



Research article

Robust adaptive masked and filtered hierarchical multi-innovation stochastic gradient identification for multivariable ARX systems with missing outputs and common colored non-Gaussian disturbances

Slim Dhahri^{1,*}, Mourad Elloumi^{2,3}, Hend Aljahani⁴, Salem Albalawi⁵, Sahar Almashaan^{5,*}, Hatem Alwardi⁵ and Foued Mtiri⁶

¹ Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka 72341, Al-Jouf, Saudi Arabia

² Faculty of Sciences of Gafsa, University of Gafsa, Gafsa, Tunisia

³ Laboratory of Sciences and Technology of Automatic Control and Computer Engineering, National School of Engineering of Sfax, Sfax University, P.O. Box 1173, Sfax 3038, Tunisia

⁴ Department of Mathematics, College of Science, Northern Border University, Arar 91431, Saudi Arabia

⁵ Department of Mathematics, College of Science, Jouf University, Sakaka, Saudi Arabia

⁶ Department of Mathematics, College of Science, King Khalid University, Abha 61413, Saudi Arabia

* **Correspondence:** Email: sdhahri@ju.edu.sa, sralmesaan@ju.edu.sa.

Abstract: Subsystem-by-subsystem recursive identification is very inefficient for an r -input m -output multivariable autoregressive with exogenous input (ARX) system since it performs multiple operations due to the common input-information vector used by all output channels. Methods that take advantage of this coupled structure such as hierarchical stochastic-gradient and hierarchical multi-innovation stochastic-gradient methods cut down on computations. Unfortunately, most of the formulations are based on full output measurements and white noise of the measurement, and this restricts their use in networked and industrial acquisition systems. In this paper, a powerful multivariable ARX system with missing output samples, corrupted by a typical first-order colored disturbance and a set of impulsive non-Gaussian innovations, was dealt with using a new stochastic-gradient hierarchical multi-innovation adaptive algorithm (RAMF-HMISG). A masked coupled identification model was initially proposed including one-step reconstructed outputs and binary output-availability variables to ensure that the autoregressive regressors are still available, while the missing samples were not included in the identification loss. A commonly used auxiliary prewhitening filter was then applied to the output, the autoregressive regressors, as well as the shared input regressor, leading to preservation of the deterministic part of the ARX model, while factually separating the filtering/reconstruction

mismatch. The correction term was based on Huber loss and the bounded score to make it more robust. A filtered reliability mask was also added as this takes into account current and previous output measurements for each filtered output sample. Based on the previous innovation energy, which was smoothed, non-circular adaptive laws were also proposed for the innovation length $p(t)$ and convergence index $\varepsilon(t)$. It was shown that a residual relation and a conservative mean-square boundedness result can be made under the conditions of bounded regressors, block excitation, stable filtering, bounded reconstruction mismatch, and stochastic-approximation step-size conditions. The analysis ensured the stability to a noise/mismatch dependent neighborhood for nonzero disturbances. Numerical simulations, ablation tests, and Monte Carlo statistics demonstrated that the proposed RAMF-HMISG algorithm outperforms other variants of the algorithm, such as zero-filling, masked-only, robust-only, and fixed filtered ones, in terms of smaller parameter-estimation errors under the simultaneous colored noise, missing outputs, and impulsive perturbations. The evaluation further covered a higher-order three-input three-output system, missing-output probabilities up to 50%, several noise levels, coloring coefficients and impulsive intensities, a component-wise ablation of the two adaptive laws, a scalability study of the adaptive mechanisms, and a numerical verification of the excitation and mismatch conditions used in the analysis. The applicable boundary of the common-filter assumption was also quantified for channel-dependent coloring.

Keywords: hierarchical identification; multi-innovation stochastic gradient; multivariable ARX model; missing outputs; Huber loss; colored noise; robust recursive identification

Mathematics Subject Classification: 62F35, 93C35, 93E10, 93E12

1. Introduction

System identification, which is a key tool in prediction, monitoring, control, and adaptive decision-making, is concerned with developing mathematical models of dynamical systems based on measured input–output data. The time-domain, frequency-domain, recursive, and stochastic-approximation methods are based on classical identification theory [1–3]. Recursive identification is very appealing for online applications as the parameter estimates are computed sequentially as new data is received [4–6]. Additionally, some related adaptive-filtering and stochastic-approximation frameworks facilitate the analysis of recursive gradient-type algorithms [7–9]. Multivariable systems are more difficult to identify as several outputs may be related to several inputs and to one another's autoregressive past. A subsystem-by-subsystem approach is straightforward, and can duplicate calculations if common information vectors are used for multiple output channels. In order to utilize data better, multi-innovation identification is performed with a finite window of the latest innovations and hierarchical identification is used to decompose large parameter-estimation problems into smaller coupled subproblems, see [10–12]. Recently, they have been extended to hierarchical recursive least-squares, hierarchical iterative, and auxiliary-model-based stochastic-gradient algorithms for multiple-input multiple-output systems [13–15]. Some recent works include the hierarchical stochastic-gradient algorithm (HSG) and the hierarchical multi-innovation stochastic-gradient algorithm (HMI-SG) for the multivariable ARX model, that take advantage of the commonality of the input-information vector for all output subsystems [16]. Partially coupled ARX systems, multivariable autoregressive

moving average (ARMA)-like systems, as well as auxiliary-model maximum-likelihood approaches for multiple-input multiple-output (MIMO) systems have been studied in related works [17–19]. The three-stage maximum-likelihood recursive least-squares (RLS) algorithm, two-stage filtering-based stochastic-gradient method, and multi-innovation forgetting-gradient algorithm for the multivariable system were also developed [20–22]. It is also worth noting that iterative and moving-data-window generalized extended gradient methods for multivariable autoregressive output-error and ARMA-type systems have been developed, based on least squares [23, 24].

But, coupled multivariable ARX identification algorithms still exist that are derived from ideal measurement assumptions. In practical networked, embedded, and industrial sensing systems, output samples can be missing due to packet loss, sensor saturation, communication failures, or pre-processing failures. Methods for missing-data estimation have been extensively investigated using expectation-maximization (EM) methods, as well as statistical missing-data theory and time-series smoothing methods [25–27]. In the case of dynamic systems, handling missing outputs is particularly challenging as it impacts not only the current innovation but also the autoregressive regressors in the future, which has led to studies on ARX identification with missing outputs and Kalman-filtering based gradient methods [28–30].

One restriction is that the disturbances are assumed to be white Gaussian. In actual acquisition systems, the sensor and communication dynamics can generate colored noise, which could lead to wrong stochastic-gradient estimates. The idea of filtering identification and an auxiliary model for colored-noise systems was demonstrated to be an effective tool [31–33]. The same notions of filtering and hierarchy have also been applied in nonlinear time-series identification and in engineering applications like battery-model parameter estimation [34–36]. Auxiliary filters designed to remove the correlation introduced by unmeasured dynamics are equally central outside identification: filter-based predefined-time optimal fault-tolerant consensus schemes for nonlinear multi-agent systems use a dedicated filter inside a reinforcement-learning dynamic-surface design in order to avoid the repeated differentiation of virtual control laws and to preserve the prescribed convergence time under actuator faults [37]. The role played by the filter there, namely, to isolate a nuisance dynamic term so that the remaining recursion keeps its nominal structure and its convergence certificate, is exactly the role played by the prewhitening filter in the present work, where it isolates the coloring mechanism so that the coupled multivariable ARX structure and the common input-information regressor survive filtering.

Real measurements can also include impulsive (or non-Gaussian) disturbances. As Huber-type losses, influence-function-based estimation, etc., have been developed to reduce the impact of abnormal innovations, robust statistics are a relevant field. Another important approach for robust non-Gaussian signal processing and nonlinear-system estimation is based on correntropy and mixed-kernel criteria [38–40]. In recent years, there has been a strong interest in recursive and distributed estimation methods that are also robust to noisy, heavy-tailed, and censored data (see, e.g., [41–43]).

1.1. Robust and missing-data recursive identification: State-of-the-art methods and the remaining gap

Because the algorithm proposed below borrows from three research lines that are usually developed separately, it is useful to summarize what each line already achieves and, more importantly, where each one stops.

(i) *Correntropy and information-theoretic criteria.* Replacing the quadratic cost by the maximum-

correntropy criterion suppresses heavy-tailed and impulsive innovations through a Gaussian kernel whose bandwidth controls the effective rejection level [38]. Generalized correntropy extends the kernel family and provides a tunable compromise between efficiency under Gaussian noise and resistance to outliers [44], and the maximum-correntropy Kalman filter transfers the idea to state estimation [45]. Related heterogeneous-kernel and Gaussian–Versoria criteria have been developed for non-Gaussian signal processing and nonlinear-system estimation [39, 40]. These criteria are, however, formulated for completely observed regression data and for white disturbances; the kernel weight is applied to an innovation that is assumed to be available at every instant, and no mechanism is provided for a regressor whose own entries are missing.

(ii) *M-estimation and robust recursive least squares.* The least mean M-estimate and recursive least M-estimate algorithms replace the squared error by a bounded-score loss and provide fast implementations together with convergence analyses in impulsive environments [46, 47]; the underlying statistical machinery is classical robust estimation [48–50]. This family is designed for a single-output adaptive filter. Applying it channel by channel to an r -input m -output ARX system destroys precisely the structural property exploited here, namely, that the input-information vector $\varphi(t)$ is shared by all output channels, and it reintroduces the m -fold duplication of computations that hierarchical identification is meant to remove.

(iii) *Expectation-maximization and likelihood-based missing-data recovery.* The EM algorithm is the reference tool for incomplete data [25, 26] and has been specialized to dynamic systems for irregularly sampled chemical processes [51], for robust maximum-likelihood estimation of multivariable models [52], for nonlinear state-space models [53], and, with a Student’s t measurement model, for ARX systems subject to outliers [54]. Smoothing- and Kalman-based treatments of missing outputs in ARX identification follow the same philosophy [27–30]. These methods are statistically efficient, but they are iterative and batch-oriented, they require a smoothing pass over the whole record at every iteration, and their per-iteration cost is quadratic in the parameter dimension. They are therefore not directly usable as an online, per-sample recursion with a fixed memory budget, which is the operating regime targeted in this paper.

The gap that follows from this survey is the following. Correntropy and M-estimation handle impulsive noise but not missing regressor entries or colored disturbances. Filtering-based hierarchical schemes handle colored disturbances but assume complete measurements and quadratic innovations. EM-type methods handle missing data but are not recursive and do not exploit the coupled multivariable structure. No existing recursive multivariable ARX identification algorithm addresses randomly missing outputs, a common colored disturbance, and impulsive non-Gaussian innovations simultaneously while retaining the shared input-information vector, a per-sample cost linear in the innovation length, and a convergence certificate. Closing this gap is the purpose of the algorithm developed below.

Based on these observations, this paper proposes a “masked and filtered” adaptive hierarchical multi-innovation stochastic-gradient algorithm with strong adaptability for a multivariable ARX system with missing output, common first-order colored disturbance, and impulsive non-Gaussian innovation. The proposed approach expands in four directions the coupled HMISG approach. First, binary output-availability masks and one-step reconstructed outputs are introduced, to keep autoregressive regressors well-defined while at the same time avoiding the inclusion of missing samples in the identification loss. Second, a common auxiliary prewhitening filter is used at the output

of the model, the autoregressive regressors and the shared input regressor, while maintaining the coupled multivariable ARX model. Third, the powerful correction is based on the precise Huber loss and its bounded score. Fourth, non-circular adaptive laws for the innovation length and for the convergence index are proposed, which balance transient convergence, steady-state accuracy, and computational burden.

The key contributions are:

- A model for modeling multivariable ARX with randomly missing output samples is proposed, in which the reconstructed outputs are only used to assure regressor continuity.
- A common first-order filtered auxiliary model is developed for the systems with common first-order colored disturbances and reconstruction/filtering mismatch is explicitly isolated.
- A Huber-score-based robust masked HMISG algorithm is designed using hierarchical identification, multi-innovation theory, filtered regressors, and missing-output reliability masks.
- An adaptive rule for the innovation length $p(t)$ and convergence index $\varepsilon(t)$ is given based on the past smoothed innovation energy, which does not involve circular tuning.
- The boundedness of the mean-square error is proved under the assumptions of bounded regressors, block excitation, stable filtering, bounded mismatch, and stochastic-approximation step-size conditions.
- The proposed method is further shown to be effective through simulation, ablation, and Monte Carlo studies, when compared to the zero-filling, masked-only, robust-only, and fixed filtered approaches.

The key mathematical contributions of the suggested RAMF-HMISG method compared to the HSG/HMISG method are summarized in Table 1.

Table 1. Methodological extensions of RAMF-HMISG over the baseline HSG/HMISG framework.

Feature	HSG/HMISG [16]	RAMF-HMISG
Output data	Complete output measurements	Missing-output mask with reconstructed regressors
Noise model	White disturbance	Common colored disturbance with auxiliary prewhitening
Innovation correction	Raw quadratic innovation	Masked Huber-score innovation
Reliability weighting	Not included	Filtered reliability mask $\gamma_{f,i}(t) = \gamma_i(t)\gamma_i(t-1)$
Innovation structure	Single innovation or fixed multi-innovation length	Adaptive multi-innovation length $p(t)$
Convergence index	Fixed ε in modified variants	Adaptive convergence index $\varepsilon(t)$
Analysis	Residual relation and complexity comparison	Residual relation plus mean-square boundedness with a mismatch term
Numerical validation	Complete-output white-noise setting	Missing outputs, colored noise, impulsive noise, ablation, and Monte Carlo tests

1.2. What is structurally new and what is an integration of known tools?

It is important to state precisely which parts of RAMF-HMISG are new constructions and which parts are established components reused without modification, because the two carry very different scientific weight.

