A Kernel-Based Approach to Data-Driven Actuator Fault Estimation[**]

Mohammad Amin Sheikhi[*] Peyman Mohajerin Esfahani
Tamás Keviczky

Delft Center for Systems and Control, Delft University of Technology,
Mekelweg 2, 2628 CD Delft, The Netherlands (e-mail:
m.a.sheikhi@tudelft.nl, p.mohajerinesfahani@tudelft.nl,
t.keviczky@tudelft.nl)

Abstract: This paper considers the problem of fault estimation in linear time-invariant systems when actuators are subject to unknown additive faults. A data-driven approach is proposed to design an inverse-system-based filter for reconstructing fault signals when the underlying fault subsystem can be either a minimum phase or non-minimum phase system. Unlike traditional two-step data-driven methods in the literature, the proposed method directly computes the filter parameters from input-output data to avoid the propagation of identification errors through an inverse operation into the fault estimates, which is the case in state-of-the-art filter designs. Furthermore, regarding out-of-sample performance of the filter, a kernel-based regularization is exploited to not only reduce the model complexity but also enable the design scheme to take advantage of available prior knowledge on the underlying system behavior. This knowledge can be incorporated into basis functions, promoting the desired solution to the optimization problem. To validate the effectiveness of the proposed method, a simulation study is conducted, demonstrating a notable reduction in estimation error compared to state-of-the-art methods.

Keywords: Fault estimation, Data-driven, Non-minimum phase systems, Kernel-based regularization.

1. INTRODUCTION

During the past decades, fault diagnosis techniques on detection and isolation tasks have been extensively studied (Hafezi et al., 2022); however, fault estimation (FE) has been less investigated, which is the main focus of this paper. On the one hand, FE is a more challenging task, requiring the diagnosis system to detect, isolate, and determine the size and shape of faults. On the other hand, the outcomes of FE would provide more informative insights for control objectives, particularly in fault-tolerant design and predictive maintenance. Model-based residual generation approaches, such as unknown input observers (Chen et al., 1996; Ghanipoor et al., 2023) and nullspace-based filters (Zhong et al., 2010; Mohajerin Esfahani and Lygeros, 2015) under certain conditions can track fault signals. However, in practice, an explicit and accurate model of the real system often is not available, which promotes the adoption of data-driven methods.

As far as fault estimation problem is concerned, it is inherently tied to inverting the underlying fault subsystem, for which the model is unavailable in data-driven applications. As a result, inversion-based filters have recently garnered attention (Wan et al., 2016; Naderi and Khorasani, 2019). The principal obstacle in this class of problems is the presence of transmission zeros, causing the system to become input-state unobservable (Kirtikar et al., 2011). This signifies that exact reconstruction is impossible within a finite-time if the underlying system possesses any transmission zero (Ansari and Bernstein, 2019; Palanthandalam-Madapusi and Bernstein, 2009). In Dong and Verhaegen (2011), a subspace-based data-driven fault estimation filter is developed based on predictor Markov parameters (MPs) (Wan et al., 2016, 2017). The main scheme is initially identifying the MPs and then proceeding with an inversion step to perform FE. These approaches are referred to as *indirect methods* in this paper. Stable system-inversion is the fundamental requirement in these methods to ensure asymptotically unbiased estimation. Most of the recent developments in data-driven fault estimation methods share the restrictive minimum phase condition of the fault subsystem; otherwise, the reconstruction error grows exponentially. Hence, achieving stable inversion of non-minimum phase (NMP) systems remains a significant challenge across various contexts, especially FE problems. In this regard, few model-based methods have been proposed in recent literature, although their applications in real scenarios are limited. In Marro and Zattoni (2010), a geometric approach is proposed in a noise-free condition, where the system matrices are assumed to be known. The model-based approach in Naderi and Khorasani (2019) is only unbiased for either step or ramp fault signals. A more general model-based solution is provided in Ansari and Bernstein (2018) by introducing a retrospective cost function for unknown input recon-

[*] Corresponding author.

[**]This work was supported by Digital Twin project with project number P18-03 of the research programme TTW-Perspective which is (partly) financed by the Dutch Research Council (NWO).

struction. Inspired by preview-based techniques in control tracking problems of NMP systems, two methods were developed recently for fault estimation in model-based (Naderi and Khorasani, 2018) and data-driven (Yu and Verhaegen, 2018) fashions. The major issues with these methods include the high sensitivity of the reconstruction error to the correct location of zeros, especially in a data-driven framework due to identification errors, as well as significant estimation delay in case the invariant zeros are close to the unit circle.

This paper addresses the fault estimation (FE) problem in non-minimum phase (NMP) systems within a data-driven framework, allowing the recovery of solutions for minimum phase systems as well. In contrast to indirect methods, a bilateral finite impulse response (FIR) filter is proposed for the fault subsystem inverse, directly parameterized by corresponding Markov parameters (MPs) of the inverse system, and is referred to as a *direct method*. This implies that identification errors bypass the inversion operation, improving the estimation performance. Moreover, the proposed scheme allows for using kernel-based regularization to shape the filter characteristics at the design stage.

