New result in robust actuator fault reconstruction with application to an aircraft

Kok Yew Ng, Chee Pin Tan, Christopher Edwards and Ye Chow Kuang

Abstract— **This paper presents a robust actuator fault reconstruction scheme for linear uncertain systems using sliding mode observers. In existing work, fault reconstruction via sliding mode is limited to either linear certain systems subject to unknown inputs, relative degree one systems or a specific class of relative degree two systems; in particular systems that have more outputs than unknown inputs, or systems whereby all position measurements are available. This paper presents a new method that is applicable to a wider class of systems with relative degree higher than one, and can also be used for systems with more unknown inputs than outputs, and systems where not all position measurements are available. The method uses two sliding mode observers in cascade. Signals from the first observer are processed and used to drive the second observer. Overall this results in actuator fault reconstruction being feasible for a wider class of systems than existing methods, in particular is useful for the application to an aircraft.**

I. Introduction

Fault detection and isolation (FDI) is an important area of research. A fault is deemed to occur when the system experiences an abnormal condition, such as a malfunction in the actuators or sensors. The fundamental purpose of an FDI scheme is to generate an alarm when a fault occurs and to identify its location. An overview of work in this area appears in [6]. The most commonly used FDI methods are observer-based where the plant output is compared with the observer output, and the discrepancy is used to form a residual [12] which then is used to determine whether a fault is present.

A useful alternative to residual generation is *fault reconstruction* [4][11], which not only detects and isolates the fault, but provides an estimate of the fault so that its shape and magnitude can be better understood and more precise corrective action can be taken. However, a fault reconstruction scheme is usually designed about a model of the system and this model usually does not perfectly represent the system as it will possess uncertainties represented as a class of disturbances within the model [9]. The disturbances could corrupt the reconstruction, producing nonzero reconstructions when there are no faults, or worse, masking the effect of a fault, producing a 'zero' reconstruction in the presence of faults. Hence, the scheme needs to be designed so that the reconstruction is robust to disturbances.

Edwards *et al.* [4] used a Sliding Mode Observer (SMO) [3] to reconstruct faults, but there was no explicit consideration of the disturbances. Tan & Edwards [14] built on this work and designed the observer using Linear Matrix Inequalities (LMIs) [1] to minimize the $\mathcal{L}_2$ gain from the disturbances to the fault reconstruction. Saif & Guan [11] aggregated the faults and disturbances to form a new 'fault' vector and used a linear observer to reconstruct this new 'fault' vector. A necessary condition in [4][14][11] is that the transfer function from the faults to the output has a relative degree of one. This limits the class of systems for which the schemes [4][14][11] are applicable. Recently, there have been developments in the area of fault reconstruction for systems with relative degree higher than one. Floquet & Barbot [5] converted the system into an 'output information' form so that existing SMO techniques could be used to reconstruct faults. However, their design does not consider disturbances and the class of systems for which the transformation is feasible is unknown. Hence, it is not easy to determine whether the algorithm suits the system under consideration. Davila *et al.*[2] used a 2nd order SMO on nonlinear mechanical systems where only position is measured. The work in [2] could be easily extended to the case of robust fault reconstruction for actuator faults occurring in the acceleration equation. However, it is applicable only to a limited class of systems as it requires *all* position signals to be measurable.

This paper presents a robust fault reconstruction method for systems with relative degree higher than one, relaxing the condition required by [4][14]. The method in this paper essentially uses two SMOs [3] in cascade. Suitable processing of the equivalent output error injection in the first observer yields the measurable output of a 'fictitious' system that is relative degree one. This means the robust fault reconstruction method in [14] is applicable to the fictitious system and a second observer is implemented on the fictitious system to generate a reconstruction of the fault that is robust to the disturbances. This approach is applicable to a wider class of systems than the methods in [4][14]. Furthermore, this paper considers robustness against disturbances and the scheme may be feasible for systems for which the method in [2] is not applicable. In terms of real engineering applications, the scheme proposed in this paper gives the benefit of requiring less sensors in the application to an aircraft.

This paper is organized as follows: §II introduces the system and states the main result, whilst §III sets up the framework for the proposed method together with existence conditions. An example to demonstrate the effectiveness of the scheme is given in §IV and finally §V makes some conclusions. The notation used throughout this paper is quite standard; in particular $\|.\|$ represents the Euclidean norm for

Ng, Tan and Kuang are with the School of Engineering, Monash University Malaysia, 2 Jalan Kolej, 46150 Petaling Jaya, Malaysia tan.chee.pin@eng.monash.edu.my

Edwards is with the Engineering Department, Leicester University, University Road, LE1 7RH UK

vectors and the induced spectral norm for matrices, and $\lambda(.)$ denotes the spectrum of a square matrix.

II. Preliminaries and Statement of The Main Result

Consider a system

$$\begin{aligned}\dot{\tilde{x}}(t) &= \tilde{A}\tilde{x}(t) + \tilde{B}u(t) + \tilde{M}f(t) + \tilde{Q}\tilde{\xi}(t) && (1)\\ y(t) &= \tilde{C}\tilde{x}(t) && (2)\end{aligned}$$

where $\tilde{x} \in \mathbb{R}^{\tilde{n}}$, $y \in \mathbb{R}^{p}, u \in \mathbb{R}^{m}$ are the states, outputs and inputs respectively with $\tilde{n} \geq p$ while $f \in \mathbb{R}^{q}$ is an unknown fault and $\tilde{\xi} \in \mathbb{R}^{h}$ is an unknown disturbance, which encapsulates all system nonlinearities. Assume $rank(\tilde{M}) = q, rank(\tilde{Q}) = h, rank(\tilde{C}) = p$ and suppose $rank(\tilde{C}\tilde{M}) = r < q < p$. Also assume that $(\tilde{A}, \tilde{C})$ is observable.

The main objective is to reconstruct f whilst being robust to $\tilde{\xi}$. Edwards *et al.*[4] have reconstructed f for the case when $\tilde{\xi} = 0$. Tan & Edwards [14] built on this early work and presented a method that minimizes the $\mathcal{L}_2$ gain from $\tilde{\xi}$ to the fault reconstruction. The fault reconstruction scheme in [4][14] is feasible if and only if the following conditions are satisfied

A1. $rank(\tilde{C}\tilde{M}) = rank(\tilde{M}) = q$

A2. Any invariant zeros of $(\tilde{A}, \tilde{M}, \tilde{C})$ are stable

Condition A1 implies that the system is relative degree one and A2 implies that the system is minimum phase. This paper proposes a method to robustly reconstruct the fault when A1 is not satisfied. The fulfilment of A1 implies that there is a certain minimum number of appropriate sensors.

Assume that the disturbance $\tilde{\xi}$ is piecewise continuous [11]

$$\dot{\tilde{\xi}}(t) = A_{\Omega}\tilde{\xi}(t) + B_{\Omega}\xi(t) \tag{3}$$

where $\xi \in \mathbb{R}^{h}$ and $A_{\Omega} \in \mathbb{R}^{h\times h}$ is stable and $B_{\Omega} \in \mathbb{R}^{h\times h}$. If $\tilde{\xi}$ is known to be a signal in the frequency region $\omega_1 < \omega < \omega_2$, then (3) can be taken to be first order filters with cut-off frequency ω_2.

Theorem 1: For the case when A1 is not satisfied, i.e. $r = rank(\tilde{C}\tilde{M}) < rank(\tilde{M}) = q$, then the fault f can be reconstructed such that the $\mathcal{L}_2$ gain from ξ to the fault reconstruction will be bounded if

B1. $rank\begin{bmatrix} \tilde{C}\tilde{A}\tilde{M} & \tilde{C}\tilde{M} \\ \tilde{C}\tilde{M} & 0 \end{bmatrix} = rank(\tilde{C}\tilde{M}) + rank(\tilde{M})$

B2. Any invariant zeros of $(\tilde{A}, \tilde{M}, \tilde{C})$ must be stable □

It is clear that B1 is less restrictive than A1. The next section will provide a constructive proof of Theorem 1.

