Available online at www.sciencedirect.com

ScienceDirect

Computers and Chemical Engineering 30 (2006) 1502–1513

Computers & Chemical Engineering

www.elsevier.com/locate/compchemeng

An overview of subspace identification

S. Joe Qin*

Department of Chemical Engineering, The University of Texas at Austin, Austin, TX 78712, USA

Received 20 February 2006; received in revised form 30 May 2006; accepted 31 May 2006

Abstract

This paper provides an overview of the state of the art of subspace identification methods for both open-loop and closed-loop systems. Practical considerations and future directions are given at the end of the paper.

Keywords: Subspace identification; Closed-loop identification; Overview; State space models; Causal model; Practice

1. Introduction

Subspace identification methods (SIM) have enjoyed tremendous development in the last 15 years in both theory and practice. SIMs offer an attractive alternative to input-output methods due to simple and general parametrization for MIMO systems (there is no linear input-output parametrization that is general enough for all linear MIMO systems, see (Katayama, 2005)). Most SIMs fall into the unifying theorem proposed by van Overschee and de Moor (1995), among which are canonical variate analysis (CVA) (Larimore, 1990), N4SID (van Overschee & de Moor, 1994), subspace splitting (Jansson & Wahlberg, 1996), and MOESP (Verhaegen & Dewilde, 1992). Based on the unifying theorem, all these algorithms can be interpreted as a singular value decomposition of a weighted matrix. The statistical properties such as consistency and efficiency have been investigated recently (Bauer, 2003; Bauer & Ljung, 2002; Gustafsson, 2002; Jansson & Wahlberg, 1998; Knudsen, 2001).

The closed-loop identification is of special interest for a large number of engineering applications. For safety reasons or quality restrictions, it is desirable that identification experiments are carried out under the closed-loop or partial closed-loop condition. As pointed out by many researchers (Ljung, 1999; Soderstrom & Stoica, 1989), the fundamental problem with closed-loop data is the correlation between the unmeasurable noise and the input. This is true for traditional closed-loop identification approaches such as the prediction error methods (PEMs) (Forssell & Ljung, 1999). It causes additional difficulty for SIMs.

Although SIM algorithms are attractive because of the state space form that is very convenient for estimation, filtering, prediction and control, several drawbacks have been recognized. In general, the estimates from SIMs are not as accurate as those from prediction error methods. Further, it is not until recently some SIMs are applicable to closed-loop identification, even though the data satisfy identifiability conditions for traditional methods such as PEMs.

Unlike PEMs, the traditional SIMs (e.g., CVA, N4SID and MOESP) are biased under closed-loop condition, which requires special treatment. Verhaegen (1993) proposed a closed-loop SIM via the identification of an overall open-loop state space model followed by a model reduction step to obtain state space representations of plant and controller. Ljung and McKelvey (1996) investigated the SIM through the classical realization path and proposed a recursive approach based on ARX model as a feasible closed-loop SIM. Formulated in an errors-in-variables (EIV) framework, Chou and Verhaegen (1997) proposed a new SIM that can be applied to closed-loop data. The algorithm has to treat white input from non-white input differently. Wang and Qin (2002) proposed the use of parity space and principal component analysis (PCA) for EIV and closed-loop identification which is applicable to correlated input excitation. Recent work of Qin and Ljung (2003a), Jansson (2003), and Chiuso and Picci (2005) analyzed SIMs with feedback, proposed several new closed-loop SIMs and provided theoretical analysis to these methods.

The purpose of this paper is to provide an overview of the state of the art in both open-loop and closed-loop SIMs. The

* Tel.: +1 512 471 4417; fax: +1 512 471 7060.
E-mail address: qin@che.utexas.edu.

0098-1354/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2006.05.045

paper starts with basic stochastic system representations and assumptions, then reviews most existing SIMs in the literature to date. Practical considerations and future directions are given to conclude the paper.

2. Models, notations, and assumptions

2.1. Stochastic state space models

A stochastic linear system can be written in the following *process* form,

$$x_{k+1} = Ax_k + Bu_k + w_k \tag{1a}$$

$$y_k = Cx_k + Du_k + v_k \tag{1b}$$

where $y_k \in R^{n_y}$, $x_k \in R^n$, $u_k \in R^{n_u}$, $w_k \in R^n$, and $v_k \in R^{n_y}$ are the system output, state, input, state noise, and output measurement noise, respectively. A, B, C and D are system matrices with appropriate dimensions.

It is well known that one can design a Kalman filter for this system to estimate the state variables if the system is observable,

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k + K(y_k - C\hat{x}_k - Du_k) \tag{2}$$

where K is the steady state Kalman gain that can be obtained from an algebraic Ricatti equation. Denoting

$$e_k = y_k - C\hat{x}_k - Du_k$$

as the innovations of the Kalman filter and ignoring the ”∧“on x_k in the rest of this paper, we have the following equivalent innovation form,

$$x_{k+1} = Ax_k + Bu_k + Ke_k \tag{3a}$$

$$y_k = Cx_k + Du_k + e_k \tag{3b}$$

where the innovation e_k is white noise and independent of past input and output data. The system described by (2) can also be represented in the predictor form,

$$x_{k+1} = A_K x_k + B_K z_k \tag{4a}$$

$$y_k = Cx_k + Du_k + e_k \tag{4b}$$

where $z_k = [u_k^T,\ y_k^T]^T$, $A_K = A - KC$, and $B_K = [B - KD, K]$.

The three model forms, that is, the process form, the innovation form, and the predictor form, all can represent the input and output data (u_k, y_k) exactly. Therefore, one has the option to use any of these forms for convenience. For example, the well-known N4SID (Overschee & Moor, 1994) algorithm uses the process form. The MOESP (Verhaegen, 1994) algorithm uses the innovation form. For the convenience of closed-loop identification, Chiuso and Picci (2005) use the predictor form.

The subspace identification problem is: given a set of input/output measurements, estimate the system matrices (A, B, C, D), Kalman filter gain K up to within a similarity transformation, and the innovation covariance matrix R_e.

There are also minor differences among these model forms. The process form and the innovation form use the process (A, B, C, D) matrices, while the predictor form uses the (A_K, B_K, C, D) matrices. Since $A_K = A - KC$ is guaranteed stable even though the original process A matrix is unstable, the predictor form is numerically advantageous for identifying both stable and unstable processes. The other two model forms may lead to ill-conditioning for unstable processes (Chiuso & Picci, 2005). However, the optimal Kalman gain K is time-varying for finite number of samples, making A_K time varying even though the original process is time invariant. This is a minor drawback of the predictor form for limited number of samples.

Based on state space description in (4), an extended state space model can be formulated as

$$y_f(k) = \bar{\Gamma}_f x_k + \bar{G}_f z_{f-1}(k) + D_f u_f(k) + e_f(k) \tag{5}$$

where the subscript f denotes the future horizon. The extended observability matrix is

$$\bar{\Gamma}_f = \begin{bmatrix} C \\ CA_K \\ \vdots \\ CA_K^{f-1} \end{bmatrix}; \quad D_f = \begin{bmatrix} D \\ D \\ \vdots \\ D \end{bmatrix}$$

$$\bar{G}_f = \begin{bmatrix} 0 & 0 & \dots & 0 \\ CB_K & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA_K^{f-2}B_K & CA_K^{f-3}B_K & \dots & CB_K \end{bmatrix}$$

where the overbar means that the matrix is composed of parameters of the predictor form.

The input and output are arranged in the following form:

$$y_f(k) = \begin{bmatrix} y_k \\ y_{k+1} \\ \vdots \\ y_{k+f-1} \end{bmatrix} \tag{6a}$$

$$z_{f-1}(k) = \begin{bmatrix} z_k \\ z_{k+1} \\ \vdots \\ z_{k+f-2} \end{bmatrix} \tag{6b}$$

$u_f(k)$ and $e_f(k)$ are formed similar to $y_f(k)$.