This paper is organized as follows: The problem is formulated in Section 2. Section 3 is devoted to the main result of the research. Simulation studies to verify the proposed method are presented in Section 4, and the conclusions are drawn in Section 5.

Notation. Throughout this paper, $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$ denote the set of integer, real, and complex numbers, respectively. In the complex plane, $\mathbb{D}$ represents the interior of the unit circle, and $\mathbb{T}$ is the unit circle. $\overline{z}$ stands for the complex conjugate, and the space $\ell_2(\mathbb{Z})$ characterizes sequences as $\sum_{k\in\mathbb{Z}}|x(k)|^2 < \infty$, where $\mathcal{R}\ell_2(\mathbb{Z})$ denotes real sequences. $\mathcal{L}_2(\mathbb{T})$ consists of all complex-valued functions defined and square-integrable on the unit circle, with $\mathcal{RL}_2$ referring to real-rational functions. The inner product of two vectors in an inner-product space $\mathcal{X}$ is shown by $\langle .,.\rangle_{\mathcal{X}}$. The symbol "$\otimes$" represents the Kronecker product, and vec(.) is the vectorization operator. "$*$" denotes convolution operation.

2. PROBLEM STATEMENT

Consider the following general linear time-invariant (LTI) system

$$\mathcal{M}: \begin{aligned} x(k+1) &= Ax(k) + B_u u(k) + B_f f(k) + B_w w(k)\,,\\ y(k) &= Cx(k) + D_u u(k) + D_f f(k) + v(k)\,, \end{aligned} \qquad (1)$$

in which $x(k) \in \mathbb{R}^n$, $y(k) \in \mathbb{R}^{n_y}$, $u(k) \in \mathbb{R}^{n_u}$, and $f(k) \in \mathbb{R}^{n_f}$ denote the state, the output measurement, the control input, and the unknown fault signal at discrete-time instant k, respectively. The disturbances are represented by the process noise $w(k) \in \mathbb{R}^{n_w}$ and the measurement noise $v(k) \in \mathbb{R}^{n_v}$, both of which are considered to be white noises with zero-mean, without loss of generality, and the corresponding bounded covariance matrix $\begin{bmatrix} w(k)\\ v(k)\end{bmatrix} \sim \begin{bmatrix} Q & S\\ S^T & R\end{bmatrix}$. It is further assumed that $w(k)$ and $v(k)$ are uncorrelated with $u(k)$ and $f(k)$. $(A, B_u, B_f, B_w, C, D_u, D_f)$ are minimal state-space realization matrices with compatible dimensions. Under detectability assumption of pair (A, C) and controllable $(A, Q^{1/2})$, the system in (1) admits the following stable one-step ahead prediction form as (Kailath et al., 2000)

$$\begin{aligned} \hat{x}(k+1) &= \tilde{A}\hat{x}(k) + \tilde{B}_u u(k) + \tilde{B}_f f(k) + Ky(k)\\ y(k) &= C\hat{x}(k) + D_u u(k) + D_f f(k) + e(k) \end{aligned} \qquad (2)$$

with $\tilde{A} = A - KC$, $\tilde{B}_u = B_u - KD_u$, $\tilde{B}_f = B_f - KD_f$. $K \in \mathbb{R}^{n\times n_y}$ represents the steady-state Kalman gain, and $e(k) \in \mathbb{R}^{n_y}$ is the zero-mean innovation process with covariance matrix Σ_e.

Definition 2.1. For the system $(A,\, B,\, C,\, D)$ with $u(k) \in \mathbb{R}^{n_u}$, $x(k) \in \mathbb{R}^n$, $y(k) \in \mathbb{R}^{n_y}$, $z_0 \in \mathbb{C}$ is called a *transmission zero* if the associated Rosenbrock system matrix loses rank, i.e.,

$$\operatorname{rank}\left(\mathbf{R}(z_0) \triangleq \begin{bmatrix} A - z_0 I & B\\ C & D\end{bmatrix}\right) < n + \min\{n_y,\, n_u\}\,.$$

Remark 2.1. The transmission zeros of the original system (1) are equivalent to its observer form (2). In fact, the system zeros are not being modified by the output feedback, which can be verified by investigating the relation between both system matrices and Definition 2.1.

$$\begin{bmatrix} \tilde{A} - zI & [\tilde{B}_u \quad \tilde{B}_f]\\ C & [D_u \quad D_f]\end{bmatrix} = \begin{bmatrix} I & -K\\ 0 & I\end{bmatrix}\begin{bmatrix} A - zI & [B_u \quad B_f]\\ C & [D_u \quad D_f]\end{bmatrix}$$

The observer form is linked to the original system through a full rank matrix that does not affect the rank of Rosenbrock matrix.

To proceed with the problem statement, it is required to introduce the notion of *left-invertibility*, which is provided in Definition 2.2.

Definition 2.2. (Left-invertibility) The system $\mathcal{M}$ defined by an $n_y \times n_u$ proper transfer function $H(z)$ is *τ-delay left invertible* if there exists an $n_u \times n_y$ proper transfer function $H_\tau^{\text{inv}}(z)$ such that $H_\tau^{\text{inv}}(z)H(z) = z^{-\tau}I_{n_u}$ for almost all $z \in \mathbb{C}$ and nonnegative integer τ. Lastly, the smallest nonnegative integer for which $H(z)$ is left invertible is the *relative degree* of the system.