III. Robust Fault Reconstruction

Lemma 1: There exist nonsingular linear transformations $\tilde{x} \mapsto T_1\tilde{x}, f \mapsto T_2 f$ such that the triple $(\tilde{A}, \tilde{M}, \tilde{C})$ from (1) - (2) in the new coordinates are given by

$$\tilde{A} = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 \\ \tilde{A}_3 & \tilde{A}_4 \end{bmatrix}, \tilde{C} = \begin{bmatrix} 0 & \tilde{T} \end{bmatrix}, \tilde{M} = \begin{bmatrix} \tilde{M}_1 \\ \tilde{M}_2 \end{bmatrix} \tag{4}$$

where $\tilde{A}_1 \in \mathbb{R}^{(\tilde{n}-p)\times(\tilde{n}-p)}, \tilde{M}_2 \in \mathbb{R}^{p\times q}$ and $\tilde{T} \in \mathbb{R}^{p\times p}$ is orthogonal. Furthermore, $\tilde{M}_1, \tilde{M}_2$ can be partitioned to be

$$\tilde{M}_1 = \begin{bmatrix} 0 & 0 \\ M_{11} & 0 \end{bmatrix}, \quad \tilde{M}_2 = \begin{bmatrix} 0 & 0 \\ 0 & M_{22} \end{bmatrix} \tag{5}$$

where M_{11}, M_{22} are invertible. In this coordinate system, $f \mapsto T_2 f = \text{col}(f_1, f_2)$ where $f_2 \in \mathbb{R}^{r}$.

Proof: See Proposition 2 in [15]. ■

In the coordinates of (4) - (5), further partition $\tilde{A}$, $\tilde{Q}$ as

$$\tilde{A} = \left[\begin{array}{cc|cc} \tilde{A}_{11} & \tilde{A}_{12} & \tilde{A}_{13} & \tilde{A}_{14} \\ \tilde{A}_{21} & \tilde{A}_{22} & \tilde{A}_{23} & \tilde{A}_{24} \\ \hline \tilde{A}_{31} & \tilde{A}_{32} & \tilde{A}_{33} & \tilde{A}_{34} \\ \tilde{A}_{41} & \tilde{A}_{42} & \tilde{A}_{43} & \tilde{A}_{44} \end{array}\right], \quad \tilde{Q} = \left[\begin{array}{c} \tilde{Q}_{11} \\ \tilde{Q}_{12} \\ \hline \tilde{Q}_{21} \\ \tilde{Q}_{22} \end{array}\right] \tag{6}$$

Combine (1) - (2) and (3) to obtain the augmented system

$$\underbrace{\begin{bmatrix} \dot{\tilde{\xi}} \\ \dot{\tilde{x}} \end{bmatrix}}_{\dot{x}} = \underbrace{\begin{bmatrix} A_{\Omega} & 0 \\ \tilde{Q} & \tilde{A} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \tilde{\xi} \\ \tilde{x} \end{bmatrix}}_{x} + \underbrace{\begin{bmatrix} 0 \\ \tilde{B} \end{bmatrix}}_{B} u + \underbrace{\begin{bmatrix} 0 \\ \tilde{M} \end{bmatrix}}_{M} f + \underbrace{\begin{bmatrix} B_{\Omega} \\ 0 \end{bmatrix}}_{Q} \xi \tag{7}$$

$$y = \underbrace{\begin{bmatrix} 0 & \tilde{C} \end{bmatrix}}_{C} \underbrace{\begin{bmatrix} \tilde{\xi} \\ \tilde{x} \end{bmatrix}}_{x} \tag{8}$$

Expanding the matrices in (7) - (8) as in (4) - (5) gives

$$A = \left[\begin{array}{ccc|cc} A_{\Omega} & 0 & 0 & 0 & 0 \\ \tilde{Q}_{11} & \tilde{A}_{11} & \tilde{A}_{12} & \tilde{A}_{13} & \tilde{A}_{14} \\ \tilde{Q}_{12} & \tilde{A}_{21} & \tilde{A}_{22} & \tilde{A}_{23} & \tilde{A}_{24} \\ \hline \tilde{Q}_{21} & \tilde{A}_{31} & \tilde{A}_{32} & \tilde{A}_{33} & \tilde{A}_{34} \\ \tilde{Q}_{22} & \tilde{A}_{41} & \tilde{A}_{42} & \tilde{A}_{43} & \tilde{A}_{44} \end{array}\right] \tag{9}$$

$$M = \left[\begin{array}{cc} 0 & 0 \\ 0 & 0 \\ M_{11} & 0 \\ \hline 0 & 0 \\ 0 & M_{22} \end{array}\right], Q = \left[\begin{array}{c} B_{\Omega} \\ 0 \\ 0 \\ \hline 0 \\ 0 \end{array}\right], C = \begin{bmatrix} 0 & \tilde{T} \end{bmatrix} \tag{10}$$

Lemma 2: The pair (A, C) from (7) - (8) is detectable.

Proof: See §VI-A in the appendix. ■

Lemma 3: Condition B1 from Theorem 1 is satisfied if and only if $\tilde{A}_{32}$ from (6) has full column rank $q - r$.

Proof: See §VI-B in the appendix. ■

Define $\bar{p} := rank\begin{bmatrix} \tilde{Q}_{21} & \tilde{A}_{31} & \tilde{A}_{32} \end{bmatrix} + r$. It follows that $\bar{p} - r \leq min\{p - r, n - p\}$ and therefore $\bar{p} \leq p$. Since condition B1 implies that $\tilde{A}_{32}$ has full column rank, then $\bar{p} - r \geq q - r$ which implies that $\bar{p} \geq q$.

Lemma 4: There exists a coordinate transformation such that A, M, Q, C from (9) - (10) have the structure below:

$$A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} = \left[\begin{array}{ccc|cc} A_{11} & A_{12} & A_{13} & A_{14} & A_{15} \\ A_{21} & A_{22} & A_{23} & A_{24} & A_{25} \\ A_{31} & A_{32} & A_{33} & A_{34} & A_{35} \\ \hline 0 & A_{42} & A_{43} & A_{44} & A_{45} \\ A_{51} & A_{52} & A_{53} & A_{54} & A_{55} \end{array}\right] \tag{11}$$

$$M = \left[\begin{array}{cc} 0 & 0 \\ 0 & 0 \\ M_{11} & 0 \\ \hline 0 & 0 \\ 0 & M_{22} \end{array}\right], Q = \left[\begin{array}{c} Q_{11} \\ Q_{12} \\ Q_{13} \\ \hline 0 \\ 0 \end{array}\right], C = \begin{bmatrix} 0 & T \end{bmatrix} \tag{12}$$

where $\begin{bmatrix} A_{42} & A_{43} \end{bmatrix} \in \mathbb{R}^{(p-r)\times(\bar{p}-r)}$ can be partitioned as

$$\begin{bmatrix} A_{42} & A_{43} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ A^{o}_{42} & A^{o}_{43} \end{bmatrix} \tag{13}$$

where $\begin{bmatrix} A_{42}^o & A_{43}^o \end{bmatrix}$ is square and invertible and $rank(\tilde{A}_{32}) = rank(A_{43}^o)$ while $T \in \mathbb{R}^{p\times p}$ is orthogonal.

Proof: See §VI-C in the appendix. ■

The canonical form in (11) - (12) is the basis for the proof of Theorem 1 which will be developed in the next section. Also partition $A_3 \in \mathbb{R}^{p\times(n-p)}$ from (11) as

$$A_3 = \begin{bmatrix} A_{311} \\ A_{312} \end{bmatrix} = \begin{bmatrix} 0 & A_{42} & A_{43} \\ A_{51} & A_{52} & A_{53} \end{bmatrix} \tag{14}$$

Assume that the unknown signals $f(t), \xi(t)$ are norm bounded by known scalars α, β so that $\|f(t)\| < \alpha,\ \|\xi\| < \beta$

The remainder of this section develops a fault estimation scheme for $f(t)$ based on a pair of SMOs.

