A Sliding Mode Observer for Infinitely Unobservable Descriptor Systems

Jeremy H.T. Ooi, Chee Pin Tan, Surya G. Nurzaman, and Kok Yew Ng

Abstract—In existing work of sliding mode observers for descriptor systems, a necessary condition is that the system must be infinitely observable. This paper presents a scheme that circumvents that condition, by reformulating the system as a reduced-order system where certain structures in the system matrix are manipulated and certain states are treated as unknown inputs. Following that, a sliding mode observer (SMO) is implemented on the reduced-order system where state and fault estimation can be achieved. Necessary and sufficient conditions of this scheme are also presented. Finally, a simulation example shows the effectiveness of the proposed scheme.

Index Terms—Estimation, fault detection, linear systems, sliding-mode observers

I. INTRODUCTION

THE descriptor system representation [1], being more general than the regular state space, can better model the dynamics of large-scale interconnected systems [2]. In addition, it has also been shown that several problems can be modelled and solved in the descriptor system framework [3]–[5]. While a perfect model is desired, there will usually be a mismatch between the actual system and the corresponding model (on which the observer design is based). This mismatch, which appears as an unknown input can diminish the observer performance. Sliding Mode Observers (SMOs) [6], developed from Sliding Mode principles [7]–[9], can robustly estimate states independently of the unknown inputs, and also estimate the unknown input. By modelling faults as unknown inputs, these observers can then be used for fault estimation [10] which can be used for fault tolerant control [11]. SMOs for descriptor systems were first developed by Yeu et al. [12] and subsequently improved in [13]. Generally, observer designs are developed under the assumption that the descriptor system has global observability [14], which comprises finite observability and infinite observability. Finite observability ensures that the invariant zeros of the system are stable while infinite observability mandates that certain states be measurable [15]. Those works [12], [13] require global observability. Though various schemes [16]–[22] have been proposed for both infinitely observable and unobservable descriptor systems, accurate fault estimation remains contentious. In [23], a variation of this work was carried out using the SMO but requires certain control inputs to be fault-free. Ooi et al. [24] presented an improved scheme whereby certain states are treated as unknown inputs.

This paper proposes a novel scheme capable of state and fault estimation for a class of infinitely unobservable descriptor systems. Firstly, certain states with infinite dynamical modes are re-expressed in terms of other states. Then, the remainder of those states are removed and augmented with the original fault to form an ‘augmented’ unknown input. This reformulation results in an infinitely observable system, on which the SMO [12] is implemented, to estimate the states and the augmented unknown input (from which the estimates of the removed states and fault could be obtained). Then, existence conditions for the scheme are investigated. Compared to [24], this scheme treats less states as unknown inputs, resulting in less unknown inputs and a less conservative scheme.

This paper is organized as follows; Section II reformulates the system and presents the observer; Section III investigates necessary and sufficient conditions; Section IV presents a simulation example; and Section V draws some conclusions.

II. PRELIMINARIES AND PROBLEM STATEMENT

Consider the following descriptor system

$$E \dot{x} = Ax + Mf, \ y = Cx$$ \hspace{1cm} (1)

where $E \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$, $f \in \mathbb{R}^q$ are the states, outputs and faults (which could also be unknown inputs) respectively. Assume $x$, $f$ are bounded by $\|x\| \leq \alpha_x$, $\|f\| \leq \alpha_f$. The bound on $f$ is a standard assumption in SMO research whereas $\alpha_x$ can normally be determined by knowing the physical properties of the system. For the case when $\alpha_x$ and $\alpha_f$ cannot be determined, refer to Remark 2 later in this section. Let $\text{rank}(E) = k < n$ and assume $p \geq q$ and that $C$, $M$ are full rank. Yeu et al. [12] developed a SMO for (1) to estimate $x$ and $f$ while Yu & Liu [13] investigated its necessary and sufficient conditions in terms of the original matrices $(E, A, M, C)$. A necessary condition is found to be

$$\text{rank} \begin{bmatrix} E^T & C \end{bmatrix} = n \hspace{1cm} (2)$$

which implies infinite observability. In this paper, we present a scheme that estimates $x$ and $f$ for a class of systems where (2) does not hold, i.e. $\text{rank}(A) = \bar{n} < n$.

Proposition 1. There exist transformations such that $x$ and $(E, A, M, C)$ can be written as

$$E = \begin{bmatrix} 0 & 0 & E_2 \\ 0 & I_{\bar{n} - p} & 0 \end{bmatrix}, \ E_2 = \begin{bmatrix} 0 \\ E_{22} \end{bmatrix}, \ \bar{n} - k \leq k + p - \bar{n}, \ C = [0 \ C_3] \hspace{1cm} (3)$$

J.H.T. Ooi, C.P. Tan, S.G. Nurzaman, and K.Y. Ng are with the School of Engineering and Advanced Engineering Platform, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway 46150, Selangor Malaysia (e-mail: tan.chee.pin@monash.edu).

This work was supported by the Fundamental Research Grant Scheme (FRGS) from the Ministry of Education Malaysia (grant code FRGS/2/2013/SG04/MUSM/02/1).
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2017.2665699, IEEE Transactions on Automatic Control

\[ A = \begin{bmatrix} A_1 & A_2 & A_3 \\ A_4 & A_5 & A_6 \\ A_7 & A_8 & A_9 \end{bmatrix} = \begin{bmatrix} A_1 & A_2 & A_3 \\ 0 & I_j & 0 \\ A_4 & A_10 & A_11 \\ A_5 & A_{13} & A_{14} \\ A_6 & A_{15} & A_{16} \\ A_7 & A_{17} & A_{18} \\ A_8 & A_{19} & A_{20} \end{bmatrix} \begin{bmatrix} n - n - j \\ j \\ n - k \\ k + p - n \end{bmatrix} \] (4)

\[ M = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix} = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{12} \\ x_2 \end{bmatrix} \begin{bmatrix} n - n - j \\ j \\ n - p \end{bmatrix} \] (5)

where \(|C_5| \neq 0, \text{rank}(E_{22}) = k + p - n\), the partitions of \(A\) and \(M\) have the same row dimensions, and the column partitions of \(E\), \(A\) are conformable to that of \(x\).

