Data-Driven Robust Receding Horizon Fault Estimation *

Yiming Wan [a], Tamas Keviczky [a], Michel Verhaegen [a], Fredrik Gustafsson [b]

[a]*Delft Center for Systems and Control, Delft University of Technology, Delft, 2628 CD, The Netherlands*

[b]*Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden*

Abstract

This paper presents a data-driven receding horizon fault estimation method for additive actuator and sensor faults in unknown linear time-invariant systems, with enhanced robustness to stochastic identification errors. State-of-the-art methods construct fault estimators with identified state-space models or Markov parameters, but they do not compensate for identification errors. Motivated by this limitation, we first propose a receding horizon fault estimator parameterized by predictor Markov parameters. This estimator provides (asymptotically) unbiased fault estimates as long as the subsystem from faults to outputs has no unstable transmission zeros. When the identified Markov parameters are used to construct the above fault estimator, zero-mean stochastic identification errors appear as model uncertainty multiplied with unknown fault signals and online system inputs/outputs (I/O). Based on this fault estimation error analysis, we formulate a mixed-norm problem for the offline robust design that regards online I/O data as unknown. An alternative online mixed-norm problem is also proposed that can further reduce estimation errors when the online I/O data have large amplitudes, at the cost of increased computational burden. Based on a geometrical interpretation of the two proposed mixed-norm problems, systematic methods to tune the user-defined parameters therein are given to achieve desired performance trade-offs. Simulation examples illustrate the benefits of our proposed methods compared to recent literature.

Key words: Data-driven methods; fault estimation; receding horizon estimation; parameter uncertainty.

1 Introduction

Model-based fault diagnosis techniques for linear dynamic systems have been well established during the past two decades [2, 5, 7, 16]. Recently, the model-based receding horizon approach has received attention because it provides a flexible framework to enhance robustness of passive fault diagnosis [34, 36] and to enable optimal input design in active fault diagnosis [28, 29, 31]. However, an explicit and accurate system model is often unknown in practice. In such situations, a conventional approach first identifies the system model from system I/O data, and then designs the model-based fault diagnosis system under various performance criteria [22, 25, 32]. Without explicitly identifying a system model, recent research efforts investigate data-driven approaches to construct a fault diagnosis system utilizing the link be-

tween system identification and the model-based fault diagnosis methods [8, 9, 30]. These recent data-driven approaches simplify the design procedure by skipping the realization of an explicit system model, while at the same time allow developing systematic methods to address the same fault diagnosis performance criteria as the existing model-based approaches.

Most recent data-driven fault diagnosis approaches for unknown linear dynamic systems can be classified into two categories. The first category, e.g., [27] and [9, 10], identifies a projection matrix known as parity space/vectors for residual generation, by exploiting the subspace identification method based on principal component analysis (SIM-PCA) [17]. However, as pointed out in [13], a model reduction step is needed to determine the projection matrix, hence leads to the nonlinear dependence of the generated residuals on the identification errors. Therefore it is difficult to guarantee the robustness of such data-driven methods to the identification errors.

The second category of data-driven fault diagnosis methods, e.g., [11], utilizes the Markov parameters (or impulse response parameters) which can be obtained in

* This paper was not presented at any IFAC meeting. Corresponding author Yiming Wan. Tel.: +31152787019; Fax: +31152786679.

Email addresses: y.wan@tudelft.nl (Yiming Wan), t.keviczky@tudelft.nl (Tamas Keviczky), m.verhaegen@tudelft.nl (Michel Verhaegen), fredrik.gustafsson@liu.se (Fredrik Gustafsson).

the first step of the predictor based subspace identification (PBSID) technique [6, 33]. It constructs residual generators parameterized by the predictor Markov parameters. The main advantage of this method is that the residual signal linearly depends on the identification errors of the predictor Markov parameters. Hence a robust scheme has been developed in [13, 14] to cope with stochastic identification errors. This benefit of robustness compared to the SIM-PCA based method in [10] is achieved at the cost of increased computational burden in incorporating past I/O data.

Most of the data-driven fault diagnosis literature mentioned above discuss only fault detection and isolation. It is much more involved to estimate/identify the fault signal in the data-driven setting. The work in [1, 26] proposed to reconstruct faults by minimizing the reconstructed squared prediction error obtained from PCA. However, this approach did not fully investigate the statistical properties of the calculated fault estimates. By investigating the link between system-inversion based fault reconstruction and the predictor Markov parameters, the method in [12] constructed fault estimators parameterized by the predictor Markov parameters. Its fault estimates are asymptotically unbiased as the estimation horizon length tends to infinity, under the condition that the underlying inverted system is stable.

One drawback of the data-driven fault estimator proposed in [12] is that it cannot be directly applied to sensor faults in an unstable open-loop plant because its underlying inverted system is unstable. Another limitation of this method is that it does not compensate for the identification errors. The robustness of fault estimation to the identification errors is critical in two situations: 1) there exist large identification errors due to small number of identification data samples or low signal-to-noise ratio in identification data; 2) multiplication of the erroneous identified matrices with online I/O data of large amplitude cannot be simply ignored.

Motivated by the above two drawbacks of the proposed method in [12], this paper develops data-driven robust fault estimation methods for additive actuator/sensor faults, utilizing the identified Markov parameters. In order to pave the way for data-driven design, we first construct a receding horizon (RH) fault estimator parameterized by the predictor Markov parameters, assuming that the predictor Markov parameters are accurately available. It gives (asymptotically) unbiased fault estimates under the condition that the subsystem from faults to outputs has no unstable transmission zeros. The above condition for unbiasedness generalizes the requirement of stable inversion in [12]. An immediate benefit is that our fault estimator can be applied to sensor faults in unstable open-loop plants as long as the above condition for unbiasedness is satisfied, whereas the proposed method in [12] cannot.

Our data-driven design parameterizes the above RH fault estimator with predictor Markov parameters identified from closed-loop data. The obtained data-driven fault estimation error is linear with regards to the stochastic identification errors of Markov parameters, although the identification errors appear as multiplicative uncertainty that couples with unknown fault signals as well as online I/O data. In order to enhance robustness to stochastic identification errors, we propose two mixed-norm fault estimators. The first one can be designed offline by regarding the online I/O data as unknown. By exploiting online I/O data in its formulated mixed-norm problem, the second robust fault estimator further reduces estimation errors when the online I/O data have large amplitudes, at the cost of increased online computational burden. Based on a geometric interpretation of the formulated mixed-norm problems, a systematic tuning method for the user-defined parameters therein is provided to achieve the desired trade-offs between estimation bias and variance. Our proposed methods can handle sensor and actuator faults either separately or simultaneously. Only the separate scenario is illustrated in detail in this paper. Exact formulas for the simultaneous scenario can be derived in a straightforward manner but are omitted for the sake of brevity.

The rest of this paper starts with the problem formulation and some preliminaries on closed-loop identification of predictor Markov parameters in Section 2. Section 3 constructs the predictor-based RH fault estimator, and analyzes its condition for unbiasedness. A data-driven nominal fault estimator is given in Section 4. Section 5 and 6 propose two mixed-norm fault estimators with enhanced robustness to identification errors. Simulation studies are finally given in Section 7.

2 Preliminaries and problem formulation

2.1 Notations

For a matrix X, its range and null space is denoted by $\mathcal{R}(X)$ and $\mathcal{N}(X)$, respectively. X^{-} represents the left inverse satisfying $X^{-}X = I$, while $X^{(1)}$ represents the generalized inverse satisfying

$$XX^{(1)}X = X. \tag{1}$$

$X^{[i]}$ represents the ith column of X. The trace of X is denoted by $\text{tr}(X)$. Let $\|X\|_F$ represent the Frobenius norm of the matrix X. The minimal eigenvalue of a symmetric matrix X is represented by $\lambda_{\min}(X)$. Let $\text{vec}(X)$ represent the column vector concatenating the columns of a matrix X. The symbol "$\otimes$" stands for Kronecker product. Let $\text{diag}(X_1, X_2, \cdots, X_n)$ denote a block diagonal matrix with $X_1, X_2, \cdots, X_n$ as its diagonal matrices.

2.2 Problem formulation

We consider linear discrete-time systems governed by the following state space model:

$$\xi(k+1) = A\xi(k) + Bu(k) + Ef(k) + Fw(k)$$
$$y(k) = C\xi(k) + Du(k) + Gf(k) + v(k). \qquad (2)$$

Here $\xi(k) \in \mathbb{R}^n$, $y(k) \in \mathbb{R}^{n_y}$, and $u(k) \in \mathbb{R}^{n_u}$ represent the state, the output measurement, and the known control input at time instant k, respectively. The process and measurement noises $w(k) \in \mathbb{R}^{n_w}$ and $v(k) \in \mathbb{R}^{n_v}$ are white zero-mean Gaussian, with covariance matrices $\mathrm{E}\left(w(k)w^{\mathrm{T}}(k)\right) = Q$, $\mathrm{E}\left(v(k)v^{\mathrm{T}}(k)\right) = R$, $\mathrm{E}\left(w(k)v^{\mathrm{T}}(k)\right) = 0$. $f(k) \in \mathbb{R}^{n_f}$ is the unknown fault signal to be estimated. A, B, C, D, E, F, G are constant real matrices, with bounded norms and appropriate dimensions.

The following assumption is standard in Kalman filtering [18] and subspace identification [6,19]:

Assumption 1 *The pair (C, A) is assumed detectable; and there are no uncontrollable modes of $\left(A, FQ^{\frac{1}{2}}\right)$ on the unit circle, where $Q^{\frac{1}{2}} \cdot \left(Q^{\frac{1}{2}}\right)^{\mathrm{T}} = Q$ is the covariance matrix of $w(k)$.*

Based on Assumption 1, the system (2) admits the one-step-ahead predictor form given by [18]

$$x(k+1) = \Phi x(k) + \tilde{B}u(k) + \tilde{E}f(k) + Ky(k)$$
$$y(k) = Cx(k) + Du(k) + Gf(k) + e(k), \qquad (3)$$

where K is the steady-state Kalman gain, $\Phi = A - KC$, $\tilde{B} = B - KD$, and $\tilde{E} = E - KG$, $\{e(k)\}$ is the zero-mean innovation process with the covariance matrix Σ_e.

We consider additive sensor or actuator faults in this paper, i.e.,

- fault of the jth sensor:

$$E = 0_{n_x \times 1}, \ G = I^{[j]}, \ \tilde{E} = -K^{[j]}; \qquad (4)$$

- fault of the lth actuator:

$$E = B^{[l]}, \ G = D^{[l]}, \ \tilde{E} = \tilde{B}^{[l]}; \qquad (5)$$

- simultaneous faults of the jth sensor and l^{th} actuator:

$$E = \left[0_{n_x \times 1} \ B^{[l]} \right], G = \left[I^{[j]} \ D^{[l]} \right], \tilde{E} = \left[-K^{[j]} \ \tilde{B}^{[l]} \right]; \qquad (6)$$

with $X^{[j]}$ representing the jth column of a matrix X.

Denote the predictor Markov parameters by

$$H_i^u = \begin{cases} D & i = 0 \\ C\Phi^{i-1}\tilde{B} & i > 0 \end{cases}, \ H_i^y = \begin{cases} 0 & i = 0 \\ C\Phi^{i-1}K & i > 0 \end{cases},$$
$$H_i^f = \begin{cases} G & i = 0 \\ C\Phi^{i-1}\tilde{E} & i > 0 \end{cases}. \qquad (7)$$

Assumption 2 *The relative degree of the fault subsystem $\left(\Phi, \tilde{E}, C, G\right)$ is τ, i.e., τ is the smallest nonnegative integer i such that $H_0^f = H_1^f = \cdots = H_{i-1}^f = 0$ and $H_i^f \neq 0$ [20]; moreover, $\mathrm{rank}\left(H_\tau^f\right) = n_f$ [12].*

Note that $\tau = 0$ for sensor faults and $\tau > 0$ for actuator faults.

The essential goals of this paper are to design a fault estimator from identification data without knowing the system matrices in (2), and moreover to robustify the fault estimator against identification errors.

Concerning the identification data, it should be noted that in practice data from faulty conditions may be seldomly available, or if recorded then without a reliable fault description [9]. Hence we make the assumption as below:

Assumption 3 *Only I/O data collected from the fault-free condition are used in our data-driven design.*

In contrast to [24] which assumes the fault signals $f(k)$ evolve according to a random walk model, no assumption is made in this paper about how the fault signals $f(k)$ vary with time.