By iterating (4) it is straightforward to derive the following relation,

$$x_k = \bar{L}_p z_p(k) + A_K^p x_{k-p} \tag{7}$$

where

$$\bar{L}_p = \begin{bmatrix} B_K & A_K B_K & \dots & A_K^{p-1} B_K \end{bmatrix} \tag{8a}$$

$$z_p(k) = \begin{bmatrix} z_{k-1}^T & z_{k-2}^T & \cdots & z_{k-p}^T \end{bmatrix}^T \tag{8b}$$

One can substitute (7) into (5) to obtain

$$y_f(k) = \bar{\Gamma}_f \bar{L}_p z_p(k) + \bar{\Gamma}_f A_K^p x_{k-p} + \bar{G}_f z_{f-1}(k) + D_f u_f(k) + e_f(k) \tag{9}$$

It is clear that the product of the observability and controllability matrices,

$$\bar{H}_{fp} \triangleq \bar{\Gamma}_f \bar{L}_p = \begin{bmatrix} CB_K & CA_K B_K & \dots & CA_K^{p-1} B_K \\ CA_K B_K & CA_K^2 B_K & \dots & CA_K^p B_K \\ \vdots & \vdots & \dots & \vdots \\ CA_K^{f-1} B_K & CA_K^f B_K & \dots & CA_K^{f+p-2} B_K \end{bmatrix} \tag{10}$$

is the Hankel matrix which contains the predictor Markov parameters. $\bar{G}_f$ also contains the predictor Markov parameters.

2.2. Assumptions

To establish the foundation of the SIM, we introduce following assumptions:

A1: The eigenvalues of $A - KC$ are strictly inside the unit circle.
A2: The system is minimal in the sense that (A, C) is observable and $(A, [B, K])$ is controllable.
A3: The innovation sequence e_k is a stationary, zero mean, white noise process with second order moment

$$E(e_i e_j^T) = R_e \delta_{ij}$$

where δ_{ij} is the Kronecker delta.
A4: The input u_k and innovation sequence e_j are uncorrelated for open-loop data, but u_k is directly related to past innovation e_k for closed-loop data.
A5: The input signal is quasi-stationary (Ljung, 1999) and is persistently exciting of order $f+p$, where f and p stand for future and past horizons, respectively, to be defined later.

From these assumptions we can relate the state space model forms to more traditional input-output models. For example, the innovation form (3) can be converted to the following input-output model,

$$y_k = [C(qI - A)^{-1} B + D] u_k + [C(qI - A)^{-1} K + I] e_k \tag{11}$$

from which the Box-Jenkins model or the ARMAX model can be recovered. Equivalently, the noise term in the Box-Jenkins model plays the role of innovation in the Kalman filter. The predictor form (4) can be converted to

$$y_k = C(qI - A_K)^{-1} B_K z_k + D u_k + e_k \tag{12}$$

Since A_K is strictly stable based on Assumption A1,

$$(qI - A_K)^{-1} = \sum_{i=1}^{\infty} A_K^i q^{-i}$$

can be truncated to a large number p and (12) reduces to

$$y_k \doteq \sum_{i=1}^{p} C A_K^i B_K z_{k-i} + D u_k + e_k \tag{13}$$

which is the well-known high-order ARX (HOARX) model used in the asymptotic methods (Ljung, 1999).

2.3. Linear regression and projections

We introduce the notation for linear regression and projections used in this paper. Given the input vector $x(k)$ and output vector $y(k)$, a linear relation

$$y(k) = \Theta x(k) + v(k)$$

can be built by collecting data for input and output variables and forming the data matrices

$$\underbrace{\begin{bmatrix} y(1) & y(2) & \dots & y(N) \end{bmatrix}}_{Y} = \Theta \underbrace{\begin{bmatrix} x(1) & x(2) & \dots & x(N) \end{bmatrix}}_{X} + V$$

where V is the matrix of noise.

By minimizing

$$J = ||Y - \Theta X||_F^2,$$

where $||\cdot||_F$ is the F-norm, we have the least squares solution

$$\hat{\Theta} = Y X^T (X X^T)^{-1}$$

The model prediction is

$$\hat{Y} = \hat{\Theta} X = Y X^T (X X^T)^{-1} X$$

Defining

$$\Pi_X = X^T (X X^T)^{-1} X$$

as the projection matrix to the row space of X, then

$$\hat{Y} = Y X^T (X X^T)^{-1} X = Y \Pi_X$$

is a projection of Y on X. The least square residual is

$$\tilde{Y} = Y - \hat{Y} = Y(I - \Pi_X) = Y \Pi_X^{\perp}$$

where

$$\Pi_X^{\perp} = I - \Pi_X = I - X^T (X X^T)^{-1} X$$

is the projection to the orthogonal complement of X. It is easy to verify that the model $\hat{Y}$ and the residual $\tilde{Y}$ are orthogonal.

Furthermore,

$$X \Pi_X = X X^T (X X^T)^{-1} X = X$$

$$X \Pi_X^{\perp} = X(I - X^T (X X^T)^{-1} X) = X - X = 0$$

For a model with two sets of input X and U with noise V

$$Y = \Gamma X + H U + V = [\,\Gamma \quad H\,] \begin{bmatrix} X \\ U \end{bmatrix} + V \tag{14}$$

we can find $[\,\Gamma \quad H\,]$ by least squares, assuming V is independent of both regressors X and U.

If we are only interested in estimating Γ, using the fact that V is independent of U, that is

$$\frac{1}{N} V U^T = \frac{1}{N} [\,v(1), \quad \dots, \quad v(N)\,] \quad [\,u(1), \quad \dots, \quad u(N)\,]^T \to 0 \quad \text{as} \quad N \to \infty$$

we have

$$Y\Pi_U^\perp = V(I - U^T(UU^T)^{-1}U) = V - VU^T(UU^T)^{-1}U$$
$$= V - \left(\frac{1}{N}VU^T\right)\left(\frac{1}{N}UU^T\right)^{-1} U \to V \quad \text{as} \quad N \to \infty$$

Therefore,

$$Y\Pi_U^\perp = (\Gamma X + HU + V)\Pi_U^\perp = \Gamma X\Pi_U^\perp + V$$

Γ can be found by regressing $Y\Pi_U^\perp$ on $X\Pi_U^\perp$ as follows,

$$\hat{\Gamma} = Y\Pi_U^\perp X^T(X\Pi_U^\perp X^T)^{-1} \tag{15}$$

where the relation $(\Pi_U^\perp)^2 = \Pi_U^\perp$ is used. It is straight-forward to show that $\hat{\Gamma}$ from (15) is identical to the least squares solution of (14). See Appendix of (van Overschee & de Moor, 1995).

2.4. General SIM procedures

Most SIMs involve some or all of the following steps:

(1) Step 1: Pre-estimation. In this step either the Markov parameters as in (13) (Jansson, 2003; Larimore, 2004; Shi & MacGregor, 2001) or the innovation sequence e_k (Qin & Ljung, 2003b) is pre-estimated from a high-order ARX (HOARX) model.
(2) Step 2: Regression or Projection. In this step a least squares regression or projection is performed to estimate one or several (up to *f*) high-order models.
(3) Step 3: Model Reduction. The high-order model identified in Step 2 is reduced to an appropriate low dimensional subspace that is observable. This step gives the estimates of Γ_f or the state sequence x_k.
(4) Step 4: Parameter Estimation. The reduced observability matrix or the realized state sequence from Step 3 is used to estimate the state space parameters *A*, *B*, *C*, *D* and *K*.
(5) Step 5: Iteration. The above steps can be iterated to improve accuracy.

Pre-estimation in Step 1 is usually designed to deal with closed-loop identification. It is also used to enforce the triangular structure of H_f and thus a causal model. Sometimes Steps 2 and 3 are done in one combined step, but they can always be written in two separate steps. Step 4 is where parametrization takes place, which is unique up to a similarity transform.

3. Open-loop subspace methods

The early developments of SIMs are applicable to open-loop identification where the input data are assumed independent of past noise, which admits no feedback. These methods include N4SID, MOESP and the CVA method without the pre-estimation step.