**Proof.** Two non-singular transformations will be used extensively: the state equation transformation - where the state equation (and hence \(E, A, M\) too) is pre-multiplied with a matrix, and the state transformation - where \(x\) is pre-multiplied with a matrix and \(E, A, C\) post-multiplied by its inverse. As rank \((C) = p\), a state transformation \(x \mapsto x_a = T_a x\), \(T_a = \left[ \begin{array}{c} N^T \\ C \end{array} \right] \) results in

\[ E \mapsto E_a = \begin{bmatrix} E_{a,1} & E_{a,2} \\ E_{a,3} & E_{a,4} \end{bmatrix}, \quad C \mapsto C_a = \begin{bmatrix} 0 & \bar{C} \end{bmatrix} \] (6)

where \(E_{a,4} \in \mathbb{R}^{p \times p}\). Since rank \((A) = n\), then decompose

\[ E_{a,1} = X_{11}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & I_{n-p} \end{bmatrix} X_{12}^{-1} \] (7)

Partition \(X_1 = \begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix} \). Then let

\[ \bar{E}_{a,2} = X_{11} \begin{bmatrix} E_{a,2} \\ E_{a,4} \end{bmatrix}, \quad \bar{E}_{a,4} = X_{12} \begin{bmatrix} E_{a,2} \\ E_{a,4} \end{bmatrix} \] (8)

resulting in rank \((\bar{E}_{a,2}) = k + p - n\). Decompose \(\bar{E}_{a,2} = X_3^{-1} E_2\) and multiply the state equation with \(\text{diag}(X_3, I_{n-p}) X_1\) and transform the state \(x_a \mapsto x_b = T_b x_a\),

\[ T_b = \begin{bmatrix} I_{n-n} & 0 & 0 \\ 0 & I_{n-p} & \bar{E}_{a,4} \\ 0 & 0 & I_p \end{bmatrix} X_2^{-1} \] (9)

such that

\[ E_a \mapsto E_b = \begin{bmatrix} 0 & 0 & E_{22} \\ 0 & 0 & I_{n-p} \end{bmatrix}, \quad C_a \mapsto C_b = \begin{bmatrix} 0 & 0 & C_{b,3} \end{bmatrix} \] (10)

where \(|C_{b,3}| \neq 0\) which is in the structure of (3). In these coordinates, let \((A, M)\) be

\[ A_b = \begin{bmatrix} A_{b,1} & A_{b,2} & A_{b,3} \\ A_{b,4} & A_{b,5} & A_{b,6} \\ A_{b,7} & A_{b,8} & A_{b,9} \end{bmatrix}, \quad M_b = \begin{bmatrix} M_{b,1} \\ M_{b,2} \\ M_{b,3} \end{bmatrix} \] (11)

Now multiply the state equation with \(T_c = \text{diag}(T_c, I_k)\), \(|T_c| \neq 0\) such that \(T_c A_b = A_c\). Let \(A_{c,1}\) be the top left \((n - n) \times (n - n)\) block of \(A_c\), and rank \((A_{c,1}) = j\). Then decompose

\[ A_{c,1} = X_5^{-1} \begin{bmatrix} 0 & 0 \\ 0 & I_j \end{bmatrix} X_6^{-1} \] (12)

and multiply the state equation with \(\text{diag}(X_5, I_n)\) and transform \(x_a \mapsto x_d = T_d x_b\), \(T_d = \text{diag}(X_6^{-1}, I_n)\) and as a result \(A\) will be transformed to have the structure in (4). Notice that the transformations \(T_c, T_d\) do not alter \(E, C\) from (10). Hence, the structures in (3) - (4) are achieved, completing the proof.

In the coordinates of (3) - (5), by treating \(x_{11}\) as an unknown input, and by eliminating \(x_{12}\), the system (1) can be re-expressed as a reduced-order system of order \(\bar{n}\) as follows:

\[ \ddot{\bar{x}} = \bar{A} \bar{x} + \bar{M} \bar{f}, \quad \bar{y} = \bar{C} \bar{x} \] (13)

\[ \ddot{E} = \begin{bmatrix} 0 & 0 \\ I_{n-p} & 0 \end{bmatrix}, \bar{C} = \begin{bmatrix} 0 & C_3 \end{bmatrix}, \] (14)

\[ \bar{A} = \begin{bmatrix} A_1 & A_{12} \\ A_{15} & A_{16} \\ A_{17} & A_{18} \end{bmatrix}, \bar{A}_7 = \begin{bmatrix} A_{10} & A_8 \end{bmatrix}, \bar{x} = \begin{bmatrix} x_{21} \\ y \end{bmatrix} \] (15)

**Remark 1.** Notice that this method of reformulating the system to (13) (treating states as unknown inputs to get a reduced-order system) is similar to the approach in [25], [26] (albeit for different systems purposes, where [25] considered sensor faults, and [26] used the reformulation for functional state estimation), and is different from other observer schemes for descriptor systems [27]-[29] which treat faults as states, resulting in higher-order systems.