A. A fault reconstruction scheme (proof of Theorem 1)

A SMO [3] for the system (7) - (8) is

$$\begin{aligned} \dot{\hat{x}}(t) &= A\hat{x}(t) + Bu(t) - G_l e_y(t) + G_n\nu \quad (15)\\ \hat{y}(t) &= C\hat{x}(t) \quad (16) \end{aligned}$$

where $\hat{x} \in \mathbb{R}^n$ is the estimate of x and $e_y = \hat{y} - y$ is the output estimation error and $G_l, G_n \in \mathbb{R}^{n\times p}$ are observer gains that are to be designed where G_n has the structure

$$G_n = \begin{bmatrix} -LT^T \\ T^T \end{bmatrix} P_o^{-1} \tag{17}$$

where $P_o \in \mathbb{R}^{p\times p}$ is symmetric positive definite (s.p.d.) and $L \in \mathbb{R}^{(n-p)\times p}$ is such that $A_1 + LA_3$ is stable. The term ν is a nonlinear discontinuous term defined by

$$\nu = -\rho\frac{e_y}{\|e_y\|},\ e_y \neq 0, \quad \rho \in \mathbb{R}_+ \tag{18}$$

Define $e := \hat{x} - x$ and $A_o = A - G_l C$. Then combine (7), (8), (15) - (16) to obtain the error system

$$\dot{e}(t) = A_o e(t) + G_n\nu - Mf(t) - Q\xi(t) \tag{19}$$

For a proper choice of G_l and a large enough choice of ρ, an ideal sliding motion takes place on $\mathcal{S} = \{e : Ce = 0\}$ [14] in finite time where the sliding motion dynamics are governed by $A_1 + LA_3$. Since from Lemma 2 the pair (A, C) is detectable, using the Popov-Hautus-Rosenbrock (PHR) test [10], it can be shown that (A_1, A_3) is detectable and so an L can always be found to make $A_1 + LA_3$ stable.

Introduce a change of coordinates $x \mapsto T_L x = \begin{bmatrix} x_1 \\ y \end{bmatrix}$ where

$$T_L = \begin{bmatrix} I_{n-p} & L \\ 0 & T \end{bmatrix}$$

Then the matrices in (11) - (12) are transformed to be

$$T_L A T_L^{-1} = \begin{bmatrix} A_1 + LA_3 & * \\ TA_3 & * \end{bmatrix}, T_L M = \begin{bmatrix} M_1 + LM_2 \\ TM_2 \end{bmatrix} \tag{20}$$

$$CT_L^{-1} = \begin{bmatrix} 0 & I_p \end{bmatrix}, T_L Q = \begin{bmatrix} Q_1 \\ 0 \end{bmatrix}, T_L G_n = \begin{bmatrix} 0 \\ P_o^{-1} \end{bmatrix} \tag{21}$$

where $x_1 \in \mathbb{R}^{n-p}$ are the 'non-output' states, and (*) are matrices that play no role in the analysis that follows. Partition (19) according to (20) and (21), and let e_1 be the estimation error of x_1. Assume that an ideal sliding motion has taken place on $\mathcal{S}$ so that $e_y = \dot{e}_y = 0$ [3][4], then the error system (19) can be re-expressed as

$$\dot{e}_1(t) = (A_1 + LA_3)e_1(t) - (M_1 + LM_2)f(t) - Q_1\xi(t) \tag{22}$$

$$T^T P_o^{-1}\nu_{eq} = -A_3 e_1(t) + M_2 f(t) \tag{23}$$

where ν_{eq} is the equivalent output error injection signal required to maintain a sliding motion and can be approximated to any degree of accuracy [4] by replacing ν with

$$\nu = -\rho\frac{e_y}{\|e_y\| + \delta} \tag{24}$$

where δ is a small positive scalar. As the term e_y is a measurable signal, the signal ν_{eq} is computable online.

Define $v := T^T P_o^{-1}\nu_{eq}$ and partition $v = \text{col}(v_1, v_2)$ where $v_2 \in \mathbb{R}^r$. Then partition (23) according to (14) as

$$\begin{aligned} v_1(t) &= -A_{311}e_1(t) \quad (25)\\ v_2(t) &= -A_{312}e_1(t) + M_{22}f_2(t) \quad (26) \end{aligned}$$

where f_2 is defined in Lemma 1. Define $Z \in \mathbb{R}^{(\bar{p}-r)\times(p-r)}$ as $Z = \begin{bmatrix} 0 & I_{\bar{p}-r} \end{bmatrix}$ and then premultiply (25) with Z to get

$$\bar{v}_1(t) := Zv_1(t) = -ZA_{311}e_1(t) \tag{27}$$

From (13) and the partitions of A_{311} in (14) it is clear that $ZA_{311} = \begin{bmatrix} 0 & A_{42}^o & A_{43}^o \end{bmatrix}$ has rank $\bar{p} - r$ as found in Lemma 4.

Now low-pass filter v_2 to produce v_f according to

$$\begin{aligned} \dot{v}_f(t) &= -A_f v_f(t) + A_f v_2(t)\\ &= -A_f v_f(t) - A_f A_{312}e_1(t) + A_f M_{22}f_2(t) \quad (28) \end{aligned}$$

where $-A_f \in \mathbb{R}^{r\times r}$ is a stable design matrix, and combine (22), (27), and (28) to get the following system

$$\begin{aligned} \dot{z}(t) &= \begin{bmatrix} \dot{e}_1(t) \\ \dot{v}_f(t) \end{bmatrix} = \mathcal{A}z(t) + \mathcal{M}f(t) + \mathcal{Q}\xi(t) \quad (29)\\ \bar{y}(t) &= \begin{bmatrix} \bar{v}_1(t) \\ v_f(t) \end{bmatrix} = \mathcal{C}z(t) \quad (30) \end{aligned}$$

where

$$\mathcal{A} = \begin{bmatrix} A_1 + LA_3 & 0 \\ -A_f A_{312} & -A_f \end{bmatrix}, \mathcal{Q} = \begin{bmatrix} -Q_1 \\ 0 \end{bmatrix} \tag{31}$$

$$\mathcal{M} = \begin{bmatrix} -(M_1 + LM_2) \\ \begin{bmatrix} 0 & A_f M_{22} \end{bmatrix} \end{bmatrix}, \mathcal{C} = \begin{bmatrix} -ZA_{311} & 0 \\ 0 & I_r \end{bmatrix} \tag{32}$$

Define a transformation $\mathcal{T} \in \mathbb{R}^{\bar{n}\times\bar{n}}$ so that $\bar{x} = \mathcal{T}z$ where

$$\mathcal{T} = \begin{bmatrix} I_{\bar{n}-r} & \tilde{L}A_f^{-1} \\ 0 & I_r \end{bmatrix}$$

and $\tilde{L}$ represents the last r columns of L. After transformation, $(\mathcal{A}, \mathcal{M}, \mathcal{C}, \mathcal{Q})$ from (29) - (30) will have the structure

$$\bar{A} = \left[\begin{array}{c|c} A_{11} & * \\ \hline A_{21} & * \\ A_{31} & * \\ -A_f A_{51} & * \end{array}\right] = \begin{bmatrix} \bar{A}_1 & \bar{A}_2 \\ \bar{A}_3 & \bar{A}_4 \end{bmatrix} \tag{33}$$

$$\bar{M} = \left[\begin{array}{cc} 0 & 0 \\ \hline 0 & 0 \\ -M_{11} & 0 \\ 0 & A_f M_{22} \end{array}\right] = \begin{bmatrix} 0 \\ \bar{M}_2 \end{bmatrix} \tag{34}$$

$$\bar{C} = \left[\begin{array}{c|ccc} 0 & A^o_{42} & A^o_{43} & * \\ 0 & 0 & 0 & I_r \end{array}\right] = \left[\begin{array}{cc} 0 & \bar{T} \end{array}\right] \tag{35}$$

$$\bar{Q} = \left[\begin{array}{c} Q_{11} \\ \hline Q_{12} \\ Q_{13} \\ 0 \end{array}\right] = \left[\begin{array}{c} \bar{Q}_1 \\ \bar{Q}_2 \end{array}\right] \tag{36}$$

where (*) are terms that play no role in the subsequent analysis. Clearly, $\bar{T}\in\mathbb{R}^{\bar{p}\times\bar{p}}$ is invertible since $\left[A^o_{42}\ \ A^o_{43}\right]$ is square and invertible. Define $\bar{M}_o$ to be the bottom q rows of $\bar{M}_2$ and is square and invertible. Thus, it is easy to verify

$$\bar{C}\bar{M} = \left[\begin{array}{cc} -A^o_{43}M_{11} & * \\ 0 & A_f M_{22} \end{array}\right]$$

