In this way the least squares problem (3.15) with singular noise covariance (3.15) is transformed into a least squares problem with full rank noise covariance (the "top" part of (A.1)) and equality constraints (the "bottom" part of (A.1)).

It is easy to show that L_I can be chosen to be a selection matrix so that $L_I^\top Y = \text{vec}(Y_t^{N+v}) := Y_I$. For future reference observe that $L_I^\top E = \text{vec}(E_t^{N+v}) := E_I$ so that (A.1) can be rewritten as

$$\begin{aligned} &Y_I = L_I^\top S^{\text{P}} \Omega^{\text{P}} + E_I, \\ &\text{s.t. } L_K^\top Y = L_K^\top S^{\text{P}} \Omega^{\text{P}}. \end{aligned} \tag{A.2}$$

Let us introduce the pair of indexes $(j, \bar{j})$ such that $v \geqslant \bar{j} > j \geqslant 0$ and define $\delta := \bar{j} - j$. Then it is easy to see that there exist matrices $L_K(j, \bar{j}, l)$ so that

$$0 = L_{K,}^\top(j, \bar{j}, l) Y = \Xi_j \begin{bmatrix} z_{t_0+\delta+l} \\ z_{t_0+\delta+1+l} \\ \vdots \\ z_{t+\delta-1+l} \end{bmatrix} - \Xi_{\bar{j}} \begin{bmatrix} z_{t_0+l} \\ z_{t_0+1+l} \\ \vdots \\ z_{t-1+l} \end{bmatrix}$$
$$+ \sum_{k=1}^{j} \Psi_{jk} z_{t+\delta-1+k+l} - \sum_{k=1}^{\bar{j}} \Psi_{\bar{j}k} z_{t-1+k+l}, \tag{A.3}$$

where for each pair $(j, \bar{j})$ with $\bar{j} > j \in [0, v-1]$, l ranges in the interval $[0, N - \delta - 1]$.

It is possible to extract exactly $Nv - (v+1)$ independent constraints (recall that L_K has rank $p(Nv - (v+1))$) of the form (A.3) by letting $j \in [0, v-1]$, $\bar{j} = j' := j + 1$, (so that $\delta = 1$) and $l \in [0, N-2]$. With these choices the constraints (A.3) can be written in the form

$$[\Xi_j \ \Psi_{j1} \ \dots \ \Psi_{jj}] Z_{[t_0+\delta, t+\delta+j-1)}^{N-1}$$
$$= [\Xi_{j'} \ \Psi_{j'1} \ \dots \ \Psi_{j'j'}] Z_{[t_0, t+\bar{j}-1)}^{N-1}. \tag{A.4}$$

Recalling that $\delta = j' - j = 1$, and defining 0_δ to be the zero matrix of size $p \times \delta(p+m)$, we can rewrite (A.4) in the form

$$[0_1 \;\; \Xi_j \;\; \Psi_{j1} \;\; \dots \;\; \Psi_{jj}] Z^{N-1}_{[t_0, t+j'-1)}$$
$$[Xi_{j'} \;\; \Psi_{j'1} \;\; \dots \;\; \Psi_{j'j'}] Z^{N-1}_{[t_0, t+j'-1)}. \tag{A.5}$$

From the assumption that the joint spectrum is coercive, it follows that, for N large enough (i.e. and hence $N-1$ large), the matrix $Z^{N-1}_{[t_0, t+j'-1)}$ is of full row rank for all possible choices of j; therefore (A.5) is equivalent to the "dual" equation for the coefficients:

$$[0_1 \;\; \Xi_j \;\; \Psi_{j1} \;\; \dots \;\; \Psi_{jj}] = [\Xi_{j'} \;\; \Psi_{j'1} \;\; \dots \;\; \Psi_{j'j'}].$$