Three ingredients are used essentially as they stand and are not claimed as contributions: the Huber loss and its bounded score [48, 49], the multi-innovation idea of correcting on a finite window of past innovations [12, 55], and the hierarchical decomposition of the coupled multivariable ARX estimation problem [10, 16].

Four elements are structurally new. First, the masked coupled model of Section 2 separates two roles that are conflated in every existing missing-data gradient scheme: a reconstructed output $\tilde{y}_i(t)$ is admitted into the regressor so that the autoregressive chain never breaks, but it is refused entry into the loss. Zero-filling and plain imputation both let the surrogate value act as an observation and therefore inject a deterministic bias that no amount of data removes; the numerical evidence for this is the persistent error floor of M-HMISG-ZF reported in Section 6. Second, the filtered reliability mask $\gamma_{f,i}(t) = \gamma_i(t)\gamma_i(t-1)$ is not a heuristic weight but the exact indicator of the event that the filtered sample $y_{f,i}(t)$ is a function of genuine measurements only; it is what makes the prewhitening step compatible with random data loss, where Remark 2.2 quantifies the information it costs. Third, the common prewhitening filter is applied simultaneously to the output, the autoregressive regressors, and the shared input regressor. This is the only choice that preserves the common input-information matrix, and hence the hierarchical structure, after filtering; a channel-wise filter would destroy it, as discussed in Remark 2.1. Fourth, the adaptive laws for $p(t)$ and $\varepsilon(t)$ are driven by the smoothed robust innovation energy of the previous instant, which makes them non-circular and keeps them inside the stochastic-approximation hypotheses used in the convergence proof.

The necessity of the two adaptive laws is not cosmetic. The multi-innovation window length controls a bias–variance compromise: a long window averages more innovations and therefore reduces the gradient variance, but it also mixes innovations evaluated at increasingly stale parameter estimates, which biases the correction during the transient. A window that is optimal at the beginning of the run is therefore not optimal at the end, and vice versa; this is the reason a fixed p cannot be simultaneously fast and accurate, and it is confirmed by the fixed- p benchmarks of Section 6.5. The convergence index plays a complementary role. The step size $a_t = \mu r^{-\varepsilon(t)}(t)$ must satisfy the stochastic-approximation conditions $\sum_t a_t = \infty$ and $\sum_t a_t^2 < \infty$, which force $\varepsilon(t) \in (1/2, 1]$; inside that interval, a value close to the lower end gives large steps and fast transients while a value close to the upper end gives small steps and a small steady-state neighborhood. Fixing ε therefore forces the designer to choose, once and for all, between transient speed and asymptotic accuracy. Making ε a decreasing function of the residual innovation energy resolves the dilemma without leaving the admissible interval, so that the boundedness result of Theorem 4.1 continues to apply verbatim. Section 4.3 develops this argument quantitatively.

Table 2 places RAMF-HMISG with respect to the algorithm families that are closest to it, including the masked and filtered HMISG variants used later as ablation baselines and the auxiliary-model maximum-likelihood, correntropy, and EM-based approaches surveyed in Section 1.1.

Table 2. Positioning of RAMF-HMISG with respect to the most closely related identification frameworks. “Coupled MIMO” means that the input-information vector is shared by all output channels instead of being duplicated per channel.

Framework	Missing outputs	Colored noise	Impulsive noise	Coupled MIMO	Adaptive p, ε	Per-sample cost
HSG / HMISG [16, 55]	No	No	No	Yes	No	linear
Masked HMISG (this paper, ablation)	Yes	No	No	Yes	No	linear
RM-HMISG (this paper, ablation)	Yes	No	Yes	Yes	No	linear
RF-HMISG (this paper, ablation)	Yes	Yes	Yes	Yes	No	linear
Auxiliary-model ML / three-stage RLS [19, 20]	No	Yes	No	Partly	No	quadratic
Correntropy / M-estimate recursions [44, 47]	No	No	Yes	No	No	linear–quadratic
EM / likelihood missing-data methods [52, 54]	Yes	Yes	Yes	Partly	No	batch, quadratic
RAMF-HMISG (proposed)	Yes	Yes	Yes	Yes	Yes	linear

The rest of this paper is structured as follows. The multivariable ARX model, the missing-output reconstruction mechanism, and the common filtered model are introduced in Section 2. The robust masked filtered HSG algorithm is presented in Section 3. In Section 4, the proposed RAMF-HMISG algorithm and adaptive laws, residual relation, and mean-square boundedness analysis are developed. Section 5 discusses computational efficiency. Numerical experiments are reported in Section 6. Lastly, Section 7 concludes this paper.

2. System description and identification models

2.1. Multivariable ARX subsystem models

Consider the r -input m -output ARX system

$$A_i(z)y_i(t) = \sum_{j=1}^r B_{ij}(z)u_j(t) + v_i(t), \quad i = 1, 2, \dots, m, \quad (2.1)$$

where $u_j(t)$ is the j th input, $y_i(t)$ is the i th output, and $v_i(t)$ is a stochastic disturbance. The polynomials in the unit backward-shift operator z^{-1} are

$$A_i(z) = 1 + a_i(1)z^{-1} + a_i(2)z^{-2} + \dots + a_i(n_i)z^{-n_i}, \quad (2.2)$$

$$B_{ij}(z) = b_{ij}(1)z^{-1} + b_{ij}(2)z^{-2} + \dots + b_{ij}(n_j)z^{-n_j}. \quad (2.3)$$

Define

$$\psi_i(t) = [-y_i(t-1), -y_i(t-2), \dots, -y_i(t-n_i)]^T \in \mathbb{R}^{n_i}, \quad (2.4)$$

$$\varphi_j(t) = [u_j(t-1), u_j(t-2), \dots, u_j(t-n_j)]^T \in \mathbb{R}^{n_j}, \quad (2.5)$$

$$\varphi(t) = [\varphi_1^T(t), \varphi_2^T(t), \dots, \varphi_r^T(t)]^T \in \mathbb{R}^n, \quad n = \sum_{j=1}^r n_j. \quad (2.6)$$

The parameter vectors are

$$\alpha_i = [a_i(1), a_i(2), \dots, a_i(n_i)]^T \in \mathbb{R}^{n_i}, \quad (2.7)$$

$$\theta_{ij} = [b_{ij}(1), b_{ij}(2), \dots, b_{ij}(n_j)]^T \in \mathbb{R}^{n_j}, \quad (2.8)$$

$$\theta_i = [\theta_{i1}^T, \theta_{i2}^T, \dots, \theta_{ir}^T]^T \in \mathbb{R}^n. \quad (2.9)$$

Then the i th subsystem can be written as

$$y_i(t) = \psi_i^T(t)\alpha_i + \varphi^T(t)\theta_i + v_i(t). \quad (2.10)$$

Equivalently,

$$y_i(t) = \phi_i^T(t)\vartheta_i + v_i(t), \quad \phi_i(t) = \begin{bmatrix} \psi_i(t) \\ \varphi(t) \end{bmatrix}, \quad \vartheta_i = \begin{bmatrix} \alpha_i \\ \theta_i \end{bmatrix}. \quad (2.11)$$

The vector $\varphi(t)$ is common to every output channel. This is the structural property that motivates the coupled hierarchical update.

2.2. Coupled identification model

Let

$$\alpha = \text{col}\{\alpha_1, \alpha_2, \dots, \alpha_m\} \in \mathbb{R}^{n_0}, \quad n_0 = \sum_{i=1}^m n_i, \quad (2.12)$$

$$\Theta = [\theta_1, \theta_2, \dots, \theta_m] \in \mathbb{R}^{n \times m}, \quad (2.13)$$

and define

$$\eta(\alpha, t) = [\psi_1^T(t)\alpha_1, \psi_2^T(t)\alpha_2, \dots, \psi_m^T(t)\alpha_m]^T. \quad (2.14)$$

Stacking the m output equations gives the coupled model

$$y(t) = \eta(\alpha, t) + \Theta^T \varphi(t) + v(t), \quad (2.15)$$

where $y(t) = [y_1(t), \dots, y_m(t)]^T$ and $v(t) = [v_1(t), \dots, v_m(t)]^T$. The autoregressive part remains channel-dependent, while the input part is represented by the matrix Θ and the common regressor $\varphi(t)$.

2.3. Missing outputs, reconstruction, and frozen-regressor convention

Let

$$\Gamma(t) = \text{diag}\{\gamma_1(t), \gamma_2(t), \dots, \gamma_m(t)\}, \quad \gamma_i(t) \in \{0, 1\}, \quad (2.16)$$

where $\gamma_i(t) = 1$ means that $y_i(t)$ is available. The measured output is denoted by

$$y_o(t) = \Gamma(t)y(t). \quad (2.17)$$

When $\gamma_i(t) = 0$, the value $y_i(t)$ is not available to the identification algorithm. Since future regressors contain past outputs, define the reconstructed output

$$\widehat{y}_i(t) = y_{o,i}(t) + [1 - \gamma_i(t)]\widehat{y}_i(t-1), \quad (2.18)$$

where

$$\widehat{y}_i(t-1) = \widehat{\psi}_i^T(t)\widehat{\alpha}_i(t-1) + \varphi^T(t)\widehat{\theta}_i(t-1), \quad (2.19)$$

with

$$\tilde{\psi}_i(t) = [-\tilde{y}_i(t-1), -\tilde{y}_i(t-2), \dots, -\tilde{y}_i(t-n_i)]^T. \quad (2.20)$$

The reconstructed value is not treated as an observed sample. It is used only to keep the next autoregressive regressor available.

Because (2.18) uses previous estimates, $\tilde{\psi}_i(t)$ is parameter-dependent through past updates. In the recursive algorithm below, the standard stochastic-approximation convention is used: all reconstructed outputs and regressors available at time t are treated as frozen, or $\mathcal{F}_{t-1}$ -measurable, during the current correction. Thus the algorithm is a pseudo-gradient or frozen-regressor stochastic approximation for the reconstructed prediction-error problem, not the exact gradient of the full missing-data likelihood. The effect of reconstruction is represented by

$$\epsilon_{r,i}(t) = \tilde{y}_i(t) - y_i(t), \quad \tilde{\psi}_i(t) = \psi_i(t) + \delta\psi_i(t), \quad (2.21)$$

where $\delta\psi_i(t)$ collects past reconstruction errors.

2.4. Common colored disturbance and robust auxiliary filter

Assume that the output disturbance is generated by the common first-order stable coloring mechanism

$$v(t) = \rho v(t-1) + \xi(t), \quad |\rho| < 1, \quad (2.22)$$

where $\xi(t)$ is a zero-mean innovation that may be non-Gaussian and impulsive. The common scalar filter is assumed in order to preserve the common input-information matrix. If each channel had a different coloring filter, then the filtered input regressor would become channel-dependent and the coupled structure would need to be reformulated.

Remark 2.1 (Applicable boundary of the common coloring assumption). Assumption (2.22) is a modeling choice with a precise structural motivation and a quantifiable price. Structural motivation: If channel i were colored by its own coefficient ρ_i , the filtered input regressor would become $\varphi_{f,i}(t) = \varphi(t) - \rho_i \varphi(t-1)$ and would therefore carry a channel index. The matrix update (4.12), which applies a single matrix $\Phi_f(p, t)$ to all m channels at once, would have to be replaced by m separate products, so both the dominant cost and the rolling-buffer memory of the input part would be multiplied by m . This is exactly the m -fold duplication that the hierarchical formulation is designed to eliminate, which is why the common-filter model is retained here. Price under mismatch: Suppose the true coloring is channel-dependent, $v_i(t) = \rho_i v_i(t-1) + \xi_i(t)$, and let $\bar{\rho}$ denote the value actually identified by (2.25). Filtering with $\bar{\rho}$ leaves the residual coloring

$$(\rho_i - \bar{\rho}) v_i(t-1), \quad (2.23)$$

which is not removed and is absorbed additively into $\Delta_i(t)$ in (2.32). Since $\mathbb{E}\{v_i^2(t-1)\} = \sigma_{\xi,i}^2/(1-\rho_i^2)$ in the stationary regime, the extra mismatch energy is

$$\mathbb{E}\{[(\rho_i - \bar{\rho}) v_i(t-1)]^2\} = \frac{(\rho_i - \bar{\rho})^2 \sigma_{\xi,i}^2}{1 - \rho_i^2}, \quad (2.24)$$

so it grows quadratically with the coloring spread and is bounded whenever the spread is bounded. Assumption 4.2 therefore remains valid with an enlarged constant $\bar{\Delta}^2$, and Theorem 4.1 continues

to guarantee mean-square boundedness, with an ultimate neighborhood that widens proportionally to $\max_i(\rho_i - \bar{\rho})^2$. The method thus degrades gracefully rather than failing: it remains applicable as long as the channel coloring coefficients are clustered, and it loses its advantage over the unfiltered robust variant only when the spread becomes comparable to the coefficients themselves. Section 6.9 measures this degradation numerically and identifies the spread beyond which prewhitening with a single coefficient stops paying off.