Based on the innovation form in (3), an extended state space model can be formulated as

$$Y_f = \Gamma_f X_k + H_f U_f + G_f E_f \tag{16}$$

where the subscript *f* denotes future horizon, respectively. The extended observability matrix is

$$\Gamma_f = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{f-1} \end{bmatrix} \tag{17}$$

and H_f and G_f are Toeplitz matrices:

$$H_f = \begin{bmatrix} D & 0 & \cdots & 0 \\ CB & D & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{f-2}B & CA^{f-3}B & \cdots & D \end{bmatrix} \tag{18a}$$

$$G_f = \begin{bmatrix} I & 0 & \cdots & 0 \\ CK & I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{f-2}K & CA^{f-3}K & \cdots & I \end{bmatrix} \tag{18b}$$

The input data are arranged in the following Hankel form:

$$U_f = \begin{bmatrix} u_k & u_{k+1} & \cdots & u_{k+N-1} \\ u_{k+1} & u_{k+2} & \cdots & u_{k+N} \\ \vdots & \vdots & \ddots & \vdots \\ u_{k+f-1} & u_{k+f} & \cdots & u_{k+f+N-2} \end{bmatrix} \tag{19a}$$

$$U_f = [\, u_f(k) \quad u_f(k+1) \quad \dots \quad u_f(k+N-1)\,] \tag{19b}$$

Similar formulations are made for Y_f and E_f. The state sequences are defined as:

$$X_k = [\, x_k, \quad x_{k+1}, \quad \dots, \quad x_{k+N-1}\,] \tag{20}$$

The Kalman state X_k is unknown, but we know that the Kalman state is estimated from past input and output data based on (7),

$$X_k = \bar{L}_p Z_p + A_K^p X_{k-p} \tag{21}$$

where $X_{k-p} = [x_{k-p}, \quad x_{k-p+1}, \quad \dots, \quad x_{k-p+N-1}]$. For a sufficiently large *p*, $A_K^p \simeq 0$. Hence, from (21) and (16),

$$Y_f = \Gamma_f \bar{L}_p Z_p + H_f U_f + G_f E_f$$
$$= H_{fp} Z_p + H_f U_f + G_f E_f \tag{22}$$

where $H_{fp} = \Gamma_f \bar{L}_p$ is the product of the process observability matrix and the predictor controllability matrix. It is analogous to $\bar{H}_{fp}$ in (10) but it is not exactly a Hankel matrix. However, it does have a reduced rank *n* which is less than the matrix dimensions.

Eqs. (16) and (22) can both be used to explain open-loop SIMs, whichever is more convenient. Under open-loop conditions, E_f is uncorrelated to U_f, that is,

$$\frac{1}{N} E_f U_f^T \to 0 \quad \text{as} \quad N \to \infty, \tag{23}$$

or

$$E_f \Pi_{U_f}^{\perp} = E_f(I - U_f^T(U_f U_f^T)^{-1} U_f) = E_f$$

Furthermore, E_f is uncorrelated to Z_p from the Kalman filter theory. Therefore,

$$\frac{1}{N} E_f Z_p^T \to 0 \quad \text{as} \quad N \to \infty \tag{24}$$

The above two relations (23) and (24) are very useful in open-loop SIMs.

The open-loop SIMs do not involve a pre-estimation step. Most of them involve only three major steps: projection or regression, model reduction, and parameter estimation. We will summarize each step in the following subsections.

3.1. SIM projections and model reduction

3.1.1. N4SID

Open-loop SIMs such as the N4SID (Overschee & Moor, 1994) first eliminate U_f by post-multiplying $\Pi_{U_f}^{\perp}$ on (22),

$$\begin{aligned} Y_f \Pi_{U_f}^{\perp} &= H_{fp} Z_p \Pi_{U_f}^{\perp} + H_f U_f \Pi_{U_f}^{\perp} + G_f E_f \Pi_{U_f}^{\perp} \\ &= H_{fp} Z_p \Pi_{U_f}^{\perp} + G_f E_f \end{aligned} \tag{25}$$

Then the noise term is removed by multiplying Z_p^T from the result of (24),

$$Y_f \Pi_{U_f}^{\perp} Z_p^T = H_{fp} Z_p \Pi_{U_f}^{\perp} Z_p^T + G_f E_f Z_p^T = H_{fp} Z_p \Pi_{U_f}^{\perp} Z_p^T$$

and

$$\hat{H}_{fp} = Y_f \Pi_{Uf}^{\perp} Z_p^T (Z_p \Pi_{Uf}^{\perp} Z_p^T)^{-1}$$

N4SID performs SVD on

$$\hat{H}_{fp} Z_p = \widehat{\Gamma_f \bar{L}_p} Z_p = USV^T \simeq U_n S_n V_n^T$$

where S_n contains the n largest singular values, and chooses $\hat{\Gamma}_f = U_n S_n^{1/2}$ as the estimated observability matrix, which is a special, balanced realization. Since Γ_f and $\bar{L}_p$ are observability and controllability matrices for different models, this is not exactly a balanced realization.

3.1.2. Regression approach

Perform least square solution to (25) by minimizing

$$J = ||Y_f \Pi_{U_f}^{\perp} - H_{fp} Z_p \Pi_{U_f}^{\perp}||_F^2,$$

$$\begin{aligned} \hat{H}_{fp} = \widehat{\Gamma_f \bar{L}_p} &= Y_f \Pi_{U_f}^{\perp} (\Pi_{U_f}^{\perp} Z_p^T)(Z_p \Pi_{U_f}^{\perp} \Pi_{U_f}^{\perp} Z_p^T)^{-1} \\ &= Y_f \Pi_{U_f}^{\perp} Z_p^T (Z_p \Pi_{U_f}^{\perp} Z_p^T)^{-1} \end{aligned} \tag{26}$$

Note that the rank of $\Gamma_f \bar{L}_p$ should be n. In the model reduction step we perform SVD,

$$\hat{H}_{fp} = \widehat{\Gamma_f \bar{L}_p} = USV^T \simeq U_n S_n V_n^T$$

and choose $\hat{\Gamma}_f = U_n S_n^{1/2}$ as the observability matrix. The observer-Kalman filter method (OKID) (Phan, Horta, Juang, & Longman, 1992) uses this approach. The MOESP algorithm (Verhaegen, 1994) uses this linear regression and performs SVD on $\hat{H}_{fp} Z_p \Pi_{U_f}^{\perp}$.

3.1.3. CVA approach

Since the coefficient matrix $H_{fp} = \Gamma_f \bar{L}_p$ in (25) is not full rank, the exact solution to (25) should be the canonical correlation analysis (CCA) which performs SVD on

$$W_r Y_f \Pi_{U_f}^{\perp} \Pi_{U_f}^{\perp} Z_p^T W_c = W_r Y_f \Pi_{U_f}^{\perp} Z_p^T W_c \simeq U_n S_n V_n^T$$

and chooses $\hat{\Gamma}_f = W_r^{-1} U_n S_n^{1/2}$ for balanced realization. In the above equation $W_r = (Y_f \Pi_{U_f}^{\perp} Y_f^T)^{-1/2}$, $W_c = (Z_p \Pi_{U_f}^{\perp} Z_p^T)^{-1/2}$. This is exactly canonical correlation analysis which extracts the n smallest angles between $Y_f \Pi_{U_f}^{\perp}$ and $Z_p \Pi_{U_f}^{\perp}$.

3.1.4. A unified formulation

van Overschee and de Moor (1995) have unified several SIMs in the open-loop case which offer insights into the relations among the SIMs. For the three SIM algorithms presented above, they are all equivalent to performing SVD on

$$W_1 \hat{H}_{fp} W_2 = U_n S_n V_n^T \tag{27}$$

where $\hat{H}_{fp}$ is the least squares estimate in (26) and for the regression approach, $W_1 = I$, $W_2 = I$, for N4SID, $W_1 = I$, $W_2 = (Z_p Z_p^T)^{1/2}$, for MOESP, $W_1 = I$, $W_2 = (Z_p \Pi_{U_f}^{\perp} Z_p^T)^{1/2}$, for CVA, $W_1 = (Y_f \Pi_{U_f}^{\perp} Y_f^T)^{-1/2}$, $W_2 = (Z_p \Pi_{U_f}^{\perp} Z_p^T)^{1/2}$.