Pre-multiply (13) with a nonsingular matrix \(R \in \mathbb{R}^{n \times n}\) and add \(V \bar{y}\) to both sides. Notice in (14) that rank \(\bar{E} = \bar{n}\) (and the system (13) is infinitely observable), thus there exist \(R\) and \(V\) such that \(R \bar{E} + V \bar{C} = \bar{I}_n\) and (13) becomes

\[ \ddot{\bar{x}} = R \bar{A} \bar{x} + R \bar{M} \bar{f} + V \bar{y} \] (15)

A SMO [12] for (13) has the form:

\[ \ddot{z} = (R \bar{A} - G_t \bar{C}) z + (-G_t (I_p - \bar{C} V) - R \bar{A} V) y - G_t \nu \] (16)

\[ \ddot{\bar{x}} = \nu - y \quad \text{z is attained in (17) into (16) yields} \] (17)

\[ \ddot{\bar{x}} = R \ddot{\bar{x}} + G_n \nu - G_t e_y + V \bar{y} \] (19)

Define \(e = \bar{x} - \bar{x}\). Then from (15) and (19), the state estimation error \(e\) (which characterizes the observer performance) can be obtained as follows:

\[ \dot{\epsilon} = (R \bar{A} - G_t \bar{C}) e + G_n \nu - R \bar{M} \bar{f} \] (20)

**Lemma 1.** If \(G_t\) and \(G_n\) are designed appropriately, and if \(\rho\) is chosen as

\[ \rho > \|(\bar{C} G_n)^{-1} \bar{C} R M\| (\alpha_x + \alpha_f) \] (21)

then sliding motion \((e_y = \bar{e}_y = 0 \Rightarrow \bar{y} = y)\) is attained in finite time, and \(x\) and \(f\) can be estimated by the observer.

**Proof.** Notice that the error equation (20) is in the same form as the Edwards-Spurgeon SMO [30]. Also, (21) implies \(\rho > \|(\bar{C} G_n)^{-1} \bar{C} R M\| \|f\|\), and using appropriate design methods.
The scheme described in this paper can be easily implemented using standard commands in MATLAB, and the observer has been found to perform satisfactorily in several works for instance [11].

Yu & Liu [13] investigated the necessary and sufficient conditions such that $G_1$ and $G_n$ exist to satisfy Lemma 1, and found them to be:

1. \[ \text{B1. rank } \begin{bmatrix} E & M \\ C & 0 \end{bmatrix} = \tilde{n} + \text{rank}(M) \]
2. \[ \text{B2. rank } \begin{bmatrix} sE - A & M \\ C & 0 \end{bmatrix} = \tilde{n} + \text{rank}(M), \forall s \in \mathbb{C}_+ \]

In addition, all components of $x_{11}$ and $f$ can be estimated if and only if

3. \[ \text{B3. rank}(M) = n - \tilde{n} + j + q \]

Note that B1 - B3 are based on (14) which depends on the transformation $T_\omega$. Hence it is important to determine the conditions in terms of the original system matrices (1) or (10) - (11) such that $T_\omega$ exists to satisfy B1 - B3. From the proof of Proposition 1, notice that after the structures in (10) - (11) have been attained, the top $n - k$ state equations undergo the transformation

\[ \Phi = \begin{bmatrix} X_5 & 0 \\ 0 & I_{n-k} \end{bmatrix} T_\omega \begin{bmatrix} \Phi_1 & \Phi_2 \\ T_\omega,2 \end{bmatrix} \]

where \( T_\omega = \begin{bmatrix} T_{\omega,1} & 0 \\ T_{\omega,2} & 0 \end{bmatrix} \). Hence it can be shown that the partitions of $A$, $M$ in (4) - (5) are

\[ \Phi_1 A_{11} = 0, \Phi_2 A_{11} X_6 = [0 I_j], T_{\omega,2} A_{11} X_6 = [A_9, A_{10}], \]
\[ M_2 = \Phi_2 M_{11}, M_3 = T_{\omega,2} M_{11}, M_4 = M_{2,1} \]

Similarly substitute (14) into B3 and use the Schur Complement to get rank $\Delta_2 = n - \tilde{n} + q$ and combining B1 and B3 yields rank $\Delta_1 = n - \tilde{n} + q$. From the proof of Proposition 1 and (23) - (24), it can be shown that

\[ \Delta_1 = \begin{bmatrix} \Phi_2 & 0 \\ T_{\omega,2} & 0 \\ I_{k+p-n} \end{bmatrix} \begin{bmatrix} A_{11} & M_{b,1} \\ A_{b,4} M_{b,2} & X_6 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ I_q \end{bmatrix} \]

Using (3) - (5), C1 is equivalent to rank $\Omega = n - \tilde{n} + q$. Suppose rank $\Omega < n - \tilde{n} + q$. From (31), it follows that rank $\Delta_1 \leq n - \tilde{n} + q$ (B1 or B3 not satisfied) which proves the necessity of C1. Then define

\[ \Omega_1 = [A_{b,1}, M_{b,1}], \Omega_2 = [A_{b,4}, M_{b,2}] \]

Now, further transformations will be performed on (10) - (11) to facilitate calculating $\Phi_1$, $\Phi_2$, and $T_{\omega,2}$. Let rank $(A_{b,1}) = \varphi$, and decompose $A_{b,1} = W_1^{-1} [\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \varphi \end{bmatrix} W_2^{-1}, \varphi \in \mathbb{R}^{\varphi \times \varphi}$, and

C2. \[ \text{rank } \begin{bmatrix} E & A & M \\ 0 & A & 0 \\ C & 0 & 0 \end{bmatrix} = n + \tilde{n} + q \]