Let $\widehat{\rho}(t)$ be a residual-based estimate. To reduce sensitivity to impulsive residuals, we use

$$\widehat{\rho}(t) = \text{proj}_{[-\rho_{\max}, \rho_{\max}]} \frac{\sum_{s=t-L_f+1}^t \omega_{\rho}(s) \widehat{v}^T(s) \widehat{v}(s-1)}{\sum_{s=t-L_f+1}^t \omega_{\rho}(s) \|\widehat{v}(s-1)\|^2 + \epsilon_f}, \quad (2.25)$$

where $0 < \rho_{\max} < 1$, L_f is the residual window, $\epsilon_f > 0$, and

$$\widehat{v}(t) = \Gamma(t)[\widetilde{y}(t) - \eta(\widehat{\alpha}(t-1), t) - \widehat{\Theta}^T(t-1)\varphi(t)]. \quad (2.26)$$

A typical choice is

$$\omega_{\rho}(s) = \min \left\{ 1, \frac{\kappa_{\rho}}{\max\{\|\widehat{v}(s)\| + \|\widehat{v}(s-1)\|, \epsilon_{\rho}\}} \right\}. \quad (2.27)$$

We define the filtered variables

$$y_f(t) = \widetilde{y}(t) - \widehat{\rho}(t-1)\widetilde{y}(t-1), \quad (2.28)$$

$$\varphi_f(t) = \varphi(t) - \widehat{\rho}(t-1)\varphi(t-1), \quad (2.29)$$

$$\psi_{f,i}(t) = \widetilde{\psi}_i(t) - \widehat{\rho}(t-1)\widetilde{\psi}_i(t-1). \quad (2.30)$$

If $\widehat{\rho}(t-1) = \rho$ and no reconstructed past value is used, then

$$y_{f,i}(t) = \psi_{f,i}^T(t)\alpha_i + \varphi_f^T(t)\theta_i + \xi_i(t). \quad (2.31)$$

In the practical case, the filtered model is written as

$$y_{f,i}(t) = \psi_{f,i}^T(t)\alpha_i + \varphi_f^T(t)\theta_i + \xi_i(t) + \Delta_i(t), \quad (2.32)$$

where $\Delta_i(t)$ contains the filter-estimation mismatch and reconstruction mismatch.

Because $y_f(t)$ uses $\widetilde{y}(t)$ and $\widetilde{y}(t-1)$, we define the filtered reliability mask

$$\gamma_{f,i}(t) = \gamma_i(t)\gamma_i(t-1). \quad (2.33)$$

Thus a filtered output sample contributes to the identification loss only when both current and previous outputs are actually observed. This reliability mask certifies the filtered output sample itself. Older lagged reconstructed values may still appear inside $\psi_{f,i}(t)$ when $n_i \geq 2$; their influence is not hidden but is included in the reconstruction/filter mismatch term $\Delta_i(t)$ and in Assumption 4.2.

Remark 2.2 (Meaning of $\gamma_{f,i}(t)$, its information cost, and softened variants). *Interpretation.* The product $\gamma_i(t)\gamma_i(t-1)$ is not a tuning weight; it is the exact indicator function of the event “the filtered sample $y_{f,i}(t) = \widetilde{y}_i(t) - \widehat{\rho}(t-1)\widetilde{y}_i(t-1)$ is a difference of two genuine measurements”. Because

prewhitening is a one-step difference, a single missing sample corrupts two consecutive filtered samples, and $\gamma_{f,i}(t)$ is precisely the certificate that neither of the two ingredients has been imputed. Admitting an uncertified sample would let a model-generated value act as data, which is the very mechanism that biases zero-filling.

Information cost. Let the availability variables be independent with $\Pr\{\gamma_i(t) = 1\} = 1 - p_m$. Then

$$\mathbb{E}\{\gamma_i(t)\} = 1 - p_m, \quad \mathbb{E}\{\gamma_{f,i}(t)\} = (1 - p_m)^2, \quad (2.34)$$

so the hard mask retains a fraction $1 - p_m$ of the samples that survive the raw mask, and the excitation level available to the algorithm degrades quadratically rather than linearly in the availability probability. The constant λ_0 of Assumption 4.3 inherits this scaling, $\lambda_0(p_m) \approx (1 - p_m)^2 \lambda_0(0)$; the empirical values reported in Section 6.12 follow this law closely. At $p_m = 0.2$, the loss is modest (80% of the observed filtered samples are kept), whereas at $p_m = 0.5$, only half of them are kept, which explains why the degradation of all masked methods accelerates beyond a dropout probability of about 30%.

Softened reliability masks. The hard constraint can be relaxed by the one-parameter family

$$\gamma_{f,i}^{(\omega)}(t) = \gamma_i(t)[\gamma_i(t-1) + (1 - \gamma_i(t-1))\omega], \quad \omega \in [0, 1], \quad (2.35)$$

which admits, with reduced weight ω , those filtered samples whose current output is measured but whose lag-one value has been reconstructed. It interpolates between the hard mask ($\omega = 0$) and the current-sample mask ($\omega = 1$), and it retains the larger fraction $\mathbb{E}\{\gamma_{f,i}^{(\omega)}\} = (1 - p_m)[(1 - p_m) + p_m\omega]$ of the data. The price is explicit: on an admitted sample with $\gamma_i(t-1) = 0$, one has $\widehat{y}_i(t-1) = \widehat{y}_i(t-1|t-2)$, hence

$$y_{f,i}(t) = [y_i(t) - \widehat{\rho}(t-1)y_i(t-1)] + \widehat{\rho}(t-1)\epsilon_{r,i}(t-1), \quad (2.36)$$

so the reconstruction error is injected into the loss with weight $\omega\widehat{\rho}(t-1)$, and the mismatch energy of Assumption 4.2 increases by $\omega^2\widehat{\rho}^2(t-1)p_m(1 - p_m)\mathbb{E}\{\epsilon_{r,i}^2\}$. The softened mask therefore trades a variance reduction, obtained from the additional samples, against a bias increase proportional to $\omega^2\rho^2$, and the balance depends on the dropout probability through the competing factors $(1 - p_m)$ and $p_m(1 - p_m)$ appearing above. Section 6.6 resolves the trade-off experimentally: at the moderate dropout probabilities for which the certified mask was designed, the two options are statistically indistinguishable, so the hard mask is retained as the default because it alone guarantees that no imputed value enters the loss, whereas beyond roughly 30% dropout, the variance term dominates and softening becomes clearly preferable.

3. Robust masked filtered HSG method

For the filtered model, define the one-step innovation

$$e_i(t) = y_{f,i}(t) - \psi_{f,i}^T(t)\widehat{\alpha}_i(t-1) - \varphi_f^T(t)\widehat{\theta}_i(t-1). \quad (3.1)$$

The Huber loss and its score are

$$\rho_\kappa(e) = \begin{cases} \frac{1}{2}e^2, & |e| \leq \kappa, \\ \kappa|e| - \frac{1}{2}\kappa^2, & |e| > \kappa, \end{cases} \quad \chi_\kappa(e) = \rho'_\kappa(e) = \begin{cases} e, & |e| \leq \kappa, \\ \kappa \operatorname{sgn}(e), & |e| > \kappa. \end{cases} \quad (3.2)$$

For numerical implementation, one may write $\chi_\kappa(e) = w_\kappa(e)e$ with

$$w_\kappa(e) = \begin{cases} 1, & |e| \leq \kappa, \\ \frac{\kappa}{|e|}, & |e| > \kappa, \end{cases} \quad (3.3)$$

or $w_\kappa(e) = \min\{1, \kappa / \max(|e|, \epsilon_0)\}$. This definition avoids the incorrect ratio $\chi_\kappa(e)/(e + \epsilon_0)$ for negative residuals.

The instantaneous robust masked loss is

$$J_1(\alpha, \Theta) = \sum_{i=1}^m \gamma_{f,i}(t) \rho_\kappa(y_{f,i}(t) - \psi_{f,i}^T(t)\alpha_i - \varphi_f^T(t)\theta_i). \quad (3.4)$$

Under the frozen-regressor convention, its subgradients at the previous estimates are

$$\frac{\partial J_1}{\partial \theta_i} = -\gamma_{f,i}(t) \varphi_f(t) \chi_\kappa(e_i(t)), \quad (3.5)$$

$$\frac{\partial J_1}{\partial \alpha_i} = -\gamma_{f,i}(t) \psi_{f,i}(t) \chi_\kappa(e_i(t)). \quad (3.6)$$

Define the robust masked scalar innovation

$$\bar{e}_i(t) = \gamma_{f,i}(t) \chi_\kappa(e_i(t)), \quad \bar{e}(t) = [\bar{e}_1(t), \dots, \bar{e}_m(t)]^T. \quad (3.7)$$

Define the normalized step size

$$a_t := \frac{\mu}{r^{\varepsilon(t)}(t)} = \mu r^{-\varepsilon(t)}(t). \quad (3.8)$$

The robust masked filtered HSG update is

$$\widehat{\Theta}(t) = \widehat{\Theta}(t-1) + a_t \varphi_f(t) \bar{e}^T(t), \quad (3.9)$$

$$\widehat{\alpha}_i(t) = \widehat{\alpha}_i(t-1) + a_t \psi_{f,i}(t) \bar{e}_i(t), \quad i = 1, 2, \dots, m, \quad (3.10)$$

$$r(t) = r(t-1) + \|\varphi_f(t)\|^2 + \sum_{i=1}^m \|\psi_{f,i}(t)\|^2 + \epsilon_r. \quad (3.11)$$

The complete procedure is summarized in Algorithm 1.

4. Robust adaptive masked filtered HMISG method

4.1. Stacked filtered variables and robust innovations

Let $p(t)$ be the innovation length and we write $p = p(t)$ when no confusion occurs. Define

$$Y_f(p, t) = [y_f(t), y_f(t-1), \dots, y_f(t-p+1)] \in \mathbb{R}^{m \times p}, \quad (4.1)$$

$$\Phi_f(p, t) = [\varphi_f(t), \varphi_f(t-1), \dots, \varphi_f(t-p+1)] \in \mathbb{R}^{n \times p}, \quad (4.2)$$

$$\Psi_{f,i}(p, t) = [\psi_{f,i}(t), \psi_{f,i}(t-1), \dots, \psi_{f,i}(t-p+1)] \in \mathbb{R}^{n_i \times p}. \quad (4.3)$$

Algorithm 1 Robust masked filtered HSG algorithm

Require: Inputs $u_j(t)$, incomplete outputs $y_{o,i}(t)$, masks $\gamma_i(t)$, Huber threshold κ , update gain μ , and initial estimates $\widehat{\Theta}(0)$ and $\widehat{\alpha}_i(0)$.

Ensure: $\widehat{\Theta}(t)$ and $\widehat{\alpha}_i(t)$.

```

1: for  $t = 1, 2, \dots, L$  do
2:   Construct  $\varphi(t)$ .
3:   for  $i = 1, 2, \dots, m$  do
4:     Compute  $\widehat{y}_i(t|t-1)$  and  $\widetilde{y}_i(t)$  from (2.18)–(2.19).
5:     Construct  $\widetilde{\psi}_i(t)$ .
6:   end for
7:   Estimate  $\widehat{\rho}(t-1)$  from (2.25); compute  $y_f(t)$ ,  $\varphi_f(t)$ , and  $\psi_{f,i}(t)$ .
8:   Compute  $e_i(t)$ ,  $\gamma_{f,i}(t)$ , and  $\bar{e}_i(t) = \gamma_{f,i}(t)\chi_\kappa(e_i(t))$ .
9:   Update  $r(t)$ , compute  $a_t$  from (3.8), and update  $\widehat{\Theta}(t)$  and  $\widehat{\alpha}_i(t)$  from (3.9)–(3.10).
10: end for

```

For each channel, define

$$Y_{f,i}(p, t) = [y_{f,i}(t), y_{f,i}(t-1), \dots, y_{f,i}(t-p+1)]^T. \quad (4.4)$$

The raw innovation vector of channel i is

$$\mathcal{E}_i(p, t) = Y_{f,i}(p, t) - \Psi_{f,i}^T(p, t)\widehat{\alpha}_i(t-1) - \Phi_f^T(p, t)\widehat{\theta}_i(t-1). \quad (4.5)$$

Collecting the rows gives $\mathcal{E}(p, t) = [\mathcal{E}_1(p, t), \dots, \mathcal{E}_m(p, t)]^T \in \mathbb{R}^{m \times p}$. The robust masked innovation vector is

$$\bar{\mathcal{E}}_i(p, t) = \begin{bmatrix} \gamma_{f,i}(t)\chi_\kappa(\mathcal{E}_i^{(0)}(p, t)) \\ \gamma_{f,i}(t-1)\chi_\kappa(\mathcal{E}_i^{(1)}(p, t)) \\ \vdots \\ \gamma_{f,i}(t-p+1)\chi_\kappa(\mathcal{E}_i^{(p-1)}(p, t)) \end{bmatrix} \in \mathbb{R}^p, \quad (4.6)$$

where $\mathcal{E}_i^{(\ell)}(p, t)$ is the $(\ell+1)$ th element of $\mathcal{E}_i(p, t)$. Let

$$\bar{\mathcal{E}}(p, t) = [\bar{\mathcal{E}}_1(p, t), \bar{\mathcal{E}}_2(p, t), \dots, \bar{\mathcal{E}}_m(p, t)]^T \in \mathbb{R}^{m \times p}. \quad (4.7)$$

The finite-window Huber loss is

$$J_2(\alpha, \Theta) = \sum_{i=1}^m \sum_{\ell=0}^{p-1} \gamma_{f,i}(t-\ell) \rho_\kappa \left(Y_{f,i}^{(\ell)} - \psi_{f,i}^T(t-\ell)\alpha_i - \varphi_f^T(t-\ell)\theta_i \right). \quad (4.8)$$

Its frozen-regressor subgradients are

$$\nabla_{\Theta} J_2 = -\Phi_f(p, t)\bar{\mathcal{E}}^T(p, t), \quad (4.9)$$

$$\frac{\partial J_2}{\partial \alpha_i} = -\Psi_{f,i}(p, t)\bar{\mathcal{E}}_i(p, t). \quad (4.10)$$

Thus the RAMF-HMISG update uses the same unambiguous step-size notation

$$a_t := \frac{\mu}{r^{\varepsilon(t)}(t)} = \mu r^{-\varepsilon(t)}(t), \quad (4.11)$$

and is given by

$$\widehat{\Theta}(t) = \widehat{\Theta}(t-1) + a_t \Phi_f(p, t) \bar{\mathcal{E}}^T(p, t), \quad (4.12)$$

$$\widehat{\alpha}_i(t) = \widehat{\alpha}_i(t-1) + a_t \Psi_{f,i}(p, t) \bar{\mathcal{E}}_i(p, t), \quad (4.13)$$

$$r(t) = r(t-1) + \|\Phi_f(p, t)\|_F^2 + \sum_{i=1}^m \|\Psi_{f,i}(p, t)\|_F^2 + \epsilon_r. \quad (4.14)$$

4.2. Non-circular adaptive laws

Let $\zeta(t)$ be a smoothed, normalized, robust innovation energy. The innovation length and convergence index are selected using $\zeta(t-1)$:

$$p(t) = \begin{cases} p_{\max}, & \zeta(t-1) > \eta_2, \\ p_{\text{mid}}, & \eta_1 < \zeta(t-1) \leq \eta_2, \\ p_{\min}, & \zeta(t-1) \leq \eta_1, \end{cases} \quad (4.15)$$

$$\varepsilon(t) = \text{proj}_{[\varepsilon_{\min}, \varepsilon_{\max}]} \left[\varepsilon_{\max} - (\varepsilon_{\max} - \varepsilon_{\min}) \frac{\zeta(t-1)}{\zeta(t-1) + \eta_\varepsilon} \right]. \quad (4.16)$$

Only after $p(t)$ has been selected and $\bar{\mathcal{E}}(p(t), t)$ has been computed is the energy updated:

$$\zeta(t) = \lambda_\zeta \zeta(t-1) + (1 - \lambda_\zeta) \frac{\|\bar{\mathcal{E}}(p(t), t)\|_F^2}{N_f(p(t), t) + \epsilon_0}, \quad (4.17)$$

where

$$N_f(p, t) = \sum_{i=1}^m \sum_{\ell=0}^{p-1} \gamma_{f,i}(t-\ell). \quad (4.18)$$

This order removes the circularity between $p(t)$ and $\zeta(t)$. The resulting procedure is summarized in Algorithm 2.