2.3 Closed-loop identification of predictor Markov parameters

Considering Assumption 3, we set $f(k) = 0$ in (2) for the identification data collected from the fault-free condition. Then with $f(k) = 0$, the predictor form (3) over the time window $[t, \cdots, t + N - 1]$ can be written into the following data equation [6,33]:

$$\mathbf{Y}_{\mathrm{id}} = C\Phi^p \mathbf{X}_{\mathrm{id}} + \Xi \mathbf{Z}_{\mathrm{id}} + \mathbf{E}_{\mathrm{id}}, \qquad (8)$$

where

$$\Xi = \left[\begin{array}{cccccc} H_p^u & H_p^y & \cdots & H_1^u & H_1^y & H_0^u \end{array} \right] \qquad (9)$$

denotes the sequence of Markov parameters $\{H_i^u\}$ and $\{H_i^y\}$ (defined in (7)) to be identified. The detailed definitions of the data matrices $\mathbf{X}_{\mathrm{id}}$, $\mathbf{Y}_{\mathrm{id}}$ and $\mathbf{Z}_{\mathrm{id}}$ can be

found in [33], and $\mathbf{E}_{\text{id}}$ is the sequence of the innovation signal in the identification data.

The least-squares (LS) estimate of the Markov parameters Ξ is

$$
\begin{aligned}
\hat{\Xi} &= \arg\min_{\Xi} \|\mathbf{Y}_{\text{id}} - \Xi\mathbf{Z}_{\text{id}}\|_F^2 = \mathbf{Y}_{\text{id}}\mathbf{Z}_{\text{id}}^- \\
&= \Xi + C\Phi^p\mathbf{X}_{\text{id}}\mathbf{Z}_{\text{id}}^- + \mathbf{E}_{\text{id}}\mathbf{Z}_{\text{id}}^-,
\end{aligned}
\tag{10}
$$

with $\mathbf{Z}_{\text{id}}^- = \mathbf{Z}_{\text{id}}^{\mathrm{T}}\left(\mathbf{Z}_{\text{id}}\mathbf{Z}_{\text{id}}^{\mathrm{T}}\right)^{-1}$. As standard assumptions for consistent identification from closed-loop data, we assume that 1) the data matrix $\mathbf{Z}_{\text{id}}$ has full row rank, and 2) either the controller has at least one-step delay or the plant model has no direct feedthrough ($D = 0$) [6, 33].

With sufficiently large p, the estimation bias $C\Phi^p\mathbf{X}_{\text{id}}\mathbf{Z}_{\text{id}}^-$ can be neglected. Then the stochastic identification errors are

$$
\Delta\hat{\Xi} = \hat{\Xi} - \Xi \approx \mathbf{E}_{\text{id}}\mathbf{Z}_{\text{id}}^-.
\tag{11}
$$

Hence according to (11), the identification errors in Markov parameters can also be written as

$$
\begin{aligned}
\Delta H_i^u &= \hat{H}_i^u - H_i^u = \mathbf{E}_{\text{id}}M_i^u, \\
\Delta H_i^y &= \hat{H}_i^y - H_i^y = \mathbf{E}_{\text{id}}M_i^y,
\end{aligned}
\tag{12}
$$

where $\hat{H}_i^u$ and $\hat{H}_i^y$ represent the estimated Markov parameters in $\hat{\Xi}$ given by (10), M_i^u and M_i^y are the corresponding blocks of $\mathbf{Z}_{\text{id}}^-$, i.e.,

$$
\mathbf{Z}_{\text{id}}^- = \left[\begin{array}{cccccc} M_p^u & M_p^y & \cdots & M_1^u & M_1^y & M_0^u \end{array}\right], \quad M_0^y = 0.
\tag{13}
$$

The innovation covariance can be estimated by [16, 19]

$$
\hat{\Sigma}_e = \text{cov}\left(\mathbf{Y}_{\text{id}} - \hat{\Xi}\mathbf{Z}_{\text{id}}\right).
\tag{14}
$$

For the sake of brevity, we shall not distinguish between the estimated innovation covariance $\hat{\Sigma}_e$ and its true value Σ_e in the rest of this paper.

3   Predictor-based receding horizon fault estimation

In this section, we will construct an RH fault estimator based on the predictor form of the system (2). Here we consider the predictor form instead of the original system model (2) in order to pave the way for data-driven design.

Consider a sliding window with a length of L sampling instants. Define stacked data vectors in this window as $\mathbf{u}_{k,L}$, $\mathbf{y}_{k,L}$, $\mathbf{f}_{k,L}$, and $\mathbf{e}_{k,L}$, respectively for the signals u, y, f, and e; e.g.,

$$
\mathbf{u}_{k,L} = \left[\begin{array}{ccc} u^{\mathrm{T}}(k_0) & \cdots & u^{\mathrm{T}}(k) \end{array}\right]^{\mathrm{T}},
\tag{15}
$$

with $k_0 = k - L + 1$. For the predictor form (3), let $\mathcal{O}_L$ denote its extended observability matrix with L block elements, and $\mathbf{T}_L^\star$ be the lower triangular block-Toeplitz matrix with L block columns and rows, with $\star$ representing u, y, or f:

$$
\mathcal{O}_L = \left[\begin{array}{c} C \\ C\Phi \\ \vdots \\ C\Phi^{L-1} \end{array}\right], \quad \mathbf{T}_L^\star = \left[\begin{array}{cccc} H_0^\star & 0 & \ldots & 0 \\ H_1^\star & H_0^\star & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ H_{L-1}^\star & H_{L-2}^\star & \cdots & H_0^\star \end{array}\right].
\tag{16}
$$

Given the I/O data over the sliding window $[k_0, k]$, the stacked residual signal $\mathbf{r}_{k,L}$ in $[k_0, k]$ can be computed by

$$
\mathbf{r}_{k,L} = \mathbf{y}_{k,L} - \mathbf{T}_L^y\mathbf{y}_{k,L} - \mathbf{T}_L^u\mathbf{u}_{k,L},
\tag{17}
$$

according to the predictor form (3). We can further write down the transitions from unknown initial state, faults and noises to the stacked residual signal $\mathbf{r}_{k,L}$ as

$$
\mathbf{r}_{k,L} = \mathcal{O}_L x(k_0) + \mathbf{T}_L^f\mathbf{f}_{k,L} + \mathbf{e}_{k,L}.
\tag{18}
$$

With Assumption 2, (18) can be simplified as

$$
\mathbf{r}_{k,L} = \underbrace{\left[\begin{array}{cc} \mathcal{O}_L & \mathbf{T}_{L,\tau}^f \end{array}\right]}_{\Psi_{L,\tau}}\underbrace{\left[\begin{array}{c} x(k_0) \\ \mathbf{f}_{k-\tau,L-\tau} \end{array}\right]}_{\mathbf{f}_{k-\tau,L-\tau}^x} + \mathbf{e}_{k,L},
\tag{19}
$$

where τ is the relative degree of the fault subsystem (A, E, C, G), $\mathbf{T}_{L,\tau}^f$ represents the first $L - \tau$ block-columns of $\mathbf{T}_L^f$ defined similar to (16), $\mathbf{f}_{k-\tau,L-\tau}$ is defined in the same way as in (15).

With (19), we can formulate the receding horizon fault estimation (RHFE) problem

$$
\min_{\mathbf{f}_{k-\tau,L-\tau}^x} \left\|\mathbf{r}_{k,L} - \Psi_{L,\tau}\mathbf{f}_{k-\tau,L-\tau}^x\right\|_{\Sigma_{e,L}^{-1}}^2
\tag{20}
$$

in the LS sense, with

$$
\Sigma_{e,L} = I_L \otimes \Sigma_e
\tag{21}
$$

denoting the covariance matrix of $\mathbf{e}_{k,L}$. It has non-unique solutions because $\Psi_{L,\tau}$ may not have full column rank. One solution to the problem (20) is

$$
\hat{\mathbf{f}}_{k-\tau,L-\tau}^x = \left(\Psi_{L,\tau}^{\mathrm{T}}\Sigma_{e,L}^{-1}\Psi_{L,\tau}\right)^{(1)}\Psi_{L,\tau}^{\mathrm{T}}\Sigma_{e,L}^{-1}\mathbf{r}_{k,L}.
\tag{22}
$$

We will show in the following theorem, however, that the last n_f entries of $\hat{\mathbf{f}}^x_{k-\tau,L-\tau}$, i.e.,

$$\hat{f}(k-\tau) = \mathcal{I}_{n_f}\hat{\mathbf{f}}^x_{k-\tau,L-\tau} \tag{23}$$

with $\mathcal{I}_{n_f} = \begin{bmatrix} 0 & I_{n_f} \end{bmatrix} \in \mathbb{R}^{n_f \times (n+n_f(L-\tau))}$, represent an (asymptotically) unbiased estimate of $f(k-\tau)$ under certain conditions. The estimation delay τ in (23) is caused by the relative degree in Assumption 2.

Theorem 1 *Let τ and ν denote the relative degree and the observability index of the fault subsystem $(\Phi, \tilde{E}, C, G)$, respectively.*

(i) *The τ-delay fault estimate $\hat{f}(k-\tau)$ defined in (23) is unbiased for all $L \geq \nu + \tau$ if and only if $(\Phi, \tilde{E}, \mathcal{O}_{\tau+1}, \mathbf{H}^f_\tau)$ has no transmission zeros, with*

$$\mathbf{H}^f_\tau = \begin{bmatrix} (H^f_0)^{\mathrm{T}} & (H^f_1)^{\mathrm{T}} & \cdots & (H^f_\tau)^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}. \tag{24}$$

(ii) *The τ-delay fault estimate $\hat{f}(k-\tau)$ is asymptotically unbiased for $L \to \infty$ if and only if all transmission zeros of $(\Phi, \tilde{E}, \mathcal{O}_{\tau+1}, \mathbf{H}^f_\tau)$ are stable.*

The proof is given in Appendix B.

Instead of including the unknown initial state as in the RHFE problem (20), the essential idea of [12] is to find a lower triangular block-Toeplitz matrix $\mathbf{T}^g_L$ such that $\mathbf{T}^g_L \cdot \mathbf{T}^f_{L,\tau} = I$ and the estimation error caused by the unknown initial state exponentially decays with L. The condition for unbiasedness in [12] requires that the inverse system related to $\mathbf{T}^g_L$ is stable. However this has several drawbacks: it does not clarify how the unbiasedness condition is related to the system property of the underlying plant; and moreover, for the case of sensor faults in an open-loop unstable plant, the method in [12] cannot find a stable left inverse matrix $\mathbf{T}^g_L$ for $\mathbf{T}^f_{L,\tau}$.

On the contrary, Theorem 1 clearly states that the condition for unbiasedness is related to the invariant zeros of the fault subsystem in the underlying plant. An immediate benefit is that our proposed RH fault estimator can ensure (asymptotically) unbiased estimates for sensor faults in an open-loop unstable plant, as long as the fault subsystem has no unstable transmission zeros.

Remark 1 *The unbiasedness condition of the τ-delay fault estimate stated in Theorem 1 has close links with the τ-delay left inversion in [15, 23] and the τ-delay input and initial-state reconstruction in [20]. However, the τ-delay left inversion in [15, 23] requires the initial state to be known a priori, while the τ-delay input and initial-state reconstruction in [20] requires observability of the pair (Φ, C) to simultaneously reconstruct the initial state*

with the unknown input. Although it seems that the RHFE problem (20) jointly estimates initial state and faults, we are actually only interested in the fault estimate without unbiased reconstruction of the unknown initial state. This is an intuitive reason why Theorem 1 can cope with the unknown initial state in the case that (Φ, C) is detectable.

Remark 2 *Theorem 1 above generalizes Theorems 1 and 2 in [35] in two aspects: 1) Theorems 1 and 2 in [35] are limited to the case $\tau = 0$, while Theorem 1 here applies to general relative degrees; 2) Theorems 1 and 2 in [35] focus on the fault estimator constructed with the original system (2), while in this work we construct in Theorem 1 the fault estimator with the predictor (3).*

It should be noted that an RHFE problem similar to (20) can also be formulated using the original system (2), see [35]. Its equivalence to our RHFE problem (20) is shown in the following theorem.

Theorem 2 *If both the original system model (2) and its predictor form (3) are accurately available, the τ-delay fault estimate $\hat{f}(k-\tau)$, computed by (22) and (23) based on the predictor form (3), is equivalent to the fault estimate proposed as Equation (15) in [35] based on the original system model (2).*

The proof of Theorem 2 is given in Appendix C. The above equivalence implies that the predictor gain K does not affect the statistics of the fault estimation error, and the condition of unbiasedness in Theorem 1 holds for the RH fault estimation using the original form.

4 Data-driven nominal receding horizon fault estimator

In this section, we will parameterize the RH fault estimator introduced in Section 3 with the predictor Markov parameters, and then provide the nominal data-driven design method without considering identification errors.