Theorem 1. (Kirtikar et al., 2011) The system $\mathcal{M}$ defined by an $n_y \times n_u$ proper transfer function $H(z) = C(zI_n - A)^{-1}B + D$ is a left invertible system if $\exists z \in \mathbb{C}$ such that the following equivalent statements hold

- rank $H(z) = n_u$ (full column rank).
- rank $\begin{bmatrix} A - zI & B\\ C & D\end{bmatrix} = n + n_u$ (full column rank).

Proof. See Theorem 1 in Kirtikar et al. (2011).

According to Theorem 1, $n_y \geq n_u$ is always a necessary condition for system $\mathcal{M}$ to be a left invertible system, i.e., the system must have at least as many outputs as inputs. As a matter of fact, this condition usually holds in most practical systems, and it is not restrictive.

Assumption 2.1. (System detectability). The fault matrices B_f and D_f satisfy the rank condition

$$\max\,\{\operatorname{rank}\,[B_f^T \quad D_f^T]\} = n_f\,,$$

which is a necessary condition for the fault *detectability.*

The fault subsystem (A, B_f, C, D_f) satisfies Assumption 2.1, and is τ-delay left invertible, where τ represents the corresponding relative degree. Furthermore, fault estimation in the presence of transmission zeros (main

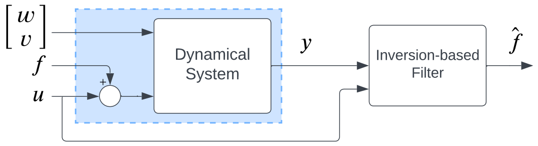

Fig. 1. The actuator fault estimation structure

focus of this paper) is a significant challenge in actuator fault scenarios. Motivated by this, we shall consider the problematic case of actuator fault throughout the rest of paper, that is,

$$j\text{-th actuator fault: } B_f = B_u^{(:,j)}\,, \quad D_f = D_u^{(:,j)}$$

where $X^{(:,j)}$ denotes the j-th column of the matrix X. For the sake of brevity in subsequent analysis, we further assume that the fault signals affect all the input component/channels, i.e., $n_f = n_u$, while recovering the general case is straightforward.

As faulty datasets are often not available in practice, the following assumption is required to be made on the collected data.

Assumption 2.2. (Dataset regularity) Fault-free input-output (I/O) data obtained from the system is available for the proposed data-driven design. Moreover, the input signal $u(k)$ is persistently exciting of sufficient order.

3. MAIN RESULT

In this section, an inverse-system-based filter (ISF) is developed for the data-driven fault estimation problem. Figure 1 illustrates the cascade configuration for which the proposed filter is designed, underscoring the requirement for a left-invertible fault subsystem.

Suppose linear maps $\mathcal{H} : \mathbb{R}^{n_u} \mapsto \mathbb{R}^{n_y}$ and $\mathcal{H}^{-1} : \mathbb{R}^{n_y} \mapsto \mathbb{R}^{n_u}$ represent the nominal system and its left inverse, respectively. In the absence of noises and disturbances, the estimation performance $f - \hat{f}$ can be quantified as

$$f - \hat{f} = (\mathrm{I} - \mathcal{H}^{-1} \circ \mathcal{H})(u + f)\,. \tag{3}$$

Based on (3), perfect fault estimation is achieved if $\mathrm{I} - \mathcal{H}^{-1} \circ \mathcal{H} = 0$. This condition depends on the accuracy of the system inversion to the true inverse of the system. The challenge is minimizing this difference when both the actual system parameters and the fault signal are unknown.

3.1 Model

Regarding the data-driven design, the problem should be parameterized, where the Markov parameters are the well-known choice to describe the linear dynamical systems. To do so, the prediction form (2) can be consistently represented by a VARX (Vector AutoRegressive with eXogenous inputs) model of sufficiently large order p (Chiuso, 2007):

$$\mathcal{A}(q^{-1})y(k) = \mathcal{B}(q^{-1})(u(k) + f(k)) + \varepsilon(k)\,, \tag{4}$$

where q^{-1} is the backward shift operator, $\mathcal{A}(q^{-1}) = \mathbb{I}_{n_y} - \sum_{i=1}^{p} M_i^y\, q^{-i}$, $\mathcal{B}(q^{-1}) = \sum_{i=0}^{p} M_i^u\, q^{-i}$, and $\varepsilon(k) \in \mathbb{R}^{n_y}$ represents noise signals as a sequence of independent and identically distributed (*i.i.d.*) multivariate random variable. M_i^u and M_i^y are associated predictor Markov Parameters (MPs) defined as follows

$$M_i^u = \begin{cases} D & i = 0 \\ C\tilde{A}^{i-1}\tilde{B}_u & i > 0 \end{cases}, \quad M_i^y = \begin{cases} 0 & i = 0 \\ C\tilde{A}^{i-1}K & i > 0 \end{cases}.$$

To develop the direct approach, we rearrange the model in (4) w.r.t. to the input signal in a fault-free condition

$$\mathcal{B}(q^{-1})u(k) = \mathcal{A}(q^{-1})y(k) + \varepsilon'(k) \tag{5}$$

with $\varepsilon'(k) = -\varepsilon(k)$ as another *i.i.d.* sequence. This implies that the inverse of a VARX model is another VARX model. There is, however, a major difference. Unlike the model in (4), the stability of the inverse model relies on the characteristics of $\mathcal{B}(q^{-1})$, while simultaneously dictating the autoregressive behavior of the stochastic noise term. For systems with NMP transmission zeros, the inverse VARX model results in unstable dynamics that has to be taken care of in the proposed scheme.