It is pointed out in (Gustafsson & Rao, 2002) that the weighting W_1 has little impact on the results and the solution to Γ_f will undo this weighting

$$\hat{\Gamma}_f = W_1^{-1} U_n S_n^{1/2}.$$

However, for finite data length W_1 can make a difference since (25) is indeed a reduced rank regression. A requirement for the weights is that W_1 is nonsingular and W_2 does not reduce rank for $H_{fp} W_2$.

3.2. Enforcing causal models

In the extended state space model (22) H_f is block-triangular, which makes the model causal. However, this information is not normally taken care of in SIMs, as pointed out in (Shi & MacGregor, 2001). While there is no problem from a consistency point of view given proper excitation of the input, known parameters are estimated from data. Shi (2001) proposes an algorithm known as CVA_{Hf} that removes the impact of future input from the future output using pre-estimated the Markov parameters and then performs sub-space projections. Shi (2001) further shows that this procedure achieves consistency. Larimore (2004) argues that the CVA_{Hf} was implemented in Adaptx and that it is efficient, but he does not discuss the impact of imperfect pre-estimates.

To avoid these problems the SIM model must not include these non-causal terms, Peternell, Scherrer, and Deistler (1996) propose a few methods to exclude these extra terms. Specifically, they recommend a two-step procedure: (i) use a conventional (unconstrained) SIM to estimate the deterministic Markov parameters $CA^{i-1}B$; and (ii) form H_f with these Markov parameters to ensure that it is lower triangular and then estimate the extended observability matrix. Qin and Ljung (2003a), Qin et al. (2005) propose a causal subspace identification method (PARSIM) which remove these non-causal terms by performing f least squares projections in parallel. To accomplish this we partition the extended state space model row-wise as follows:

$$Y_f = \begin{bmatrix} Y_{f1} \\ Y_{f2} \\ \vdots \\ Y_{ff} \end{bmatrix}; \quad Y_i \triangleq \begin{bmatrix} Y_{f1} \\ Y_{f2} \\ \vdots \\ Y_{fi} \end{bmatrix}; \quad i = 1, 2, \ldots, f \tag{28}$$

where $Y_{fi} = [\, y_{k+i-1} \quad y_{k+i} \quad \ldots \quad y_{k+N+i-2}\,]$. Partition U_f and E_f in a similar way to define U_{fi}, U_i, E_{fi} and E_i, respectively, for $i = 1, 2, \ldots, f$. Denote further

$$\Gamma_f = \begin{bmatrix} \Gamma_{f1} \\ \Gamma_{f2} \\ \vdots \\ \Gamma_{ff} \end{bmatrix} \tag{29a}$$

$$H_{fi} \triangleq [\, CA^{i-2}B \quad \ldots \quad CB \quad D\,] \tag{29b}$$

$$G_{fi} \triangleq [\, CA^{i-2}K \quad \ldots \quad CK \quad I\,] \tag{29c}$$

where $\Gamma_{fi} = CA^{i-1}$. We have the following equations by partitioning (22),

$$Y_{fi} = \Gamma_{fi}\bar{L}_p Z_p + H_{fi} U_i + G_{fi} E_i \tag{30}$$

for $i = 1, 2, \ldots, f$. Note that each of the above equations is guaranteed causal. Now we have the following parallel PARSIM algorithm.

4. Parallel PARSIM (PARSIM-P)

(1) Perform the following LS estimates, for $i = 1, 2, \ldots, f$,

$$\left[\, \widehat{\Gamma_{fi}\bar{L}_p} \quad \hat{H}_{fi} \,\right] = Y_{fi} \begin{bmatrix} Z_p \\ U_i \end{bmatrix}^{\dagger} \tag{31}$$

where $[\cdot]^{\dagger}$ is the Moore-Penrose pseudo-inverse. Stack $\widehat{\Gamma_{fi}\bar{L}_p}$, together to obtain $\hat{\Gamma}_f \bar{L}_p$ as

$$\begin{bmatrix} \widehat{\Gamma_{f1}\bar{L}_p} \\ \widehat{\Gamma_{f2}\bar{L}_p} \\ \vdots \\ \widehat{\Gamma_{ff}\bar{L}_p} \end{bmatrix} = \widehat{\Gamma_f \bar{L}_p} \tag{32}$$

(2) Perform SVD for the following weighted matrix

$$W_1 \left(\widehat{\Gamma_f \bar{L}_p} \right) W_2 \simeq U_n S_n V_n^T \tag{33}$$

where W_1 is nonsingular and $\bar{L}_p W_2$ does not lose rank. U_n, S_n and V_n are associated to the n largest singular values. For CVA weighting we can choose $W_1 = (Y_f \Pi_{Uf}^{\perp} Y_f^T)^{-1/2}$ and $W_2 = (Z_p \Pi_{Uf}^{\perp} Z_p^T)^{1/2}$ We choose

$$\hat{\Gamma}_f = W_1^{-1} U_n S_n^{1/2} \tag{34}$$

from which the estimate of A and C can be obtained (Verhaegen, 1994).

(3) The estimate of B and D is discussed in the end of this section.

4.1. Estimating A and C from Γ_f

In the subspace identification literature A and C are extracted from Γ_f by choosing $f \geq n+1$, making Γ_{f-1} also full column rank. Denoting

$$\Gamma_f^{2:f} = \Gamma_f(n_y + 1 : n_y f, :)$$

which is the bottom $(f-1)$ block rows of Γ_f, we have

$$\Gamma_f^{2:f} = \Gamma_{f-1} A$$

Therefore,

$$\hat{A} = \hat{\Gamma}_{f-1}^{\dagger} \hat{\Gamma}_f^{2:f}$$

The estimate of C is simply

$$\hat{C} = \hat{\Gamma}_f(1 : n_y, :).$$

4.2. Estimation of K

A simple method to estimate the Kalman filter gain K is to extract it from $\bar{L}_p$. From the unified expression (27), we have:

$$W_1 \hat{\Gamma}_f \hat{\bar{L}}_p W_2 = U_n S_n V_n^T$$

and $W_1 \hat{\Gamma}_f$ is chosen a $U_n S_n^{1/2}$,

$$\hat{\bar{L}}_p W_2 = S_n^{1/2} V_n^T$$

Therefore,

$$\hat{\bar{L}}_p = S_n^{1/2} V_n^T W_2^{-1}$$

Note that $\bar{L}_p$ is the extended controllability matrix of the predictor. Similar to the extraction of $\hat{A}$ and $\hat{C}$ from $\hat{\Gamma}_f$, we can extract $\hat{A}_K$ and $[\, \hat{B}_K \quad \hat{K}\,]$.

Another approach to estimating K is to extract it from G_f (Qin et al., 2005).

From (22) we have

$$Y_f \Pi^{\perp}_{\begin{bmatrix} Z_p \\ U_f \end{bmatrix}} = G_f E_f \Pi^{\perp}_{\begin{bmatrix} Z_p \\ U_f \end{bmatrix}} = G_f E_f \tag{35}$$

since E_f is not correlated with Z_p and U_f in open-loop. Performing QR decomposition,

$$\begin{bmatrix} Z_p \\ U_f \\ Y_f \end{bmatrix} = \begin{bmatrix} R_{11} & & \\ R_{21} & R_{22} & \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \end{bmatrix} \tag{36}$$

then

$$R_{33} Q_3 = G_f E_f \tag{37}$$

Denoting $e_k = F e_k^*$ such that $cov(e_k^*) = I$, from Assumption A3 we have $FF^T = R_e$. Using this notation we have

$$G_f E_f = G_f^* E_f^* \tag{38}$$

where

$$G_f^* = \begin{bmatrix} F & 0 & \dots & 0 \\ CKF & F & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{f-2}KF & CA^{f-3}KF & \cdots & F \end{bmatrix} \in \Re^{n_y f \times n_y f}$$

From Eqs. (37) and (38) and using the fact that Q_3 is an orthonormal matrix, we choose

$$\hat{E}_f^* = Q_3 \tag{39a}$$

$$\hat{G}_f^* = R_{33} \tag{39b}$$

Denoting the first block column of G_f^* by G_{f1}^*,

$$G_{f1}^* = \begin{bmatrix} F \\ CKF \\ \vdots \\ CA^{f-2}KF \end{bmatrix} = \begin{bmatrix} I_{n_y} & 0 \\ 0 & \hat{\Gamma}_{f-1} \end{bmatrix} \begin{bmatrix} F \\ KF \end{bmatrix} \tag{40}$$

KF and F can be estimated as

$$\begin{bmatrix} \hat{F} \\ \widehat{KF} \end{bmatrix} = \begin{bmatrix} I_{n_y} & 0 \\ 0 & \hat{\Gamma}_{f-1} \end{bmatrix}^{\dagger} G_{f1}^* \tag{41}$$

Finally,

$$\hat{K} = \widehat{(KF)} \hat{F}^{-1} \tag{42}$$

and

$$\hat{R}_e = \hat{F} \hat{F}^T \tag{43}$$

4.3. Determination of B, D

Qin et al. (2005) give an optimal approach to estimate B and D and the initial state using A, C, K and F for the general innovation form. Since the initial state is estimated this step does not introduce a bias for finite p.