In order to construct the LS fault estimator (22), we first need to construct the block-Toeplitz matrices $\mathbf{T}^u_L$, $\mathbf{T}^y_L$, and $\mathbf{T}^f_L$ from the predictor Markov parameters according to (16). Then, we need the extended observability matrix $\mathcal{O}_L$. One possible approach is to identify $\mathcal{O}_L$ from the block-Hankel matrix

$$\mathbf{H}^o_{L,m} = \begin{bmatrix} H^u_1 & H^u_2 & \cdots & H^u_m \\ H^u_2 & H^u_3 & \cdots & H^u_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ H^u_L & H^u_{L+1} & \cdots & H^u_{L+m-1} \end{bmatrix} \tag{25}$$

through a model reduction step [33]. But this model reduction step would make the fault estimation error depend nonlinearly on the identification errors. In order to

avoid this difficulty, we substitute $\mathcal{O}_L x(k_0) = \mathbf{H}^o_{L,m}\zeta_m$ into (19) by exploiting the following property:

$$\mathcal{R}\left(\mathcal{O}_L\right) = \mathcal{R}\left(\mathbf{H}^o_{L,m}\right) \qquad (26)$$

for $m \geq n$. Then (19) can be rewritten as

$$\mathbf{r}_{k,L} = \underbrace{\left[\begin{array}{cc}\mathbf{H}^o_{L,m} & \mathbf{T}^f_{L,\tau}\end{array}\right]}_{\Upsilon_{L,\tau}}\underbrace{\left[\begin{array}{c}\zeta_m \\ \mathbf{f}_{k-\tau,L-\tau}\end{array}\right]}_{\mathbf{f}^\zeta_{k-\tau,L-\tau}} + \mathbf{e}_{k,L}, \qquad (27)$$

where $\mathbf{T}^f_{L,\tau}$ consists of the first $L-\tau$ block-columns of $\mathbf{T}^f_L$ defined in (16). By doing so, the fault estimation error becomes linear with regards to the identification errors, as shown later in (43). Based on (27), an LS problem similar to (20) can be formulated, and one solution is

$$\hat{\mathbf{f}}^\zeta_{k-\tau,L-\tau} = \left(\Upsilon^{\mathrm{T}}_{L,\tau}\Sigma^{-1}_{e,L}\Upsilon_{L,\tau}\right)^{(1)}\Upsilon^{\mathrm{T}}_{L,\tau}\Sigma^{-1}_{e,L}\mathbf{r}_{k,L}. \qquad (28)$$

Similarly to (23), we obtain the fault estimate

$$\hat{f}(k-\tau) = \mathcal{I}_{n_f}\hat{\mathbf{f}}^\zeta_{k-\tau,L-\tau}, \qquad (29)$$

with $\mathcal{I}_{n_f} = \left[\begin{array}{cc}0 & I_{n_f}\end{array}\right] \in \mathbb{R}^{n_f \times (n_u \cdot m + n_y(L-\tau))}$.

Theorem 3 *The sufficient and necessary condition for unbiased estimation in Theorem 1 applies to the fault estimate defined in (28)-(29).*

The proof is given in Appendix D.

Combining (17), (28), and (29) yields the RH fault estimator as below:

$$\hat{f}(k-\tau) = \mathcal{G}_{\mathrm{n}}\mathbf{r}_{k,L} = \mathcal{G}_{\mathrm{n}}\left[\begin{array}{ccc}I & -\mathbf{T}^y_L & -\mathbf{T}^u_L\end{array}\right]\left[\begin{array}{c}\mathbf{y}_{k,L} \\ \mathbf{u}_{k,L}\end{array}\right], \qquad (30)$$

$$\mathcal{G}_{\mathrm{n}} = \mathcal{I}_{n_f}\left(\Upsilon^{\mathrm{T}}_{L,\tau}\Sigma^{-1}_{e,L}\Upsilon_{L,\tau}\right)^{(1)}\Upsilon^{\mathrm{T}}_{L,\tau}\Sigma^{-1}_{e,L}, \qquad (31)$$

where $\mathcal{G}_{\mathrm{n}}$ represents the nominal RH fault estimator based on the residual signal $\mathbf{r}_{k,L}$.

Without considering the identification errors, the data-driven design of nominal RH fault estimator can now be summarized in Algorithm 1. For the sake of brevity, we do not list the estimated fault Markov parameters $\hat{H}^f_i$ and their estimation errors for simultaneous sensor and actuator faults, because they can be straightforwardly derived similarly to (32) and (33). Thus all our proposed algorithms in this paper can be directly extended to deal with simultaneous sensor and actuator faults.

Algorithm 1 Data-driven nominal RH fault estimation

1) Collect identification data from the fault-free condition, and form the data matrices $\mathbf{Y}_{\mathrm{id}}$ and $\mathbf{Z}_{\mathrm{id}}$ with sufficiently large p [33].
2) Compute the sequence of Markov parameters $\hat{\Xi}$ and the innovation covariance $\hat{\Sigma}_e$ via (10) and (14); extract the identified Markov parameters $\hat{H}^u_i$ and $\hat{H}^y_i$ from $\hat{\Xi}$ according to (9); and extract $\hat{H}^f_i$ according to (4)-(7):
   - for jth sensor faults:

   $$\hat{H}^f_i = -(\hat{H}^y_i)^{[j]} \text{ for } i > 0, \text{ and } \hat{H}^f_0 = I^{[j]}; \qquad (32)$$

   - or for lth actuator faults:

   $$\hat{H}^f_i = (\hat{H}^u_i)^{[l]} \ (i \geq 0). \qquad (33)$$

3) Select sufficiently large L. Construct the estimates of $\Sigma_{e,L}$ in (21), $\mathbf{T}^y_L, \mathbf{T}^u_L, \mathbf{T}^f_L$ in (16), $\mathbf{H}^o_{L,m}$ in (25), and $\Upsilon_{L,\tau}$ in (27) as $\hat{\Sigma}_{e,L}, \hat{\mathbf{T}}^y_L, \hat{\mathbf{T}}^u_L, \hat{\mathbf{T}}^f_L, \hat{\mathbf{H}}^o_{L,m}$, and $\hat{\Upsilon}_{L,\tau}$ by using $\hat{\Sigma}_e$ and the identified Markov parameters $\{\hat{H}^u_i, \hat{H}^y_i, \hat{H}^f_i\}$. Form $\hat{\mathbf{T}}^f_{L,\tau}$ with the first $L - \tau$ block-columns of $\hat{\mathbf{T}}^f_L$.
4) Compute the nominal fault estimator according to (30) and (31).

---

5 Data-driven robust receding horizon fault estimation

The data-driven nominal design in Algorithm 1 might give biased fault estimates due to errors in the identified Markov parameters. To address this problem, this section proposes an offline robust design which regards the online I/O data as unknown in the design stage.

5.1 Data-driven robust design

Since the Markov parameters related to faults are extracted from $\hat{H}^u_i$ or $\hat{H}^y_i$ via (33) or (32), the identification errors of $\hat{H}^f_i$ can be expressed as

$$\Delta H^f_i = \mathbf{E}_{\mathrm{id}} M^f_i, \qquad (34)$$

where

$$M^f_i = \begin{cases} (M^u_i)^{[j]} & \text{for faults of the } j\text{th actuator} \\ -(M^y_i)^{[j]} & \text{for faults of the } j\text{th sensor} \end{cases} \qquad (35)$$

with M^u_i and M^y_i defined in (12)-(13).

With (12) and (34), the estimated matrices $\hat{\mathbf{T}}^y_L$, $\hat{\mathbf{T}}^u_L$,

$\hat{\mathbf{T}}_{L,\tau}^f$, $\hat{\mathbf{H}}_{L,m}^o$ and $\hat{\Upsilon}_{L,\tau}$ in Algorithm 1 can be written as

$$\hat{\mathbf{H}}_{L,m}^o = \mathbf{H}_{L,m}^o + \bar{\mathbf{E}}_{\mathrm{id}}\bar{\mathbf{M}}_{L,m}^o, \quad \hat{\mathbf{T}}_L^y = \mathbf{T}_L^y - \bar{\mathbf{E}}_{\mathrm{id}}\bar{\mathbf{M}}_L^y, \quad (36)$$

$$\hat{\mathbf{T}}_L^u = \mathbf{T}_L^u + \bar{\mathbf{E}}_{\mathrm{id}}\bar{\mathbf{M}}_L^u, \quad \hat{\mathbf{T}}_{L,\tau}^f = \mathbf{T}_{L,\tau}^f + \bar{\mathbf{E}}_{\mathrm{id}}\bar{\mathbf{M}}_{L,\tau}^f, \quad (37)$$

$$\hat{\Upsilon}_{L,\tau} = \Upsilon_{L,\tau} + \bar{\mathbf{E}}_{\mathrm{id}}\bar{\mathbf{M}}_{\Upsilon}, \quad (38)$$

where $\bar{\mathbf{M}}_{L,m}^o$ is the block-Hankel matrix constructed with $M_1^u, M_2^u, \cdots, M_{L+m-1}^u$ similarly to $\mathbf{H}_{L,m}^o$ in (25), $\bar{\mathbf{M}}_L^\star$ is the block-Toeplitz matrix constructed with $M_0^\star, M_1^\star, \cdots, M_{L-1}^\star$ similarly to $\mathbf{T}_L^\star$ in (16) with $\star$ representing u, y, or f,

$$\bar{\mathbf{E}}_{\mathrm{id}} = \mathrm{diag}\underbrace{(\mathbf{E}_{\mathrm{id}}, \mathbf{E}_{\mathrm{id}}, \cdots, \mathbf{E}_{\mathrm{id}})}_{L\ blocks}, \quad (39)$$

$$\bar{\mathbf{M}}_{\Upsilon} = \begin{bmatrix} \bar{\mathbf{M}}_{L,m}^o & \bar{\mathbf{M}}_{L,\tau}^f \end{bmatrix}, \quad (40)$$

and $\bar{\mathbf{M}}_{L,\tau}^f$ consists of the first $L - \tau$ block-columns of $\bar{\mathbf{M}}_L^f$.

Based on (36)-(38), we can write down the residual signal $\hat{\mathbf{r}}_{k,L}$ considering identification errors according to (17)-(19) and (27):

$$\begin{aligned} \hat{\mathbf{r}}_{k,L} &= \mathbf{y}_{k,L} - \hat{\mathbf{T}}_L^y \mathbf{y}_{k,L} - \hat{\mathbf{T}}_L^u \mathbf{u}_{k,L} \\ &= \Upsilon_{L,\tau}\mathbf{f}_{k-\tau,L-\tau}^\zeta + \mathbf{e}_{k,L} + \left(\mathbf{T}_L^y - \hat{\mathbf{T}}_L^y\right)\mathbf{y}_{k,L} \\ &\quad + \left(\mathbf{T}_L^u - \hat{\mathbf{T}}_L^u\right)\mathbf{u}_{k,L} \\ &= \left(\hat{\Upsilon}_{L,\tau} - \bar{\mathbf{E}}_{\mathrm{id}}\bar{\mathbf{M}}_{\Upsilon}\right)\mathbf{f}_{k-\tau,L-\tau}^\zeta + \mathbf{e}_{k,L} \\ &\quad - \bar{\mathbf{E}}_{\mathrm{id}}\underbrace{\begin{bmatrix} -\bar{\mathbf{M}}_L^y & \bar{\mathbf{M}}_L^u \end{bmatrix}}_{\bar{\mathbf{M}}_L^z}\underbrace{\begin{bmatrix} \mathbf{y}_{k,L} \\ \mathbf{u}_{k,L} \end{bmatrix}}_{\mathbf{z}_{k,L}}. \end{aligned} \quad (41)$$

Similarly to $\mathcal{G}_{\mathrm{n}}$ in (30), let the matrix $\mathcal{G}$ denote the τ-delay fault estimator based on the residual $\hat{\mathbf{r}}_{k,L}$, i.e.,

$$\hat{f}(k-\tau) = \mathcal{G}\hat{\mathbf{r}}_{k,L}. \quad (42)$$

It follows from (41) that the fault estimation error is

$$\begin{aligned} \Delta f(k-\tau) &= \hat{f}(k-\tau) - \mathcal{I}_{n_f}\mathbf{f}_{k-\tau,L-\tau}^\zeta \\ &= \underbrace{\left(\mathcal{G}\hat{\Upsilon}_{L,\tau} - \mathcal{G}\bar{\mathbf{E}}_{\mathrm{id}}\bar{\mathbf{M}}_{\Upsilon} - \mathcal{I}_{n_f}\right)}_{\mathcal{T}_f(\mathcal{G})}\mathbf{f}_{k-\tau,L-\tau}^\zeta \\ &\quad - \underbrace{\mathcal{G}\bar{\mathbf{E}}_{\mathrm{id}}\bar{\mathbf{M}}_L^z}_{\mathcal{T}_z(\mathcal{G})}\mathbf{z}_{k,L} + \mathcal{G}\mathbf{e}_{k,L} \end{aligned}$$

$$(43)$$

where $\mathcal{I}_{n_f}$ is defined in (29). It can be seen that $\bar{\mathbf{E}}_{\mathrm{id}}$ appears as multiplicative uncertainty coupled with the true augmented fault signal $\mathbf{f}_{k-\tau,L-\tau}^\zeta$ and the online I/O data $\mathbf{z}_{k,L}$.