3.2 Inverse-system-based filter

We aim at providing a stable ISF to be used for fault estimation in the presence of NMP zeros. To this end, stable inversion techniques are developed in which the central idea involves treating the unstable internal dynamics as stable noncausal operators in order to produce bounded responses (George et al., 1999). The noncausality viewpoint implies that the trajectory is known ahead of time over a preview horizon. Since the proposed filter parameters are derived from only the fault-free I/O data, this information is available to the design scheme and is not a limitation. In turn, the preview window would naturally cause the estimation delay that needs to be minimized. In order to take advantage of preview-based methods, the following assumption has to be made.

Assumption 3.1. (Transmission zero)
The system(A, B, C, D) has no transmission zeros on the unit circle, i.e.,

$$\operatorname{rank} \begin{bmatrix} A - zI & B \\ C & D \end{bmatrix} = n + n_u\,, \quad \forall z \in \mathbb{C}\,,\ |z| = 1\,.$$

The inverse of any linear system that satisfies Assumption 3.1 can be described using the following two-sided representation based on the Laurent Series:

$$\mathcal{H}^{-1}: \quad \mathcal{H}^{\text{inv}}(q, \mathbf{H}) = \sum_{i=-\infty}^{\infty} H_i\, q^{-i}\,; \quad i \in \mathbb{Z}\,, \tag{6}$$

where $\mathcal{H}^{\text{inv}}(q, \mathbf{H})$ represents a bilateral impulse response expansion of the nominal left inverse system $\mathcal{H}^{-1}$. In this representation, $H_i \in \mathbb{R}^{n_u \times n_y}$ is a matrix of expansion coefficients in the impulse response sequence $\mathbf{H} = \{H_i\}_{i=-\infty}^{\infty}$, and the time shift operator q allows the manipulation of multi-dimensional signals either forward or backward in time.

Given the model (6), the input-output relation in a noise-free condition can be written as

$$u(k) = (\mathbf{H} * y)(k) = \sum_{i=-\infty}^{\infty} H_i\, y(k-i)\,; \quad i \in \mathbb{Z}\,, \tag{7}$$

where the input signal at time k is a convolution of the output signal with the bilateral Markov parameter sequence

$\mathbf{H}$. Therefore, negative time-delays $i \in \{\dots, -1, 0, 1, \dots\}$ are acceptable, making the model noncausal and dependent on both past and future input data at each instant. As a result, the fault estimation filter parameterized by $\mathbf{H}$ can be proposed as

$$\text{ISF filter: } f(k) = \sum_{i=-\infty}^{\infty} H_i u^{\text{ISF}}(k-i) - u(k)\,; \quad i \in \mathbb{Z}\,, \tag{8}$$

with u^{ISF} as the input of the filter. Hence, the filter design boils down to estimating the sequence $\mathbf{H}$ from healthy data. Let the set of fault-free measurement data be denoted by $\mathcal{D}_{\text{healthy}} = \{u(k), y(k)\}_{k=1}^{N}$. To estimate the parameters of bilateral model (6) in terms of mean-squared error (MSE) criterion, the following optimization problem can be considered

$$\hat{\mathbf{H}} = \underset{\{H_i\}_{i=-\infty}^{\infty} \in \mathbb{R}^{n_u \times n_y}}{\arg\min} \frac{1}{N} \sum_{k=1}^{N} \left\| u(k) - \sum_{i=-\infty}^{\infty} H_i\, y(k-i) \right\|_2^2$$

This problem can be reformulated as

$$\hat{\theta} = \underset{\{\theta_i\}_{i=-\infty}^{\infty} \in \mathbb{R}^{n_u n_y}}{\arg\min} \frac{1}{N} \sum_{k=1}^{N} \left\| u(k) - \sum_{i=-\infty}^{\infty} (y^T(k-i) \otimes \mathbb{I}_{n_u})\, \theta_i \right\|_2^2 \tag{9}$$

with $\theta_i = \text{vec}\,(H_i)$. Examining (9) shows that an infinite dimensional optimization problem should be solved to retrieve the parameters of model (6). This is because the VARX model (5) is described by a two-sided *infinite impulse response* (IIR) model, wherein the set of corresponding Markov parameters has no compact support. Consequently, achieving the exact inverse of the system from a finite number of noisy measurements is an ill-posed problem, leading to high estimation variance. To address ill-posedness and overfitting, regularization should be introduced into the regression problem. Kernel-based regularization is a technique commonly employed in the machine learning theory to robustify the prediction model performance against unseen data. Moreover, prior knowledge on the system behavior can be incorporated into the decision variables through this framework (Pillonetto et al., 2014). This approach can capture aspects including stability, smoothness, noncausality, and model complexity.