From the innovation form of the system we have:

$$x_{k+1} = A_K x_k + B_K u_k + K_{y_k} \tag{44}$$

The process output can be represented as

$$\begin{aligned} y_k = {} & C(qI - A_K)^{-1} x_0 + [C(qI - A_K)^{-1} B_K + D] u_k \\ & + C(qI - A_K)^{-1} K_{y_k} + e_k \end{aligned} \tag{45}$$

or

$$\begin{aligned} y_k & = [I - C(qI - A_K)^{-1} K]_{y_k} \\ & = C(qI - A_K)^{-1} x_0 + [C(qI - A_K)^{-1} B_K + D] u_k + e_k \end{aligned} \tag{46}$$

using $e_k = F e_k^*$ where e_k^* has an identity covariance matrix, and defining

$$\tilde{y}_k = F^{-1} [I - C(qI - A_K)^{-1} K] y_k \tag{47a}$$

$$G(q) = F^{-1} C(qI - A_K)^{-1} \tag{47b}$$

$$D^* = F^{-1} D \tag{47c}$$

we obtain,

$$\begin{aligned} \tilde{y}_k & = G(q) B_K u_k + D^* u_k + G(q) x_0 \delta_k + e_k^* \\ & = G(q) \otimes u_k^T \mathrm{vec}(B_K) + I_{n_y} \otimes u_k^T \mathrm{vec}(D^*) + G(q) x_0 \delta_k + e_k^* \end{aligned} \tag{48}$$

where $\mathrm{vec}(B_K)$ and $\mathrm{vec}(D^*)$ are vectorized B_K and D^* matrices along the rows. δ_k is the Kronecker delta function. Now $\mathrm{vec}(B_K)$, $\mathrm{vec}(D^*)$ and x_0 can be estimated using least squares from the above equation. The B, D matrices can be backed out as:

$$\hat{D} = F \hat{D}^* \tag{49a}$$

$$\hat{B} = \hat{B}_K + K \hat{D} \tag{49b}$$

4.4. Estimating all parameters from the state

An alternative approach is to estimate all model parameters from the state sequence. With the estimate of $\hat{\hat{L}}_p$ we have

$$\hat{x}_k \simeq \hat{\hat{L}}_p z_p(k)$$

From (3) we obtain

$$\begin{aligned} [\hat{C} \quad \hat{D}] & = arg\ \min \left\{ \sum_{k=1}^{N} \left\| y_k - [C \quad D] \begin{bmatrix} \hat{x}_k \\ u_k \end{bmatrix} \right\|^2 \right\} \hat{e}_k \\ & = y_k - [\hat{C} \quad \hat{D}] \begin{bmatrix} \hat{x}_k \\ \hat{u}_k \end{bmatrix} \end{aligned}$$

$$[\hat{A} \quad \hat{B} \quad \hat{K}] = arg\ \min \left\{ \sum_{k=1}^{N} \left\| x_k - [A \quad B \quad K] \begin{bmatrix} \hat{x}_k \\ u_k \\ \hat{e}_k \end{bmatrix} \right\|^2 \right\}$$

For more detail see (Ljung & McKelvey, 1996; Overschee & Moor, 1994), and (Chiuso & Picci, 2005).

As one can see from the above illustration, two major paths for open-loop SIMs are to use the estimates of Γ_f or x_k to further estimate the model parameters. However, no results are available as to which path leads to a better model.

5. Closed-loop SIMs

In order to identify a state space model with closed-loop data, a couple of closed-loop subspace identification methods (SIMs) have been proposed in the last decade (Ljung & McKelvey, 1996; van Overschee & de Moor, 1997; Verhaegen, 1993). More recent work is presented in (Jansson, 2003; Qin & Ljung, 2003b), which has been regarded as a significant advance in subspace identification of feedback systems (Chiuso & Picci, 2005). The consistency of the algorithms has been investigated in (Chiuso & Picci, 2005; Lin, Qin, & Ljung, 2004).

Due to the feedback control the future input is correlated with past output measurement or past noise, making the traditional SIMs biased. That is, the last two terms of (22) are correlated for closed-loop systems. Therefore, most of the closed-loop SIMs try to decouple these two terms. The SIMPCA methods proposed in (Wang & Qin, 2002) and a later modification in (Huang, Ding, & Qin, 2005) move $H_f U_f$ to the LHS and use principal component analysis on the joint input/output data simultaneously. The observer/Kalman filter identification (OKID) algorithm (Phan et al., 1992), which is not traditionally known as SIMs, does not use an extended future horizon, therefore is free from the bias problem. These are some of the closed-loop SIMs which do not require special manipulations.

Most closed-loop SIMs involve four or five of the steps outlined in Section II-D. Based on the notation in Section II-A, we have four different approaches to estimate the model parameters:

(1) Estimate the Markov parameters from a high-order ARX (HOARX) model, form the Hankel matrix H_{fp}, then perform SVD on H_{fp} to estimate A_K, B_K and C. (OKID, Phan et al., 1992);
(2) Estimate the Markov parameters from a high-order ARX model, form $\bar{G}_f$, then estimate $\bar{\Gamma}_f \bar{L}_p$ from (9) and perform SVD to estimate A_K, B_K and C (SSARX, Jansson, 2003; CVA, Larimore, 2004); and
(3) Partition (9) row-wise into f separate sub-problems, enforce causal relations similar to (Qin & Ljung, 2003a), estimate $\bar{\Gamma}_f \bar{L}_p$ (or $\bar{L}_p z_p(k)$ as the state vector), and then estimate A, B, C and D. (WFA, Chiuso & Picci, 2004; Chiuso & Picci, 2005).
(4) Pre-estimate the innovation E_f from a HOARX and use (22) to estimate the state space model (Qin & Ljung, 2003b).

Since (22) is actually composed of f block rows in each term and the first block row gives an estimate of the innovation, Qin and Ljung (2003b) propose an innovation estimation method (IEM) that partitions (22) into f block rows and uses the estimated innovation from previous block rows to further estimate model parameters of the next block row sequentially. An alternative method known as IEM1 (Lin et al., 2004) estimates the innovation from the first block row and then treats $\hat{e}_k$ as known to estimate other model parameters. The SSARX approach proposed in (Jansson, 2003) uses the predictor form (4) and pre-estimates a high-order ARX model to decouple the correlation between U_f and E_f. The well-known CVA algorithm proposed by Larimore (1990) actually pre-estimates H_f using a high-order ARX and then move $\hat{H}_f U_f$ to the LHS of (22). Shi and MacGregor (2001) also use this technique.

Inspired from the SSARX approach, Chiuso and Picci (2005) give a variation known as the whitening filter approach (WFA) that uses the predictor model form and carry out multi-stage projections row by row. In each block row projection causality is strictly enforced, similar to (Qin et al., 2005). No pre-estimation is involved but the projections have to be done block-row wise to decouple noise from control input. In the rest of this section we briefly introduce these closed-loop SIMs.