We regard $\mathbf{f}_{k-\tau,L-\tau}^\zeta$ and $\mathbf{z}_{k,L}$ as unknown but energy bounded. Hence $\mathbf{f}_{k-\tau,L-\tau}^\zeta$ and $\mathbf{z}_{k,L}$ in the first two terms of (43) lead to an estimation bias, while the online innovation signal $\mathbf{e}_{k,L}$ in the third term causes zero mean, stochastic estimation errors. We would like to reduce the estimation bias by minimizing the matrix 2-norms $\|\mathcal{T}_s(\mathcal{G})\|_2$ ($s = f, z$), and at the same time minimize the Frobenius norm $\mathrm{tr}\left(\mathcal{G}\Sigma_{e,L}\mathcal{G}^{\mathrm{T}}\right)$ by using the available innovation covariance $\Sigma_{e,L}$. These three objectives are formulated by the following mixed-norm problem:

$$\begin{aligned} \mathcal{G}_{\mathrm{r,off}} &= \arg\min_{\mathcal{G}} \ \mathrm{tr}\left(\mathcal{G}\Sigma_{e,L}\mathcal{G}^{\mathrm{T}}\right) \\ &\text{s.t. } \bar{\mathbb{E}}\left(\mathcal{T}_s(\mathcal{G})\mathcal{T}_s^{\mathrm{T}}(\mathcal{G})\right) \leq \gamma_s^2 I, \ s = f, z \end{aligned} \quad (44)$$

where the matrix $\mathcal{G}$ denotes the τ-delay fault estimator (42), $\bar{\mathbb{E}}$ denotes mathematical expectation over the identification innovations $\bar{\mathbf{E}}_{\mathrm{id}}$, $\gamma_f > 0$ and $\gamma_z > 0$ are the user-defined parameters to achieve a trade-off between estimation error variance and bias. Note that the matrix 2-norms $\|\mathcal{T}_s(\mathcal{G})\|_2$ ($s = f, z$) are affected by the stochastic identification innovations $\bar{\mathbf{E}}_{\mathrm{id}}$ according to (43), hence their mathematical expectations are used in (44). Note also that it is straightforward to prove $\bar{\mathbb{E}}\left(\mathcal{T}_s^{\mathrm{T}}(\mathcal{G})\mathcal{T}_s(\mathcal{G})\right) \leq \gamma_s^2 I$ holds if and only if $\bar{\mathbb{E}}\left(\mathcal{T}_s(\mathcal{G})\mathcal{T}_s^{\mathrm{T}}(\mathcal{G})\right) \leq \gamma_s^2 I$ in (44) holds. Here we use $\bar{\mathbb{E}}\left(\mathcal{T}_s(\mathcal{G})\mathcal{T}_s^{\mathrm{T}}(\mathcal{G})\right)$ in (44), because it brings a clear geometrical interpretation for parameter tuning as explained later in Section 5.2. With the tedious but straightforward derivations summarized in Appendix E, the above problem (44) can be explicitly written as

$$\mathcal{G}_{\mathrm{r,off}} = \arg\min_{\mathcal{G}} \ \mathrm{tr}\left(\mathcal{G}\Sigma_{e,L}\mathcal{G}^{\mathrm{T}}\right) \quad (45a)$$

$$\text{s.t. } \begin{bmatrix} \mathcal{G} & \mathcal{I}_{n_f} \end{bmatrix} \begin{bmatrix} \Pi_f & -\hat{\Upsilon}_{L,\tau} \\ -\hat{\Upsilon}_{L,\tau}^{\mathrm{T}} & I_{n_f} \end{bmatrix} \begin{bmatrix} \mathcal{G}^{\mathrm{T}} \\ \mathcal{I}_{n_f}^{\mathrm{T}} \end{bmatrix} \leq \gamma_f^2 I \quad (45b)$$

$$\mathcal{G}\Pi_z\mathcal{G}^{\mathrm{T}} \leq \gamma_z^2 I, \quad (45c)$$

with Π_f and Π_z defined in (E.5) and (E.6), respectively. The mixed-norm problem (45) can be easily transformed into an equivalent semi-definite programming (SDP) problem that can be solved efficiently [3]. Since the optimization problem (45) is determined only by the identification data and does not involve any online I/O data, it can be solved offline to obtain the robust fault estimator denoted as $\mathcal{G}_{\mathrm{r,off}}$.

5.2 Parameter tuning using geometric interpretation

Next, we will provide a systematic method to tune the two user-defined parameters γ_f^2 and γ_z^2 by using a geometric interpretation of the mixed-norm problem (45).

With some matrix manipulations, we can see that the constraints (45b) and (45c) define two ellipsoids

$$\Omega_f = \left\{ \mathcal{G} \,\middle|\, (\mathcal{G} - \mathcal{G}_0) \, \Pi_f \, (\mathcal{G} - \mathcal{G}_0)^{\mathrm{T}} \leq \mathcal{G}_0 \Pi_f \mathcal{G}_0^{\mathrm{T}} - I + \gamma_f^2 I \right\},$$
(46)

$$\Omega_z = \left\{ \mathcal{G} \,\middle|\, \mathcal{G} \Pi_z \mathcal{G}^{\mathrm{T}} \leq \gamma_z^2 I \right\},$$
(47)

respectively, with $\mathcal{G}_0 = \mathcal{I}_{n_f} \hat{\Upsilon}_{L,\tau}^{\mathrm{T}} \Pi_f^{-1}$. Since the objective function (45a) can be regarded as a measure of the distance from $\mathcal{G}$ to the origin $0_{n_f \times (n_y \cdot L)}$, the optimization problem (45) is equivalent to finding the point $\mathcal{G}_{\mathrm{r,off}}$ in the set $\Omega_f \bigcap \Omega_z$ that is closest to the origin, as shown in Fig. 1.

First, we would like to find the region of γ_f^2 and γ_z^2 so that the optimization problem (45) is feasible and non-trivial. In the case that the origin $0_{n_f \times (n_y \cdot L)} \in \Omega_f \bigcap \Omega_z$, we would have the trivial solution $\mathcal{G}_{\mathrm{r,off}} = 0_{n_f \times (n_y \cdot L)}$ which makes no sense for fault estimation. Hence $0_{n_f \times (n_y \cdot L)} \notin \Omega_f$ and $\Omega_f \neq \varnothing$ are both required, which implies the region of γ_f^2 as below according to (46):

$$1 - \lambda_{\min} \left(\mathcal{G}_0 \Pi_f \mathcal{G}_0^{\mathrm{T}} \right) = \gamma_{f,\min}^2 \leq \gamma_f^2 < 1.$$
(48)

For a given γ_f^2 satisfying (48), we solve the following optimization problem

$$\left\{ \mathcal{G}_{\min}, \gamma_{z,\min}^2 \right\} = \arg \min_{\mathcal{G}, \gamma_z^2} \; \gamma_z^2$$
$$\text{s.t. (45b) and (45c)}$$
(49)

whose solution gives the minimal γ_z^2, referred to as $\gamma_{z,\min}^2$, that ensures $\Omega_f \bigcap \Omega_z \neq \varnothing$. Therefore, we should select $\gamma_z^2 \in \left[\gamma_{z,\min}^2, \infty \right)$ to ensure feasibility of the optimization problem (45). The ellipsoid $\Omega_{z,\min}$ in Fig. 1 represents the ellipsoid Ω_z with $\gamma_z^2 = \gamma_{z,\min}^2$, and its intersection with the ellipsoid Ω_f includes only the single point $\mathcal{G}_{\min}$.

By discarding the constraint (45c) from the problem (45) and fixing γ_f^2 at the same given value as in (49), we formulate another problem

$$\mathcal{G}_1 = \arg \min_{\mathcal{G}} \; \mathrm{tr} \left(\mathcal{G} \Sigma_{e,L} \mathcal{G}^{\mathrm{T}} \right)$$
$$\text{s.t. (45b)}$$
(50)

Because the optimal solution $\mathcal{G}_1$ gives the shortest distance from the origin to the ellipsoid Ω_f, and moreover $0_{n_f \times (n_y \cdot L)} \notin \Omega_f$, the solution $\mathcal{G}_1$ must lie at the boundary of the ellipsoid Ω_f, as shown in Fig. 1. Define $\gamma_{z,1}^2 = \lambda_{\max} \left(\bar{\mathbb{E}} \left(\mathcal{T}_z \left(\mathcal{G}_1 \right) \mathcal{T}_z^{\mathrm{T}} \left(\mathcal{G}_1 \right) \right) \right)$. Let the ellipsoid $\Omega_{z,1}$ in Fig. 1 represent the set Ω_z with $\gamma_z^2 = \gamma_{z,1}^2$, and it has the solution $\mathcal{G}_1$ at its boundary.

Similarly to the above obtained solution $\mathcal{G}_1$ of the problem (50), the solution $\mathcal{G}_{\mathrm{r,off}}$ of the problem (45) also lies at the boundary of the ellipsoid Ω_f. This allows the three terms of the fault estimation error in (43) to be explained using Fig. 1:

1) The bias related to the first term $\mathcal{T}_f \left(\mathcal{G} \right) \mathbf{f}_{k-\tau, L-\tau}^{\zeta}$ is determined by the size of the ellipsoid Ω_f;
2) The bias related to the second term $\mathcal{T}_z \left(\mathcal{G} \right) \mathbf{z}_{k,L}$ is determined by the size of the ellipsoid $\Omega_z \left(\mathcal{G}_{\mathrm{r,off}} \right)$ with $\mathcal{G}_{\mathrm{r,off}}$ lying on its boundary, i.e., the ellipsoid Ω_z with $\gamma_z^2 = \lambda_{\max} \left(\bar{\mathbb{E}} \left(\mathcal{T}_z \left(\mathcal{G}_{\mathrm{r,off}} \right) \mathcal{T}_z^{\mathrm{T}} \left(\mathcal{G}_{\mathrm{r,off}} \right) \right) \right)$;
3) The fault estimation error variance related to the third term $\mathcal{G} \mathbf{e}_{k,L}$ is represented by the distance from the origin to the optimal solution $\mathcal{G}_{\mathrm{r,off}}$.

With the above basic geometric interpretation, we can analyze the performance trade-offs of the robust fault estimator $\mathcal{G}_{\mathrm{r,off}}$ when tuning $\gamma_f^2 \in \left[\gamma_{f,\min}^2, 1 \right)$ and $\gamma_z^2 \in \left[\gamma_{z,\min}^2, \infty \right)$, as shown in Table 1. First, we fix γ_f^2 and tune γ_z^2. In this case, the ellipsoid Ω_f is fixed, which makes the first bias term in the first two rows of Table 1 remain constant. With the fixed γ_f^2, by increasing γ_z^2 from $\gamma_{z,\min}^2$ towards $\gamma_{z,1}^2$, the intersection set $\Omega_f \bigcap \Omega_z$ becomes larger, and the optimal solution $\mathcal{G}_{\mathrm{r,off}}$ moves from the point $\mathcal{G}_{\min}$ along the boundary of the ellipsoid Ω_f towards the point $\mathcal{G}_1$. When we further increase γ_z^2 for $\gamma_z^2 \geq \gamma_{z,1}^2$, the optimal solution $\mathcal{G}_{\mathrm{r,off}}$ of the problem (45) would remain located at the point $\mathcal{G}_1$, because $\mathcal{G}_1$ satisfies both constraints (45b) and (45c) and gives the shortest distance to the origin according to the problem (50). Therefore, the size of the ellipsoid $\Omega_z \left(\mathcal{G}_{\mathrm{r,off}} \right)$, which determines the second estimation bias term in the first two rows of Table 1, monotonically increases for $\gamma_z^2 \in \left[\gamma_{z,\min}^2, \gamma_{z,1}^2 \right)$ and remains constant for $\gamma_z^2 \in \left[\gamma_{z,1}^2, \infty \right)$. The distance from the origin to $\mathcal{G}_{\mathrm{r,off}}$, which determines the fault estimation error variance in the first two rows of Table 1, monotonically decreases for $\gamma_z^2 \in \left[\gamma_{z,\min}^2, \gamma_{z,1}^2 \right)$ and remains constant for $\gamma_z^2 \in \left[\gamma_{z,1}^2, \infty \right)$. For the third row of Table 1, we tune γ_f^2 and select a sufficiently large value of γ_z^2 that ensures the problem (45) to be feasible. With γ_f^2 increasing, the size of the ellipsoid Ω_f, which determines the first bias term in the third row of Table 1, monotonically increases. Meanwhile, the optimal solution $\mathcal{G}_{\mathrm{r,off}}$, which lies at the boundary of the ellipsoid Ω_f, moves closer to the origin. Therefore, both the second bias term and the fault estimation error variance in the third row of Table 1, which are determined by the size of the ellipsoid $\Omega_z \left(\mathcal{G}_{\mathrm{r,off}} \right)$ and the distance from the origin to the point $\mathcal{G}_{\mathrm{r,off}}$, monotonically decrease.