3.3 Kernel-based regularization

In order to introduce regularization, we limit the hypothesis space to a class of reproducing kernel Hilbert space (RKHS) $\mathscr{H}_K$ that can be uniquely characterized by a positive semi-definite *kernel* function K. In this regards, we penalize the feasible solutions that do not align with the prior knowledge by including a regularization term associated with the induced norm on $\mathscr{H}_K$ in (9)

$$\hat{\theta}^{\text{reg}} = \underset{\theta \in \mathscr{H}_K}{\arg\min} \frac{1}{N} \sum_{k=1}^{N} \left\| u(k) - \sum_{i=-\infty}^{\infty} (y^T(k-i) \otimes \mathbb{I}_{n_u})\, \theta_i \right\|_2^2 + \mu \left\|\theta\right\|_{\mathscr{H}_K}^2 . \tag{10}$$

in which $\theta = \begin{bmatrix} \theta_{-\infty}^T & \cdots & \theta_{\infty}^T \end{bmatrix}^T$ is the column stack of θ_i, and the positive scalar $\mu > 0$ is the regularization parameter to create a bias/variance balance. An immediate result of adopting an RKHS function space is that the infinite-dimensional problem (10) has a finite-dimensional solution according to the representer theorem (Pillonetto et al., 2014). In the proposed optimization problem (10), kernels also accounts for noncausal terms in the impulse response expansion (7) for general NMP systems. Non-causal kernels are recently introduced in Blanken and Oomen (2020), where negative indices are allowed for kernels as well. Henceforth, we define the (i, j)-th element of the *kernel* matrix $[\mathbf{K}]_{ij}$ corresponds to the real-valued kernel function K as $K(t_i, t_j; \alpha) : \mathbb{Z} \times \mathbb{Z} \mapsto \mathbb{R}$. The kernel matrix $\mathbf{K}(\alpha)$ is a positive semi-definite matrix ($\mathbf{K}(\alpha) \succeq 0$) parameterized in terms of α, a set of hyper-parameters that depends on the type of kernels in use.

Rational orthonormal basis functions (ROBF) kernels have been demonstrated to be effective in describing a wide range of linear dynamical systems, as well as estimating impulse responses (Van den Hof et al., 2000). Since the proposed fault estimation filter is formulated on the basis of bilateral impulse responses, ROBF-based kernels are considered for the ISF. These basis functions are essentially a network of interconnected transfer functions in a cascading fashion. Each basis function has an IIR sequence that enables the design scheme to effectively capture slow dynamics with a relatively small number of bases. These advantages motivate the use of ROBF recently presented in Blanken and Oomen (2020), which is an extension of the so-called Takenaka-Malmquist basis functions for noncausal systems. Let $G(z) \in \mathcal{RL}_2^{n_u \times n_y}(\mathbb{T})$ denotes the left inverse system. Under Assumption 3.1, $G(z)$ can be decomposed with a complete set of scalar orthonormal basis functions for $\mathcal{RL}_2(\mathbb{T})$ as $G(z) = \sum_{i=-\infty}^{\infty} h_i \psi_i(z)$, where $h_i \in \mathbb{R}^{n_u \times n_y}$ is a sequence of matrices whose (m, n)-th entry $h_i^{[m,n]}$ corresponds to the m-th input and n-th output channel $G^{[m,n]}(z)$ as $h_i^{[m,n]} = \langle \psi_i, G^{[m,n]} \rangle_{\mathcal{L}_2}$. In this regards, we define noncausal ROBF kernels as

$$\psi_i(z) = \begin{cases} \psi_i^{\text{c}}(z) = \dfrac{\sqrt{1 - |\lambda_{\text{c},\,i}|^2}}{z - \lambda_{\text{c},\,i}} \displaystyle\prod_{j=1}^{i-1} \dfrac{1 - \overline{\lambda_{\text{c},\,j}}\, z}{z - \lambda_{\text{c},\,j}} & i > 0, \\ \psi_i^{\text{ac}}(z) = \dfrac{\sqrt{1 - |\lambda_{\text{ac},\,i}|^2}}{1 - \overline{\lambda_{\text{ac},\,i}}\, z} \displaystyle\prod_{j=-1}^{i+1} \dfrac{z - \lambda_{\text{ac},\,j}}{1 - \overline{\lambda_{\text{ac},\,j}}\, z} & i < 0, \end{cases}$$