**Proof.** First partition from (4) $A_1 = \begin{bmatrix} 0 \\ \varphi \end{bmatrix}$. Then substitute (14) into B1 to get

\[ \text{rank } \begin{bmatrix} A_9 & M_3 - A_{10} M_2 \\ A_{13} M_4 - A_{14} M_2 \\ A_{17} M_5 - A_{18} M_2 \end{bmatrix} = \text{rank } \begin{bmatrix} A_9 & M_3 - A_{10} M_2 \\ A_{13} M_4 - A_{14} M_2 \\ A_{17} M_5 - A_{18} M_2 \end{bmatrix} \]

The LHS can be re-expressed as:

\[ \begin{bmatrix} A_9 & M_3 - A_{10} M_2 \\ A_{13} M_4 - A_{14} M_2 \\ A_{17} M_5 - A_{18} M_2 \end{bmatrix} = \begin{bmatrix} A_9 & M_3 - A_{10} \\ A_{13} M_4 - A_{14} \\ A_{17} M_5 - A_{18} \end{bmatrix} [0 M_2] \]

Using the Schur Complement yields

\[ \text{rank } \begin{bmatrix} A_9 & M_3 - A_{10} \\ A_{13} M_4 - A_{14} \\ A_{17} M_5 - A_{18} \end{bmatrix} = \text{rank } \begin{bmatrix} A_9 & M_3 - A_{10} \\ A_{13} M_4 - A_{14} \\ A_{17} M_5 - A_{18} \end{bmatrix} - j \]

Repeat (26) - (27) for the RHS of (25) to get

\[ \text{rank } \begin{bmatrix} A_9 & M_3 - A_{10} \\ A_{13} M_4 - A_{14} \\ A_{17} M_5 - A_{18} \end{bmatrix} = \text{rank } \begin{bmatrix} A_9 & M_3 - A_{10} \\ A_{13} M_4 - A_{14} \\ A_{17} M_5 - A_{18} \end{bmatrix} - j \]

Substitute (27) - (28) into (25) to get

\[ \text{rank } \begin{bmatrix} A_9 & M_3 - A_{10} \\ A_{13} M_4 - A_{14} \\ A_{17} M_5 - A_{18} \end{bmatrix} = \text{rank } \begin{bmatrix} A_9 & M_3 - A_{10} \\ A_{13} M_4 - A_{14} \\ A_{17} M_5 - A_{18} \end{bmatrix} - j \]

Rearrange the rows and columns, compare with (4) - (5) and it is obvious that B1 is equivalent to

\[ \text{rank } \begin{bmatrix} A_{12} & M_2 \\ A_{14} & M_2 \end{bmatrix} = \text{rank } \begin{bmatrix} A_{12} & M_2 \\ A_{14} & M_2 \end{bmatrix} \]

Likewise substitute (14) into B3 and use the Schur Complement to get rank $\Delta_2 = n - \tilde{n} + q$ and combining B1 and B3 yields rank $\Delta_1 = n - \tilde{n} + q$. From the proof of Proposition 1 and (23) - (24), it can be shown that

\[ \Delta_1 = \begin{bmatrix} \Phi_2 & 0 \\ T_{\omega,2} & 0 \\ I_{k+p-n} \end{bmatrix} \begin{bmatrix} A_{b,1}, M_{b,1} \\ A_{b,4} M_{b,2} \end{bmatrix} X_6 \begin{bmatrix} 0 \\ 0 \\ I_q \end{bmatrix} \]

III. **MAIN RESULT**

Here the Schur Complement will be used extensively, where $Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$. Suppose rank $\Delta_1 \leq n - \tilde{n} + q$ (B1 or B3 not satisfied) which proves the necessity of C1. Then define

\[ \Omega_1 = [A_{b,1}, M_{b,1}], \Omega_2 = [A_{b,4}, M_{b,2}] \]
Define $n = \frac{M_1}{M_2}$, where $\text{rank}(M_1) = r$, $\text{rank}(M_2) = v$. Then further decompose

$$
\begin{bmatrix}
W_3 & 0 \\
0 & W_4
\end{bmatrix}
\begin{bmatrix}
\frac{M_1}{M_2}
\end{bmatrix}
W_5 =
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & I_r & 0 & 0 \\
0 & 0 & 0 & I_v
\end{bmatrix}
\tag{33}
$$

where $W_3, W_4$, and $W_5$ are invertible. Now define invertible matrices

$$
\tilde{T}_a = \begin{bmatrix} W_3 & 0 \\ 0 & W_4 \end{bmatrix}W_1, \quad \tilde{T}_b = \begin{bmatrix} W_2 & 0 \\ 0 & I_q \end{bmatrix} \begin{bmatrix} I_n - \tilde{n} & 0 \\ \tilde{n} & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\tag{34}
$$

and it can be shown that

$$
\Omega_1 \mapsto \tilde{T}_a \Omega_1 \tilde{T}_b = 
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & I_r & 0 & 0 & 0 \\
0 & 0 & I_v & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\tag{35}
$$

where $W_6 = (W_3 \Omega_l)^{-1}$. Suppose that $C1$ holds (since it is necessary), then $\Omega_l$ is full rank, and due to the structure of $\Omega_l$ in (35), there exists an invertible matrix $\tilde{T}_c = \begin{bmatrix} I_{n-k} & 0 \\ \tilde{T}_{c3} & \tilde{T}_{c4} \end{bmatrix}$ to be pre-multiplied with $\Omega_l$, resulting in $\Omega_2 \mapsto (\tilde{T}_{c3} \Omega_l + \tilde{T}_{c4} \Omega_2) \tilde{T}_b$ such that

$$
(\tilde{T}_{c3} \Omega_l + \tilde{T}_{c4} \Omega_2) \tilde{T}_b = 
\begin{bmatrix}
I_n - \varphi & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \tilde{A}_{42} \\
0 & 0 & I_{r-v} & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\tag{36}
$$

where $\varphi = k + p - n + \varphi - q + r$ and $\tilde{A}_{42} \in \mathbb{R}^{v \times v}$ is invertible.