We summarize the data-driven robust design in Algorithm 2. The nominal design $\mathcal{G}_{\mathrm{n}}$ obtained from Algorithm 1 can be used as a benchmark for tuning γ_f^2 and γ_z^2 in Step 2 of Algorithm 2. For example, compared to the nominal design, the robust design achieves smaller aver-

Table 1
Trade-offs between fault estimation bias and error variance of the robust fault estimator $\mathcal{G}_{\mathrm{r,off}}$ at time instant k when tuning user-defined parameters γ_f^2 and γ_z^2 in (45): "Constant", "↗", and "↘" means that the performance criterion in the corresponding column remains constant, monotonically increases, and monotonically decreases with regard to the user-defined parameter specified in the corresponding row, respectively.

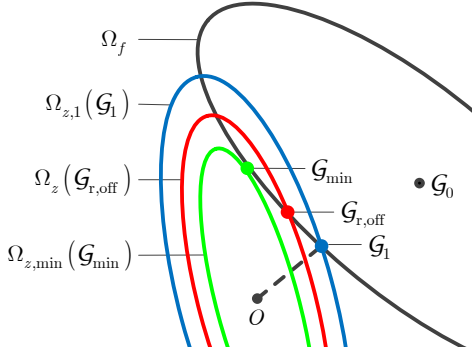| User-defined parameters	First bias term $\bar{\mathbb{E}} \left\| \mathcal{T}_f \left(\mathcal{G}_{\mathrm{r,off}} \right) \mathbf{f}_{k-\tau, L-\tau}^{\zeta} \right\|_2^2$	Second bias term $\bar{\mathbb{E}} \left\| \mathcal{T}_z \left(\mathcal{G}_{\mathrm{r,off}} \right) \mathbf{z}_{k,L} \right\|_2^2$	Variance $\mathrm{tr} \left(\mathcal{G}_{\mathrm{r,off}} \Sigma_{e,L} \mathcal{G}_{\mathrm{r,off}}^{\mathrm{T}} \right)$
$\gamma_z^2 \in \left[\gamma_{z,\min}^2, \gamma_{z,1}^2 \right]$	Constant	↗	↘
$\gamma_z^2 \in \left[\gamma_{z,1}^2, \infty \right)$	Constant	Constant	Constant
$\gamma_f^2 \in \left[\gamma_{f,\min}^2, 1 \right)$	↗	↘	↘

Fig. 1. Geometric interpretation of the mixed-norm problem (45): the constraints (45b) and (45c) define the ellipsoid Ω_f centered at $\mathcal{G}_0$ and the ellipsoid Ω_z centered at the origin O, respectively. Lying at the boundary of the ellipsoid Ω_f, the optimal solution $\mathcal{G}_{\mathrm{r,off}}$ gives the shortest distance measured by the objective function (45a) from the origin to the intersection set $\Omega_f \bigcap \Omega_z$. With $\gamma_z^2 = \gamma_{z,\min}^2$, the ellipsoid Ω_z becomes $\Omega_{z,\min} (\mathcal{G}_{\min})$ in green which intersects with the ellipsoid Ω_f at a single point $\mathcal{G}_{\min}$. At the boundary of the ellipsoid Ω_f, $\mathcal{G}_1$ gives the shortest distance from the origin to the ellipsoid Ω_f. The ellipsoids $\Omega_{z,1} (\mathcal{G}_1)$ in blue and $\Omega_z (\mathcal{G}_{\mathrm{r,off}})$ in red represent the ellipsoids Ω_z with $\mathcal{G}_1$ and $\mathcal{G}_{\mathrm{r,off}}$ lying at the boundary, respectively.

aged worst-case bias if $\gamma_s^2 \leq \lambda_{\max} \left(\bar{\mathbb{E}} \left(\mathcal{T}_s \left(\mathcal{G}_{\mathrm{n}} \right) \mathcal{T}_s^{\mathrm{T}} \left(\mathcal{G}_{\mathrm{n}} \right) \right) \right)$ $(s = f, z)$.

Algorithm 2 Data-driven robust RH fault estimation
1) Complete the steps 1-3 in Algorithm 1; compute M_i^u, M_i^y, and M_i^f according to (13) and (35).
2) Tune $\gamma_f^2 \in \left[\gamma_{f,\min}^2, 1 \right)$ and $\gamma_z^2 \in \left[\gamma_{z,\min}^2, \infty \right)$ according to the performance trade-offs shown in Table 1, where $\gamma_{f,\min}^2$ and $\gamma_{z,\min}^2$ are obtained from the optimization problems (48) and (49) respectively.
3) Solve the problem (45) to compute the robust RH fault estimator $\mathcal{G}_{\mathrm{r,off}}$.

---

6 Data-driven robust receding horizon fault estimation with online optimization

The online I/O data is regarded as unknown in Algorithm 2. In order to better exploit the available online data, this section proposes an online mixed-norm optimization approach. This can further reduce the estimation errors when the online I/O data have large amplitudes, at the expense of increased computational burden.

6.1 Online mixed-norm problem

With the notation

$$\bar{\beta}_{k,L} = \bar{\mathbf{M}}_L^z \mathbf{z}_{k,L}, \tag{51}$$

we divide $\bar{\beta}_{k,L}$ into L row blocks as in

$$\bar{\beta}_{k,L} = \begin{bmatrix} \beta_{k,1}^{\mathrm{T}} & \beta_{k,2}^{\mathrm{T}} & \cdots & \beta_{k,L}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \tag{52}$$

with $\beta_{k,i} \in \mathbb{R}^N$. Then the term $\mathcal{G}\bar{\mathbf{E}}_{\mathrm{id}}\bar{\mathbf{M}}_L^z \mathbf{z}_{k,L}$ in (43) can be rewritten as

$$\mathcal{G}\bar{\mathbf{E}}_{\mathrm{id}}\bar{\mathbf{M}}_L^z \mathbf{z}_{k,L} = \mathcal{G}\bar{\mathbf{E}}_{\mathrm{id}}\bar{\beta}_{k,L}$$
$$= \mathcal{G} \begin{bmatrix} \mathbf{E}_{\mathrm{id}}\beta_{k,1} \\ \mathbf{E}_{\mathrm{id}}\beta_{k,2} \\ \vdots \\ \mathbf{E}_{\mathrm{id}}\beta_{k,L} \end{bmatrix} = \mathcal{G} \underbrace{\begin{bmatrix} \beta_{k,1}^{\mathrm{T}} \otimes I_{n_y} \\ \beta_{k,2}^{\mathrm{T}} \otimes I_{n_y} \\ \vdots \\ \beta_{k,L}^{\mathrm{T}} \otimes I_{n_y} \end{bmatrix}}_{\Gamma_{k,L}} \mathrm{vec} \left(\mathbf{E}_{\mathrm{id}} \right) \tag{53}$$

according to the property of Kronecker product [4]. Based on (53), the estimation error in (43) becomes

$$\Delta f(k - \tau) = \mathcal{T}_f \left(\mathcal{G} \right) \mathbf{f}_{k-\tau, L-\tau}^{\zeta} - \mathcal{G}\Gamma_{k,L}\mathrm{vec} \left(\mathbf{E}_{\mathrm{id}} \right) + \mathcal{G}\mathbf{e}_{k,L}. \tag{54}$$

Then the statistics of $\mathrm{vec} \left(\mathbf{E}_{\mathrm{id}} \right)$, i.e.,

$$\mathbb{E} \left(\mathrm{vec} \left(\mathbf{E}_{\mathrm{id}} \right) \mathrm{vec} \left(\mathbf{E}_{\mathrm{id}} \right)^{\mathrm{T}} \right) = I_N \otimes \Sigma_e,$$

can be exploited to evaluate the fault estimation error variance. Therefore, we formulate the following optimization problem similarly to (44):

$$\mathcal{G}_{\mathrm{r,on}} = \arg \min_{\mathcal{G}} \; \mathrm{tr} \left(\mathcal{G} \Sigma_{e,L} \mathcal{G}^{\mathrm{T}} + \mathcal{G} \Gamma_{k,L} \left(I_N \otimes \Sigma_e \right) \Gamma_{k,L}^{\mathrm{T}} \mathcal{G}^{\mathrm{T}} \right)$$
$$\mathrm{s.t.} \; \bar{\mathbb{E}} \left(\mathcal{T}_f \left(\mathcal{G} \right) \mathcal{T}_f^{\mathrm{T}} \left(\mathcal{G} \right) \right) \leq \gamma_f^2 I$$
(55)

with the user-defined parameter γ_f. The constraint in the above optimization problem (55) can be explicitly written as (45b). The optimization problem (55) has to be solved at each time instant to update the robust fault estimator $\mathcal{G}_{\mathrm{r,on}}$ because $\Gamma_{k,L}$ in the cost function is determined by the online I/O data according to (51)-(53).

6.2 Parameter tuning using geometric interpretation

Since the online mixed-norm problem (55) has the structure similar to that of the offline mixed-norm problem (45), the performance trade-offs by tuning γ_f in (55) are also similar to that explained in Table 1.

The proposed data-driven robust fault estimation with online optimization is summarized in Algorithm 3. In order to reduce the computational burden of online optimization, the problem (55) is implemented only if the estimation bias of the offline designed fault estimator is larger than a user-defined threshold α, as shown in Step 2 of Algorithm 3.

The offline designed fault estimator $\mathcal{G}_{\mathrm{r,off}}$ from Algorithm 2 can be used as a benchmark for tuning γ_f^2 in Step 2.2 of Algorithm 3. For example, compared to $\mathcal{G}_{\mathrm{r,off}}$, the online optimization (55) achieves smaller averaged worst-case bias if $\gamma_f^2 \leq \lambda_{\max} \left(\bar{\mathbb{E}} \left(\mathcal{T}_f \left(\mathcal{G}_{\mathrm{r,off}} \right) \mathcal{T}_f^{\mathrm{T}} \left(\mathcal{G}_{\mathrm{r,off}} \right) \right) \right)$.

Algorithm 3 Data-driven robust RH fault estimation with online optimization

---

1) Follow Algorithm 2 to compute the offline designed fault estimator $\mathcal{G}_{\mathrm{r,off}}$.
2) If $\lambda_{\min} \left(\bar{\mathbb{E}} \left(\mathcal{T}_z \left(\mathcal{G}_{\mathrm{r,off}} \right) \mathcal{T}_z^{\mathrm{T}} \left(\mathcal{G}_{\mathrm{r,off}} \right) \right) \right) \| \mathbf{z}_{k,L} \|_2^2 > \alpha$ (α is a user-defined threshold), the online optimization in the following steps is implemented; otherwise, the offline designed estimator $\mathcal{G}_{\mathrm{r,off}}$ is used.
   2.1) Compute $\Gamma_{k,L}$ according to (51)-(53).
   2.2) Tune $\gamma_f^2 \in \left[\gamma_{f,\min}^2, 1 \right)$ similarly to Step 2 of Algorithm 2, with $\gamma_{f,\min}^2$ defined in (48).
   2.3) Solve the problem (55) to compute the robust RH fault estimator $\mathcal{G}_{\mathrm{r,on}}$.

---

7 Simulation studies

Consider a continuous-time linearized vertical take-off and landing (VTOL) aircraft model that has been studied in [12–14, 16]:

$$\dot{x}_c(t) = A_c x_c(t) + B_c u_c(t),$$
$$y_c(t) = C_c(t),$$
$$A_c = \begin{bmatrix} -0.0366 & 0.0271 & 0.0188 & -0.4555 \\ 0.0482 & -1.01 & 0.0024 & -4.0208 \\ 0.1002 & 0.3681 & -0.707 & 1.42 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$
$$B_c = \begin{bmatrix} 0.4422 & 0.1761 \\ 3.5446 & -7.5922 \\ -5.52 & 4.49 \\ 0 & 0 \end{bmatrix}, \; C_c = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}.$$

With a sampling rate of 0.5 seconds, the discrete-time model (2) is obtained, with $D = 0$ and $F = I_4$. The process and measurement noises, $w(k)$ and $v(k)$, are assumed to be zero mean white noises, respectively with covariances of $Q = 0.16 I_4$ and $R = 0.64 I_4$.