in which $\{\lambda_{\text{c},\,i}\}_{i \in \mathbb{Z}_+}, \{\lambda_{\text{ac},\,i}\}_{i \in \mathbb{Z}_-} \subset \mathbb{D}$ are sequences of generating poles that play the role of hyper-parameters in the kernel design, i.e., $\alpha = \{\lambda_{\text{c}}, \lambda_{\text{ac}}\}$. To guarantee real-valued impulse responses, the set of poles should appear as either real numbers or pairs of complex conjugate numbers. For the sake of completeness of basis, the conditions I)$\sum_{i=1}^{\infty}(1 - |\lambda_{\text{c},\,i}|) = \infty$, II)$\sum_{i=-\infty}^{-1}(1 - |\lambda_{\text{ac},\,i}|) = \infty$ have to be satisfied. Direct feedthrough can be included in basis functions by setting $\psi_0^{\text{ac}} = 1$, and $\psi_0^{\text{ac}} = 0$ otherwise. The proposed noncausal kernels are orthonormal with respect to the unit circle in the sense that $\frac{1}{2\pi} \oint_{\mathbb{T}} \psi_i(e^{j\omega})\, \overline{\psi_j(e^{j\omega})}\, \text{d}\omega = \delta_{i,\,j}$. $\{\psi_i^{\text{c}}\}_{i>0}$ and $\{\psi_i^{\text{ac}}\}_{i \leq 0}$ form causal and anti-causal basis functions, respectively. As we aim at estimating impulse responses of the inverse system, the corresponding basis functions in time-domain can be computed by taking the inverse z-transform $\mathrm{Z}^{-1}\{.\}$ of the ROBFs with respect to the valid region of convergence (ROC). Therefore, $\phi_i(k) = \mathrm{Z}^{-1}\{\psi_i(z)\}$ denotes the associated basis function in the time-domain. Since $\{\psi_i\}_{i=-\infty}^{\infty}$ is a complete basis of Hilbert space $\mathcal{RL}_2$, a similar argument applies to the time-domain basis functions $\{\phi_i\}_{i=-\infty}^{\infty}$ to verify that they are complete in $\mathcal{R}\ell_2(\mathbb{Z})$

and orthonormal: $\sum_{k=-\infty}^{\infty}\phi_i(k)\,\overline{\phi_j(k)}=\delta_{i,j}$. Accordingly, the bilateral impulse response $g(k)\in\mathcal{R}\ell_2(\mathbb{Z})$ associated with $G(z)$ can be equivalently expressed by $g(k)=\sum_{i=-\infty}^{\infty}h_i\phi_i(k)$. Hence, the following kernel function is built

$$K_{\mathrm{OBF}}(t_i,t_j;\{\lambda_{\mathrm{c}},\lambda_{\mathrm{ac}}\})=\sum_{i=-\infty}^{\infty}\phi_i(t_i)\,\overline{\phi_i(t_j)} \tag{11}$$

It can be shown that it is a reproducing kernel for the RKHS space spanned by $\{\phi_i\}_{i=-\infty}^{\infty}$. In practice, a finite number of basis functions is used for the estimation while accepting a certain level of error, which can be made arbitrarily small by increasing the model order. To ensure a specified level of approximation error, the required preview time is determined by the closest NMP zero to the unit circle. Denote by n_{ac} and n_{c} the number of anti-causal and causal terms in (6), respectively. Then define the stacked vector of data over N measurements $\mathcal{D}_{\mathrm{healthy}}$ as

$$U_N=\left[u^T(n_c+1)\ u^T(n_c+2)\ \cdots\ u^T(N-n_{ac})\right]^T,$$

and the block-Toeplitz structure of output data

$$\mathcal{T}_N^y=\begin{bmatrix} y^T(n_c+n_{ac}+1) & y^T(n_c+n_{ac}) & \cdots & y^T(1)\\ y^T(n_c+n_{ac}+2) & y^T(n_c+n_{ac}+1) & \cdots & y^T(2)\\ \vdots & \vdots & \ddots & \vdots\\ y^T(N) & y^T(N-1) & \cdots & y^T(N-n_c-n_{ac})\end{bmatrix},$$

$$X=\mathcal{T}_N^y\otimes\mathbb{I}_{n_u}.$$

Consider $\theta=\mathrm{vec}([H_{-n_{\mathrm{ac}}}\cdots H_0\cdots H_{n_{\mathrm{c}}})])\in\mathbb{R}^{n_\theta}$ with $n_\theta=n_yn_u(n_{ac}+n_c+1)$. Using (10) and the kernel matrix $\mathbf{K}(\alpha)\in\mathbb{R}^{n_\theta\times n_\theta}$ associated with (11), the model parameters are the solution of the following optimization problem

$$\begin{aligned}\hat{\theta}^{\mathrm{reg}}&=\underset{\theta\in\mathbb{R}^{n_\theta}}{\arg\min}\ \frac{1}{N}\,\|U_N-X\,\theta\|_2^2+\frac{\mu}{N}\,\theta^T\mathbf{K}(\alpha)^{-1}\theta\,,\\ &=(\mathbf{K}(\alpha)X^TX+\mu\,\mathbb{I}_{n_\theta})^{-1}\mathbf{K}(\alpha)X^TU_N\,.\end{aligned} \tag{12}$$

In case the kernel matrix is rank deficient, the optimization problem can be modified based on the singular value decomposition of kernel matrix, as suggested in Pillonetto et al. (2014)(Remark 1). The poles of rational OBF $\{\lambda_{\mathrm{c}},\lambda_{\mathrm{ac}}\}$ can be designed to approximate the true dynamics of the inverse system $\mathcal{H}^{-1}$. Since the inverse fault subsystem is estimated by solving the optimization problem (12), the fault estimator of $f(k)$ is given without performing any further operation as

$$\hat{f}(k)=\sum_{i=-n_{\mathrm{ac}}}^{n_{\mathrm{c}}}\hat{H}_i^{\theta}\,u^{\mathrm{ISF}}(k-i)-u(k)\,. \tag{13}$$

4. SIMULATION RESULTS

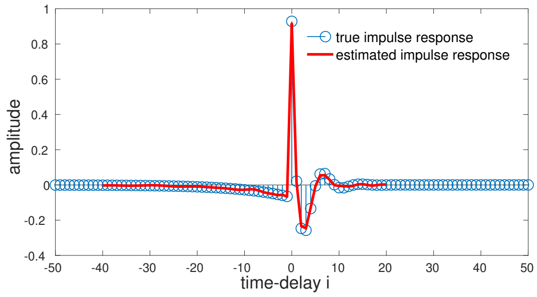

Fig. 2. The bilateral impulse responses estimation.