As mentioned above this should hold for each pair (j, j'), $v > j \geqslant 0$; this is equivalent to the following constraints on the estimated coefficients:

$$\begin{bmatrix} \hat{\Xi}_1^{P_{opt}} & \hat{\Psi}_{11}^{P_{opt}} \end{bmatrix} = \begin{bmatrix} 0_1 & \hat{\Xi}_0^{P_{opt}} \end{bmatrix},$$
$$\begin{bmatrix} \hat{\Xi}_2^{P_{opt}} & \hat{\Psi}_{22}^{P_{opt}} & \hat{\Psi}_{21}^{P_{opt}} \end{bmatrix} = \begin{bmatrix} 0_2 & \hat{\Xi}_0^{P_{opt}} \end{bmatrix},$$
$$\vdots$$
$$\begin{bmatrix} \hat{\Xi}_v^{P_{opt}} & \hat{\Psi}_{vv}^{P_{opt}} & \dots & \hat{\Psi}_{v1}^{P_{opt}} \end{bmatrix} = \begin{bmatrix} 0_v & \hat{\Xi}_0^{P_{opt}} \end{bmatrix}. \tag{A.6}$$

For convenience, let us define $\Xi_0 = [\Phi_{t-t_0} \;\; \dots \;\; \Phi_2 \;\; \Phi_1]$. Using the constraints above, some extra algebra will show that (A.2) can be written in the form

$$Y_I = \begin{bmatrix} \Phi_{t-t_0} & \dots & \Phi_1 & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \dots & 0 & \Phi_{t-t_0} & \dots & \Phi_1 \end{bmatrix} \mathrm{vec}(Z_{t_0}^{N+v}) + E_I. \tag{A.7}$$

A more compact expression of (A.7) is

$$Y_t^{N+v} = [\Phi_{t-t_0} \;\; \dots \;\; \Phi_2 \;\; \Phi_1] Z^{N+v}_{[t_0,t)} + E_t^{N+v}. \tag{A.8}$$

The "optimal" (Markov) solution to (A.7) is obtained by pre-whitening the residual vector E_I, which can be obtained pre-multiplying by $(I \otimes \Lambda^{-1/2})$ both sides of (A.7) or, equivalently, pre-multiplying both sides of (A.8) by $\Lambda^{-1/2}$. It is a simple calculation to check that, indeed, solving in the least squares sense

$$\Lambda^{-1/2} Y_t^{N+v} \simeq \Lambda^{-1/2} [\Phi_{t-t_0} \;\; \dots \;\; \Phi_2 \;\; \Phi_1] Z^{N+v}_{[t_0,t)}$$

is equivalent to solving (4.10); this implies that $\hat{\Xi}_0^{P_{opt}} = [\hat{\Phi}_{t-t_0} \;\; \dots \;\; \hat{\Phi}_2 \;\; \hat{\Phi}_1]$; conditions (4.11) and (4.12) can then be obtained using the constraints in (A.6), which completes the proof. □

References

Bauer, D. (2001). Order estimation for subspace methods. *Automatica, 37*, 1561–1573.

Bauer, D. (2005a). Asymptotic properties of subspace estimators. *Automatica, 41*, 359–376.

Bauer, D. (2005b). Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs. *Journal of Time Series Analysis*, *26*, 631–668.

Bauer, D., & Jansson, M. (2000). Analysis of the asymptotic properties of the MOESP type of subspace algorithms. *Automatica*, *36*, 497–509.

Bauer, D., & Ljung, L. (2002). Some facts about the choice of the weighting matrices in Larimore type of subspace algorithm. *Automatica*, *38*, 763–773.

Chiuso, A. (2007a). Some insights on the choice of the future horizon in CCA-type subspace algortihms, *Proc. of the ACC 2007*, to appear, available at ⟨http://www.dei.unipd.it/∼chiuso⟩.

Chiuso, A. (2007b). On the relation between CCA and predictor based subspace identification, *IEEE Transactions on Automatic Control*, to appear, available at ⟨http://www.dei.unipd.it/∼chiuso⟩.

Chiuso, A. (2006a). Asymptotic equivalence of certain closed-loop subspace identification methods. *Proceedings of SYSID, 2006* Australia: Newcastle.

Chiuso, A. (2006b). Asymptotic variance of closed-loop subspace identification algorithms. *IEEE Transactions on Automatic Control*, *51*(8), 1299–1314.

Chiuso, A. (2006c). The role of vector autoregressive modeling in subspace identification. *Proceedings of CDC 2006*, San Diego, USA, December 2006.