4.3. Design rationale, admissible ranges, and transferability of the adaptive laws

The rules (4.15)–(4.17) are not arbitrary switching heuristics. Each of their ingredients is dictated either by the error recursion established in Section 4 or by the hypotheses under which Theorem 4.1 is proved, and the resulting design is scale free. This subsection makes that statement explicit.

(a) *Why the innovation length must be time-varying.* The parameter-error recursion (4.54) reads $\bar{\vartheta}_i(t) = [I - a_t G_i(t)] \bar{\vartheta}_i(t-1) + a_t d_i(t)$. Both the contraction generator $G_i(t)$ of (4.40) and the driving term $d_i(t)$ of (4.52) are sums of $p(t)$ windowed contributions. When the samples inside the window are weakly correlated, $\lambda_{\min} \mathbb{E}\{G_i(t)\} \approx p \bar{\lambda}_G$ grows linearly in p , whereas $\mathbb{E}\{\|d_i(t)\|^2\} \approx p C_d (\bar{\sigma}^2 + \bar{\Delta}^2)$ also

grows linearly in p because the score terms are conditionally centered and therefore add in variance, not in amplitude. Balancing the two effects in (4.60) gives a stationary level

$$V_\infty \approx \frac{C a_t C_d (\bar{\sigma}^2 + \bar{\Delta}^2)}{\bar{\lambda}_G}, \quad (4.19)$$

which is, to first order, independent of p , while the contraction rate per step, $a_t p \bar{\lambda}_G$, is proportional to p . A long window therefore buys transient speed at almost no asymptotic price—but only as long as the first-order approximation holds. It breaks down for large p for two reasons: the samples inside a long window become strongly correlated, so $\lambda_{\min} \mathbb{E}\{G_i\}$ saturates while the noise contribution keeps accumulating, and every term of the window is evaluated at the single estimate $\hat{\theta}_i(t-1)$, so the oldest innovations in the window are stale by up to $p-1$ update steps. The consequence is a clear prescription: Use a long window while the parameter error dominates the innovation, and shorten it once the innovation is dominated by noise and mismatch. Since $\zeta(t-1)$ is precisely a smoothed estimate of the normalized robust innovation energy, it is the natural switching variable, and (4.15) is the coarsest rule that realizes the prescription.

(b) *Why the convergence index must lie in a narrow interval and why it must move inside it.* Assumption 4.4 requires $1/2 < \varepsilon_{\min} \leq \varepsilon(t) \leq \varepsilon_{\max} \leq 1$. The lower limit is not a design preference: With $r(t)$ growing linearly, as established in (4.63), the condition $\sum_t a_t^2 < \infty$ holds if and only if $\varepsilon(t) > 1/2$, and the condition $\sum_t a_t = \infty$ holds if and only if $\varepsilon(t) \leq 1$. Any admissible interval is thus a subinterval of $(1/2, 1]$. Inside it, small ε yields large steps, fast transients, and a wider ultimate neighborhood, while large ε yields the opposite; values approaching 1 make a_t decay like $1/t$ and stall the transient of a gradient-type recursion. Two guidelines follow: Keep a safety margin below the boundary, $\varepsilon_{\min} = 1/2 + \delta$ with $\delta \approx 0.05$, and keep the interval narrow, $\varepsilon_{\max} - \varepsilon_{\min} \approx 0.1$, so that a switch changes a_t by the bounded factor $r^{\varepsilon_{\max} - \varepsilon_{\min}}(t)$ rather than by orders of magnitude. The choice $[\varepsilon_{\min}, \varepsilon_{\max}] = [0.56, 0.66]$ used in Section 6 is the instance of these guidelines. Section 6.11 shows that the penalty for misplacing this interval is strongly asymmetric: shifting it down toward the admissible limit $1/2$ costs little, whereas shifting it up toward 1 degrades the final error by more than an order of magnitude, because the step size then decays like $1/t$ and the transient stalls. This is why the recommended interval sits just above the lower admissible limit rather than in the middle of $(1/2, 1]$.

(c) *Why the thresholds are transferable to other problems.* The switching variable is bounded by a quantity fixed by the designer. Indeed, $\gamma_{f,i} \in \{0, 1\}$ gives $\gamma_{f,i}^2 = \gamma_{f,i}$, and $|\chi_k(\cdot)| \leq \kappa$, so

$$\|\bar{\mathcal{E}}(p, t)\|_F^2 = \sum_{i=1}^m \sum_{\ell=0}^{p-1} \gamma_{f,i}(t-\ell) \chi_k^2(\cdot) \leq \kappa^2 N_f(p, t), \quad (4.20)$$

and therefore, since (4.17) is a convex combination,

$$0 \leq \zeta(t) \leq \max\{\zeta(0), \kappa^2\} \quad \text{for all } t, \quad \text{so that } 0 \leq \zeta(t) \leq \kappa^2 \text{ whenever } \zeta(0) \leq \kappa^2. \quad (4.21)$$

Initializing at $\zeta(0) = \kappa^2$, as is done throughout this paper, therefore confines the energy variable to a known interval whose length is set by the Huber threshold alone, independent of the plant gain, of the signal units and of the system dimension. This is what makes the adaptive laws portable: the thresholds should be specified as fractions of κ^2 ,

$$\eta_1 \approx 0.1\kappa^2, \quad \eta_2 \approx 0.5\kappa^2, \quad (4.22)$$

and the shape constant as $\eta_\varepsilon \approx 0.25\kappa^2$. With $\kappa = 1$, these guidelines return exactly the values $\eta_1 = 0.12$, $\eta_2 = 0.55$, and $\eta_\varepsilon = 0.25$ used in the simulations, and they transfer unchanged to a different system once κ has been set from the residual scale by the rule of Section 4.6. The same reasoning fixes the window candidates in terms of the channel parameter dimension $d_i = n_i + n$ rather than in absolute terms: the window must be long enough for the windowed Gram matrix to be informative, yet short enough to avoid staleness, which gives the band $p \in [d_i/2, 2d_i]$. For the example of Section 6, $d_i = 6$ and the band is $[3, 12]$, inside which the candidate set $\{p_{\min}, p_{\text{mid}}, p_{\max}\} = \{5, 8, 10\}$ lies. Section 6.11 shows that candidate sets from the middle of that band upward perform equivalently, that only a set sitting at its very bottom loses accuracy, and that enlarging the candidates well beyond the band brings no further benefit while increasing the per-sample cost linearly. The upper end of the candidate set is therefore a computational-budget decision rather than an accuracy optimum, and the economical set used here is chosen on that basis.

Remark 4.1 (Role of λ_ζ , boundary behavior of $\zeta(t)$, and the nature of the feedback). Three points deserve a separate statement.

(i) *The feedback is causal, not circular, and it is bounded.* The quantity $\zeta(t)$ computed from the window actually used at time t does influence the selection of $p(t+1)$ and $\varepsilon(t+1)$. This is a causal feedback loop, and it is deliberate: it is what allows the algorithm to react to a change of regime. Circularity, in the sense that would make the update implicit and the analysis invalid, would require $p(t)$ to depend on a quantity that itself depends on $p(t)$ at the same instant; the ordering imposed after (4.16) excludes this. The loop cannot destabilize the recursion either, because by (4.21), its state is confined to the bounded interval $[0, \max\{\zeta(0), \kappa^2\}]$ for every admissible input, its output $p(t)$ is confined to the finite set $\{p_{\min}, p_{\text{mid}}, p_{\max}\}$, while $\varepsilon(t)$ is confined to $[\varepsilon_{\min}, \varepsilon_{\max}]$ by the projection in (4.16). Assumption 4.4 therefore holds for any value of λ_ζ , η_1 , and η_2 , and Theorem 4.1 is insensitive to their tuning; these constants affect how quickly the algorithm converges, not whether it remains bounded.

(ii) *Choice of the smoothing coefficient.* The recursion (4.17) is a first-order low-pass filter of effective memory $N_\zeta = 1/(1 - \lambda_\zeta)$ samples. Two opposite requirements fix it. The memory must be long enough for a single impulsive innovation not to trigger a spurious switch, which requires N_ζ to cover several outlier inter-arrival times, $N_\zeta \gtrsim 2/p_o$, where p_o is the outlier probability; and it must be short enough to track the transient, $N_\zeta \ll L$. For $p_o = 2\%$, this gives $N_\zeta \gtrsim 100$ at the lower end and $N_\zeta \leq L/20 = 250$ at the upper end for $L = 5000$, and the value $\lambda_\zeta = 0.98$, that is, $N_\zeta = 50$, sits slightly below the conservative lower limit while remaining well-damped in practice. Section 6.11 reports the measured sensitivity over $\lambda_\zeta \in [0.90, 0.995]$.

(iii) *Behavior at the thresholds.* Because ζ is a convex combination, consecutive values satisfy $|\zeta(t) - \zeta(t-1)| \leq (1 - \lambda_\zeta)\kappa^2$, so the switching variable cannot jump across the admissible interval and the number of threshold crossings per unit time is bounded by $(1 - \lambda_\zeta)\kappa^2 / \min\{\eta_2 - \eta_1, \eta_1\}$. Chattering is therefore structurally limited, and it is further attenuated by the smoothing. The two saturated regimes are benign. If $\zeta(t)$ never exceeds η_2 , the law degenerates to a fixed window p_{mid} or $p_{\min}$; if it never falls below η_2 , the law degenerates to the fixed long window $p_{\max}$. Both degeneracies are still covered by Assumption 4.4, which only requires $p_{\min} \leq p(t) \leq p_{\max}$, so no boundary case escapes the analysis. The practically relevant consequence is that misspecified thresholds cannot break the algorithm; they can only make it behave like one of its fixed-window special cases. The fraction of time spent at each level, reported in Section 6.11, is the diagnostic that reveals such a degeneracy.

Algorithm 2 RAMF-HMISG algorithm**Require:** $p_{\min}, p_{\text{mid}}, p_{\max}, \kappa, \mu, \varepsilon_{\min}, \varepsilon_{\max}$, and incomplete data.**Ensure:** $\widehat{\Theta}(t)$ and $\widehat{\alpha}_i(t)$.

- 1: Initialize $\widehat{\Theta}(0), \widehat{\alpha}_i(0), r(0) = 1$, and $\zeta(0) > 0$.
- 2: **for** $t = 1, 2, \dots, L$ **do**
- 3: Reconstruct $\widetilde{y}_i(t)$ and form $\widetilde{\psi}_i(t)$.
- 4: Estimate $\widehat{\rho}(t-1)$ and compute the filtered variables.
- 5: Select $p(t)$ and $\varepsilon(t)$ from (4.15)–(4.16) using $\zeta(t-1)$.
- 6: Construct $Y_f(p, t), \Phi_f(p, t)$, and $\Psi_{f,i}(p, t)$.
- 7: Compute the raw innovations $\mathcal{E}_i(p, t)$ and robust masked innovations $\bar{\mathcal{E}}_i(p, t)$.
- 8: Update $r(t)$, compute a_t from (4.11), and update $\widehat{\Theta}(t)$ and $\widehat{\alpha}_i(t)$ from (4.12)–(4.13).
- 9: Update $\zeta(t)$ from (4.17).
- 10: **end for**

4.4. Residual relation

Define the post-update raw residual vector of channel i by

$$\widehat{V}_i(p, t) = Y_{f,i}(p, t) - \Psi_{f,i}^T(p, t)\widehat{\alpha}_i(t) - \Phi_f^T(p, t)\widehat{\theta}_i(t). \quad (4.23)$$

The following lemma is not introduced only as an algebraic identity. It explains exactly how the RAMF-HMISG correction changes the whole innovation window after one update. In particular, it shows that the post-update residual equals the pre-update raw innovation minus a Gram-matrix correction generated by the stacked filtered regressors. This identity is used below to identify the effective information matrix appearing in the error recursion and to justify the small-gain condition on the step size.