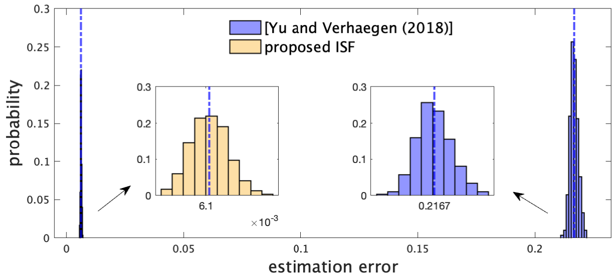

Fig. 3. Out-of-sample histogram of fault estimation error, where the I/O data has not been used during training.

To show the effectiveness of the proposed method compared to the state-of-the-art method (Yu and Verhaegen, 2018), a Monte Carlo experiment is provided. Consider a linear dynamical system with the following minimal realization of state-space representation $A=[2,1,0;-1.23,0,1;0.216,0,0]$, $B=[-0.1,0.370,-0.334]^T$, $C=[1,0,0]$, $D=1$. This system possesses an NMP zero at $z=1.1$ while the rest of zeros are inside the unit circle ($z=0.5\pm0.5i$). The fault matrix is $B_f=B_u$, and $B_w=\mathbb{I}_3$ for the process noise. The process and measurement noises are zero mean white noises generated from Gaussian distributions as $w(k)\sim\mathcal{N}(0,\,0.01\mathbb{I}_3)$ and $v(k)\sim\mathcal{N}(0,\,0.01)$, respectively. In addition, the fault signal in these experiments is following the profile as $f(k)=\{0\ \texttt{if}\ k\le 0,\ \sin(0.1\pi k)\ \texttt{if}\ 50<k\le 200,\ 2\ \texttt{if}\ k>200\}$. To design the fault estimation filter, a data set of healthy I/O samples with $N=1000$ is collected. The Signal-to-Noise Ratio (SNR) is kept at the high level of $\mathrm{SNR}=20\,\mathrm{dB}$. A bilateral impulse response model for the inverse system will be estimated using (10). The proposed model orders are chosen as $n_{\mathrm{c}}=20$ and $n_{\mathrm{ac}}=40$, and the OBF hyper parameters are designed according to $\{\lambda_{\mathrm{c},\,i}\}=\texttt{linspace}(0.4,0.6,n_{\mathrm{c}})$ and $\{\lambda_{\mathrm{ac},\,i}\}=\texttt{linspace}(0.8,0.9,n_{\mathrm{ac}})$ together with a nonzero direct feedthrough. The true two-sided expansion of the inverse system and its estimation is provided in Figure 2. A VARX model of order $3L$ with $L=50$ is identified for the system with VAF(variance-accounted-for)$=\%97.25$ using the receding horizon filter proposed in Yu and Verhaegen (2018). The filter performance is evaluated based on the normalized error $e=\|f-\hat{f}\|^2/\|f\|^2$. The Monte Carlo simulation involves 300 runs with *i.i.d.* realizations for process and measurement noises. In each run, the FE filter is performing on real-time operating data resulting from a multi-step input signal, and the comparison has been made in Figure 3. Despite a perfect VARX identification, the method in Yu and Verhaegen (2018) exhibits a clear bias in the estimation, whereas the proposed ISF filter leads to an almost unbiased estimation. This result resonates with the fact that nonlinear propagation of identification errors via an inversion operation can introduce a noticeable bias in indirect approaches, which is addressed in the proposed method by directly designing the filter parameters. The estimation bias due to truncation can be controlled through regularization trade-off. Figure 4 presents the fault reconstruction performance for one specific noise realization. In this simulation, we also demonstrate that receding horizon filter (Yu and Verhaegen, 2018) based on the true MPs is unbiased. In addition, the proposed filter offers two more advantages over the recent literature. First, the ISF filter order $(=60)$ is much lower than the receding horizon

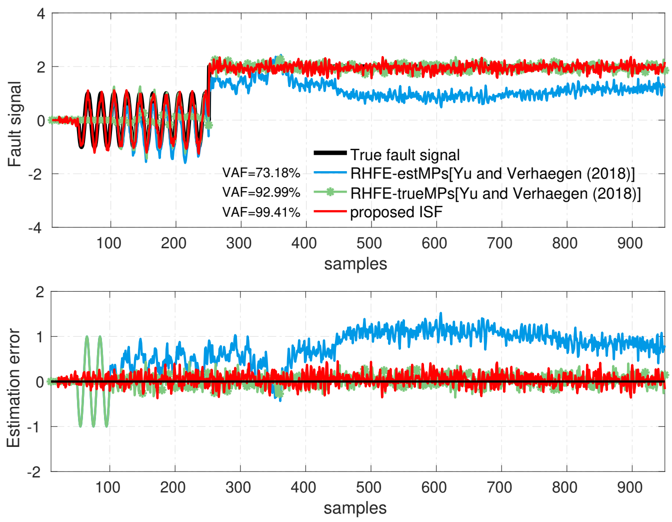

Fig. 4. The fault estimation performance.

filter (= 151), implying less estimation delay. Second, unlike indirect approaches, the proposed method does not require any inversion operation, making it more favorable for online applications.