Since the open-loop plant is unstable, an empirical stabilizing output feedback controller is used [12], i.e.,

$$u(k) = - \begin{bmatrix} 0 & 0 & -0.5 & 0 \\ 0 & 0 & -0.1 & -0.1 \end{bmatrix} y(k) + \eta(k), \qquad (56)$$

where $\eta(k)$ is the reference signal.

In the identification experiment, the reference signal $\eta(k)$ is zero-mean white noise with the covariance of $\mathrm{diag}\,(1, 1)$, which ensures persistent excitation. We collect $N = 1000$ data samples from the identification experiment. In the identification algorithm, the past horizon is selected as $p = 10$.

The considered fault cases include:

- Actuator faults: $E = B$, $G = D$,
- Sensor faults: $E = 0_{4 \times 2}$, $G = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}^{\mathrm{T}}$.

The case of simultaneous actuator and sensor faults is not included here, because all the considered algorithms can be applied to the simultaneous scenario in a straightforward way, and their performance comparisons are the same as in the case of separate actuator or sensor faults.

The simulated fault signals in both fault cases are the same:

$$f(k) = \begin{cases} \begin{bmatrix} 0 & 0 \end{bmatrix}^{\mathrm{T}}, & 0 \leq k \leq 50, \\ \begin{bmatrix} \sin(0.1\pi k) & 1 \end{bmatrix}^{\mathrm{T}}, & k > 50. \end{cases}$$

We will compare the following fault estimation methods:

- Alg0: the RH fault estimator using accurate Markov parameters, described in Section 4.
- DONG: the method proposed by [12].
- Alg1: the data-driven nominal RH fault estimator $\mathcal{G}_{\mathrm{n}}$ proposed in Algorithm 1;

- Alg2: the data-driven robust RH fault estimator $\mathcal{G}_{\mathrm{r,off}}$ proposed in Algorithm 2; in Step 3 of Algorithm 2, we select $\gamma_f^2 = \lambda_{\max}\left(\bar{\mathbb{E}}\left(\mathcal{T}_f\left(\mathcal{G}_{\mathrm{n}}\right)\mathcal{T}_f^{\mathrm{T}}\left(\mathcal{G}_{\mathrm{n}}\right)\right)\right)$, and

$$\gamma_z^2 = 0.5\left(\gamma_{z,\min}^2 + \gamma_{z,1}^2\right). \qquad (57)$$

- Alg3: the data-driven robust RH fault estimator $\mathcal{G}_{\mathrm{r,on}}$ with online optimization, proposed in Algorithm 3; in Step 2 of Algorithm 3, we select $\alpha = 300$ as the threshold to determine whether or not the online optimization should be implemented; γ_f^2 is set to the same value as in Alg2.

We select the estimation horizon length $L = 30$ for the considered five algorithms.

In order to show the necessity of compensating for the identification errors, we make the identification-error-effect term $\mathcal{T}_z\left(\mathcal{G}\right)\cdot\mathbf{z}_{k,L}$ in (43) significantly large by setting $\eta(k) = 15$. Fault estimates from the above five algorithms are illustrated in Fig. 2, and the distributions of their fault estimation errors are shown in Fig. 3. By using accurate Markov parameters, Alg0 achieves unbiased fault estimation in both fault scenarios. Note that DONG cannot be directly applied to sensor faults in the unstable open-loop VTOL model [12], hence it is not included in Fig. 2 and 3(b) for sensor faults. Because of neglecting the effect of identification errors, both Alg1 and DONG yield estimation biases even larger than the amplitude of true faults. In comparison, Alg2 obtains its robustness to identification error by solving an offline mixed-norm problem, as shown in Fig. 3(a). However, the poor performance of Alg2 in our sensor fault case (Fig. 3(b)) shows the limitation of neglecting the online availability of I/O data in the offline mixed-norm problem. Compared to Alg2, Alg3 significantly reduces estimation bias, as shown in Fig. 3(b), by formulating an online mixed-norm problem to exploit online I/O data. This performance improvement is achieved at the cost of higher online computational burden. When implemented with YALMIP [21] in the MATLAB2011b environment, on a computer with a 3.4 GHz processor and 8 GB RAM, the averaged and peak computational time per sample of Alg3 are 1.70s and 2.05s for the estimation horizon length $L = 30$, while those of Alg2 are 8.37×10^{-6}s and 3.17×10^{-5}s respectively. We will investigate the computational efficiency of Alg3 for real-time implementation in future work.

In order to illustrate the performance trade-offs of Alg2, we set γ_z^2 as in (57) and tune γ_f^2 under the condition of different reference signals $\eta(k)$. Fig. 4 shows how the fault estimation bias, error variance and root mean square error (RMSE) vary with γ_f^2, which can be explained as follows using Table 1. According to the fault estimation error analysis in (43), the fault estimation bias is related to both $\mathcal{T}_f\left(\mathcal{G}_{\mathrm{r,off}}\right)\mathbf{f}_{k-\tau,L-\tau}^{\zeta}$
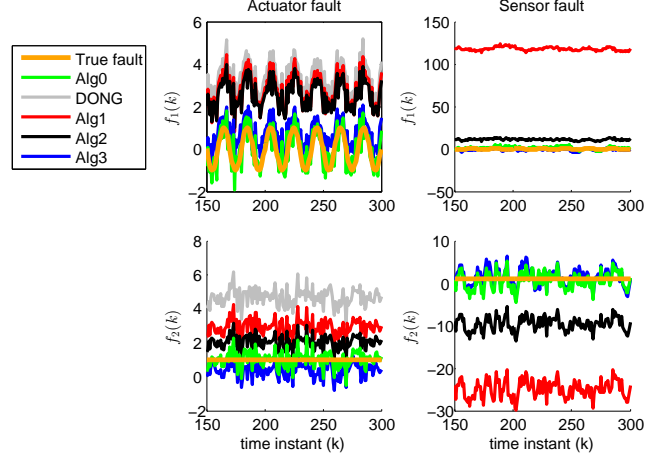

Fig. 2. True fault signal and fault estimates from different algorithms.
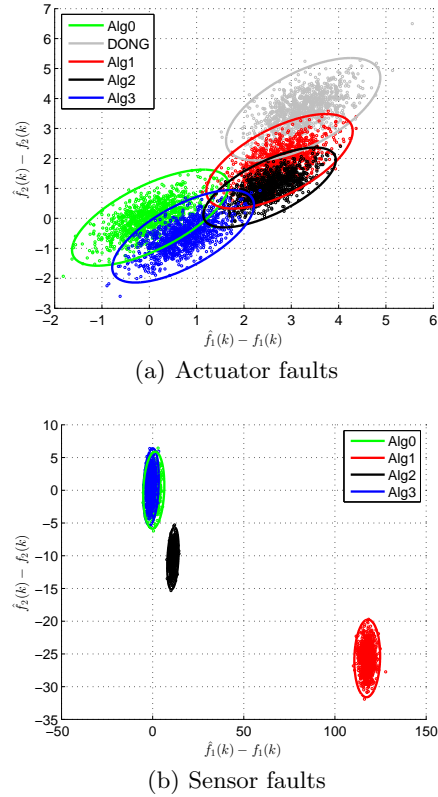

(a) Actuator faults

(b) Sensor faults

Fig. 3. Distribution of fault estimation errors when $\eta(k) = 15$. Circles: 1000 estimation errors by different fault estimation algorithms based on 1000 online I/O data samples. Ellipses: the 3σ-contour of the approximated two-dimensional Gaussian distribution of the 1000 estimation errors, i.e., the contour at $\left[\hat{f}(k) - f(k)\right]^{\mathrm{T}}\mathrm{cov}^{-1}\left(\hat{f}(k)\right)\left[\hat{f}(k) - f(k)\right] = 3$.

and $\mathcal{T}_z\left(\mathcal{G}_{\mathrm{r,off}}\right)\mathbf{z}_{k,L}$. For $\eta(k) = 0$ or $\eta(k) = 1$, the online I/O data $\mathbf{z}_{k,L}$ have small amplitude, thus the total estimation bias is dominated by the bias related

to $\mathcal{T}_f\left(\mathcal{G}_{\mathrm{r,off}}\right)\mathbf{f}_{k-\tau,L-\tau}^{\zeta}$ which monotonically increases with γ_f^2 according to the third row of Table 1. This explains the fault estimation bias curves for $\eta(k)=0$ and $\eta(k)=1$ in Fig. 4. For $\eta(k)=2$, the online I/O data $\mathbf{z}_{k,L}$ have relatively large amplitudes, hence for relatively small values of γ_f^2 the total estimation bias is dominated by the bias related to $\mathcal{T}_z\left(\mathcal{G}_{\mathrm{r,off}}\right)\mathbf{z}_{k,L}$ which monotonically decreases with γ_f^2, and for relatively large values of γ_f^2 the total estimation bias is dominated by the bias related to $\mathcal{T}_f\left(\mathcal{G}_{\mathrm{r,off}}\right)\mathbf{f}_{k-\tau,L-\tau}^{\zeta}$ which monotonically increases with γ_f^2, according to the third row of Table 1. This explains the fault estimation bias curve for $\eta(k)=2$ in Fig. 4. The monotonic decrease of the fault estimation error variances with γ_f^2 can be directly explained with the third row of Table 1. As the objective function of the optimization problem (45), the fault estimation error variance $\mathrm{tr}\left(\mathcal{G}_{\mathrm{r,off}}\Sigma_{\mathrm{e,L}}\mathcal{G}_{\mathrm{r,off}}^{\mathrm{T}}\right)$ for different reference signals $\eta(k)$ is the same because it does not depend on the reference signal $\eta(k)$. Combining the increase of fault estimation bias and the decrease of fault estimation error variance with γ_f^2, there exist the optimal $\gamma_{f,*}^2\in\left(\gamma_{f,\mathrm{min}}^2,1\right)$ such that the RMSE achieves its minimal value, as can be seen in Fig. 4. It is also shown that the minimal RMSE is achieved at a larger value of $\gamma_{f,*}^2$ when the amplitude of $\eta(k)$ increases, because the online I/O data have larger amplitudes with larger $\eta(k)$, thus the decrease of the bias related to $\mathcal{T}_z\left(\mathcal{G}_{\mathrm{r,off}}\right)\mathbf{z}_{k,L}$ dominates the fault estimation bias. Based on the above insights, we can anticipate how the estimation performance of Alg2 varies with different γ_z^2 for a fixed γ_f^2, as well as the performance trade-offs of Alg3. Their performance curves are not plotted due to the space limitation.

From the simulation results with different lenghts L of the estimation horizon (omitted for the sake of brevity), it can be seen that the fault estimation bias and variance of Alg0, Alg1, Alg2, and Alg3 decrease with the increasing length L of the estimation horizon. Straightforward proof of this observation can be derived for Alg0 using accurate Markov parameters (following Section 3.4.3 of [18]), whereas analytical proof is difficult for Alg1, Alg2, and Alg3 that rely on the identified Markov parameters contaminated with identification errors.

8   Conclusions

This paper has investigated data-driven fault estimation and its robustness against stochastic identification errors. First, we proposed an RH fault estimator that can be parameterized with the predictor Markov parameters. Its condition for unbiasedness generalizes that of a recently reported data-driven fault estimation method. An immediate benefit is that our proposed method can
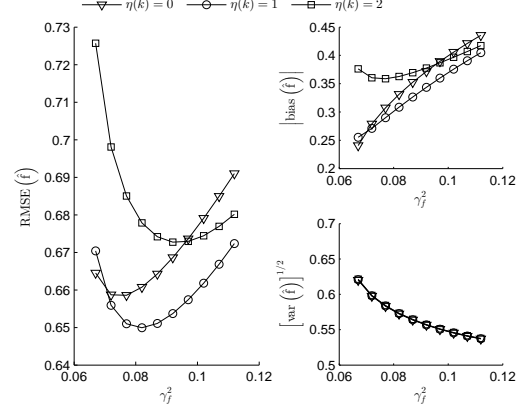

Fig. 4. Estimation performance of Alg2 when tuning γ_f^2 under different reference signal $\eta(k)$

be applied to sensor faults of an unstable open-loop plant which could not be directly addressed previously. In the formulated RH fault estimator, the identification errors appear as multiplicative model uncertainty coupled with the unknown faults and the online I/O data. Then, two mixed-norm problems were formulated to enhance robustness. One can be solved offline by regarding the online I/O data as unknown signals. The other further reduces estimation errors for larger I/O data by exploiting their online availability in the mixed-norm problem, and it requires online optimization. Based on geometric interpretations of the mixed-norm problems, systematic methods were given to tune the user-defined parameters therein. Comparisons using a simulated aircraft model illustrated the advantages and the effectiveness of our proposed method.