By construction M_{11},M_{22} are invertible, and from Lemma 4, $rank(A^o_{43}) = rank(\tilde{A}_{32}) = q-r$, hence $\bar{C}\bar{M}$ is full rank. Lemma 2 shows B1 implies that $rank(\tilde{A}_{32}) = q - r$, hence B1 implies that $\bar{C}\bar{M}$ is full rank. It is shown in Lemma 5 in §VI-D that the zeros of $(\bar{A}, \bar{M}, \bar{C})$ are given by the zeros of the original system $(\tilde{A}, \tilde{M}, \tilde{C})$ from (1) - (2) together with $\lambda(A_\Omega)$. Since B2 assumes that $(\tilde{A}, \tilde{M}, \tilde{C})$ has stable zeros, the system $(\bar{A}, \bar{M}, \bar{C})$ has stable zeros. Hence, the system in (29) - (30) meets the necessary and sufficient conditions of the reconstruction method in [14], if and only if B1 and B2 are satisfied.

Since $\bar{y}$ is measurable, the approach from [14] will be used to design the secondary SMO based on (29) - (30) to reconstruct f whilst being robust to ξ. From (33) - (36), $\bar{M}_{22}, \bar{T}$ are both invertible. Therefore, $(\bar{A}, \bar{M}, \bar{C})$ is already in the coordinates where the robustness analysis in [14] is carried out, hence no further coordinate transformations are required.[1] The proposed observer for (29) - (30) in the coordinates of (33) - (36) is

$$\dot{\hat{\bar{x}}}(t) = \bar{A}\hat{\bar{x}}(t) - \bar{G}_l\bar{e}_y(t) + \bar{G}_n\bar{\nu}, \quad \hat{\bar{y}}(t) = \bar{C}\hat{\bar{x}}(t) \tag{37}$$

where $\bar{e}_y := \hat{\bar{y}} - \bar{y}$. Again, $\bar{G}_l, \bar{G}_n \in \mathbb{R}^{\bar{n}\times\bar{p}}$ are observer gains to be designed, where $\bar{G}_n$ has the structure (in the coordinates of (33) - (34) and (35) - (36))

$$\bar{G}_n = \left[\begin{array}{c} -\bar{L}\bar{T}^{-1} \\ \bar{T}^{-1} \end{array}\right] \bar{P}_o^{-1}, \ \bar{L} = \left[\begin{array}{cc} \bar{L}_1 & 0 \end{array}\right] \tag{38}$$

where $\bar{P}_o \in \mathbb{R}^{\bar{p}\times\bar{p}}$ is s.p.d., $\bar{L} \in \mathbb{R}^{(\bar{n}-\bar{p})\times\bar{p}}$, $\bar{L}_1 \in \mathbb{R}^{(\bar{n}-\bar{p})\times(\bar{p}-q)}$ and $\bar{\nu}$ is a discontinuous term defined by

$$\bar{\nu} = -\bar{\rho}\frac{\bar{e}_y}{\|\bar{e}_y\|} \quad \text{where} \quad \bar{\rho} \in \mathbb{R}_+ \tag{39}$$

For an appropriate choice of $\bar{G}_l$ and a large enough choice of $\bar{\rho}$, it can be shown that an ideal sliding motion takes place on $\bar{\mathcal{S}} = \left\{\bar{e} : \bar{C}\bar{e} = 0\right\}$ in finite time where $\bar{e} := \hat{\bar{x}} - \bar{x}$. A detailed discussion on the design aspects is given in §III-B.

Define a signal $\hat{f}(t) := \bar{W}\bar{T}^{-1}\bar{P}_o^{-1}\bar{\nu}_{eq}$ where $\bar{W} := \left[\begin{array}{cc} \bar{W}_1 & \bar{M}_o^{-1} \end{array}\right]$ with $\bar{W}_1 \in \mathbb{R}^{q\times(\bar{p}-q)}$ and $\bar{\nu}_{eq}$ is the term required to maintain the sliding motion. The term $\bar{\nu}_{eq}$ can be calculated online in the same way that ν_{eq} in (24) is computed. When a sliding mode motion has taken place on $\bar{\mathcal{S}}$, it can be shown that $\hat{f}(t) = f(t) + G(s)\xi(t)$ where

$$G(s){:=}\bar{W}\bar{A}_3\left(sI - (\bar{A}_1{+}\bar{L}\bar{A}_3)\right)^{-1}(\bar{Q}_1{+}\bar{L}\bar{Q}_2){+}\bar{W}\bar{Q}_2 \tag{40}$$

Therefore, it is clear that $\hat{f}$ will capture f as well as a dynamic function of ξ. If there is no uncertainty then $\bar{Q}_1 = 0, \bar{Q}_2 = 0$ and so $G(s) = 0$ and perfect reconstruction of f by $\hat{f}$ is obtained. ■

B. Design of observers

In this paper, the observers will be designed using LMIs.

For the design of the primary observer (15) - (16), G_l is calculated such that the following inequality is satisfied

$$PA_o{+}A_o^TP < 0 \ \text{ where } P{=}\begin{bmatrix} P_1 & P_1L \\ L^TP_1 & T^TP_oT{+}L^TP_1L \end{bmatrix} \tag{41}$$

where $P_1 \in \mathbb{R}^{(n-p)\times(n-p)}$ is s.p.d. so that a stable sliding motion can take place on $\mathcal{S}$. Then the matrices L and P_o can be calculated from P, and G_n can be calculated from (17).

Key observation: Notice that $G(s)$ from (40) is unaffected by the elements of L because of the structures of the partitions in (33) - (36). *This means $G(s)$ is unaffected by the design parameters of the primary observer and thus can be designed using any method as long as P and G_l satisfy the inequality in (41).* ♯

In this paper, the primary observer will be designed using the method in [13]. Define the following decision variable

$$P_{lmi} = \left[\begin{array}{cc} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{array}\right] > 0$$

where $P_{11} \in \mathbb{R}^{(n-p)\times(n-p)}, P_{22} \in \mathbb{R}^{p\times p}$ are s.p.d. and P_{lmi} has the same structure as P in (41). Define another symmetric decision variable $X \in \mathbb{R}^{n\times n}$. The algorithm in [13] can be summarized as: Minimize $trace(X)$ subject to the following inequalities

$$\left[\begin{array}{cc} P_{lmi}A + A^TP_{lmi} - C^TV_2^{-1}C & P_{lmi} \\ P_{lmi} & -V_1^{-1} \end{array}\right] < 0 \tag{42}$$

$$\left[\begin{array}{cc} -P_{lmi} & I_n \\ I_n & -X \end{array}\right] < 0 \tag{43}$$

where $V_1 \in \mathbb{R}^{n\times n}, V_2 \in \mathbb{R}^{p\times p}$ are s.p.d. weighting matrices to be chosen by the designer to tune the observer gains. The LMI Toolbox will return values for the decision variables P_{lmi} and X, and the following parameters can be calculated

$$G_l{=}P_{lmi}^{-1}C^TV_2^{-1}, \ L{=}P_{11}^{-1}P_{12}, \ P_o{=}T^T(P_{22}{-}P_{12}^TL)T \tag{44}$$

and G_n can be calculated as in (17). The choice of G_l in (44) together with (42) ensures (41) is satisfied.