**Remark 4.** The structure in (36) is obtained essentially from $\Omega_2$ and a linear combination of the rows of $\Omega_1$. The first $n-\tilde{n}-\varphi$ columns of $A_{b,4}$ and the first $q-r-v$ columns of $M_{b,2}$ are independent, since the corresponding columns of $A_{b,1}$ and $M_{b,1}$ are zero (so that $\Omega_l$ is f.c.r.) and a re-arrangement of rows (which can be achieved via QR decomposition) gives the form in (36). Then the next $\varphi-v$ columns of $A_{b,4}$ and the remaining columns of $M_{b,2}$ can be made zero by adding a linear combination of $\Omega_l$ (since the corresponding columns are independent). Finally, the last $v$ columns of $A_{b,4}$ are also independent of other columns of $\Omega_2$ so that $\Omega_l$ is f.c.r. and hence the rows can be re-arranged (using QR decomposition) to get the structure $\tilde{A}_{42}$ in (36).

Using (3) - (5), the LHS of $C2$ can be found to be $r$. Choose $j = \varphi$. Then the appropriate choice of $\Phi_2$ and $T_{o,2}$ to satisfy (24) are

$$
\frac{\Phi_2}{T_{o,2}} = \begin{bmatrix} \tilde{T}_{11} & \tilde{T}_{12} & \tilde{I}_j \end{bmatrix} \begin{bmatrix} \tilde{n} - k \end{bmatrix}
\tag{37}
$$

Define $\tilde{X} = \begin{bmatrix} 0 \\ v \end{bmatrix}$ and $\tilde{A}_{o} = [0 \tilde{A}_{42}]$. Then expand $\Xi_1$ (in (31)) using (35) - (37) and it loses rank i.f.f. the following loses rank:

$$
\Xi_{1,2} = \begin{bmatrix} I_j & \tilde{T}_{12} & \tilde{X} \\ \tilde{T}_{23} & \tilde{T}_{22} & \tilde{T}_{23}X \\ \tilde{A}_o & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{n} - k \\ v \end{bmatrix}
\tag{38}
$$

Using (3) - (5), $C2$ is equivalent to $r \leq \tilde{n} - k$. For $\Xi_1$ to be f.c.r., $\Xi_{1,2}$ (and hence $\tilde{T}_{22}$ too) must be f.c.r., and thus $\Xi_{1,2}$ cannot have more columns than rows, resulting in $r \leq \tilde{n} - k$ which shows the necessity of $C2$. Now let

$$
[\tilde{T}_{21} \tilde{T}_{22} \tilde{T}_{23}] = [\tilde{T}_{21} \tilde{T}_{22} \tilde{T}_{23}] \begin{bmatrix} \tilde{n} - k - r \end{bmatrix}
\tag{39}
$$

where $\gamma = \tilde{n} - k - r$. Choose $\tilde{T}_{23} = 0$, $\tilde{T}_{26} = 0$, and $\Xi_1$ is f.c.r. Hence $B1$ and $B3$ are satisfied, proving sufficiency of $C1$ - $C2$.

**Remark 5.** In [24], $\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}$ is treated as an unknown input, while for this work only $x_{11}$ is treated as an unknown input. Recall that $q \leq p \leq n$. Since only $x_{11}$ is treated as an unknown input, compared with [24], this scheme can work with fewer output measurements and it is less likely for the aforementioned condition to be violated, thus rendering the proposed design scheme applicable to a wider class of systems compared to [24].

**Theorem 2.** To satisfy $B2$, a necessary condition is

**C3.** $(E, A, M, C)$ is minimum phase

If $\gamma > 0$, then $C3$ is also sufficient; otherwise a sufficient condition is

**C4.** $(A_x, A_{b,52})$ is detectable, where

$$
A_x = A_{b,5} - W_3 W_2^{-1} W_1
\tag{40}
$$

and formal interpretations of the components of $A_x$ will be given in the proof below.

**Proof.** From (35) - (36), first partition

$$
\begin{bmatrix} A_{b,5} & A_{b,4} & M_{b,2} \\ A_{b,52} & 0 & 0 \\ M_{b,21} & 0 & 0 \\ A_{b,13} & A_{b,22} & A_{b,12} \end{bmatrix}
\tag{41}
$$

Next introduce

$$
W_2 = \begin{bmatrix} 0 & A_{b,22} \\ A_{b,23} & A_{b,51} \end{bmatrix}, W_3 = \begin{bmatrix} A_{b,5} & 0 \\ A_{b,21} & A_{b,12} \end{bmatrix}
\tag{42}
$$

Denote $R(E, A, M, C) = \begin{bmatrix} sE - A & M \\ 0 & C \end{bmatrix}$ which is the Rosenbrock matrix of $(E, A, M, C)$ and any zero of $(E, A, M, C)$ will cause it to lose rank. It can then be shown using (10) - (11) and (41) that the Rosenbrock matrix $R(E_{b}, A_{b}, M_{b}, C_{b})$ loses rank i.f.f. the following loses rank:

$$
R_{o,1}(s) = \begin{bmatrix} -A_{b,21} & 0 \\ -A_{b,52} & sI_{n-p} - A_{b,8} \end{bmatrix}
\tag{43}
$$