Acknowledgements

References

[1] C. F. Alcala and S. J. Qin. Reconstruction-based contribution for process monitoring. *Automatica*, 45:1593–1600, 2009.

[2] M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki. *Diagnosis and Fault-Tolerant Control*. Springer, Berlin Heidelberg, 2 edition, 2006.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, 2004.

[4] J. Brewer. Kronecker products and matrix calculus in system theory. *IEEE Transactions on Automatic Control*, 25:772–781, 1978.

[5] J. Chen and R. Patton. *Robust Model-Based Fault Diagnosis for Dynamic Systems*. Kluwer Academic, Norwell, MA, 1999.

[6] A. Chiuso. The role of vector autoregressive modeling in predictor based subspace identification. *Automatica*, 43:1034–1048, 2007.

[7] S. X. Ding. *Model-Based Fault Diagnosis Techniques: Design Scheme, Algorithms, and Tools.* Springer-Verlag, London, 2 edition, 2013.

[8] S. X. Ding. *Data-Driven Design of Fault Diagnosis and Fault-Tolerant Control Systems.* Springer-Verlag, London, 2014.

[9] S. X. Ding. Data-driven design of monitoring and diagnosis systems for dynamic processes: a review of subspace technique based schemes and some recent results. *Journal of Process Control*, 24:431–449, February 2014.

[10] S. X. Ding, P. Zhang, A. Naik, E. Ding, and B. Huang. Subspace method aided data-driven design of fault detection and isolation systems. *Journal of Process Control*, 19:1496–1510, 2009.

[11] J. Dong. *Data Driven Fault Tolerant Control: a Subspace Approach.* PhD thesis, Delft University of Technology, 2009.

[12] J. Dong and M. Verhaegen. Identification of fault estimation filter from I/O data for systems with stable inversion. *IEEE Transactions on Automatic Control*, 57:1347–1361, 2012.

[13] J. Dong, M. Verhaegen, and F. Gustafsson. Robust fault detection with statistical uncertainty in identified parameters. *IEEE Transactions on Signal Processing*, 60:5064–5076, 2012.

[14] J. Dong, M. Verhaegen, and F. Gustafsson. Robust fault isolation with statistical uncertainty in identified parameters. *IEEE Transactions on Signal Processing*, 60:5556–5561, 2012.

[15] S. Gillijns. *Kalman Filtering Techniques for System Inversion and Data Assimilation.* PhD thesis, Katholieke University Leuven, 2007.

[16] F. Gustafsson. *Adaptive Filtering and Change Detection.* Wiley, New York, 2001.

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

[18] T. Kailath, A. Sayed, and B. Hassibi. *Linear Estimation.* Prentice-Hall, Englewood Cliffs, NJ, 2000.

[19] T. Katayama. *Subspace Methods for System Identification.* Springer-Verlag, London, 2005.

[20] S. Kirtikar, H. Palanthandalam-Madapusi, E. Zattoni, and D. S. Bernstein. *l*-delay input and initial-state reconstruction for discrete-time linear systems. *Circuits, Systems, and Signal Processing*, 30:233–262, 2011.

[21] J. Lofberg. YALMIP: a toolbox for modeling and optimization in matlab. In *Proc. 2004 IEEE International Symposium on Computer Aided Control Systems Design*, pages 284–289, 2004.

[22] S. Manuja, S. Narasimhan, and S. C. Patwardhan. Unknown input modeling and robust fault diagnosis using black box observers. *Journal of Process Control*, 19:25–37, 2009.

[23] J. L. Massey and M. K. Sain. Inverses of linear sequential circuits. *IEEE Transactions on Automatic Control*, 17:330–337, 1968.

[24] S. H. Park, P. S. Kim, O. Kwon, and W. H. Kwon. Estimation and detection of unknown inputs using optimal FIR filter. *Automatica*, 36:1481–1488, 2000.

[25] S. C. Patwardhan and S. L. Shah. From data to diagnosis and control using generalized orthonormal basis filters. Part I: development of state observers. *Journal of Process Control*, 15:819–835, 2005.

[26] S. J. Qin. Data-driven fault detection and diagnosis for complex industrial processes. In *Proceedings of IFAC Safeprocess Symposium*, pages 1115–1125, 2009.

[27] S. J. Qin and W. Li. Detection and identification of faulty sensors in dynamic processes. *AIChE Journal*, 47:1581–1593, 2001.

[28] D. M. Raimondo, R. D. Braatz, and J. K. Scott. Active fault diagnosis using moving horizon input design. In *Proc. European Control Conference*, pages 3131–3136, Zurich, Switzerland, 2013.

[29] D. M. Raimondo, G. R. Marseglia, R. D. Braatz, and J. K. Scott. Fault-tolerant model predictive control with active fault isolation. In *Proc. 2nd International Conference on Control and Fault-Tolerant Systems*, pages 444–449, Nice, France, 2013.

[30] E. L. Russel, L. Chiang, and R. D. Braatz. *Data-Driven Techniques for Fault Detection and Diagnosis in Chemical Processes.* Springer-Verlag, London, 2000.

[31] M. Simandl, I. Puncochar, and J. Kralovec. Rolling horizon for active fault detection. In *Proc. 44th IEEE Conference on Decision and Control / European Control Conference*, pages 3789–3794, Seville, Spain, 2005.

[32] S. Simani, S. Fantuzzi, and R. Patton. *Model-Based Fault Diagnosis in Dynamic Systems Using Identification Techniques.* Springer-Verlag, London, 2003.

[33] G. van der Veen, J. W. van Wingerden, M. Bergamasco, M. Lovera, and M. Verhaegen. Closed-loop subspace identification methods: an overview. *IET Control Theory and Applications*, 7:1339–1358, 2012.

[34] Y. Wan, W. Dong, H. Wu, and H. Ye. Integrated fault detection system design for linear discrete time-varying systems with bounded power disturbances. *International Journal of Robust and Nonlinear Control*, 23:1781–1802, 2013.

[35] Y. Wan, T. Keviczky, and M. Verhaegen. Moving horizon least-squares input estimation for linear discrete-time stochastic systems. In *Proc. IFAC World Congress*, pages 3483–3488, Cape Town, South Africa, 2014.

[36] Z. Zhang and I. M. Jaimoukha. On-line fault detection and isolation for linear discrete-time uncertain systems. *Automatica*, 50:513–518, 2014.

[37] K. Zhou, J. Doyle, and K. Glover. *Robust and Optimal Control.* Prentice Hall, Upper Saddle River, New Jersey, 1996.

A Lemmas for Theorem 1

Lemma 1 *Define $x_e(0) \in \mathbb{R}^n$, $f_e(i) \in \mathbb{R}^{n_f}$, and $r_e(i) \in \mathbb{R}^{n_y}$ $(i \geq 0)$ as the initial state, input and output signal of the fault subsystem $(\Phi, \tilde{E}, C, G)$, respectively. There exists a non-zero initial state $x_e(0)$ such that $r_e(0) = r_e(1) = \cdots = r_e(L) = 0$ for all $L \geq \nu + \tau$, if and only if*

(i) $\mathcal{O}_\tau x_e(0) = 0$;
(ii) the system

$$\begin{cases} x_e(k+1) = \underbrace{\left[\Phi - \tilde{E}\left(H_\tau^f\right)^- C\Phi^\tau\right]}_{K_d} x_e(k) \\ r_e(k) = \left[I - H_\tau^f\left(H_\tau^f\right)^-\right]C\Phi^\tau x_e(k) \end{cases} \quad (A.1)$$

is unobservable;

(iii) the inputs $\{f_e(i)\}$ take the form

$$f_e(i) = - \left(H_\tau^f\right)^- C\Phi^\tau K_d^i x_e(0). \qquad \text{(A.2)}$$

In Lemma 1, $r_e(0) = \cdots = r_e(\tau - 1) = 0$ is ensured because of the condition (i) and the zero Markov matrices $H_0^f, H_1^f, \cdots, H_{\tau-1}^f$ according to Assumption 2, while $r_e(\tau) = \cdots = r_e(L) = 0$ is ensured by the conditions (ii) and (iii). Lemma 1 can be proved by slightly modifying Lemmas A.1 and A.2 in [20].

From Lemma 1 we can see that perfect reconstruction of system inputs $\{f_e(i)\}$ from system outputs $\{r_e(i)\}$ is impossible if the unobservable input signal (A.2) is non-zero. Hence, next, we will investigate the link between the unobservable input signal (A.2) and the system property of $(\Phi, \tilde{E}, C, G)$.

By setting $i = 0$, (A.2) becomes

$$f_e(0) = - \left(H_\tau^f\right)^- C\Phi^\tau x_e(0). \qquad \text{(A.3)}$$

Then, according to the condition (i) and the unobservability of the system (A.1), there must exist a scalar λ and a non-zero $x_e(0)$ such that [37]

$$\begin{bmatrix} K_d - \lambda I \\ \mathcal{O}_\tau \\ \left[I - H_\tau^f\left(H_\tau^f\right)^-\right] C\Phi^\tau \end{bmatrix} x_e(0) = \begin{bmatrix} \Phi - \lambda I & \tilde{E} \\ \mathcal{O}_\tau & 0 \\ C\Phi^\tau & H_\tau^f \end{bmatrix} \begin{bmatrix} x_e(0) \\ f_e(0) \end{bmatrix}$$
$$= \begin{bmatrix} \Phi - \lambda I & \tilde{E} \\ \mathcal{O}_{\tau+1} & \mathbf{H}_\tau^f \end{bmatrix} \begin{bmatrix} x_e(0) \\ f_e(0) \end{bmatrix} = 0, \qquad \text{(A.4)}$$

where $\mathbf{H}_\tau^f$ defined in (24) equals to $\begin{bmatrix} 0 \\ H_\tau^f \end{bmatrix}$ because $H_0^f, H_1^f, \cdots, H_{\tau-1}^f$ are zero matrices according to Assumption 2. With (A.3) and $(K_d - \lambda I)x_e(0) = 0$ in (A.4), we can rewrite $f_e(i)$ in (A.2) as

$$f_e(i) = \lambda^i f_e(0). \qquad \text{(A.5)}$$

The above analysis indicates that the unobservable inputs $\{f_e(i) = \lambda^i f_e(0)\}$ are determined by the invariant zero λ of $(\Phi, \tilde{E}, \mathcal{O}_{\tau+1}, \mathbf{H}_\tau^f)$, as shown in the following lemma:

Lemma 2 *Considering the non-zero initial state $x_e(0)$ in Lemma 1, there are two types of the invariant zeros λ of the fault subsystem $(\Phi, \tilde{E}, \mathcal{O}_{\tau+1}, \mathbf{H}_\tau^f)$ in (A.4): 1) λ is an unobservable mode, then (A.4) implies $f_e(0) = 0$, thus the input signal $\{f_e(i) = \lambda^i f_e(0)\}$ is constantly zero; 2) λ is a transmission zero, then $f_e(0) \neq 0$, thus the unobservable input signal $\{f_e(i) = \lambda^i f_e(0)\}$ is non-zero.*

Lemma 2 directly extends Lemmas 1 and 2 in [35] which considers only the case $\tau = 0$.

B Proof of Theorem 1

A solution $\hat{\mathbf{f}}_{k-\tau,L-\tau}^x$ to the problem (20) satisfies

$$\Psi_{L,\tau}^{\mathrm{T}} \Sigma_{e,L}^{-1} \Psi_{L,\tau} \hat{\mathbf{f}}_{k-\tau,L-\tau}^x = \Psi_{L,\tau}^{\mathrm{T}} \Sigma_{e,L}^{-1} \mathbf{r}_{k,L}. \qquad \text{(B.1)}$$

Let $\Delta\mathbf{f}_{k-\tau,L-\tau}^x = \hat{\mathbf{f}}_{k-\tau,L-\tau}^x - \mathbf{f}_{k-\tau,L-\tau}^x$ denote the estimation error. By substituting (19) into (B.1), we have

$$\Psi_{L,\tau}^{\mathrm{T}} \Sigma_{e,L}^{-1} \Psi_{L,\tau} \Delta\mathbf{f}_{k-\tau,L-\tau}^x = \Psi_L^{\mathrm{T}} \Sigma_{e,L}^{-1} \mathbf{e}_{k,L},$$

which implies $\Psi_{L,\tau}^{\mathrm{T}} \Sigma_{e,L}^{-1} \Psi_{L,\tau} \mathrm{E}\left(\Delta\mathbf{f}_{k-\tau,L-\tau}^x\right) = 0$ by taking expectations on both sides. Therefore, the unbiasedness condition of the estimate in (23) reduces to the analysis of the linear equation

$$\Psi_{L,\tau} \mathrm{E}\left(\Delta\mathbf{f}_{k-\tau,L-\tau}^x\right) = 0 \qquad \text{(B.2)}$$

since $\mathcal{N}\left(\Psi_{L,\tau}^{\mathrm{T}} \Sigma_{e,L}^{-1} \Psi_{L,\tau}\right) = \mathcal{N}\left(\Psi_{L,\tau}\right)$.