5.1. Innovation estimation method

Partitioning the last term of (30) into two parts, we obtain

$$Y_{fi} = \Gamma_{fi} \bar{L}_p Z_p + H_{fi} U_i + G_{fi}^{-} E_{i-1} + E_{fi} \tag{50}$$

where

$$G_{fi}^{-} = [\, CA^{i-2}K \quad \ldots \quad CK \,].$$

For $i = 1$, (50) becomes,

$$Y_{f1} = C\bar{L}_P Z_P + DU_1 + E_{f1} \tag{51}$$

which is a high-order ARX model. Typically $D = 0$ in (51). Hence, (51) is suitable for closed-loop data since E_{f1} is always uncorrelated of past data Z_p. In the case that $D \neq 0$, there must be a delay in the feedback loop, making U_1 uncorrelated with E_{f1}. As a consequence, we can obtain unbiased estimates of $C\bar{L}_P$, D, and E_{f1} from (51) using closed-loop data.

The innovation estimation method proposed in (Lin, Qin, & Ljung, 2006) (IEM1) uses (51) to pre-estimate $\hat{E}_{f1}$, then form $\hat{E}_{i-1}$ by using the shift structure of E_{i-1}, and replace E_{i-1} in (50) using $\hat{E}_{f1}$ to estimate $\Gamma_{fi} \bar{L}_p$. Since the only error term in (50) is E_{fi} which is "future" relative to U_i, it is suitable for closed-loop data.

The innovation estimation method proposed in (Lin et al., 2004; Qin & Ljung, 2003b) (IEM) involves estimating innovation sequence repeatedly row-wise and estimating Γ_f through a weighted singular value decomposition. A, B, C, D and K can also be obtained as illustrated in the previous section.

5.2. SIMs with pre-estimation

For convenience we assume $D = 0$ to simplify the presentation. Suppose that p is chosen large enough so that $A_K^p \simeq 0$, (9) can be written as

$$y_f(k) = \bar{\Gamma}_f \bar{L}_p z_p(k) + \bar{G}_f z_{f-1}(k) + e_f(k) \tag{52}$$

Due to feedback $e_f(k)$ is correlated with $z_{f-1}(k)$. Since $\bar{G}_f$ contains the Markov parameters of the predictor form, Jansson (2003), Shi and MacGregor (2001) and Larimore (2004) pre-estimate $\bar{G}_f$ (or part of $\bar{G}_f$ that is related to $u_f(k)$) from a high-order ARX model (13). Then, the estimate $\bar{G}_f$ is used to define a new vector

$$\tilde{y}_f(k) = y_f(k) - \hat{\bar{G}}_f z_{f-1}(k) = \bar{\Gamma}_f \bar{L}_p z_p(k) + e_f(k)$$

Now the error term $e_f(k)$ is uncorrelated with past data $z_p(k)$, making it suitable for closed-loop data. The model coefficient $\bar{\Gamma}_f \bar{L}_p$ can be estimated using least squares and then SVD or weighted SVD is performed to obtain $\hat{\bar{\Gamma}}_f$. Alternatively, one can perform CCA between $\tilde{y}(k)$ and $z_p(k)$ to obtain $\hat{\bar{\Gamma}}_f$ and $\hat{\bar{L}}_p$ in one step, leading to the CVA approach in (Larimore, 2004).

5.3. Whitening filter approach

Chiuso and Picci (2005) observe that one does not need to pre-estimate $\bar{G}_f$ if the triangle structure of $\bar{G}_f$ is exploited. Partitioning (52) row-wise and denoting

$$\bar{\Gamma}_{fi} = C A_K^{i-1},$$

$$\bar{G}_{fi} = [\, C A_K^{i-2} B_K \quad C A_K^{i-3} B_K \quad \dots \quad C B_K \,],$$

$$z_{i-1}(k) = [\, z_k^T \quad z_{k+1}^T \quad \dots \quad z_{k+i-2}^T \,]^T,$$

the ith row of (52) is

$$y_{k+i-1} = \bar{\Gamma}_{fi} \bar{L}_p z_p(k) + \bar{G}_{fi} z_{i-1}(k) + e_{k+i-1} \quad \text{for } i = 1, 2, \dots, f. \tag{53}$$

Using least squares $\bar{\Gamma}_{fi} \bar{L}_p$ can be estimated for $i = 1, 2, \dots f$, which then form $\widehat{\bar{\Gamma}_f \bar{L}_p}$. Two subsequent options can be used in the model reduction step similar to the open-loop SIM procedure in the previous section. The first one is to perform SVD or weighted SVD on $\widehat{\bar{\Gamma}_f \bar{L}_p}$ to obtain $\widehat{\bar{\Gamma}_f}$, then estimate model parameters from $\widehat{\bar{\Gamma}_f}$. The other option is to form

$$\bar{\Gamma}_f X_k = \bar{\Gamma}_f \bar{L}_p Z_p \simeq \widehat{\bar{\Gamma}_f \bar{L}_p Z_p}$$

and perform SVD to obtain the state sequence X_k, from which the process A, B, C, D, and K are estimated (Chiuso & Picci, 2005).

5.4. Summary of closed-loop SIMs

Subspace identification methods are difficult to apply to closed-loop data because of the use of an extended future horizon that introduces correlation between inputs and past noise. To avoid this correlation several methods such as CVA and SSARX use pre-estimation to separate these two terms. The SIMPCA algorithm avoids the correlation by using the parity space instead of the observability subspace. Interestingly, the extended future horizon is not a necessary requirement for the projection or regression step of SIMs. It is only necessary to extend the order of the Hankel matrix, from which the observability matrix is reduced. The OKID (Phan et al., 1992) and the SMARX (Ljung & McKelvey, 1996) do not require the extended future horizon for the regression step. See (Qin & Ljung, 2006) for more discussions. The closed-loop SIMs can be summarized in Fig. 1.

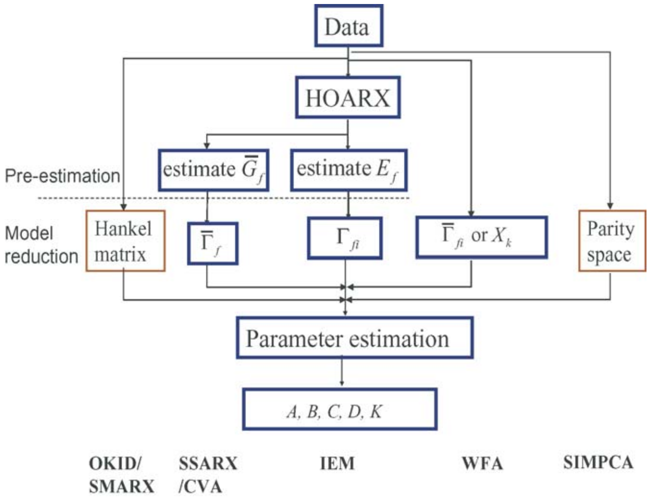

Fig. 1. Closed-loop SIMs comparison.

It is interesting to compare the innovation estimation method and the whitening filter approach. They all partition the extended state space row-wise and utilize a multi-stage least squares method to estimate system matrices. The innovation estimation method starts from a state space model in innovations form, while the whitening filter approach is based on a state space model in predictor form.

The IEM, CVA and SIMPCA use the process A matrix to form the observability matrix, while the WFA, OKID, and SSARX use the predictor matrix A_K. For open-loop unstable systems the whitening filter approach can be numerically advantageous, as demonstrated in (Chiuso & Picci, 2005). However, for bounded systems such as stable or integrating systems, this advantage disappears. For limited data length where K is time varying, it is better to use process A matrix.

The major difference between closed-loop SIMs and open-loop SIMs is in estimating the observability subspace Γ_f or $\bar{\Gamma}_f$. The remaining steps to estimating model parameters are essentially the same.