Then, from (14), the Rosenbrock matrix $R(\tilde{E}, \tilde{A}, \tilde{M}, \tilde{C})$ loses rank i.f.f. the following loses rank:

$$
\tilde{R}_{o,1}(s) = \begin{bmatrix} -\tilde{A}_{11} & -\tilde{A}_{10} \tilde{A}_{7} & A_{9} & M_{1} - A_{10} M_{2} \\ -\tilde{A}_{15} & -\tilde{A}_{14} \tilde{A}_{7} & A_{13} & M_{4} - A_{14} M_{12} \\ sI_{n-p} - (A_{19} - A_{18} \tilde{A}_{7}) & A_{17} & M_{5} - A_{18} M_{12} \end{bmatrix}
\tag{44}
$$

Use the Schur Complement, accordingly re-arrange the rows and columns, and it can be shown that $\tilde{R}_{o,1}(s)$ loses rank if
and only if the following loses rank:

$$\tilde{R}_{o,2}(s) = \begin{bmatrix} -A_7 & I_j & 0 & M_2 \\ -A_{11} & A_{10} & A_9 & M_1 \\ -A_{15} & A_{14} & A_{13} & M_4 \\ sI_{n-p} - A_{19} & A_{18} & A_{17} & M_5 \end{bmatrix}$$  \hspace{1cm} (45)

where the following relationship holds:

$$\tilde{R}_{o,2}(s) = \begin{bmatrix} \Phi_2 & 0 \\ T_{o,2} & 0 \\ 0 & I_k \end{bmatrix} R_{o,1}(s)$$  \hspace{1cm} (46)

showing that rank \((R(\tilde{E}, \tilde{A}, \tilde{M}, \tilde{C})) \leq \text{rank} \((R(E, A, M, C))\) and proves the necessity of C3. Now choose (from (37) and (39)) \(T_{11} = 0, T_{12} = 0, T_{22} = 0, T_{24} = 0, T_{25} \) to be invertible, and \(T_{21} \) to be full row rank (f.r.r.); these comply with the earlier requirements of \(T_{22} \) being f.c.r., and \([\Phi_2 \ T_{o,2}] \) to be f.r.r. Hence (37) becomes

$$\begin{bmatrix} \Phi_2 \\ T_{o,2} \end{bmatrix} = \begin{bmatrix} \Phi_2 & 0 \\ T_{o,2} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & I_1 \\ T_{21} & 0 & 0 \\ 0 & T_{25} & 0 \end{bmatrix} I.$$  \hspace{1cm} (47)

Now substitute (43) and (47) into (46) to get

$$\tilde{R}_{o,2}(s) = \begin{bmatrix} 0 & 0 & I_j & 0 \\ I_7 & 0 & 0 & 0 \\ 0 & T_{25} & 0 & 0 \\ 0 & 0 & 0 & I_k \end{bmatrix} \begin{bmatrix} -\tilde{T}_{21}A_{b,21} & 0 \\ -\tilde{W}_1 & \tilde{W}_2 \\ -A_{b,52} & 0 \\ sI_{n-p} - A_{b,8} \end{bmatrix} \tilde{R}_{o,3}(s)$$  \hspace{1cm} (48)

Recall that \(\text{rank}(\Omega) = n - \tilde{n} + q \) and that \(\Omega = \left[ \begin{array}{c} \Omega_1 \\ \Omega_2 \end{array} \right] \). From (35) - (36) and (41), it is clear that \(W_2 \in \mathbb{R}^{(n-\tilde{n}+q) \times (n-\tilde{n}+q)}\) is invertible. Define

$$I = \begin{bmatrix} I_{n-p+q} & 0 & 0 \\ 0 & I_{k} & 0 \\ \Pi & 0 & I_{n-p} \end{bmatrix}$$  \hspace{1cm} (49)

where \(\Pi = W_{1}W_{2}^{-1}\). It can then be shown that

$$\tilde{R}_{o,3}(s) = \begin{bmatrix} \Phi_1_2 \\ T_{o,2} \end{bmatrix} = \begin{bmatrix} I_7 & 0 \\ 0 & T_{21} \end{bmatrix} \begin{bmatrix} -\tilde{T}_{21}A_{b,21} & 0 \\ -\tilde{W}_1 & \tilde{W}_2 \\ -A_{b,52} & 0 \\ sI_{n-p} - A_{b,8} \end{bmatrix}$$  \hspace{1cm} (50)

Since \(W_2 \) is invertible, then \(\tilde{R}_{o,3}(s) \) loses rank i.f.f. the following matrix loses rank:

$$\tilde{R}_{o,4}(s) = \begin{bmatrix} -\Sigma \\ sI_{n-p} - A_x \end{bmatrix}, \quad \Sigma = \begin{bmatrix} -\tilde{T}_{21}A_{b,21} \\ -\tilde{W}_1 \\ -A_{b,52} \\ sI_{n-p} - A_x \end{bmatrix}$$  \hspace{1cm} (51)

where the unobservable modes of \((\Sigma, A_x)\) are the zeros of \((\tilde{E}, \tilde{A}, \tilde{M}, \tilde{C})\). Pre-multiply \(\tilde{R}_{o,5}(s)\) with \(\begin{bmatrix} I_{7+\xi} & 0 \\ 0 & H^{-1} \end{bmatrix}\), and post-multiply with \(H\), where the columns of \(H\) are the eigenvectors of \(A_x\) and hence \(H^{-1}A_xH\) is diagonal. Therefore, a zero of \((\tilde{E}, \tilde{A}, \tilde{M}, \tilde{C})\) which is an unobservable mode of \((\Sigma H, H^{-1}A_x H)\) will be the element of \(H^{-1}A_x H\) where the corresponding column of \(\Sigma H\) is zero. Now revisit the zeros of \((E, A, M, C)\) which are given by the values of \(s\) that make \(R_{o,1}(s)\) in (43) lose rank. It can be shown that

$$R_{o,1}(s) = \begin{bmatrix} I_{n-k-r-j} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} -A_{b,21} \\ -\tilde{W}_1 \\ -A_{b,52} \\ sI_{n-p} - A_x \end{bmatrix}$$  \hspace{1cm} (52)