Lemma 4.1 (Post-update raw residual relation). *For the RAMF-HMISG update (4.12)–(4.13), the post-update residual (4.23) satisfies*

$$\widehat{V}_i(p, t) = \mathcal{E}_i(p, t) - a_t \left[\Psi_{f,i}^T(p, t)\Psi_{f,i}(p, t) + \Phi_f^T(p, t)\Phi_f(p, t) \right] \bar{\mathcal{E}}_i(p, t). \quad (4.24)$$

Proof. Fix an output channel i and a time instant t . For compactness, write $a_t = \mu/r^{\varepsilon(t)}(t) = \mu r^{-\varepsilon(t)}(t)$, $\Phi_f = \Phi_f(p, t)$, $\Psi_i = \Psi_{f,i}(p, t)$, $\mathcal{E}_i = \mathcal{E}_i(p, t)$, and $\bar{\mathcal{E}}_i = \bar{\mathcal{E}}_i(p, t)$. From the i th column of the matrix update (4.12),

$$\widehat{\theta}_i(t) = \widehat{\theta}_i(t-1) + a_t \Phi_f \bar{\mathcal{E}}_i. \quad (4.25)$$

Similarly, the autoregressive-parameter update (4.13) gives

$$\widehat{\alpha}_i(t) = \widehat{\alpha}_i(t-1) + a_t \Psi_i \bar{\mathcal{E}}_i. \quad (4.26)$$

Substituting (4.26) and (4.25) into the definition of the post-update residual yields

$$\widehat{V}_i(p, t) = Y_{f,i}(p, t) - \Psi_i^T \left[\widehat{\alpha}_i(t-1) + a_t \Psi_i \bar{\mathcal{E}}_i \right] - \Phi_f^T \left[\widehat{\theta}_i(t-1) + a_t \Phi_f \bar{\mathcal{E}}_i \right] \quad (4.27)$$

$$= \left[Y_{f,i}(p, t) - \Psi_i^T \widehat{\alpha}_i(t-1) - \Phi_f^T \widehat{\theta}_i(t-1) \right] - a_t \Psi_i^T \Psi_i \bar{\mathcal{E}}_i - a_t \Phi_f^T \Phi_f \bar{\mathcal{E}}_i. \quad (4.28)$$

The bracketed term is exactly the raw innovation vector $\mathcal{E}_i(p, t)$ defined in (4.5). Hence

$$\widehat{V}_i(p, t) = \mathcal{E}_i(p, t) - a_t \left(\Psi_i^T \Psi_i + \Phi_f^T \Phi_f \right) \bar{\mathcal{E}}_i(p, t), \quad (4.29)$$

which is (4.24). Notice that the matrix in brackets has dimension $p \times p$ and is positive semidefinite. Therefore the update subtracts from the raw innovation a correction lying in the span of the stacked filtered regressors. In the ideal noiseless and unmasked quadratic case, $\bar{\mathcal{E}}_i = \mathcal{E}_i$ and the relation reduces to the familiar residual-shrinking identity

$$\widehat{V}_i(p, t) = \left[I_p - a_t \left(\Psi_i^T \Psi_i + \Phi_f^T \Phi_f \right) \right] \mathcal{E}_i(p, t), \quad (4.30)$$

which explains why the step-size restriction must control the largest eigenvalue of the stacked Gram matrix. This also explains why the lemma is useful in the boundedness theorem: the same Gram structure generates the effective information matrix in the parameter-error recursion. The proof is complete. $\square$

4.5. Mean-square boundedness

Let

$$\Phi_{a,i}(t) = \begin{bmatrix} \psi_{f,i}(t) \\ \varphi_f(t) \end{bmatrix}, \quad \widehat{\vartheta}_i(t) = \begin{bmatrix} \widehat{\alpha}_i(t) \\ \widehat{\theta}_i(t) \end{bmatrix}, \quad \vartheta_i = \begin{bmatrix} \alpha_i \\ \theta_i \end{bmatrix}, \quad \widetilde{\vartheta}_i(t) = \widehat{\vartheta}_i(t) - \vartheta_i. \quad (4.31)$$

For a window of length $p(t)$, define the augmented regressor matrix

$$\Phi_{a,i}(p, t) = [\Phi_{a,i}(t), \Phi_{a,i}(t-1), \dots, \Phi_{a,i}(t-p(t)+1)]. \quad (4.32)$$

The filtered model (2.32) can be equivalently written as

$$y_{f,i}(s) = \Phi_{a,i}^T(s) \vartheta_i + \xi_i(s) + \Delta_i(s), \quad (4.33)$$

where $\Delta_i(s)$ is zero in the ideal case of exact filtering and no reconstruction error.

Assumption 4.1. *The filtered regressors are bounded: $\|\Phi_{a,i}(t)\| \leq \bar{\phi}$ for all i, t . The innovation sequence is conditionally symmetric, or equivalently satisfies the Huber-score martingale condition*

$$\mathbb{E}\{\chi_\kappa(\xi_i(t)) \mid \mathcal{F}_{t-1}\} = 0, \quad \mathbb{E}\{\chi_\kappa^2(\xi_i(t)) \mid \mathcal{F}_{t-1}\} \leq \bar{\sigma}^2. \quad (4.34)$$

The masks and regressors are adapted to the filtration $\mathcal{F}_t$ and the conditional moments appearing below are well-defined.

Assumption 4.2. *The aggregate reconstruction and filter-estimation mismatch satisfies*

$$\mathbb{E}\{\|\Delta_{a,i}(t)\|^2 \mid \mathcal{F}_{t-1}\} \leq \bar{\Delta}^2 \quad (4.35)$$

for all i, t , where $\Delta_{a,i}(t)$ collects the windowed terms generated by $\Delta_i(t)$, the error $\widehat{\rho}(t) - \rho$, and older reconstructed lagged outputs inside $\psi_{f,i}(t)$. No consistency of $\widehat{\rho}(t)$ is claimed; its effect is absorbed by this bounded mismatch term.

Remark 4.2 (Why the aggregate mismatch is bounded). Subsuming the reconstruction error and the filter-estimation error into a single bounded term is standard, but it should not be accepted on convention alone. The intuition is that every constituent of $\Delta_i(t)$ is the difference between a quantity computed from finitely many bounded signals and the same quantity computed from their exact counterparts, so that boundedness follows from boundedness of the ingredients and does not require any of the auxiliary estimators to converge. Making this explicit, we substitute (2.28)–(2.30) and (2.21) into (2.32). The mismatch splits into three physically distinct contributions,

$$\Delta_i(t) = \underbrace{[\rho - \widehat{\rho}(t-1)]v_i(t-1)}_{\text{filter-estimation error}} + \underbrace{\epsilon_{r,i}(t) - \widehat{\rho}(t-1)\epsilon_{r,i}(t-1)}_{\text{reconstruction error in the output}} - \underbrace{[\delta\psi_i(t) - \widehat{\rho}(t-1)\delta\psi_i(t-1)]^\top \alpha_i}_{\text{reconstruction error in the regressor}}, \quad (4.36)$$

and each term is bounded for an elementary reason. First, the factor $\rho - \widehat{\rho}(t-1)$ never exceeds $2\rho_{\max} < 2$ because (2.25) projects the estimate onto $[-\rho_{\max}, \rho_{\max}]$ and the true coefficient is stable. This is the precise sense in which no consistency of $\widehat{\rho}$ is needed: even a completely wrong filter estimate produces a bounded, not an unbounded, perturbation. It is multiplied by $v_i(t-1)$, whose second moment is $\sigma_{\xi,i}^2/(1-\rho^2) < \infty$ because the coloring mechanism (2.22) is stable and the innovation has finite variance. Second, $\epsilon_{r,i}(t) = [1 - \gamma_i(t)][\widehat{y}_i(t|t-1) - y_i(t)]$ vanishes identically on observed samples and, on missing samples, is the difference of two quantities that are individually bounded: $y_i(t)$ has a finite second moment because the plant is stable and the input is bounded, and $\widehat{y}_i(t|t-1)$ is a bounded bilinear function of bounded regressors and of estimates confined to a compact set by the projection used in the implementation. Third, $\delta\psi_i(t)$ merely stacks at most n_i past values of $\epsilon_{r,i}$, so it inherits the same bound, and it is contracted by the fixed vector α_i . Consequently, $\bar{\Delta}^2$ is finite under the same conditions that are already imposed elsewhere in the paper—stable plant and coloring filter, bounded inputs, finite innovation variance, projected $\widehat{\rho}$, and bounded estimates—so Assumption 4.2 restates those conditions compactly rather than adding a new restriction. Two consequences are worth recording. The bound is uniform in t but far from tight: all three terms shrink as the estimates improve, since the convergence of $\widehat{\rho}$ to a stable value reduces the first, and a smaller prediction error reduces the second and the third. Section 6.12 measures this decay and shows that the mismatch energy falls by at least a factor of four between the transient and the steady state, except under heavy impulsive contamination, where it remains essentially constant because the impulsive component propagates into the reconstructed outputs and does not diminish as the estimates improve. In the ideal limit of exact filtering and complete data, all three terms vanish identically, so $\bar{\Delta}^2 = 0$ and the last term of (4.46) disappears, in agreement with the final statement of Theorem 4.1.

Assumption 4.3. For each output channel i , the reliable filtered regressors satisfy the block persistent-excitation condition

$$\mathbb{E} \left\{ \sum_{\ell=0}^{p(t)-1} \gamma_{f,i}(t-\ell) \Phi_{a,i}(t-\ell) \Phi_{a,i}^\top(t-\ell) \middle| \mathcal{F}_{t-1} \right\} \geq \lambda_0 I \quad (4.37)$$

for some $\lambda_0 > 0$. Moreover, the Huber generalized slope is not degenerate on average: if

$$\chi_\kappa(w-z) - \chi_\kappa(w) = -q(w,z)z, \quad q(w,z) \in [0,1], \quad (4.38)$$

then the corresponding window slopes satisfy an effective excitation bound

$$\mathbb{E}\{G_i(t) \mid \mathcal{F}_{t-1}\} \geq \lambda_G I, \quad \lambda_G > 0, \quad (4.39)$$

where

$$G_i(t) = \sum_{\ell=0}^{p(t)-1} q_i(t-\ell) \gamma_{f,i}(t-\ell) \Phi_{a,i}(t-\ell) \Phi_{a,i}^T(t-\ell). \quad (4.40)$$

Because $q_i(t-\ell) \in [0, 1]$, $G_i(t)$ is positive semidefinite and, by the boundedness of the regressors and $p(t)$, there is a constant $\bar{G} > 0$ such that $\|G_i(t)\| \leq \bar{G}$.

Remark 4.3 (When is Assumption 4.3 satisfied in practice?). Assumption 4.3 has two parts, and they degrade very differently under missing data and impulsive noise. Separating them shows that the assumption is mild in the operating regime considered in this paper.

Part one: Block excitation of the masked filtered regressors. If the availability process $\{\gamma_i(t)\}$ is independent of the input and of the disturbance—the standard model for packet loss and sensor dropout—then the mask factorizes out of the expectation in (4.37). Using $\gamma_{f,i}(t-\ell) = \gamma_i(t-\ell)\gamma_i(t-\ell-1)$ and $\Pr\{\gamma_i = 1\} = 1 - p_m$,

$$\mathbb{E}\left\{\sum_{\ell=0}^{p-1} \gamma_{f,i}(t-\ell) \Phi_{a,i}(t-\ell) \Phi_{a,i}^T(t-\ell)\right\} = (1 - p_m)^2 \mathbb{E}\left\{\sum_{\ell=0}^{p-1} \Phi_{a,i}(t-\ell) \Phi_{a,i}^T(t-\ell)\right\}, \quad (4.41)$$

so that

$$\lambda_0(p_m) = (1 - p_m)^2 \lambda_0(0). \quad (4.42)$$

Two conclusions follow. Excitation is never destroyed by random dropout: λ_0 stays strictly positive for every $p_m < 1$, and the condition reduces to the usual requirement that the unmasked filtered regressors be persistently exciting, which holds whenever the inputs are persistently exciting of order $\max_j n_j$, the plant and the coloring filter are stable, and $|\widehat{\rho}(t)| \leq \rho_{\max} < 1$ so that the filtering operation is invertible and cannot annihilate a spectral line. Excitation is, however, quadratically penalized, which is the analytical counterpart of the information count (2.34) and the reason why performance deteriorates markedly beyond a dropout probability of about 30%. At the nominal 20% used here, (4.42) predicts a retention of 64% of the nominal excitation.

Part two: Non-degeneracy of the average Huber slope. By (4.38), $q_i(s) = 1$ whenever both arguments of the score lie in its linear region, and $q_i(s) < 1$ only on samples where the residual is truncated. Truncation occurs on genuinely impulsive samples, which arrive with probability p_o , and on the tail of the nominal innovation, with probability $\nu = \Pr\{|\xi_i| > \kappa \mid \text{no outlier}\}$. Hence $\mathbb{E}\{q_i(s)\} \geq 1 - p_o - \nu$ and, combining with (4.41),

$$\lambda_G \geq (1 - p_o - \nu) \lambda_0 = (1 - p_o - \nu)(1 - p_m)^2 \lambda_0(0). \quad (4.43)$$

The condition $\lambda_G > 0$ therefore fails only if the Huber threshold is chosen so small that almost every residual is truncated, that is, if κ is far below the residual scale. This is exactly what the tuning rule $\kappa \approx 1.5\widehat{\sigma}_e$ of Section 4.6 prevents. With the nominal settings $p_m = 20\%$, $p_o = 2\%$, and $\kappa = 1$, the tail probability ν is negligible, so (4.43) predicts $\lambda_G \gtrsim (1 - p_o)\lambda_0 \approx 0.98\lambda_0$; the values measured directly from the simulation in Section 6.12 give a ratio λ_G/λ_0 of 0.99 at the nominal setting, in close agreement with that prediction, and still above 0.94 under 8% impulsive contamination and at 50% missing outputs, so the robust score costs at most a few percent of the available excitation. Assumption 4.3 is thus comfortably satisfied in the regime studied here, and (4.42)–(4.43) indicate precisely where it would start to fail: extreme dropout, $p_m \rightarrow 1$, or a threshold set far below the noise level.