The rest of the proof follows the intuitive arguments below. According to Lemma 1, (A.5), and the definition of $\mathbf{f}_{k-\tau,L-\tau}^x$ in (19), there are three scenarios:

1) When $(\Phi, \tilde{E}, \mathcal{O}_{\tau+1}, \mathbf{H}_\tau^f)$ has no invariant zeros, the non-zero initial state $x_e(0)$ in Lemma 1 does not exist according to (A.4), thus (B.2) implies $\mathrm{E}\left(\Delta\mathbf{f}_{k-\tau,L-\tau}^x\right) = 0$, i.e., unbiased fault estimation.

2) When $(\Phi, \tilde{E}, \mathcal{O}_{\tau+1}, \mathbf{H}_\tau^f)$ has invariant zeros, (B.2) implies that for each invariant zero λ, the expected error of the τ-delay fault estimate $\hat{f}(k - \tau)$ is

$$\mathrm{E}\left(\Delta f(k - \tau)\right) = \lambda^{L-\tau-1} \mathrm{E}\left(\Delta f(k_0)\right) \qquad \text{(B.3)}$$

in the estimation horizon $[k_0, k]$ ($k_0 = k - L + 1$).

 2.1) If all the invariant zeros of $(\Phi, \tilde{E}, \mathcal{O}_{\tau+1}, \mathbf{H}_\tau^f)$ correspond to unobservable modes, it follows from the case 1) in Lemma 2 that the expected estimation error (B.3) is zero because $\mathrm{E}\left(\Delta f(k_0)\right) = 0$.

 2.2) If transmission zeros exist but are all stable, i.e., $|\lambda| < 1$, it follows from the case 2) in Lemma 2 that $\mathrm{E}\left(\Delta f(k_0)\right) \neq 0$ and the expected estimation error (B.3) asymptotically reduced to zero as L goes to infinity.

The scenarios 1) and 2.1) correspond to the case (i) of Theorem 1, and the scenario 2.2) corresponds to the case (ii) of Theorem 1.

C Proof of Theorem 2

For the original system model (2), the extended output equation in the time window $[k_0, k]$ is

$$\mathbf{y}_{k,L} = \mathcal{O}_L x(k_0) + \mathscr{T}_L^u \mathbf{u}_{k,L} + \mathscr{T}_L^f \mathbf{f}_{k,L} + \mathscr{T}_L^w \mathbf{w}_{k,L} + \mathbf{v}_{k,L},$$
$$\tag{C.1}$$

where $\mathcal{O}_L$, $\mathscr{T}_L^u$, $\mathscr{T}_L^f$, and $\mathscr{T}_L^w$ are defined in the same way as $\mathcal{O}_L$ and $\mathbf{T}_L^u$ in (16). According to (C.1), we can rewrite (17) and (18) as

$$\begin{aligned}
\mathbf{r}_{k,L} &= (I - \mathbf{T}_L^y)(\mathbf{y}_{k,L} - \mathscr{T}_L^u \mathbf{u}_{k,L}) \\
&= (I - \mathbf{T}_L^y)\left(\mathcal{O}_L x(k_0) + \mathscr{T}_L^f \mathbf{f}_{k,L} + \mathscr{T}_L^w \mathbf{w}_{k,L} + \mathbf{v}_{k,L}\right) \\
&= (I - \mathbf{T}_L^y)\underbrace{\left[\mathcal{O}_L \ \mathscr{T}_{L,\tau}^f\right]}_{\check{\Psi}_{L,\tau}} \mathbf{f}_{k-\tau,L-\tau}^x + \mathbf{e}_{k,L}.
\end{aligned}$$
$$\tag{C.2}$$

by following the relation between the original system model (2) and its predictor form (3). Similarly to $\mathbf{T}_{L,\tau}^f$ in (19), $\mathscr{T}_{L,\tau}^f$ in (C.2) consists of the first $L - \tau$ block-columns of $\mathscr{T}_L^f$.

Define $\check{\mathbf{r}}_{k,L} = \mathbf{y}_{k,L} - \mathscr{T}_L^u \mathbf{u}_{k,L}$ and

$$\check{\Sigma}_L = \text{cov}\left(\mathscr{T}_L^w \mathbf{w}_{k,L} + \mathbf{v}_{k,L}\right).$$

Comparing (19) with (C.2) leads to

$$\begin{aligned}
\mathbf{r}_{k,L} &= (I - \mathbf{T}_L^y)\check{\mathbf{r}}_{k,L}, \quad \Psi_{L,\tau} = (I - \mathbf{T}_L^y)\check{\Psi}_{L,\tau}, \\
\Sigma_{e,L} &= (I - \mathbf{T}_L^y)\check{\Sigma}_L (I - \mathbf{T}_L^y)^{\text{T}}.
\end{aligned}$$
$$\tag{C.3}$$

Then by substituting (C.3) into (22), the estimate of $\mathbf{f}_{k-\tau,L-\tau}^x$ becomes

$$\hat{\mathbf{f}}_{k-\tau,L-\tau}^x = \left(\check{\Psi}_{L,\tau}^{\text{T}} \check{\Sigma}_L^{-1} \check{\Psi}_{L,\tau}\right)^{(1)} \check{\Psi}_{L,\tau}^{\text{T}} \check{\Sigma}_L^{-1} \check{\mathbf{r}}_{k,L}, \tag{C.4}$$

which is actually the LS estimate proposed in [35] based on the original system model (2).

D Proof of Theorem 3

Split $\mathbf{T}_{L,\tau}^f$ into two blocks as $\left[\check{\mathbf{T}}_{L,\tau}^f \ \tilde{\mathbf{T}}_{L,\tau}^f\right]$, with $\check{\mathbf{T}}_{L,\tau}^f$ consisting of the first $L - \tau - 1$ block-columns of $\mathbf{T}_{L,\tau}^f$, and $\tilde{\mathbf{T}}_{L,\tau}^f$ consisting of the last block-column of $\mathbf{T}_{L,\tau}^f$. With these notations, unbiased fault estimation can be proved by showing that $\tilde{\mathbf{T}}_{L,\tau}^f \text{E}(\Delta f(k-\tau)) = 0$ because $\tilde{\mathbf{T}}_{L,\tau}^f$ has full column rank according to Assumption 2.

According to (26), the following two expressions are equivalent:

$$\varepsilon \in \mathcal{R}\left(\left[\mathcal{O}_L \ \check{\mathbf{T}}_{L,\tau}^f\right]\right) \bigcap \mathcal{R}\left(\tilde{\mathbf{T}}_{L,\tau}^f\right), \tag{D.1}$$

$$\varepsilon \in \mathcal{R}\left(\left[\mathbf{H}_{L,m}^o \ \check{\mathbf{T}}_{L,\tau}^f\right]\right) \bigcap \mathcal{R}\left(\tilde{\mathbf{T}}_{L,\tau}^f\right). \tag{D.2}$$

Since the two sufficient conditions for (asymptotic) unbiasedness in Theorem 1 imply $\varepsilon = 0$ and $\varepsilon \to 0$ ($L \to \infty$) for (D.1), it then follows from the equivalence between (D.1) and (D.2) that the sufficient conditions in Theorem 1 also imply $\varepsilon = 0$ and $\varepsilon \to 0$ ($L \to \infty$) for (D.2), or equivalently, $\mathcal{R}\left(\tilde{\mathbf{T}}_{L,\tau}^f\right) = \{0\}$ and $\mathcal{R}\left(\tilde{\mathbf{T}}_{L,\tau}^f\right) \to \{0\}$ ($L \to \infty$). Therefore we can conclude that the sufficient conditions in Theorem 1 imply (asymptotically) unbiased fault estimation for (D.2). Similarly, we can prove the necessary condition for the (asymptotically) unbiased fault estimation.

E Computation of $\bar{\mathbb{E}}\left(\mathcal{T}_s(\mathcal{G})\mathcal{T}_s^{\text{T}}(\mathcal{G})\right)$

By dividing $\bar{\mathbf{M}}_\Upsilon$ in (40) into L row blocks as

$$\bar{\mathbf{M}}_\Upsilon = \left[\mathbf{M}_{\Upsilon,1}^{\text{T}} \ \mathbf{M}_{\Upsilon,2}^{\text{T}} \ \cdots \ \mathbf{M}_{\Upsilon,L}^{\text{T}}\right]^{\text{T}}, \tag{E.1}$$

with $\mathbf{M}_{\Upsilon,i} \in \mathbb{R}^{n_f \times (m \cdot n_u + (L-\tau)n_f)}$, we define $\mathbf{P}_\Upsilon$ as

$$\mathbf{P}_\Upsilon = \begin{bmatrix} \text{tr}(\mathbf{M}_{\Upsilon,1}\mathbf{M}_{\Upsilon,1}^{\text{T}}) & \text{tr}(\mathbf{M}_{\Upsilon,1}\mathbf{M}_{\Upsilon,2}^{\text{T}}) & \cdots & \text{tr}(\mathbf{M}_{\Upsilon,1}\mathbf{M}_{\Upsilon,L}^{\text{T}}) \\ \text{tr}(\mathbf{M}_{\Upsilon,2}\mathbf{M}_{\Upsilon,1}^{\text{T}}) & \text{tr}(\mathbf{M}_{\Upsilon,2}\mathbf{M}_{\Upsilon,2}^{\text{T}}) & \cdots & \text{tr}(\mathbf{M}_{\Upsilon,2}\mathbf{M}_{\Upsilon,L}^{\text{T}}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{tr}(\mathbf{M}_{\Upsilon,L}\mathbf{M}_{\Upsilon,1}^{\text{T}}) & \text{tr}(\mathbf{M}_{\Upsilon,L}\mathbf{M}_{\Upsilon,2}^{\text{T}}) & \cdots & \text{tr}(\mathbf{M}_{\Upsilon,L}\mathbf{M}_{\Upsilon,L}^{\text{T}}) \end{bmatrix}.$$
$$\tag{E.2}$$

$\mathbf{P}_z$ is defined similarly to (E.2), by dividing $\bar{\mathbf{M}}_L^z$ in (41) into L row blocks as in (E.1). Then,

$$\bar{\mathbb{E}}\left(\mathcal{T}_f(\mathcal{G})\mathcal{T}_f^{\text{T}}(\mathcal{G})\right) = \left[\mathcal{G} \ \mathcal{I}_{n_f}\right] \begin{bmatrix} \Pi_f & -\hat{\Upsilon}_{L,\tau} \\ -\hat{\Upsilon}_{L,\tau}^{\text{T}} & I_{n_f} \end{bmatrix} \begin{bmatrix} \mathcal{G}^{\text{T}} \\ \mathcal{I}_{n_f}^{\text{T}} \end{bmatrix},$$
$$\tag{E.3}$$

$$\bar{\mathbb{E}}\left(\mathcal{T}_z(\mathcal{G})\mathcal{T}_z^{\text{T}}(\mathcal{G})\right) = \mathcal{G}\Pi_z\mathcal{G}^{\text{T}} \tag{E.4}$$

with

$$\begin{aligned}
\Pi_f &= \hat{\Upsilon}_{L,\tau}\hat{\Upsilon}_{L,\tau}^{\text{T}} + \bar{\mathbb{E}}\left(\bar{\mathbf{E}}_{\text{id}}\bar{\mathbf{M}}_\Upsilon \bar{\mathbf{M}}_\Upsilon^{\text{T}}\bar{\mathbf{E}}_{\text{id}}^{\text{T}}\right) \\
&= \hat{\Upsilon}_{L,\tau}\hat{\Upsilon}_{L,\tau}^{\text{T}} + \mathbf{P}_\Upsilon \otimes \Sigma_e,
\end{aligned} \tag{E.5}$$

$$\Pi_z = \bar{\mathbb{E}}\left(\bar{\mathbf{E}}_{\text{id}}\bar{\mathbf{M}}_L^z(\bar{\mathbf{M}}_L^z)^{\text{T}}\bar{\mathbf{E}}_{\text{id}}^{\text{T}}\right) = \mathbf{P}_z \otimes \Sigma_e. \tag{E.6}$$