6. Simulation example

To demonstrate how SIM works in closed-loop case, we use the example in (Verhaegen, 1993). The model of the plant is given in transfer function form:

$$\frac{10^{-3}(0.95q^4 + 12.99q^3 + 18.59q^2 + 3.30q - 0.02)}{q^5 - 4.4q^4 + 8.09q^3 - 7.83q^2 + 4q - 0.86} \tag{54}$$

The output disturbance of the plant is a zero-mean white noise with standard deviation 1/3 filtered by the linear filter

$$F_1(q) = \frac{0.01(2.89q^2 + 11.13q + 2.74)}{q^3 - 2.7q^2 + 2.61q - 0.9}$$

The controller is

$$F(q) = \frac{(0.61q^4 - 2.03q^3 + 2.76q^2 - 1.83q + 0.49)}{q^4 - 2.65q^3 + 3.11q^2 - 1.75q + 0.39}$$

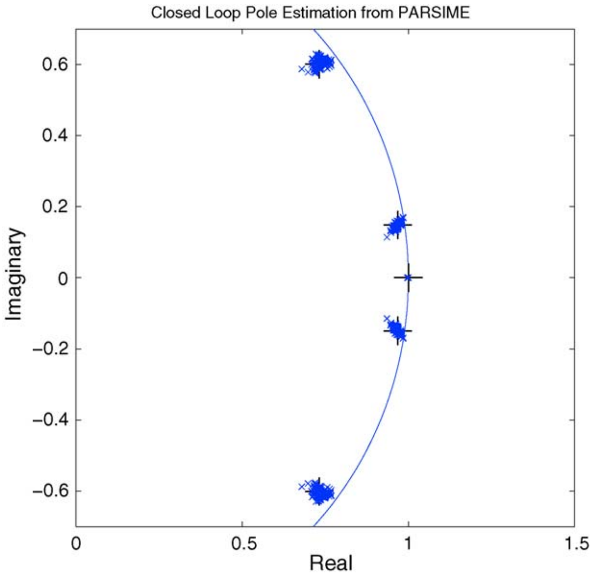

Fig. 2. The eigenvalues of estimated A matrix using IEM: (×) estimated pole, (+) system pole.

The feedback mechanism is

$$u_k = -F(q)y_k + r_k$$

where r_k is a zero-mean white noise sequence with standard deviation 1. We take the number of data points $j = 1200$, and generate 100 data set, each time with the same reference input r_k but with different noise sequence e_k. We choose $f = p = 20$ for "innovation estimation" approaches, and $f = p = 30$ for "whitening filter" approaches. In our simulation, we observe that to obtain unbiased estimation the "whitening filter" approach needs larger f and p than the "innovation estimation" approach.

The pole estimation results for the closed-loop experiments are shown in Figs. 2–5. From the results we can see that all the methods can provide consistent estimates. The "whitening filter" approach produces the worst results, but there is no general statement we can make from this specific example. The SSARX has one pair of outlying poles, but this could happen to other methods due to the actual noise based on our experience.

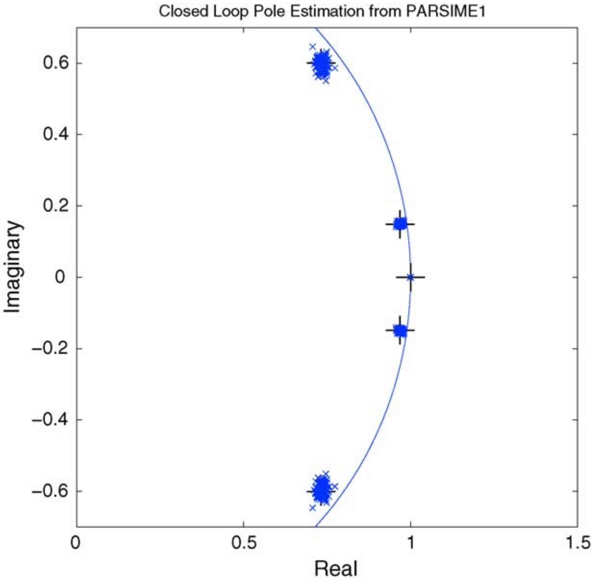

Fig. 3. The eigenvalues of estimated A matrix using IEM1: (×) estimated pole, (+) system pole.

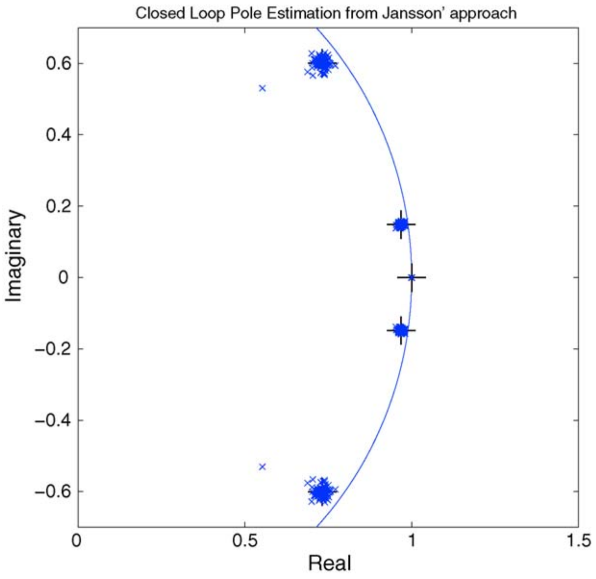

Fig. 4. The eigenvalues of estimated A matrix using SSARX: (×) estimated pole, (+) system pole.

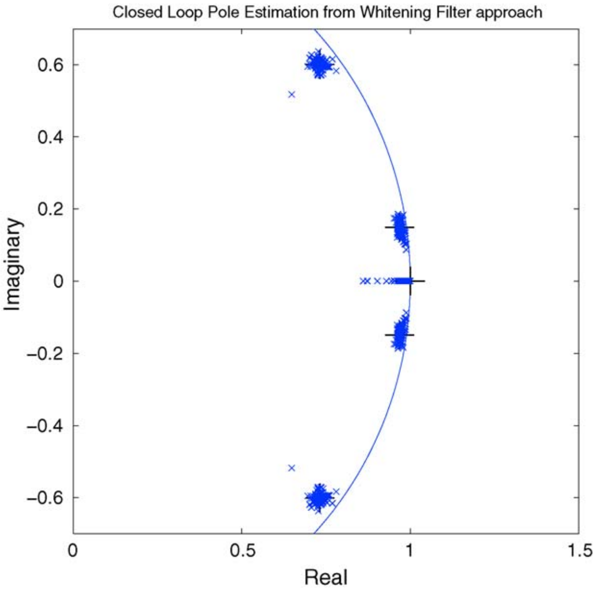

Fig. 5. The eigenvalues of estimated A matrix using WFA: (×) estimated pole, (+) system pole.

7. Further discussions and conclusions

7.1. Statistical properties

The statistical properties such as consistency and efficiency of SIMs have been investigated recently (Bauer, 2003, 2005; Bauer & Ljung, 2002; Chiuso & Picci, 2004; Gustafsson, 2002;

Jansson & Wahlberg, 1998; Knudsen, 2001; Larimore, 1996). Consistency is concerned with the bias of the estimates while efficiency is concerned with variance. All these variants are shown to be generically consistent. For some special cases, it has also been shown that CVA gives statistical efficiency and/or gives the lowest variance among available weighting choices. Simulations also seem to indicate that CVA may have better variance properties in overall comparisons, see, e.g. (Ljung, 2003).

While most SIMs are consistent, few if any can achieve the efficiency of the maximum likelihood estimate (MLE). For open-loop SIMs Bauer and Ljung (2002) show that SIMs with the CVA type of weighting are optimal for the special case of white noise input perturbation and $f \rightarrow \infty$. They also show that the variance of the estimates improves as f increases. For closed-loop SIMs only variance expressions of the estimates are available (Chiuso & Picci, 2005).

7.2. Practical considerations

The popularity of SIM in industry has increased tremendously in recent years (Larimore, 2004; Zhao & Harmse, 2006). One of the reasons behind the rapid adoption of SIMs in practice is the simplicity of SIMs and the inherent characteristics of multivariable industrial control problems, such as model predictive control (MPC) problems. While significant progress has been made in the analysis and understanding of SIMs, the following issues are still standing to some extent.