Assumption 4.4. *The adaptive variables satisfy*

$$p_{\min} \leq p(t) \leq p_{\max}, \quad \frac{1}{2} < \varepsilon_{\min} \leq \varepsilon(t) \leq \varepsilon_{\max} \leq 1. \quad (4.44)$$

The design gain satisfies the small-gain restriction

$$0 < \mu < \min \left\{ \frac{1}{\bar{G}}, \frac{\lambda_G}{4\bar{G}^2} \right\}. \quad (4.45)$$

Theorem 4.1 (Mean-square boundedness of the RAMF-HMISG recursion). *Under Assumptions 4.1–4.4, there exist positive constants C_1, C_2, C_3, C_4 such that*

$$\mathbb{E} \left[\left\| \widehat{\Theta}(t) - \Theta \right\|_F^2 + \sum_{i=1}^m \left\| \widehat{\alpha}_i(t) - \alpha_i \right\|^2 \right] \leq C_1 \exp(-C_2 \mathcal{S}_t) \left[\left\| \widehat{\Theta}(0) - \Theta \right\|_F^2 + \sum_{i=1}^m \left\| \widehat{\alpha}_i(0) - \alpha_i \right\|^2 \right] + C_3 \bar{\sigma}^2 + C_4 \bar{\Delta}^2, \quad (4.46)$$

where

$$\mathcal{S}_t = \sum_{s=1}^t \frac{\mu}{r^{\varepsilon(s)}(s)}. \quad (4.47)$$

Consequently, the estimates are mean-square bounded. The decay factor is exponential in the accumulated step size $\mathcal{S}_t$, not necessarily exponential in the physical time t . If $r(t)$ is linear in t , the decay is polynomial for $\varepsilon(t) = 1$ and stretched-exponential for $1/2 < \varepsilon(t) < 1$. If $\bar{\sigma}^2 = \bar{\Delta}^2 = 0$, the bound implies convergence of the mean-square estimation error to zero.

Proof. The proof is given in detail because the recursion contains masking, filtering, Huber scoring, and a multi-innovation window. The argument is a standard stochastic-approximation boundedness proof under the stated effective-excitation assumptions [5, 6, 9].

For any sample s inside the window, (4.33) and (4.31) give

$$\mathcal{E}_i^{(\ell)}(p, t) = y_{f,i}(s) - \Phi_{a,i}^T(s) \widehat{\vartheta}_i(t-1), \quad s = t - \ell, \quad (4.48)$$

$$= -\Phi_{a,i}^T(s) \widetilde{\vartheta}_i(t-1) + \xi_i(s) + \Delta_i(s). \quad (4.49)$$

Thus the deterministic part of the raw innovation is a negative projection of the parameter error, while the stochastic part is the filtered innovation plus the reconstruction/filter mismatch.

The Huber score is monotone and globally Lipschitz with a generalized derivative in $[0, 1]$. Therefore, for each s , there is a generalized slope $q_i(s) \in [0, 1]$ such that, with $w_i(s) = \xi_i(s) + \Delta_i(s)$ and $z_i(s) = \Phi_{a,i}^T(s) \widetilde{\vartheta}_i(t-1)$,

$$\chi_\kappa(w_i(s) - z_i(s)) = \chi_\kappa(w_i(s)) - q_i(s)z_i(s). \quad (4.50)$$

Multiplying by the filtered reliability mask and stacking over the window gives

$$\Phi_{a,i}(p, t) \bar{\mathcal{E}}_i(p, t) = -G_i(t) \widetilde{\vartheta}_i(t-1) + d_i(t), \quad (4.51)$$

where $G_i(t)$ is defined in (4.40) and

$$d_i(t) = \sum_{\ell=0}^{p(t)-1} \gamma_{f,i}(t-\ell) \Phi_{a,i}(t-\ell) \chi_\kappa(\xi_i(t-\ell) + \Delta_i(t-\ell)). \quad (4.52)$$

This is the precise form behind the informal “robust noise score plus mismatch” term.

The update of the augmented vector $\widehat{\vartheta}_i(t)$ is exactly

$$\widehat{\vartheta}_i(t) = \widehat{\vartheta}_i(t-1) + a_t \Phi_{a,i}(p, t) \bar{\mathcal{E}}_i(p, t), \quad \text{where } a_t \text{ is defined in (4.11).} \quad (4.53)$$

Substituting (4.51) into (4.53) yields

$$\widetilde{\vartheta}_i(t) = [I - a_t G_i(t)] \widetilde{\vartheta}_i(t-1) + a_t d_i(t). \quad (4.54)$$

The Gram structure in $G_i(t)$ is the same structure exposed by Lemma 4.1. The lemma therefore connects the residual correction identity to the contraction term in the estimation-error dynamics; the precise relation between the two, including the sense in which they coincide when no residual is truncated, is established in Remark 4.4.

Since $\|\Phi_{a,i}(t)\| \leq \bar{\phi}$, $p(t) \leq p_{\max}$, $0 \leq \gamma_{f,i} \leq 1$, and $|\chi_\kappa(x)| \leq \kappa$, the vector $d_i(t)$ has a finite second moment. More explicitly, using $(a+b)^2 \leq 2a^2 + 2b^2$, the martingale score condition, and Assumption 4.2, there is a constant $C_d > 0$, independent of t , such that

$$\mathbb{E}\{\|d_i(t)\|^2 \mid \mathcal{F}_{t-1}\} \leq C_d(\bar{\sigma}^2 + \bar{\Delta}^2). \quad (4.55)$$

The conditional mean of the pure-noise part is zero by Assumption 4.1. The possible bias caused by $\Delta_i(t)$ is bounded by Assumption 4.2; when a cross term with $\widetilde{\vartheta}_i(t-1)$ appears, it is controlled by Young’s inequality and absorbed into the contraction for sufficiently small μ , with a remaining term proportional to $\bar{\Delta}^2$.

Taking squared norms in (4.54) gives

$$\|\widetilde{\vartheta}_i(t)\|^2 = \|[I - a_t G_i(t)] \widetilde{\vartheta}_i(t-1)\|^2 + 2a_t \widetilde{\vartheta}_i^T(t-1) [I - a_t G_i(t)]^T d_i(t) + a_t^2 \|d_i(t)\|^2. \quad (4.56)$$

Because $G_i(t)$ is positive semidefinite and $\|G_i(t)\| \leq \bar{G}$,

$$\|[I - a_t G_i(t)] x\|^2 \leq x^T [I - 2a_t G_i(t) + a_t^2 \bar{G} G_i(t)] x. \quad (4.57)$$

The small-gain condition (4.45) and $a_t \leq \mu$ imply $a_t \bar{G} \leq 1$. Hence, after conditioning on $\mathcal{F}_{t-1}$ and using (4.39),

$$\mathbb{E}\left\{\|[I - a_t G_i(t)] \widetilde{\vartheta}_i(t-1)\|^2 \mid \mathcal{F}_{t-1}\right\} \leq (1 - a_t \lambda_G) \|\widetilde{\vartheta}_i(t-1)\|^2. \quad (4.58)$$

The cross term in (4.56) is zero for the conditionally centered score part. For the mismatch part, Young’s inequality gives, for any $0 < \eta < \lambda_G$,

$$2a_t \left| \widetilde{\vartheta}_i^T(t-1) \mathbb{E}\{d_i(t) \mid \mathcal{F}_{t-1}\} \right| \leq a_t \eta \|\widetilde{\vartheta}_i(t-1)\|^2 + C_\eta a_t \bar{\Delta}^2. \quad (4.59)$$

Choosing $\eta = \lambda_G/2$ and using (4.55), there are constants $c_i > 0$ and $C_i > 0$ such that

$$\mathbb{E}\{\|\widetilde{\vartheta}_i(t)\|^2\} \leq (1 - c_i a_t) \mathbb{E}\{\|\widetilde{\vartheta}_i(t-1)\|^2\} + C_i a_t^2 \bar{\sigma}^2 + C_i (a_t^2 + a_t) \bar{\Delta}^2. \quad (4.60)$$

This inequality is intentionally conservative. It separates the zero-mean stochastic score, whose contribution is scaled by a_t^2 , from the bounded deterministic mismatch, whose contribution produces an ultimate term proportional to $\bar{\Delta}^2$.

Define the global Lyapunov sequence

$$V(t) = \sum_{i=1}^m \mathbb{E}\{\|\tilde{\vartheta}_i(t)\|^2\} = \mathbb{E}\left[\|\widehat{\Theta}(t) - \Theta\|_F^2 + \sum_{i=1}^m \|\widehat{\alpha}_i(t) - \alpha_i\|^2\right]. \quad (4.61)$$

Let $c = \min_i c_i$ and absorb all constants into C . Summing (4.60) over i gives

$$V(t) \leq (1 - ca_t)V(t-1) + Ca_t^2\bar{\sigma}^2 + C(a_t^2 + a_t)\bar{\Delta}^2. \quad (4.62)$$

From (4.14), Assumptions 4.1 and 4.4, and $\epsilon_r > 0$, there exist constants $0 < c_r < C_r$ such that

$$c_r t \leq r(t) \leq C_r t, \quad t \geq 1. \quad (4.63)$$

Indeed, the lower bound follows from the accumulated ϵ_r term, and the upper bound follows from the bounded regressors and bounded window length. Therefore

$$\sum_{t=1}^{\infty} a_t = \infty, \quad \sum_{t=1}^{\infty} a_t^2 < \infty, \quad (4.64)$$

because $\varepsilon(t) \leq 1$ gives $a_t \geq \mu/(C_r t)$ up to a constant, whereas $\varepsilon(t) \geq \varepsilon_{\min} > 1/2$ gives $a_t^2 \leq C/t^{2\varepsilon_{\min}}$.

Iteration of the scalar recursion. Applying (4.62) recursively yields

$$\begin{aligned} V(t) &\leq \left[\prod_{s=1}^t (1 - ca_s) \right] V(0) + C\bar{\sigma}^2 \sum_{s=1}^t a_s^2 \prod_{j=s+1}^t (1 - ca_j) \\ &\quad + C\bar{\Delta}^2 \sum_{s=1}^t (a_s^2 + a_s) \prod_{j=s+1}^t (1 - ca_j). \end{aligned} \quad (4.65)$$

Using $1 - x \leq e^{-x}$ for $x \geq 0$,

$$\prod_{s=1}^t (1 - ca_s) \leq \exp\left(-c \sum_{s=1}^t a_s\right) = \exp(-c\mathcal{S}_t). \quad (4.66)$$

The series involving a_s^2 is bounded because $\sum_s a_s^2 < \infty$. The series involving a_s and the exponential tail is bounded by a constant multiple of $1/c$; this is the standard discrete convolution estimate for a stable scalar stochastic-approximation recursion. Consequently, there exist constants $C_1, C_2, C_3, C_4 > 0$ such that (4.46) holds. If $\bar{\sigma}^2 = \bar{\Delta}^2 = 0$, the last two terms vanish and $V(t) \rightarrow 0$ because $\mathcal{S}_t \rightarrow \infty$. The proof is complete. $\square$

Remark 4.4 (The $p \times p$ Gram matrix of Lemma 4.1 versus the gain matrix $G_i(t)$ of the error recursion). The correction matrix appearing in the residual identity (4.24) has dimension $p \times p$, whereas the contraction generator of the parameter-error recursion (4.54) has dimension $(n_i + n) \times (n_i + n)$ and carries the Huber slopes. The two are neither equal nor of the same size, and it is worth stating precisely how they are related, because the relation determines how residual shrinkage and parameter-error shrinkage differ during the transient.

They are the two Gram matrices of one and the same object. Stacking the windowed augmented regressors as in (4.32) gives $\Phi_{a,i}(p, t) \in \mathbb{R}^{(n_i+n) \times p}$, and a direct comparison with (4.2)–(4.3) shows that the bracket of (4.24) is exactly

$$\Psi_{f,i}^T(p, t) \Psi_{f,i}(p, t) + \Phi_f^T(p, t) \Phi_f(p, t) = \Phi_{a,i}^T(p, t) \Phi_{a,i}(p, t) \in \mathbb{R}^{p \times p}, \quad (4.67)$$

while (4.40) can be written compactly as

$$G_i(t) = \Phi_{a,i}(p, t) D_i(t) \Phi_{a,i}^T(p, t) \in \mathbb{R}^{(n_i+n) \times (n_i+n)}, \quad D_i(t) = \text{diag}\{q_i(t - \ell) \gamma_{f,i}(t - \ell)\}_{\ell=0}^{p-1}. \quad (4.68)$$

The first is the Gram matrix in data space, indexed by the p samples of the innovation window; the second is the weighted Gram matrix in parameter space, indexed by the $n_i + n$ unknowns of channel i . A rectangular matrix always possesses these two Gram matrices, they always have different sizes, and they always have the same rank, $\text{rank } \Phi_{a,i}^T \Phi_{a,i} = \text{rank } \Phi_{a,i} = \text{rank } \Phi_{a,i} \Phi_{a,i}^T \leq \min\{p, n_i + n\}$. There is therefore no inconsistency: the same finite set of excited directions is contracted in both recursions, and only the coordinates in which they are expressed, and the weighting applied to them, differ.