Since \(W_2 \) is invertible, then \(\tilde{R}_{o,2}(s) \) loses rank i.f.f. the following matrix loses rank:

$$\tilde{R}_{o,3}(s) = \begin{bmatrix} -A_{b,21} \\ -A_{b,52} \\ sI_{n-p} - A_x \end{bmatrix}$$  \hspace{1cm} (53)

If \(C3\) holds, it means that if \(H^{-1}A_xH\) has any positive elements, then the corresponding columns of \([A_{b,21} \ A_{b,52}]\) will be nonzero; if \(\gamma > 0\), then \(\tilde{T}_{21} \) exists and it will always be possible that the columns of \(\Sigma H\) (corresponding to positive elements of \(H^{-1}A_x H\)) are non-zero, thus proving the sufficiency of \(C3\). If \(\gamma = 0\), then \(\tilde{T}_{21} \) does not exist; from (43), the Rosenbrock matrix \(R(\tilde{E}, \tilde{A}, \tilde{M}, \tilde{C})\) loses rank i.f.f. the following matrix loses rank:

$$\tilde{R}_{o,1}(s) = I \begin{bmatrix} -\tilde{W}_1 & \tilde{W}_2 \\ -A_{b,52} & 0 \\ sI_{n-p} - A_x \end{bmatrix}$$  \hspace{1cm} (54)

Since \(W_2 \) is invertible, then \(\tilde{R}_{o,2}(s) \) loses rank i.f.f. the following matrix loses rank:

$$\tilde{R}_{o,3}(s) = \begin{bmatrix} -A_{b,52} \\ sI_{n-p} - A_x \end{bmatrix}$$  \hspace{1cm} (55)

which is \(\tilde{R}_{o,5}(s)\) when \(\tilde{T}_{21} \) does not exist. The values of \(s\) that make \(\tilde{R}_{o,3}(s)\) lose rank are the unobservable modes of \((A_x, A_{b,52})\) and hence the sufficiency of \(C4\) is proven. ■

An appropriate choice for \(\Phi_1\) is \([T_{x1} \ T_{x2} \ 0]\) (partitioned conformably with (37)) and \(T_{x1} \) is chosen such that \([T_{x1} \ T_{x2}]\) (and \(\Phi\) as well) is invertible. Then for the original system (10) - (11), post-multiply \(\text{diag}(\Phi, I_k)\) with the state equation transformation \(\tilde{T}_x = \text{diag}(T_n, I_k)\), and let \(X_0 = T_{lb1}\) where \(\text{diag}(T_{lb1}, T_{lb2}) = \tilde{T}_b\), all of which do not alter \(E, C\) in (10). Meanwhile, the transformation \(\tilde{T}_c\) (in the proof of Theorem 1) is not required in the algorithm design.

Remark 6. The works in [27]-[29], [32], [33] use linear observers, which can estimate faults only if they appear in the output equation; for faults that appear only in the state equation (which is the case considered in (1) of this paper), asymptotic estimation cannot be achieved. For the case of faults in the output equation, the following sub-section will demonstrate that our scheme is still applicable.

A. Extension to the case of sensor faults estimation

Consider the system (1) with sensor fault as follows

$$E\tilde{x} = Ax + Mf, \quad y = Cx + Nf_s$$  \hspace{1cm} (56)
where \( f_s \in \mathbb{R}^h \) is the sensor fault, \( N \in \mathbb{R}^{p \times h} \), and \( \text{rank}(N) = h \) where \( h \leq p \). Define an invertible matrix \( K \) such that
\[
KN = \begin{bmatrix} 0 \\ N_2 \end{bmatrix}
\]  
(57)
where \( N_2 \in \mathbb{R}^{h \times h} \) is invertible. Thus scaling the output \( y \) through the matrix \( K \) (effectively pre-multiplying the output equation of (56) with \( K \)) yields
\[
Ky = \begin{cases} 
\dot{y}_1 = C_1 x \\
\dot{y}_2 = C_2 x + N_2 f_s
\end{cases}
\]  
(58)
Consider a measurable signal \( \epsilon \in \mathbb{R}^h \) generated by \( y_2 \) as follows
\[
\dot{\epsilon} = -A_f \epsilon + A_f y_2
\]  
(59)
where \(-A_f \in \mathbb{R}^{h \times h} \) is a stable matrix. Substituting from (58) into (59) yields
\[
\dot{\epsilon} = -A_f \epsilon + A_f C_2 x + A_f N_2 f_s
\]  
(60)
Combining (56) and (60) forms an augmented system as follows:
\[
\begin{bmatrix} E_0 \ I_h \\
\bar{A}_a \\
\bar{C}_a
\end{bmatrix} \begin{bmatrix} \dot{x} \\
\epsilon
\end{bmatrix} = \begin{bmatrix} A & 0 \\
A_f C_2 & -A_f \\
C_1 & 0 \end{bmatrix} \begin{bmatrix} x \\
\epsilon
\end{bmatrix} + \begin{bmatrix} M \\
0 \\
0 \end{bmatrix} \begin{bmatrix} f_s \\
\epsilon
\end{bmatrix}
\]  
(61)
Alternatively, if \( f_s = f \), the system (56) becomes
\[
E \dot{x} = Ax + M f, \ y = C x + N f
\]  
(62)
where \( N \in \mathbb{R}^{p \times q} \) and \( \text{rank}(N) = q \). By repeating the steps from (56) - (60), the augmented system (61) then becomes:
\[
\begin{bmatrix} E_0 \ I_h \\
\bar{A}_a \\
\bar{C}_a
\end{bmatrix} \begin{bmatrix} \dot{x} \\
\epsilon
\end{bmatrix} = \begin{bmatrix} A & 0 \\
A_f C_2 & -A_f \\
C_1 & 0 \end{bmatrix} \begin{bmatrix} x \\
\epsilon
\end{bmatrix} + \begin{bmatrix} M \\
0 \end{bmatrix} \begin{bmatrix} f \\
\epsilon
\end{bmatrix}
\]  
(63)
Equations (61) and (63) are in the same form as (1). Hence the proposed observer (16) - (17), transformations in Section II, and the analysis in Section III are applicable to \( (E_a, \bar{A}_a, \bar{M}_a, \bar{C}_a) \) (for both cases (61) and (63)) and therefore the augmented fault (containing sensor faults) can be estimated using the method in Lemma 1. Thus, the proposed observer in this paper is also applicable for the case of sensor faults.