Remark: The observer (15) - (17) is slightly different compared to existing work as L in [3][14] is forced to have a special structure. The observer defined in (15) - (16) treats all the unknown signals col(ξ, f) as an 'unmatched' disturbance, because in general M and Q are not matched to G_n, i.e. $rank\left[\begin{array}{ccc} G_n & M & Q \end{array}\right] > rank(G_n)$. As L is unconstrained, the observer (15) - (17) can be considered to be a modified Utkin observer [16] with the additional term G_le_y. ♯

[1] However, there is a slight difference in that $\bar{T}$ is invertible but not necessarily orthogonal as in [14]. This is of no major consequence as will be shown in the analysis which follows.

The observer in (37) will be designed to satisfy

$$\bar{P}\bar{A}_o+\bar{A}_o^T\bar{P} < 0 \text{ where } \bar{P}=\begin{bmatrix}\bar{P}_1 & \bar{P}_1\bar{L}\\ \bar{L}^T\bar{P}_1 & \bar{T}^T\bar{P}_o\bar{T}+\bar{L}^T\bar{P}_1\bar{L}\end{bmatrix} \quad (45)$$

where $\bar{A}_o = \bar{A}-\bar{G}_l\bar{C}$ and $\bar{L}$ is given in (38). In particular, the design algorithm in [14] will be used, where the objective is to minimize the $\mathcal{L}_2$ gain of $G(s)$. *The design of the secondary observer is crucial to the quality of the reconstruction.* Again an LMI method will be used.

Define the s.p.d. variables $\bar{P}_{lmi}=\begin{bmatrix}\bar{P}_{11} & \bar{P}_{12}\\ \bar{P}_{12}^T & \bar{P}_{22}\end{bmatrix}, \bar{P}_{12}=\begin{bmatrix}\bar{P}_{121} & 0\end{bmatrix}$ where $\bar{P}_{11} \in \mathbb{R}^{(\bar{n}-\bar{p})\times(\bar{n}-\bar{p})}, \bar{P}_{22} \in \mathbb{R}^{\bar{p}\times\bar{p}}, \bar{P}_{121} \in \mathbb{R}^{(\bar{n}-\bar{p})\times(\bar{p}-q)}$. Also, define other decision variables $\bar{\gamma} \in \mathbb{R}$ and $\bar{W}_1 \in \mathbb{R}^{q\times(\bar{p}-q)}$. Notice that the structure of $\bar{P}_{12}$ causes $\bar{P}_{lmi}$ to have the same structure as $\bar{P}$ in (45).

The design in [14] can be summarized as follows: Minimize $\bar{\gamma}$ subject to the following inequalities

$$\begin{bmatrix}\bar{P}_{11}\bar{A}_1+\bar{A}_1^T\bar{P}_{11}+\bar{P}_{12}\bar{A}_3+\bar{A}_3^T\bar{P}_{12}^T & * & *\\ -(\bar{P}_{11}\bar{Q}_1+\bar{P}_{12}\bar{Q}_2)^T & -\bar{\gamma}I_h & *\\ -\bar{W}\bar{A}_3 & \bar{W}\bar{Q}_2 & -\bar{\gamma}I_q\end{bmatrix} < 0 \quad (46)$$

$$\begin{bmatrix}\bar{P}_{lmi}\bar{A}+\bar{A}^T\bar{P}_{lmi}-\bar{\gamma}_o\bar{C}^T(\bar{D}_d\bar{D}_d^T)^{-1}\bar{C} & * & *\\ -\bar{B}_d^T\bar{P} & -\bar{\gamma}_oI_{\bar{p}+h} & *\\ \bar{E} & \bar{H} & -\bar{\gamma}_oI_q\end{bmatrix} < 0 \quad (47)$$

where (*) are terms that make (46) - (47) symmetric. The fixed matrices are $\bar{B}_d := \begin{bmatrix} 0 & \bar{Q}\end{bmatrix}, \bar{D}_d := \begin{bmatrix}\bar{D}_1 & 0\end{bmatrix}$ whilst $\bar{H} = \begin{bmatrix} 0 & \bar{W}\bar{Q}_2\end{bmatrix}$ where $\bar{D}_1 \in \mathbb{R}^{\bar{p}\times\bar{p}}$ and $\bar{\gamma}_o$ are user-specified parameters to tune $\bar{G}_l, \bar{G}_n$. After the LMI solver returns the values of $\bar{W}, \bar{P}$, $\bar{G}_l$ can be calculated as $\bar{G}_l = \bar{\gamma}_o^{-1}\bar{P}_{lmi}^{-1}\bar{C}^T(\bar{D}_d\bar{D}_d^T)^{-1}$ and with $\bar{G}_n$ as in (38). This algorithm ensures inequality (45) is satisfied and the $\mathcal{L}_2$ gain from ξ to $\hat{f}$ is bounded by $\bar{\gamma}$. By a proper choice of a large enough $\bar{\rho}$, an ideal sliding motion in the secondary observer takes place on $\bar{\mathcal{S}}$ [14]. The secondary observer now treats f as the matched fault (since its distribution matrix is 'matched' to $\bar{G}_n$, i.e. $rank\begin{bmatrix}\bar{G}_n & \bar{M}\end{bmatrix}=rank(\bar{G}_n)$) and ξ as the unmatched disturbance.

Remark: The matrix H in (49), associated with condition B1, is formed from Markov parameters and is system realization independent. Intuitively it is related to the system $(\tilde{A}, \tilde{M}, \tilde{C})$ having relative degree two since for example if

$$\tilde{A} = \begin{bmatrix} 0 & 1\\ 0 & 0\end{bmatrix}, \ \tilde{M} = \begin{bmatrix} 0\\ 1\end{bmatrix}, \ \tilde{C} = \begin{bmatrix} 1 & 0\end{bmatrix}$$

then B1 is satisfied although $\tilde{C}\tilde{M} = 0$. ♯

IV. An Example

A 7th order model of an aircraft [8] will be used to verify the method proposed in this paper. In the notation of (1) - (2), the matrices that describe the system are as follows

$$\tilde{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0\\ 0 & -0.2 & 1.5 & 0 & 0 & -0.7 & 0\\ 0 & -1 & -2.1 & 0 & 0 & 0 & 0\\ 0 & 0.2 & -5.2 & -1 & 0 & 0.3 & -1.1\\ 0 & 0.5 & 0 & 0 & -4 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & -20 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & -25\end{bmatrix}$$

where the states are the bank angle, yaw rate, roll rate, sideslip angle, washed-out filter state, rudder deflection, aileron deflection, and the inputs are the rudder command and the aileron command. Assume that the bank angle, yaw rate and role rate are measurable, and that the first actuator is faulty. Therefore the matrices $\tilde{C}$ and $\tilde{M}$ are

$$\tilde{C} = \begin{bmatrix} I_3 & 0_{3\times4}\end{bmatrix}, \tilde{M} = \begin{bmatrix} 0_{1\times5} & 20 & 0\end{bmatrix}^T$$