In the linear untruncated case, the two coincide exactly. Suppose the whole window is reliable and no residual is truncated, so that $\gamma_{f,i}(t - \ell) = 1$ and $q_i(t - \ell) = 1$ for all ℓ , hence $D_i(t) = I_p$ and $\bar{\mathcal{E}}_i = \mathcal{E}_i$. Then $G_i(t) = \Phi_{a,i} \Phi_{a,i}^T$, and the two Gram matrices share the same nonzero spectrum: if $\Phi_{a,i}^T \Phi_{a,i} u = \lambda u$ with $\lambda > 0$, then $w = \Phi_{a,i} u / \sqrt{\lambda}$ is a unit eigenvector of $\Phi_{a,i} \Phi_{a,i}^T$ for the same λ . More than a spectral coincidence, the two recursions become the same operator viewed through $\Phi_{a,i}^T$, because of the elementary intertwining identity

$$[I_p - a_t \Phi_{a,i}^T \Phi_{a,i}] \Phi_{a,i}^T = \Phi_{a,i}^T [I_{n_i+n} - a_t \Phi_{a,i} \Phi_{a,i}^T]. \quad (4.69)$$

Substituting the noiseless form $\mathcal{E}_i(p, t) = -\Phi_{a,i}^T \tilde{\vartheta}_i(t - 1)$ from (4.49) into (4.24) and using (4.69) gives $\widehat{V}_i(p, t) = -\Phi_{a,i}^T [I - a_t G_i(t)] \tilde{\vartheta}_i(t - 1)$, that is, the post-update residual is the image under $\Phi_{a,i}^T$ of the post-update parameter error. Residual and parameter error then shrink by identical factors $1 - a_t \lambda$ along corresponding singular directions, which is the sense in which the two structures become consistent in the untruncated case.

Here is where the transient behaviors diverge. Outside that case, $D_i(t) \leq I_p$ with strict inequality on truncated or unreliable samples, so

$$G_i(t) = \Phi_{a,i} D_i \Phi_{a,i}^T \leq \Phi_{a,i} \Phi_{a,i}^T, \quad \lambda_{\min}(G_i(t)) \geq \min_{\ell} q_i(t - \ell) \lambda_{\min}(\Phi_{a,i} \Gamma_i \Phi_{a,i}^T), \quad (4.70)$$

and (4.69) no longer applies. The practical consequence is asymmetric and, at first sight, counter-intuitive: the residual identity (4.24) always subtracts a correction generated by the unweighted Gram matrix, whereas the parameter error is contracted only by the down-weighted one. During the transient, when the parameter error is large, the innovations are large, the Huber slopes q_i are well below one, and the parameter error therefore contracts strictly more slowly than the residual appears to shrink. Monitoring the residual during the transient consequently overstates the progress actually made in parameter space, and this is precisely why the relative parameter error $\delta(t)$ of (6.4), rather than a residual norm, is used as the performance measure throughout Section 6. Missing data acts in the same direction through the factor $\gamma_{f,i}$, which deletes rows of D_i and lowers $\lambda_{\min}(G_i)$ by the factor quantified in (4.42). As the estimates improve, the residuals fall inside the linear region on all but the genuinely

impulsive samples, $D_i(t)$ converges to $\text{diag}\{\gamma_{f,i}(t - \ell)\}$, and the two structures become consistent again on the reliable part of the window. The measured slope ratios of Section 6.12, which exceed 0.94 in every configuration tested, quantify how small the residual discrepancy is once the transient is over.

Remark 4.5 (Tightness of the ultimate bound and the role of the mismatch term). The bound (4.46) contains two very different error contributions, and separating their asymptotic weight clarifies both what the theorem guarantees and how conservative it is.

Asymptotic size of each term. Setting $V(t) = V(t - 1) = V_\infty$ in the one-step inequality (4.62) gives the stationary level

$$V_\infty \leq \frac{C}{c} \left[a_t \bar{\sigma}^2 + (1 + a_t) \bar{\Delta}^2 \right] \xrightarrow{a_t \rightarrow 0} \frac{C}{c} \bar{\Delta}^2. \quad (4.71)$$

The stochastic score contributes through a_t^2 and is therefore annihilated by the vanishing step size: the term $C_3 \bar{\sigma}^2$ in (4.46) is a transient budget, not a floor. The mismatch contributes through a_t and survives the division by ca_t : the term $C_4 \bar{\Delta}^2$ is the genuine ultimate floor of the recursion. Since c is proportional to λ_G , combining (4.71) with (4.43) exhibits the dependence of the floor on the operating conditions,

$$V_\infty \lesssim \frac{C \bar{\Delta}^2}{(1 - p_o - \nu)(1 - p_m)^2 \lambda_0(0)}, \quad (4.72)$$

so the attainable accuracy degrades quadratically in the availability probability and linearly in the contamination level, while it is proportional to the mismatch energy. This is the quantitative answer to the question of how the reconstruction and filter mismatch limit the estimation accuracy: it does not affect the rate of the exponential factor $\exp(-C_2 S_t)$, it fixes the level at which that decay stops.

Sources of conservatism. Four steps of the proof are deliberately loose. Young's inequality applied with $\eta = \lambda_G/2$ sacrifices a factor two in the contraction; the estimate $\|G_i(t)\| \leq \bar{G}$ replaces the whole spectrum of $G_i(t)$ by its largest value, which is pessimistic whenever the windowed Gram matrix is well-conditioned; the cross term is bounded by its worst case although its zero-mean part vanishes exactly; and, most importantly, $\bar{\Delta}^2$ is a uniform-in-time bound, whereas the three constituents identified in (4.36) all decrease as $\hat{\rho}(t)$ settles and the prediction error shrinks. The measurements of Section 6.12 show that the mismatch energy falls by at least a factor of four between the early transient and the steady state in all but the most heavily contaminated configuration, so a bound built on $\sup_t \bar{\Delta}^2(t)$ overestimates the true floor by a comparable factor. Theorem 4.1 should accordingly be read as a qualitative stability certificate—the estimates remain in a noise- and mismatch-dependent neighborhood whose size is controlled by quantities the designer can influence—and not as a sharp prediction of the achievable error; the observed relative errors reported in Section 6 are well inside the region it delimits.

4.6. Implementation and tuning guidelines

The algorithm exposes a number of constants, and a practitioner needs to know which of them matter, in what order they should be fixed, and what a safe range is. The constants fall into three groups. The first group has a direct effect on accuracy and must be tuned: the Huber threshold κ and the gain μ . The second group shapes the adaptation and admits scale-free default rules: $p_{\min}$, p_{mid} , $p_{\max}$, $\varepsilon_{\min}$, $\varepsilon_{\max}$, η_1 , η_2 , η_ε , and λ_ζ . The third group consists of numerical safeguards that only prevent division by zero or overflow and may be left at their default values over many orders of magnitude: ϵ_r , ϵ_0 , ϵ_f , and ϵ_p . A workable procedure is the following.

- (1) *Scale first.* Run a few hundred samples with $\kappa = \infty$ and $\varepsilon = 0.6$, and record the median absolute residual $\widehat{\sigma}_e = 1.4826 \text{ median}|e_i(t)|$. All remaining thresholds are expressed in units of $\widehat{\sigma}_e$.
- (2) *Huber threshold.* Take $\kappa \approx 1.5\widehat{\sigma}_e$. Values in $[1, 2]\widehat{\sigma}_e$ retain most of the Gaussian efficiency while bounding the score; values far below $\widehat{\sigma}_e$ truncate almost every residual and, by (4.43), erode the effective excitation λ_G .
- (3) *Gain.* Increase μ from a small value until the transient no longer accelerates. The theoretical ceiling is the small-gain condition (4.45); in practice, the useful range is broad, and the safeguard $a_t \leq a_{\max}$ used in the implementation makes the recursion insensitive to an overestimate of μ . Note that μ is not dimensionless: Because $r(t)$ accumulates regressor energy, the appropriate μ scales with the data, so it must be re-tuned when the signal units change.
- (4) *Window candidates.* Use the band $p \in [d_i/2, 2d_i]$ with $d_i = n_i + n$, and place the three candidates near its lower end, its middle, and its upper end.
- (5) *Convergence index.* Use $[\varepsilon_{\min}, \varepsilon_{\max}] = [0.55, 0.65]$, or any narrow interval strictly inside $(1/2, 1]$.
- (6) *Switching thresholds and smoothing.* Use (4.22), that is, $\eta_1 \approx 0.1\kappa^2$, $\eta_2 \approx 0.5\kappa^2$, $\eta_\varepsilon \approx 0.25\kappa^2$, and λ_ζ such that $1/(1 - \lambda_\zeta)$ covers a few tens of samples. Verify afterward that the fraction of time spent at each window level is not degenerate, as explained in Remark 4.1.
- (7) *Filter window.* Choose L_f of the order of a few tens of samples and $\rho_{\max}$ slightly below one. Both are non-critical, as the sensitivity results of Section 6.11 show.

Table 3 collects these rules. The corresponding numerical values used in the simulations are listed in Section 6.1, and the measured sensitivity of the final error to each constant is reported in Section 6.11.

Table 3. Tuning guidelines for the constants of the RAMF-HMISG algorithm.

Constant	Role	Recommended rule	Criticality
κ	robust score threshold	$\approx 1.5\widehat{\sigma}_e$ (robust residual scale)	high
μ	update gain	largest value that does not degrade the steady state; see (4.45)	high
$p_{\min}, p_{\text{mid}}, p_{\max}$	innovation-window candidates	inside $[d_i/2, 2d_i]$, $d_i = n_i + n$	moderate
$\varepsilon_{\min}, \varepsilon_{\max}$	convergence index range	narrow interval strictly inside $(1/2, 1]$	moderate
η_1, η_2	window switching thresholds	$0.1\kappa^2$ and $0.5\kappa^2$	low
η_ε	shape constant of (4.16)	$\approx 0.25\kappa^2$	low
λ_ζ	energy smoothing	$1/(1 - \lambda_\zeta)$ of the order of $10\text{--}10^2$ samples	low
$L_f, \kappa_\rho, \rho_{\max}$	robust filter estimator	$L_f \sim 10^2$, $\kappa_\rho \sim \widehat{\sigma}_e$, $\rho_{\max} < 1$	low
$\epsilon_r, \epsilon_0, \epsilon_f, \epsilon_\rho$	numerical safeguards	any small positive value	negligible

5. Computational efficiency comparison

Let $n_0 = \sum_{i=1}^m n_i$. The dominant RAMF-HMISG operations are the products $\Phi_f(p, t)\bar{\mathcal{E}}^T(p, t)$ and $\Psi_{f,i}(p, t)\bar{\mathcal{E}}_i(p, t)$, plus the construction of masks, scores, and filtered variables. The order is

$$O(p(t)mn + p(t)n_0 + p(t)m + n + n_0). \quad (5.1)$$

The rolling-buffer memory requirement is

$$O(mn + n_0 + p(t)m + p(t)n + p(t)n_0). \quad (5.2)$$

A subsystem RLS implementation propagates m covariance matrices of sizes $(n + n_i) \times (n + n_i)$, with dominant order $O(\sum_i (n + n_i)^2)$ and memory of the same quadratic order.

5.1. Scalability of the adaptive mechanisms

The orders listed in Table 4 are stated in terms of p , m , n , and n_0 ; to discuss scalability, they must be resolved in terms of the system size. Consider the representative family $m = r$ with $n_i = n_j = q$ for all i, j , so that $n = mq$ and $n_0 = mq$. Substituting into the expressions above gives, for the proposed algorithm,

$$\text{time } O(p m^2 q), \quad \text{memory } O(m^2 q + p m q), \quad (5.3)$$

whereas the subsystem RLS implementation requires

$$\text{time } O\left(\sum_i (n + n_i)^2\right) = O(m^3 q^2), \quad \text{memory } O(m^3 q^2). \quad (5.4)$$

The proposed recursion is therefore quadratic in the number of channels where a covariance-based method is cubic, and the gap widens with q . This is the structural reason why the hierarchical formulation remains attractive for larger multivariable systems.

Table 4. Computational characteristics of recursive identification algorithms.

Algorithm	Main update	Memory requirement	Dominant operation order
Subsystem RLS	covariance gain update for each output	$\sum_i (n + n_i)^2$	$O\left(\sum_i (n + n_i)^2\right)$
HMISG	multi-innovation gradient update	$mn + n_0 + p(m + n + n_0)$	$O(pmn + pn_0)$
RM-HMISG	masked Huber-score update	$mn + n_0 + p(m + n + n_0)$	$O(pmn + pn_0 + pm)$
RAMF-HMISG	masked, robust, filtered, and adaptive update	$mn + n_0 + p(m + n + n_0)$	$O(pmn + pn_0 + pm + n + n_0)$

The more delicate question is whether the additional mechanisms of RAMF-HMISG—the robust score, the reliability mask, the residual-based filter estimate, and the two adaptive laws—become a bottleneck as the system grows. They do not, and the reason is visible in the orders. Selecting $p(t)$ and $\varepsilon(t)$ costs $O(1)$ per sample, updating $\zeta(t)$ costs $O(pm)$, and evaluating the Huber score and the mask costs $O(pm)$. The only mechanism whose cost is not negligible is the filter estimate (2.25), which sweeps a residual window of length L_f and costs $O(L_f m)$ per sample. The total overhead is thus $O(pm + L_f m)$, which is linear in m , against a core cost that is quadratic in m . Consequently, the relative overhead behaves like

$$\frac{\text{overhead}}{\text{core}} = O\left(\frac{p + L_f}{p m q}\right) \quad (5.5)$$

and decreases as the system grows. The additional computational burden of the adaptive machinery is therefore most visible on small systems—exactly the two-input two-output case studied in Section 6, where it is largest—and becomes progressively less significant on the larger problems for which the question of scalability is actually raised. Table 5 and Figure 1 confirm this prediction experimentally: measured over channel counts from two to six with $q = 3$, the ratio between the execution time of

RAMF-HMISG and that of RM-HMISG falls monotonically, from more than two at the smallest size to about one and a half at the largest, while the proposed method remains the more accurate of the two at every dimension tested. If the overhead must be reduced further, (5.5) indicates the correct lever: the filter estimate may be refreshed every N_p samples instead of every sample, which divides the dominant overhead term by N_p without affecting the update itself, since $\hat{\rho}$ varies slowly by construction.