5. CONCLUSION

This paper presents a data-driven fault estimation method for linear systems, including those with unstable transmission zeros. The inverse system is represented by bilateral impulse responses, leading to a mixed causal and non-causal model. Unlike existing inversion-based approaches, this method directly estimates the inverse fault subsystem from input-output data, preventing the propagation of identification errors through the inversion operation. Moreover, it employs kernel-based regularization with ROBF kernels to reduce filter order and enhance performance on unseen data. Simulation studies confirm the superior performance of this filter.

REFERENCES

Ansari, A. and Bernstein, D.S. (2018). Input estimation for nonminimum-phase systems with application to acceleration estimation for a maneuvering vehicle. *IEEE Transactions on Control Systems Technology*, 27(4), 1596–1607.

Ansari, A. and Bernstein, D.S. (2019). Deadbeat unknown-input state estimation and input reconstruction for linear discrete-time systems. *Automatica*, 103, 11–19.

Blanken, L. and Oomen, T. (2020). Kernel-based identification of non-causal systems with application to inverse model control. *Automatica*, 114, 108830.

Chen, J., Patton, R.J., and Zhang, H.Y. (1996). Design of unknown input observers and robust fault detection filters. *International Journal of control*, 63(1), 85–105.

Chiuso, A. (2007). The role of vector autoregressive modeling in predictor-based subspace identification. *Automatica*, 43(6), 1034–1048.

Dong, J. and Verhaegen, M. (2011). Identification of fault estimation filter from I/O data for systems with stable inversion. *IEEE Transactions on Automatic Control*, 57(6), 1347–1361.

George, K., Verhaegen, M., and Scherpen, J.M. (1999). Stable inversion of mimo linear discrete time nonminimum phase systems. In *7th Mediterranean Conference on Control and Automation (MED99) Haifa, Israel*, 267–281.

Ghanipoor, F., Murguia, C., Mohajerin Esfahani, P., and van de Wouw, N. (2023). Robust fault estimators for nonlinear systems: An ultra-local model design. *arXiv preprint arXiv:2305.14036*.

Hafezi, H., Bakhtiari, A., and Khaki-Sedigh, A. (2022). Design and implementation of a fault-tolerant controller using control allocation techniques in the presence of actuators saturation for a vtol octorotor. *Robotica*, 40(9), 3057–3076.

Kailath, T., Sayed, A.H., and Hassibi, B. (2000). *Linear estimation*. Prentice Hall.

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

Marro, G. and Zattoni, E. (2010). Unknown-state, unknown-input reconstruction in discrete-time nonminimum-phase systems: Geometric methods. *Automatica*, 46(5), 815–822.

Mohajerin Esfahani, P. and Lygeros, J. (2015). A tractable fault detection and isolation approach for nonlinear systems with probabilistic performance. *IEEE Transactions on Automatic Control*, 61(3), 633–647.

Naderi, E. and Khorasani, K. (2018). Inversion-based output tracking and unknown input reconstruction of square discrete-time linear systems. *Automatica*, 95, 44–53.

Naderi, E. and Khorasani, K. (2019). Unbiased inversion-based fault estimation of systems with non-minimum phase fault-to-output dynamics. *IET Control Theory & Applications*, 13(11), 1629–1638.

Palanthandalam-Madapusi, H.J. and Bernstein, D.S. (2009). A subspace algorithm for simultaneous identification and input reconstruction. *International Journal of Adaptive Control and Signal Processing*, 23(12), 1053–1069.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3), 657–682.

Van den Hof, P., Wahlberg, B., Heuberger, P., Ninness, B., Bokor, J., and e Silva, T.O. (2000). Modelling and identification with rational orthogonal basis functions. *IFAC Proceedings Volumes*, 33(15), 445–455.

Wan, Y., Keviczky, T., and Verhaegen, M. (2017). Fault estimation filter design with guaranteed stability using markov parameters. *IEEE Transactions on Automatic Control*, 63(4), 1132–1139.

Wan, Y., Keviczky, T., Verhaegen, M., and Gustafsson, F. (2016). Data-driven robust receding horizon fault estimation. *Automatica*, 71, 210–221.

Yu, C. and Verhaegen, M. (2018). Data-driven fault estimation of non-minimum phase lti systems. *Automatica*, 92, 181–187.

Zhong, M., Ding, S.X., Han, Q.L., and Ding, Q. (2010). Parity space-based fault estimation for linear discrete time-varying systems. *IEEE Transactions on Automatic Control*, 55(7), 1726–1731.