Suppose that $\tilde{A}$ is imprecisely known and that there exists parametric uncertainty. The state equation will then become

$$\dot{\tilde{x}} = (\tilde{A} + \triangle\tilde{A})x + \tilde{B}u + \tilde{M}f \quad (48)$$

where $\triangle\tilde{A}$ is the discrepancy between $\tilde{A}$ and its actual value. For simplicity let $u \equiv 0$. Notice that the first, fifth, sixth and seventh rows of $\tilde{A}$ do not contain any uncertainty due to the nature of the state equations. Hence, any parametric uncertainty will appear in the second, third and fourth rows of $\tilde{A}$. Let the actual value of the system matrix be

$$\tilde{A} + \triangle\tilde{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0\\ 0 & -0.2 & 1.7 & 0 & 0 & -0.7 & 0\\ 0 & -1 & -2.2 & 0 & 0 & 0 & 0\\ 0 & 0.2 & 0 & -1 & 0 & 0.4 & -1.2\\ 0 & 0.5 & 0 & 0 & -4 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & -20 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & -25\end{bmatrix}$$

then (48) can be placed in the same framework as (1) - (2) using

$$\triangle\tilde{A}\tilde{x}=\tilde{Q}\tilde{\xi}=\underbrace{\begin{bmatrix} 0_{1\times3}\\ I_3\\ 0_{3\times3}\end{bmatrix}}_{\tilde{Q}}\underbrace{\begin{bmatrix} 0 & 0 & 0 & 0.1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & -0.1\\ 0 & 0 & 0 & -0.1 & 0 & 0 & 0\end{bmatrix}\tilde{x}}_{\tilde{\xi}}$$

where $\tilde{\xi}$ is generated by $\tilde{x}$, which is in turn generated by f. Assuming f is bounded, then $\tilde{x}$ and $\tilde{\xi}$ will also be bounded since $\tilde{A} + \triangle\tilde{A}$ is stable. If f and $\tilde{\xi}$ are augmented to form a new 'fault' vector [11], this would result in the new 'fault' signal having 4 components. The number of outputs in this system is only 3, resulting in a 'more faults than outputs' scenario, and hence the method in [5] is not applicable. The FDI literature based on unknown input observers (UIOs) is also not applicable (because A1 and A2 are typically required [11]). Notice that all faults and disturbances appear in states 2, 3, 4 and 6. For the method in [2] to be applicable, the integral of the states 2, 3, 4 and 6 would need to be measurable. However, in this system, of the four states, only the sideslip angle is measured. The remainder are not measured and hence the method in [2] is not applicable. Also notice that $\tilde{C}\tilde{M} = 0 \Rightarrow r = 0 < q$, and hence the existing sliding mode methods [4][14] cannot be used to reconstruct the fault.

A. Observer design

It can be easily verified that B1 and B2 are satisfied. Hence, the method proposed in this paper can be used.

The disturbance $\tilde{\xi}$ is assumed to have a frequency content $\omega < 10\ rad/s$, resulting in $A_\Omega = -10I_3, B_\Omega = 10I_3$.

From the coordinate transformation in Lemma 1,

$$\tilde{A} = \left[\begin{array}{cccc|ccc} -25 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1.1 & -1 & 0 & 0.3 & 5.2 & -0.2 & 0 \\ 0 & 0 & -4 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & -20 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2.1 & -1 & 0 \\ \hline 0 & 0 & 0 & 0.7 & 1.5 & -0.2 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

$$\tilde{Q} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T$$

which shows that $\tilde{A}_{32}$ is full rank, and C1 is fulfilled. Also, $rank \left[\begin{array}{ccc} \tilde{Q}_{21} & \tilde{A}_{31} & \tilde{A}_{32} \end{array} \right] = 3$, which means that $\bar{p} = 3$.

To design the primary observer, $V_1 = 100I_{10}, V_2 = I_3$ were chosen. The following gain matrices were obtained

$$G_l{=}G_n{=}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0.3 \\ 0.3 & 0 & 0.1 \\ 0 & 0 & 0 \\ -4.6 & 0.4 & 3.4 \\ 0 & 0.4 & 0 \\ 0 & 0.2 & 0 \\ -0.5 & 0.2 & 10.9 \\ 0.4 & 13.3 & 0.2 \\ 10.8 & 0.4 & -0.5 \end{bmatrix}, P_o{=}\begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}$$

The secondary observer was designed using $\bar{D}_1 = 10I_3, \bar{\gamma}_o = 100$ and the following gains were obtained

$$\bar{G}_l{=}\bar{G}_n{=}\begin{bmatrix} 0.2 & 0 & 0.1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ -0.1 & 0 & -0.1 \\ 0 & -0.1 & 0 \\ -0.1 & 0 & -2.9 \end{bmatrix}, \bar{P}_o{=}\begin{bmatrix} 8.4 & -0.1 & -0.4 \\ -0.1 & 13.4 & 0 \\ -0.4 & 0 & 0.7 \end{bmatrix}$$

$$\bar{W}\bar{T}^{-1}\bar{P}_o^{-1} = \left[\begin{array}{ccc} 0.0518 & 0.0005 & -0.0777 \end{array} \right]$$

The gains above provide an $\mathcal{L}_2$ bound of $\bar{\gamma} = 1.3131$.

B. Simulation results

In the following simulations, the parameters associated with ν for the primary observer were chosen as $\rho = 100, \delta = 10^{-5}$ while for the secondary observer $\bar{\rho} = 100, \bar{\delta} = 10^{-5}$. A fault was induced in the first actuator. Figure 1 shows the fault and its reconstruction, where the left subfigure is the fault and the right subfigure is the reconstruction. It can be clearly seen that $\hat{f}$ provides a good estimate of f, despite the fact that $\triangle\tilde{A}$ causes a disturbance that could corrupt the reconstruction. The observer gains are calculated such that the reconstruction is least affected by the disturbances in an $\mathcal{L}_2$ sense.

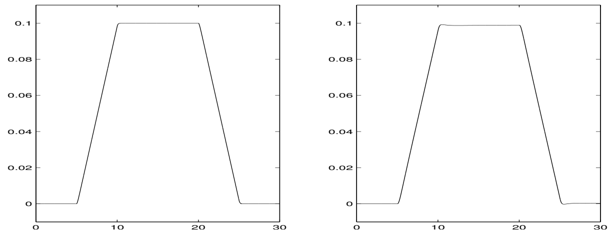

Fig. 1. The left subfigure is the fault, the right subfigure is its reconstruction.

V. Conclusion

This paper has proposed a new scheme for robust fault reconstruction in uncertain systems which is applicable to a wider class of systems compared to existing work: Specifically the approach is applicable to systems with relative degree greater than one. The application/practical benefit of the proposed method is that less sensors are required for FDI. The method proposed in this paper uses two SMOs in cascade; the equivalent output error injection term from the first observer is processed to form the measurable output of a fictitious system. Then a secondary observer is implemented for the fictitious system such that the fault can be reconstructed using existing methods. An aircraft model has shown the validity of the proposed scheme.

References

[1] S.P. Boyd, L. El-Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in Systems and Control Theory*. SIAM: Philadelphia, 1994.

[2] J Davila, L. Fridman, and A. Levant. Second-order sliding mode observer for mechanical systems. *IEEE Trans. Automatic Control*, 50:1785–1789, 2005.

[3] C. Edwards and S.K. Spurgeon. On the development of discontinuous observers. *Int. Journal of Control*, 59:1211–1229, 1994.

[4] C. Edwards, S.K. Spurgeon, and R.J. Patton. Sliding mode observers for fault detection and isolation. *Automatica*, 36:541–553, 2000.

[5] T. Floquet and J.P. Barbot. An observability form for linear systems with unknown inputs. *Int. J. Control*, 79:132–139, 2006.

[6] P.M. Frank. Fault diagnosis in dynamic systems using analytical and knowledge based redundancy - a survey and some new results. *Automatica*, 26:459–474, 1990.

[7] P. Gahinet, A. Nemirovski, A.J. Laub, and M. Chilali. *LMI Control Toolbox, Users Guide*. The MathWorks, Inc., 1995.

[8] B.S. Heck, S.V. Yallapragada, and M.K.H. Fan. Numerical methods to design the reaching phase of output feedback variable structure control. *Automatica*, 31:275–279, 1995.

[9] R.J. Patton and J. Chen. Optimal unknown input distribution matrix selection in robust fault diagnosis. *Automatica*, 29:837–841, 1993.

[10] H.H. Rosenbrock. *State space and multivariable theory*. John-Wiley, New York, 1970.

[11] M. Saif and Y. Guan. A new approach to robust fault detection and identification. *IEEE Trans. Aerospace and Electronic Systems*, 29:685–695, 1993.

[12] L.C. Shen and P.L. Hsu. Robust design of fault isolation observers. *Automatica*, 34:1421–1429, 1998.

[13] C.P. Tan and C. Edwards. An LMI approach for designing sliding mode observers. *Int. Journal of Control*, 74:1559–1568, 2001.