Chiuso, A.,& Picci, G. (2003a). Constructing the state of random processes with feedback. *Proceedings of the IFAC international symposium on system identification (SYSID)*, Rotterdam, August 2003.

Chiuso, A., & Picci, G. (2003b). Geometry of oblique splitting, minimality and hankel operators. *Lecture notes in control and information sciences*, (Vol. 286, pp. 85–124). Berlin: Springer.

Chiuso, A., & Picci, G. (2004a). The asymptotic variance of subspace estimates. *Journal of Econometrics*, *118*(1–2), 257–291.

Chiuso, A., & Picci, G. (2004b). Asymptotic variance of subspace methods by data orthogonalization and model decoupling: A comparative analysis. *Automatica*, *40*(10), 1705–1717.

Chiuso, A., & Picci, G. (2004c). On the ill-conditioning of subspace identification with inputs. *Automatica*, *40*(4), 575–589.

Chiuso, A., & Picci, G. (2005a). Consistency analysis of some closed-loop subspace identification methods. *Automatica*, *41*(3), 377–391.

Chiuso, A., & Picci, G. (2005b). Prediction error vs. subspace methods in closed-loop identification, *Proceedings of the 16th IFAC world congress*, Prague, July 2005.

Dahlén, A., & Scherrer, W. (2004). The relation of CCA subspace method to a balanced reduction of an autoregressive model. *Journal of Econometrics*, *118*(1–2), 293–312.

Ferguson, T. (1996). *A course in large sample theory*. London: Chapman and Hall.

Golub, G. H., & Van Loan, C. R. (1989). *Matrix computation*. 2nd ed., Baltimore, MD: The Johns Hopkins University Press.

Granger, C. W. J. (1963). Economic processes involving feedback. *Information and Control*, *6*, 28–48.

Hannan, E. J., & Deistler, M. (1988). *The statistical theory of linear systems*. New York: Wiley.

Hong, Z., Harmse, M., Guiver, J., & Canney, W. (2006). Subspace identification in industrial MPC applications—a review of recent progress and industrial experience (i). *Proceedings of SYSID 2006*, Newcastle, Australia.

Jansson, M. (2000). Asymptotic variance analysis of subspace identification methods. *Proceedings of SYSID 2000* CA: Santa Barbara.

Jansson, M. (2003). Subspace identification and ARX modeling. *Proceedings of SYSID 2003*, Rotterdam.

Jansson, M. (2005). A new subspace identification method for open and closed loop data. *Proceedings of the 16th IFAC World Congress*, Prague.

Katayama, T., Kawauchi, H., & Picci, G. (2005). Subspace identification of closed loop systems by orthogonal decomposition. *Automatica, 41*, 863–872.

Katayama, T., Tanaka, H., & Enomoto, T. (2005). A simple subspace identification method of closed-loop systems using orthogonal decomposition. *Proceedings of the 16th IFAC world congress*, Prague.

Kuersteiner, G. M. (2005). Automatic inference for infinite order vector autoregressions. *Econometric Theory*, *21*, 85–115.

Larimore, W.E. (1983). System identification, reduced-order filtering and modeling via canonical variate analysis. *Proceedings of the American control conference* (pp. 445–451).

Larimore, W. E. (2004). Large sample efficiency for ADAPTX subspace identification with unknown feedback. *Proceedings of IFAC DYCOPS'04* MA, USA: Boston.

Lindquist, A., & Picci, G. (1996). Canonical correlation analysis, approximate covariance extension and identification of stationary time series. *Automatica*, *32*, 709–733.

Ljung, L. (1997). *System identification, theory for the user*. Englewood Chiffs, NJ: Prentice Hall.

Ljung, L., & McKelvey, T. (1995). Subspace identification from closed loop data. *Technical report* LiTH-ISY-R-1752.

Ljung, L., & McKelvey, T. (1996). Subspace identification from closed loop data. *Signal Processing*, *52*(2), 209–216.

Onodera, K., Emoto, G., & Qin, S.J. (2006). A new subspace identification method for closed loop systems. *Proceedings of SYSID 2006*, Newcastle, Australia.

Peternell, K. (1995). Subspace methods for subspace identification, *Ph.D. thesis*, Technical University of Vienna.