IV. NUMERICAL EXAMPLE

Consider a chemical mixing tank [12] with matrices:
\[
E = \begin{bmatrix} 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{bmatrix}, \quad M = \begin{bmatrix} 0.1 & 0 & 0.02 \\
0 & 1 & 0 \\
0 & 0 & 0.02 \\
0 & 0 & 1
\end{bmatrix}
\]  
(64)
\[
A = \begin{bmatrix} -0.375 & -0.0667 & 0 & 0 \\
0 & -1 & 0 & 0 \\
0.3 & 0.0535 & -0.5 & -0.04 \\
0 & 0 & 0 & -1
\end{bmatrix}
\]
It was found that \( n = 4 \) and \( \bar{n} = 3 \), indicating infinite unobservability. By inspection, C1 and C2 hold. With regard to the stability of invariant zeros, the necessity of C3 and sufficiency of C4 hold since \( \gamma = 0 \). All control inputs are faulty and therefore the method in [23] cannot be used. Meanwhile, as \( p = q \), the observer scheme in [24] cannot work. The state equation \( T_{se} \) and state \( T_{st} \) transformations were respectively chosen as
\[
T_{se} = \begin{bmatrix} 0 & 1 & 0 & 0 \\
0 & 0 & -1 & 0 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0
\end{bmatrix}, \quad T_{st} = \begin{bmatrix} 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]  
(65)
to arrive at (10) and (35). Then a choice of \( \Phi = \begin{bmatrix} 0 & 1 \\
1 & 0
\end{bmatrix} \) returned (4). From (64), notice that \( x_{11} \) is non-existent and \( x_{12} \in \mathbb{R}^1 \). Also, \( \text{rank} \left[ \bar{E}^T \ C^T \right]^T = \bar{n} = 3 \), and B1 - B3 hold, indicating that the SMO by [12] can be used to estimate \( \bar{x} \) and \( \bar{f} \). Using (14) resulted in
\[
R = \begin{bmatrix} 0 & 0 & 1 \\
0 & -0.5 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1
\end{bmatrix}, \quad V = \begin{bmatrix} 0 \\
0.5 \\
0 \\
1
\end{bmatrix}
\]  
(66)
The observer and sliding motion poles were respectively chosen as \((-0.5, -13, -15)\) and \((-0.5)\). Then using the design method in [30] yielded
\[
G_l = \begin{bmatrix} 25.25 & 1.3958 \\
13.05 & 0.0216
\end{bmatrix}, \quad G_n = \begin{bmatrix} 2 & 0.1 \\
1 & 0
\end{bmatrix}
\]  
(67)
The system has an initial condition of \( x = (7, 0, 2, 0) \) and the SMO’s was set to zero, while \( f_1 = 0.2 \sin(0.2t) \), \( f_2 = 0.2 \sin(0.5t + 0.25\pi) \). Figures 1 - 2 shows the outputs \( y \) and their estimates, where the estimate is visually identical to \( y \), demonstrating that sliding motion occurs almost instantaneously (at 0.06 seconds). Figures 3 - 4 show the states \( x_{12} \) and \( x_2 \) and their estimates, whilst Figures 5 - 6 show the components of \( f \) and their estimates, where it can be seen that the proposed observer scheme successfully estimates the states and faults.

Fig. 1. The first component of \( y \) (solid) and its estimate (dashed); inset - a zoomed-in version to show convergence.

Next, zero-mean white noise (of standard deviation 0.06) was injected into the sensors \( y \). Figures 7 - 10 show the estimates of \( x \) and \( f \) which still shows accurate estimation, hence the proposed observer is still applicable with sensor noise.
V. CONCLUSION

This paper has presented a SMO-based scheme that enables state and fault estimation for a class of infinitely unobservable descriptor systems. The system was re-expressed as a reduced-order infinitely observable system by reformulating certain states in terms of other states, and treating certain states as unknown inputs. The SMO was then implemented on the reduced-order system; and the existence conditions were investigated. It was found that the existence conditions of the proposed scheme are more relaxed than that of previous works, and thus is applicable to a wider class of systems. A simulation example demonstrated the effectiveness of the scheme.

REFERENCES


Fig. 6. The second component of the fault $f$ (solid) and its estimate (dashed).

Fig. 7. $x_{12}$ (solid) and its estimate (dashed) with noise present at the sensors.

Fig. 8. $x_2$ (solid) and its estimate (dashed) with noise present at the sensors.

Fig. 9. The first component of the fault $f$ (solid) and its estimate (dashed) with noise present at the sensors.

Fig. 10. The second component of the fault $f$ (solid) and its estimate (dashed) with noise present at the sensors.