[14] C.P. Tan and C. Edwards. Sliding mode observers for robust detection and reconstruction of actuator and sensor faults. *Int. Journal of Robust and Nonlinear Control*, 13:443–463, 2003.

[15] C.P. Tan and C. Edwards and Y.C. Kuang. Robust sensor fault reconstruction using right eigenstructure assignment. *Proc. of the 3rd IEEE Int. Workshop on Electronic Design, Test and Applications*, 5 pages, 2006.

[16] V.I. Utkin. *Sliding Modes in Control Optimization*. Springer-Verlag, Berlin, 1992.

VI. Appendix: Proofs

A. *Proof of Lemma 2*

Let (A,C) be partitioned as in (9) - (10). By performing the Popov-Hautus-Rosenbrock (PHR) rank test [10] on (A,C) and from the fact that $(\tilde{A},\tilde{C})$ is observable if and only if $rank\begin{bmatrix} sI-\tilde{A}_1 \\ \tilde{A}_3 \end{bmatrix} = \tilde{n}-p$ for all $s \in \mathbb{C}$, it is clear that the unobservable modes of (A,C) are given by $\lambda(A_\Omega)$. By assumption A_Ω is a stable matrix and thus, (A,C) is detectable. ■

B. *Proof of Lemma 3*

Define
$$H := \begin{bmatrix} \tilde{C}\tilde{A}\tilde{M} & \tilde{C}\tilde{M} \\ \tilde{C}\tilde{M} & 0 \end{bmatrix} \tag{49}$$

Therefore from (4) - (5),

$$\begin{aligned} H &= \begin{bmatrix} \tilde{T} & 0 \\ 0 & \tilde{T} \end{bmatrix} \begin{bmatrix} \tilde{A}_3\tilde{M}_1 + \tilde{A}_4\tilde{M}_2 & \tilde{M}_2 \\ \tilde{M}_2 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \tilde{T} & 0 \\ 0 & \tilde{T} \end{bmatrix} \left[\begin{array}{cc|cc} \tilde{A}_{32}M_{11} & \tilde{A}_{34}M_{22} & 0 & 0 \\ \tilde{A}_{42}M_{11} & \tilde{A}_{44}M_{22} & 0 & M_{22} \\ \hline 0 & 0 & 0 & 0 \\ 0 & M_{22} & 0 & 0 \end{array}\right] \end{aligned}$$

It is clear that $rank(H) = rank(M_{22}) + rank(M_{22}) + rank(\tilde{A}_{32}M_{11})$. Then it follows that $rank(H) = r + r + rank(\tilde{A}_{32}) = rank(\tilde{C}\tilde{M}) + r + rank(\tilde{A}_{32})$ since M_{11}, M_{22} are square and invertible. Therefore, as $rank(\tilde{M}) = q$, B1 holds if and only if $rank(\tilde{A}_{32}) = q - r$. ■

C. *Proof of Lemma 4*

In the coordinates of (9) - (10), define

$$R_1 = \begin{bmatrix} \tilde{Q}_{21} & \tilde{A}_{31} & \tilde{A}_{32} \end{bmatrix},\ R_2 = \begin{bmatrix} 0 & 0 & M_{11}^T \end{bmatrix}^T \tag{50}$$

Therefore M in (10) is
$$\begin{bmatrix} R_2^T & 0 & 0 \\ 0 & 0 & M_{22}^T \end{bmatrix}^T$$

Recall that $rank(R_1) = \bar{p} - r$ and $\tilde{A}_{32}$ has rank $q - r$. Let $X_3 \in \mathbb{R}^{(n-p)\times(n-p)}$ and $X_4 \in \mathbb{R}^{(p-r)\times(p-r)}$ be orthogonal matrices such that

$$X_4 R_1 X_3^T = \left[\begin{array}{c|cc} 0 & A_{a,42} & A_{a,43} \end{array}\right] = \left[\begin{array}{c|c} 0 & 0 \\ 0 & U \end{array}\right] \tag{51}$$

where $U \in \mathbb{R}^{(\bar{p}-r)\times(\bar{p}-r)}$ is invertible. Then define a nonsingular change of coordinates $T_3 \in \mathbb{R}^{n\times n}$ where

$$T_3 = \left[\begin{array}{c|cc} X_3 & 0 & 0 \\ \hline 0 & X_4 & 0 \\ 0 & 0 & I_r \end{array}\right]$$

and apply it to A, M, Q, C in (9) - (10) to obtain

$$A_a = \begin{bmatrix} A_{a,1} & A_{a,2} \\ A_{a,3} & A_{a,4} \end{bmatrix}, M_a = \begin{bmatrix} M_{a,1} \\ M_{a,2} \end{bmatrix}, Q_a = \begin{bmatrix} Q_{a,1} \\ 0 \end{bmatrix} \tag{52}$$

$$C_a = \begin{bmatrix} 0 & T_a \end{bmatrix} \tag{53}$$

Further partition

$$A_{a,1} = \begin{bmatrix} A_{a,11} & A_{a,12} & A_{a,13} \\ A_{a,21} & A_{a,22} & A_{a,23} \\ A_{a,31} & A_{a,32} & A_{a,33} \end{bmatrix}, A_{a,3} = \begin{bmatrix} 0 & A_{a,42} & A_{a,43} \\ A_{a,51} & A_{a,52} & A_{a,53} \end{bmatrix} \tag{54}$$

$$Q_{a,1} = \begin{bmatrix} Q_{a,11} \\ Q_{a,12} \\ Q_{a,13} \end{bmatrix}, M_{a,1} = \begin{bmatrix} M_{a,11} & 0 \\ M_{a,12} & 0 \\ M_{a,13} & 0 \end{bmatrix}, M_{a,2} = \begin{bmatrix} 0 & 0 \\ 0 & M_{22} \end{bmatrix} \tag{55}$$

where T_a is still orthogonal. It is easy to show that $R_1R_2 = \tilde{A}_{32}M_{11}$. Since the matrix $\tilde{A}_{32}$ has full column rank $q-r$ and $det(M_{11}) \neq 0$, then $rank(R_1R_2) = q - r$. Clearly $R_1R_2 = R_1X_3^{-1}X_3R_2$ from (51) can be expanded to be

$$\begin{aligned} \underbrace{X_4^{-1}\begin{bmatrix} 0 & A_{a,42} & A_{a,43} \end{bmatrix}}_{R_1X_3^{-1}} \underbrace{\begin{bmatrix} M_{a,11} \\ M_{a,12} \\ M_{a,13} \end{bmatrix}}_{X_3R_2} &= X_4^{-1}\begin{bmatrix} A_{a,42} & A_{a,43} \end{bmatrix}\begin{bmatrix} M_{a,12} \\ M_{a,13} \end{bmatrix} \\ &= X_4^{-1}\begin{bmatrix} 0 \\ U \end{bmatrix}\begin{bmatrix} M_{a,12} \\ M_{a,13} \end{bmatrix} \end{aligned}$$

Since X_4 is orthogonal and $\bar{p} > q$, it follows that

$$rank(\tilde{A}_{32}) = q-r \Rightarrow rank(R_1R_2) = q-r \Rightarrow rank\begin{bmatrix} M_{a,12} \\ M_{a,13} \end{bmatrix} = q-r$$

Define two nonsingular matrices $X_5 \in \mathbb{R}^{(q-r)\times(\bar{p}-r)}$ and $X_6 \in \mathbb{R}^{(\bar{p}-r)\times(\bar{p}-r)}$ so that

$$X_5\begin{bmatrix} M_{a,12} \\ M_{a,13} \end{bmatrix} = I_{q-r},\ X_6\begin{bmatrix} M_{a,12} \\ M_{a,13} \end{bmatrix} = \begin{bmatrix} 0 \\ M_{11} \end{bmatrix}$$