Peternell, K., Scherrer, W., & Deistler, M. (1996). Statistical analysis of novel subspace identification methods. *Signal Processing*, *52*, 161–178.

Porat, B., & Friedlander, B. (1985). Asymptotic accuracy of ARMA parameter estimation methods based on sample covariances. *Proceedings of the seventh IFAC SYSID*, York.

Qin, S.J., & Ljung, L. (2003a). Closed-loop subspace identification with innovation estimation. *Proceedings of SYSID 2003*, Rotterdam.

Qin, S.J., & Ljung, L. (2003b). Parallel QR implementation of subspace identification with parsimonious models. *Proceedings of SYSID 2003*, Rotterdam.

Qin, S.J., & Ljung, L. (2006). On the role of future horizon in closed-loop subspace identification. *Proceedings of SYSID 2006*, Newcastle, Australia.

Rao, C. R. (1973). Representations of the best linear unbiased estimators in the Gauss–Markov model with a singular dispersion matrix. *Journal of Multivariate Analysis*, *3*, 276–292.

Shi, F., & MacGregor, J.F. (2001). A framework for subspace identification. *Proceedings of IEEE ACC*, Arlington, VA.

Söderström, T., & Stoica, P. (1989). *System identification*. Englewood Cliffs, NJ: Prentice-Hall.

Van Overschee, P., & De Moor, B. (1993). Subspace algorithms for the stochastic identification problem. *Automatica*, *29*, 649–660.

Van Overschee, P., & De Moor, B. (1994). N4SID: Subspace algorithms for the identification of combined deterministic—stochastic systems. *Automatica*, *30*, 75–93.

Van Overschee, P., & De Moor, B. (1997). Closed loop subspace systems identification. *Proceedings of 36th IEEE conference on decision and control*, San Diego, CA (pp. 1848–1853).

Verhaegen, M. (1993). Application of a subspace model identification technique to identify lti systems operating in closed-loop. *Automatica*, *29*, 1027–1040.

Wahlberg, B. (1989). Estimation of autoregressive moving-average models via high order autoregressive approximations. *Journal of Time Series Analysis*, *10*(3), 283–299.

Walker, A. M. (1961). Large-sample estimation of parameters for moving average models. *Biometrika*, *48*(3/4), 343–357.

Werner, H. J., & Yapar, C. (1996). On inequality constrained generalized least squares selections in the general possibly singular Gauss–Markov model: A projector theoretical approach. *Linear Algebra and its Applications*, *237/238*(1–3), 359–393 (Special issue honoring Calyampudi Radhakrishna Rao).

Zhu, Y. (2006). *System identification for process control: Recent experiences and outlook*. Newcastle, Australia: Plenary Lecture delivered at SYSID 2006.

Alessandro Chiuso currently holds an Associate Professor position with the Dipartimento di Tecnica e Gestione dei Sistemi Industriali, Università di Padova.

He received the "Laurea" degree summa cum laude in Telecommunication Engineering from the University of Padova in July 1996 and the Ph.D. degree (Dottorato di ricerca) in System Engineering from the University of Bologna in 2000. From March 2001 to February 2006 he has been "Ricercatore" (Assistant Professor) with the Dipartimento di Ingegneria dell' Informazione, Università di Padova. He has held visiting positions with the Deparment of Electrical Engineering, Washington University, St. Louis (USA) (visiting research scholar, August 1998–June 1999), with the Department Mathematics, Royal Institute of Technology, Sweden (EU-TMR Post-Doctoral fellow, March 2000–July 2000), with the Department of Computer Science, University of California Los Angeles (Visiting Researcher, July 2001).

Dr. Chiuso is member of the Editorial Board of IET Control Theory and Applications and serves as an Associate Editor on the IEEE Control System Society Conference Editorial Board since 2004; he is also member of the IEEE CSS Technical Committee on Identification and Adaptive Control and a reviewer for Mathematical Reviews of the AMS.

His research interests are mainly in Estimation, Identification Theory and Applications (subspace methods, stochastic realization, non-linear estimation, hybrid systems), Computer Vision (structure from motion, texture and gait analysis). Further information can be found at the personal web page http://www.dei.unipd.it/∼chiuso.