1. Optimal input design for SIMs. Since SIMs are related to high-order ARX, low-order input excitations such as sinusoidal signals are not very suitable for SIMs. Most SIMs achieve favorable results when the input is white or close to white. For industrial applications closed-loop testing and simultaneous multivariable testing are preferred (Zhu & Van Den Bosch, 2000).
2. Connection to asymptotic methods. Since SIMs are closely related to HOARX with model reduction using state space models, it is natural to probe the connection to the asymptotic methods, which has enjoyed surprising success in industry (Zhu, 1998). The two types of methods essentially perform the same first step, which is HOARX, see (Qin & Ljung, 2006). The only difference is in the model reduction step: SIMs perform model reduction in time domain, while the asymptotic methods do it in frequency domain.
3. Time delay and order estimation. To improve the accuracy of the final model one must estimate the time delay. This is a trivial task for SISO models, but can be difficult for MIMO processes. The order estimation is also an important problem.
4. Zero model responses. For MIMO processes an input-output pair can have zero responses, while the overall system is interacting. For this case it is desirable to identify input-output channels that have zero responses and keep them zero in the model parametrization. While this is easily done for input-output models, such as FIR models in DMC practice, it is not trivial for state space models. Including parameters in the model that are known to be zero usually increases the variance of the model estimates.
5. Confidence intervals. It is desirable to be able to estimate the confidence intervals for the estimated models. This is also true for SIMs. A possible approach is to derive the model confidence intervals based on the variance estimates of the model parameters (Chiuso & Picci, 2005).
6. Disturbance models: use or do not use? It is generally true that correlated disturbances happen to the process to be identified even during the data collection phase. Therefore, it is usually a good idea to identify the process model and the disturbance model. However, most industrial practice does not use the identified disturbance model. One rationale behind this is that the disturbance characteristics change very often. However, without using a disturbance model, the power of Kalman filtering is ignored. An important issue is to decide whether the disturbance model is representative for most of the disturbance scenarios, that is, whether the process is "fully" excited in the disturbance channel.
7. Model quality assessment. It is important to assess the model quality both during the identification phase and during on-line use. Due to the time-varying nature of industrial processes, on-line model assessment is necessary to determining whether model re-identification is needed. The assessment task includes process model assessment and disturbance model assessment.

7.3. Conclusions

Subspace identification methods have enjoyed rapid development for both closed-loop systems and open-loop processes. The attractive features include simple parametrization for MIMO systems and robust noniterative numerical solutions. These features lead to their rapid adoption in industry. There are, however, many unsolved issues in both statistical analysis and practical considerations. Future research should be focused on further understanding of the statistical properties and resolving the practical issues.

Acknowledgements

Financial support from Natural Science Foundation under CTS-9985074, National Science Foundation of China under an Overseas Young Investigator Award (60228001), and the Texas-Wisconsin Modeling and Control Consortium are gratefully acknowledged.

References

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

Bauer, D. (2003). Subspace methods. In *Proceedings of the 13th IFAC SYSID Symposium* (pp. 763–773).

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

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

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

Chou, C., & Verhaegen, M. (1997). Subspace algorithms for the identification of multivariable dynamic errors-in-variables models. *Automatica*, *33*(10), 1857–1869.

Forssell, U., & Ljung, L. (1999). Closed-loop identification revisited. *Automatica*, *35*, 1215–1241.

Gustafsson, T. (2002). Subspace-based system identification: Weighting and pre-filtering of instruments. *Automatica*, *38*, 433–443.

Gustafsson, T., & Rao, B. D. (2002). Statistical analysis of subspace-based estimation of reduced-rank linear regression. *IEEE Transactions on Signal Processing*, *50*, 151–159.

Huang, B., Ding, S. X., & Qin, S. J. (2005). Closed-loop subspace identification: an orthogonal projection approach. *Journal of Process Control*, *15*, 53–66.

Jansson, M. (2003). Subspace identification and ARX modelling. In *Proceedings of the 13th IFAC SYSID Symposium*.

Jansson, M., & Wahlberg, B. (1996). A linear regression approach to state-space subspace system. *Signal Processing*, *52*, 103–129.

Jansson, M., & Wahlberg, B. (1998). On consistency of subspace methods for system identification. *Automatica*, *34*(12), 1507–1519.

Katayama, T. (2005). *Subspace methods for system identification*. Springer.

Knudsen, T. (2001). Consistency analysis of subspace identification methods based on a linear regression approach. *Automatica*, *37*, 81–89.

Larimore, W. E. (1990). Canonical variate analysis in identification, filtering and adaptive control. In *Proceedings of the 29th Conference on Decision and Control* (pp. 596–604).

Larimore, W. E. (1996). Statistical optimality and canonical variate analysis system identification. *Signal Processing*, *52*, 131–144.

Larimore, W. E. (2004). Large sample efficiency for adaptx subspace system identification with unknown feedback, In *Proc. IFAC DYCOPS'04*.

Lin, W., Qin, S. J., & Ljung, L. (2004). On consistency of closed-loop subspace identification with innovation estimation. In *43rd IEEE Conference on Decision and Control* (pp. 2195–2200).

Lin, W., Qin, S. J., & Ljung, L. (2006). A framework for closed-loop subspace identification with innovation estimation. *Revised for Automatica*.

Ljung, L. (1999). *System identification: Theory for the user*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Ljung, L. (2003). Aspects and experiences of user choices in subspace identification methods, In *13th IFAC Symposium on System Identification, 2003*. pp. 1802–1807.

Ljung, L., & McKelvey, T. (1996). Subspace identification. *Signal Processing*, *52*, 209–215.

van Overschee, P., & de Moor, B. (1995). A unifying theorem for three subspace system identification algorithms. *Automatica*, *31*(12), 1853–1864.

Overschee, P. V., & Moor, B. D. (1994). N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, *30*(1), 75.

Peternell, K., Scherrer, W., & Deistler, M. (1996). Statistical analysis of novel subspace identification methods. *Signal Processing*, *52*, 161–177.

Phan, M., Horta, L. G., Juang, J. -N., & Longman, R. W. (1992). Improvement of observer/kalman filter identification (OKID) by residual whitening, In *AIAA Guidance, Navigation and Control Conference*, Hilton Head, South Carolina.

Qin, S. J., & Ljung, L. (2003a) Parallel QR implementation of subspace identification with parsimonious models, In *IFAC Symposium on System Identification*.

Qin, S. J., & Ljung, L. (2003b). Closed-loop subspace identification with innovation estimation. In *Proceedings of the 13th IFAC SYSID Symposium* (pp. 887–892).

Qin, S. J., Lin, W., & Ljung, L. (2005), A novel subspace identification approach with enforced causal models, *Automatica*, *41*, 2043–2053.

Qin, S. J., & Ljung, L. (2006). *On the role of future horizon in closed-loop subspace identification*. Accepted by the 14th IFAC Symposium on System Identification.

Shi, R. (2001). *Subspace identification methods for dynamic process modeling*, Ph.D. dissertation, MacMaster University.

Shi, R., & MacGregor, J. F. (2001). A framework for subspace identification methods. In *Proceeding of ACC* (pp. 3678–3683).

Soderstrom, T., & Stoica, P. (1989). *System identification*. London: Prentice-Hall.

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 system identification. In *Proceedings of the 36th Conference on Decision and Control, 1997*. pp. 1848–1853.

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

Verhaegen, M. (1994). Identification of the deterministic part of MIMO state space models given in innovations form from input-output data. *Automatica*, *30*(1), 61–74.

Verhaegen, M., & Dewilde, P. (1992). Subspace model identification, part i: The output-error state-space model identification class of algorithms. *International Journal of Control*, *56*, 1187–1210.

Wang, J., & Qin, S. J. (2002). A new subspace identification approach based on principal component analysis. *Journal of Process Control*, *12*, 841–855.

Zhao, H., & Harmse, M. (2006). *Subspace identification in industrial APC application-a review of recent process and industrial experience*. Accepted by the 14th IFAC Symposium on System Identification, 2006.

Zhu, Y. (1998). Multivariable process identification for mpc: the asymptotic method and its applications. *Journal of Process Control*, *8*, 101–115.

Zhu, Y., & Van Den Bosch, P. P. J. (2000). Optimal closed-loop identification test design for internal model control. *Automatica*, *36*, 1237–1241.