Then introduce the final change of coordinates

$$T_4 = \left[\begin{array}{cc|c} I_{n-\bar{p}+r-p} & -M_{a,11}X_5 & 0 \\ 0 & X_6 & 0 \\ \hline 0 & 0 & I_p \end{array}\right]$$

so that A_a, M_a, Q_a, C_a are transformed to be

$$A_b = \begin{bmatrix} A_{b,1} & A_{b,2} \\ A_{b,3} & A_{b,4} \end{bmatrix}, M_b = \begin{bmatrix} M_{b,1} \\ M_{b,2} \end{bmatrix}, C_b = \begin{bmatrix} 0 & T_a \end{bmatrix}, Q_b = \begin{bmatrix} Q_{b,1} \\ 0 \end{bmatrix} \tag{56}$$

where

$$A_{b,3} = \begin{bmatrix} 0 & A_{b,42} & A_{b,43} \\ A_{b,51} & A_{b,52} & A_{b,53} \end{bmatrix}, M_{b,1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ M_{11} & 0 \end{bmatrix}, M_{b,2} = \begin{bmatrix} 0 & 0 \\ 0 & M_{22} \end{bmatrix}$$

and from (51)

$$\begin{bmatrix} A_{b,42} & A_{b,43} \end{bmatrix} = \begin{bmatrix} 0 \\ UX_6^{-1} \end{bmatrix}$$

By defining the nonsingular transformation matrix $T_5 := T_4T_3$ and partitioning

$$UX_6^{-1} = \begin{bmatrix} A_{42}^o & A_{43}^o \end{bmatrix} \tag{57}$$

where $A_{43}^o \in \mathbb{R}^{(\bar{p}-r)\times(q-r)}$, A_b, M_b, Q_b, C_b and their partitions are in the same form as A, M, Q, C in (11) - (12) in Lemma 4. To prove that $rank(A_{43}^o) = rank(\tilde{A}_{32})$, define

$$X_7 = \begin{bmatrix} I_{n-\bar{p}-p+r} & -M_{a,11}X_5 \\ 0 & X_6 \end{bmatrix}$$

From the coordinate transformations T_3, T_4 and by observing the structures of A_b and M_b, from (51) and (57),

$$X_4 R_1 X_3^{-1} X_7^{-1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & A_{42}^o & A_{43}^o \end{bmatrix}$$

$$X_7 X_3 R_2 = \begin{bmatrix} 0 \\ 0 \\ M_{11} \end{bmatrix} \Rightarrow X_4 R_1 R_2 = \begin{bmatrix} 0 \\ A_{43}^o M_{11} \end{bmatrix}$$

Recalling that $rank(R_1 R_2) = rank(\tilde{A}_{32}) = q - r$, and since X_4 and M_{11} are invertible, $rank(A_{43}^o) = rank(\tilde{A}_{32})$. ■

D. Lemma 5 and its proof

Lemma 5: The zeros of $(\bar{A}, \bar{M}, \bar{C})$ are identical to the zeros of $(\tilde{A}, \tilde{M}, \tilde{C})$ together with the eigenvalues of A_Ω.

Proof: The Rosenbrock system matrix [10] of $(\bar{A}, \bar{M}, \bar{C})$ is given by

$$E_{a,1}(s) = \begin{bmatrix} sI - \bar{A} & \bar{M} \\ \bar{C} & 0 \end{bmatrix}$$

and the zeros of a system are the values of s that cause its Rosenbrock matrix to lose normal rank. From (33) - (36), $E_{a,1}(s)$ can be expanded to be

$$E_{a,1}(s) = \begin{bmatrix} sI - \bar{A}_1 & -\bar{A}_2 & 0 \\ -\bar{A}_3 & sI - \bar{A}_4 & \bar{M}_2 \\ 0 & \bar{T} & 0 \end{bmatrix}$$

Since $\bar{T}$ has full rank, it is clear that $E_{a,1}(s)$ loses rank if and only if the following matrix loses rank

$$E_{a,2}(s) := \begin{bmatrix} sI - \bar{A}_1 & 0 \\ -\bar{A}_3 & \bar{M}_2 \end{bmatrix}$$

Substituting for $\bar{A}_1, \bar{A}_3, \bar{M}_2$ from (33) - (36), $E_{a,2}(s)$ can be expanded to be

$$E_{a,2}(s) = \left[\begin{array}{c|cc} sI - A_{11} & 0 & 0 \\ \hline -A_{21} & 0 & 0 \\ -A_{31} & -M_{11} & 0 \\ A_f A_{51} & 0 & A_f M_{22} \end{array}\right]$$

It is then obvious to see that $E_{a,2}(s)$ loses rank if and only if $E_{a,3}(s)$ loses rank where

$$E_{a,3}(s) := \begin{bmatrix} sI - A_{11} \\ -A_{21} \end{bmatrix}$$

From the PHR rank test [10], the values of s that make $E_{a,3}(s)$ lose rank are the unobservable modes of (A_{11}, A_{21}).

The zeros of (A, M, C) are given by the values of s that cause the following matrix to lose rank

$$E_{b,1}(s) = \begin{bmatrix} sI - A & M \\ C & 0 \end{bmatrix}$$

From (11) - (12), $E_{b,1}(s)$ is

$$E_{b,1}(s) = \left[\begin{array}{cc|c} sI - A_1 & -A_2 & M_1 \\ -A_3 & sI - A_4 & M_2 \\ \hline 0 & T & 0 \end{array}\right]$$

Since T is orthogonal, then $E_{b,1}(s)$ loses rank if and only if $E_{b,2}(s)$ loses rank where

$$E_{b,2}(s) = \begin{bmatrix} sI - A_1 & M_1 \\ -A_3 & M_2 \end{bmatrix}$$

Substituting for A_1, A_3, M_1, M_2 from (11) - (12),

$$E_{b,2}(s) = \left[\begin{array}{ccc|cc} sI - A_{11} & -A_{12} & -A_{13} & 0 & 0 \\ -A_{21} & sI - A_{22} & -A_{23} & 0 & 0 \\ -A_{31} & -A_{32} & sI - A_{33} & M_{11} & 0 \\ \hline 0 & -A_{42} & -A_{43} & 0 & 0 \\ -A_{51} & -A_{52} & -A_{53} & 0 & M_{22} \end{array}\right]$$

From Lemma 4, $\begin{bmatrix} A_{42} & A_{43} \end{bmatrix}$ has a special structure and since M_{11}, M_{22} are square and invertible, $E_{b,2}(s)$ loses rank if and only if $E_{b,3}(s)$ loses rank where

$$E_{b,3}(s) = \begin{bmatrix} sI - A_{11} \\ -A_{21} \end{bmatrix}$$

which loses rank if and only if s is an unobservable mode of (A_{11}, A_{21}). This shows that $(\bar{A}, \bar{M}, \bar{C})$ and (A, M, C) have the same zeros.

By using the partitions of (A, M, C) in (9) - (10), it can be easily shown the Rosenbrock matrix of (A, M, C) loses rank if and only if the following matrix $E_{c,1}(s)$ loses rank

$$E_{c,1}(s) = \begin{bmatrix} sI - A_\Omega & 0 & 0 \\ -\tilde{Q}_{11} & sI - \tilde{A}_{11} & -\tilde{A}_{12} \\ -\tilde{Q}_{21} & -\tilde{A}_{31} & -\tilde{A}_{32} \end{bmatrix}$$

It is clear that $E_{c,1}(s)$ loses rank when $s = \lambda(A_\Omega)$ or when s is a zero of $(\tilde{A}_{11}, \tilde{A}_{12}, \tilde{A}_{31}, \tilde{A}_{32})$. Then, by finding the Rosenbrock matrix of $(\tilde{A}, \tilde{M}, \tilde{C})$ using the partitions in (4) - (6), it can be proven that the zeros of $(\tilde{A}, \tilde{M}, \tilde{C})$ are the zeros of $(\tilde{A}_{11}, \tilde{A}_{12}, \tilde{A}_{31}, \tilde{A}_{32})$.

Hence, it is proven that the zeros of $(\bar{A}, \bar{M}, \bar{C})$ are the zeros of $(\tilde{A}, \tilde{M}, \tilde{C})$ and $\lambda(A_\Omega)$. ■