Table 5. Scalability in the system dimension. Systems with $m = r$ channels and $n_i = n_j = 3$ are identified from $L_s = 1000$ samples with 20% missing outputs. The “Ratio” column is the execution time of RAMF-HMISG divided by that of RM-HMISG and measures the cost of the filtering and adaptation mechanisms; it decreases with the dimension, as predicted by (5.5). Because the number of unknowns grows while the record length is held fixed, the errors of both methods grow with the dimension; the comparison between columns, not the absolute level, is the quantity of interest. The table shows the mean over 3 runs.

Size $m \times r$	Unknowns	RM-HMISG (s)	RAMF-HMISG (s)	Ratio	RM-HMISG δ (%)	RAMF-HMISG δ (%)
2×2	18	1.35	3.09	2.29	15.5	11.1
3×3	36	1.74	3.73	2.14	26.4	15.7
4×4	60	2.27	4.40	1.94	59.8	37.2
5×5	90	2.87	4.82	1.68	64.5	35.0
6×6	126	3.20	4.94	1.54	77.1	52.5

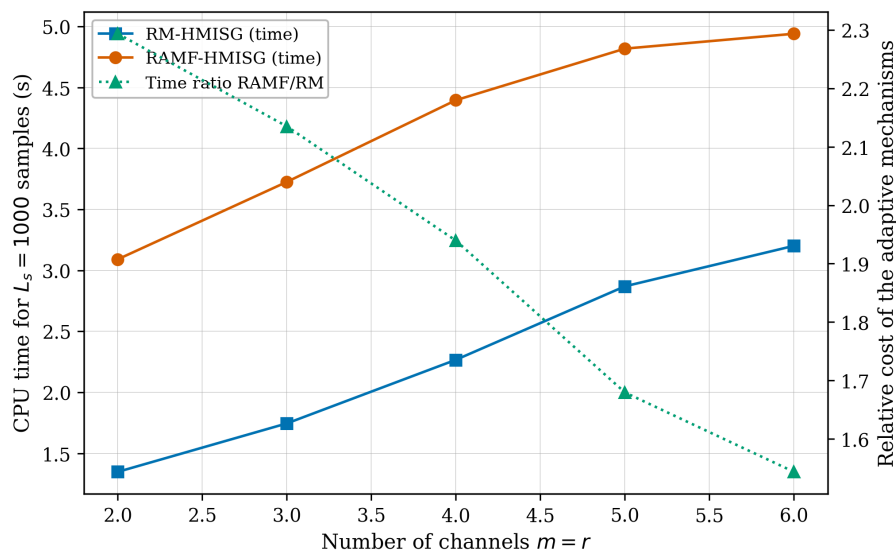


Figure 1. Execution time of RM-HMISG and RAMF-HMISG for $L_s = 1000$ samples as a function of the number of channels and their ratio. The ratio measures the cost of the filtering and adaptation mechanisms relative to the core recursion and decreases with the system size, as predicted by (5.5).

The second scalability axis is the innovation length. Table 6 reports execution time and final accuracy for a three-input three-output system as p is increased from 5 to 40 at fixed dimension.

The execution time grows markedly more slowly than p itself because the per-sample work also contains terms independent of the window, while the accuracy improves with diminishing returns. This is the practical form of the trade-off analyzed in Section 4.3: enlarging the window buys transient speed at a linear price in computation and with a saturating benefit, which is precisely the reason the proposed algorithm selects the window online rather than committing to a single conservative value. The adaptive rule spends the long window only while the innovation energy justifies it and falls back to $p_{\min}$ afterward, so it obtains the accuracy of a long-window method at an average cost closer to that of a short-window one.

Table 6. Accuracy–complexity trade-off in the innovation length, for a three-input three-output system with $n_i = n_j = 3$ and a fixed window. Execution time is normalized by the value at $p = 5$. This table shows the mean over 3 runs of length $L_s = 1000$.

p	CPU time (s)	Relative time	$\delta(L_s)$ (%)
5	2.99	1.00	30.87
10	3.79	1.27	22.21
20	5.12	1.71	16.46
40	7.21	2.41	14.51

6. Simulation experiments

6.1. Simulation system and settings

Consider a two-input two-output ARX system with $m = 2$, $r = 2$, $n_1 = n_2 = 2$, and two input lags. The true polynomials are

$$A_1(z) = 1 - 0.30z^{-1} - 0.10z^{-2}, \quad A_2(z) = 1 - 0.40z^{-1} - 0.20z^{-2}, \quad (6.1)$$

$$B_{11}(z) = 0.60z^{-1} + 1.20z^{-2}, \quad B_{12}(z) = 1.40z^{-1} + 1.70z^{-2}, \quad (6.2)$$

$$B_{21}(z) = 0.70z^{-1} + 1.18z^{-2}, \quad B_{22}(z) = 1.35z^{-1} + 1.50z^{-2}. \quad (6.3)$$

The inputs are independent zero-mean unit-variance persistently exciting sequences. The colored noise is $v_i(t) = 0.45v_i(t-1) + \xi_i(t)$, where $\xi_i(t)$ is Gaussian with variance 0.25^2 and impulsive contamination. The missing-output probability is 20% in the main experiment.

The true coefficients are collected in Tables 7 and 8 lists the complete set of simulation and algorithmic constants, including the auxiliary constants of the adaptive laws and of the robust filter estimator, so that every experiment reported below is reproducible from the table alone.

Table 7. True parameters of the two-input two-output ARX system used in the simulation example.

$a_1(1)$	$a_1(2)$	$b_{11}(1)$	$b_{11}(2)$	$b_{12}(1)$	$b_{12}(2)$	$a_2(1)$	$a_2(2)$	$b_{21}(1)$	$b_{21}(2)$	$b_{22}(1)$	$b_{22}(2)$
-0.30000	-0.10000	0.60000	1.20000	1.40000	1.70000	-0.40000	-0.20000	0.70000	1.18000	1.35000	1.50000

Table 8. Simulation and algorithmic settings. The lower block lists the auxiliary constants of the adaptive laws, of the robust filter estimator, and of the numerical safeguards; the switching thresholds are expressed as fractions of κ^2 in accordance with the scale-free guideline (4.22).

Item	Value
Data length	$L_e = 5000$
Input signals	two independent zero-mean unit-variance persistent excitation sequences
Colored noise	$v_i(t) = 0.45v_i(t-1) + \xi_i(t)$
Innovation noise	$\xi_i(t) \sim \mathcal{N}(0, 0.25^2)$ with impulsive contamination
Outlier probability	2% in each output channel
Missing-output probability	20% in each output channel
Reliability mask	$\gamma_{f,i}(t) = \gamma_i(t)\gamma_i(t-1)$ for filtered samples
Robust score	Huber score with $\kappa = 1$
Auxiliary filter	robust residual-based $\widehat{\rho}(t)$, $\rho_{\max} = 0.9$, $L_f = 80$
Proposed update gain	$\mu = 1.6$
Adaptive innovation length	$p(t) \in \{5, 8, 10\}$ selected from $\zeta(t-1)$
Adaptive convergence index	$\varepsilon(t) \in [0.56, 0.66]$ selected from $\zeta(t-1)$
<i>Auxiliary constants of the adaptive laws, of the robust filter estimator, and numerical safeguards</i>	
Window switching thresholds	$\eta_1 = 0.12$, $\eta_2 = 0.55$ (i.e., $0.12\kappa^2$ and $0.55\kappa^2$)
Shape constant of (4.16)	$\eta_\varepsilon = 0.25$
Energy smoothing factor	$\lambda_\zeta = 0.98$ (memory $1/(1 - \lambda_\zeta) = 50$ samples)
Robust weight scale of (2.25)	$\kappa_\rho = 1$
Filter-estimate smoothing	$\lambda_p = 0.985$
Increment floor of $r(t)$	$\epsilon_r = 0.0001$
Normalization floor of $\zeta(t)$	$\epsilon_0 = 1e - 09$
Denominator floor of (2.25)	$\epsilon_f = 1e - 10$
Floor inside $\omega_p(s)$	$\epsilon_p = 1e - 12$
Step-size safeguard	$a_t \leq a_{\max} = 0.025$
Initialization	$r(0) = 20$, $\zeta(0) = 1$, all parameter estimates = 0.001

The relative parameter-estimation error is

$$\delta(t) = \sqrt{\frac{\|\widehat{\Theta}(t) - \Theta\|_F^2 + \sum_{i=1}^m \|\widehat{\alpha}_i(t) - \alpha_i\|^2}{\|\Theta\|_F^2 + \sum_{i=1}^m \|\alpha_i\|^2}} \times 100\%. \quad (6.4)$$

The algorithms compared are zero-filling M-HMISG (M-HMISG-ZF), masked HMISG, robust masked HMISG (RM-HMISG), robust filtered HMISG with fixed p and ε (RF-HMISG), and the proposed RAMF-HMISG.

6.2. Parameter-estimation results

Table 9 shows that the RAMF-HMISG estimates approach the true parameters as t increases. The final relative error is below 1% even though the output data are incomplete and the disturbance is colored and impulsive.

Table 9. RAMF-HMISG estimates and relative errors under colored noise, missing outputs, and impulsive disturbances.

t	$a_1(1)$	$a_1(2)$	$b_{11}(1)$	$b_{11}(2)$	$b_{12}(1)$	$b_{12}(2)$
100	-0.50979	-0.05033	0.61980	0.70041	0.97495	1.12649
200	-0.46186	0.01758	0.61526	0.85751	1.17220	1.31090
500	-0.36275	-0.01818	0.56637	1.08412	1.37242	1.50158
1000	-0.33798	-0.06586	0.58789	1.16244	1.38961	1.59176
2000	-0.33096	-0.09562	0.59777	1.19232	1.38085	1.65343
5000	-0.28490	-0.10480	0.59278	1.19502	1.40951	1.69304
True values	-0.30000	-0.10000	0.60000	1.20000	1.40000	1.70000

t	$a_2(1)$	$a_2(2)$	$b_{21}(1)$	$b_{21}(2)$	$b_{22}(1)$	$b_{22}(2)$	$\delta(\%)$
100	-0.64393	0.00545	0.51961	0.66328	1.24031	1.09268	32.7628
200	-0.58581	-0.09856	0.54381	0.88802	1.33791	1.17479	21.9404
500	-0.47477	-0.10028	0.66795	1.03625	1.36373	1.30355	10.4651
1000	-0.44517	-0.16870	0.66663	1.10232	1.34552	1.36238	5.9127
2000	-0.42454	-0.17603	0.69633	1.15186	1.34449	1.43851	2.6947
5000	-0.39319	-0.20280	0.69202	1.17019	1.34740	1.49405	0.7487
True values	-0.40000	-0.20000	0.70000	1.18000	1.35000	1.50000	0.0000

Figure 2 shows the benefit of each correction. Zero-filling is highly biased because missing samples are treated as artificial zeros. Masking removes this bias. The Huber score improves robustness to spikes. The filtered and adaptive proposed method gives the best final accuracy.

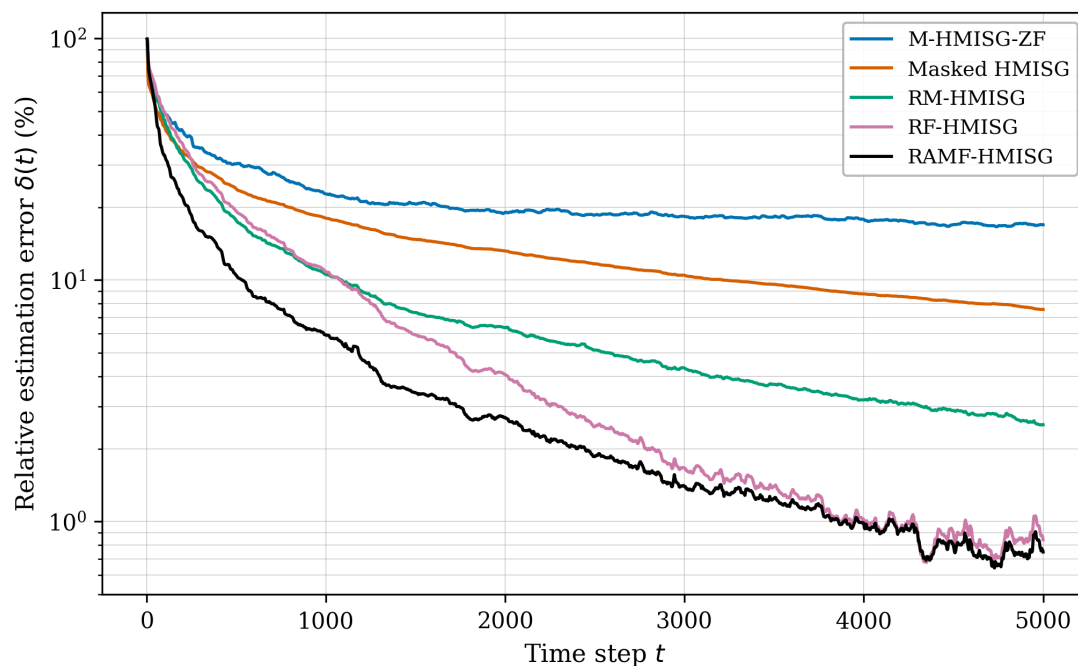


Figure 2. Relative parameter-estimation errors of the compared algorithms.

Figures 3–6 show that the autoregressive and input-related parameters converge together. The estimates are affected by the transient reconstructed regressors during the initial phase, but the final estimates are close to the true values.

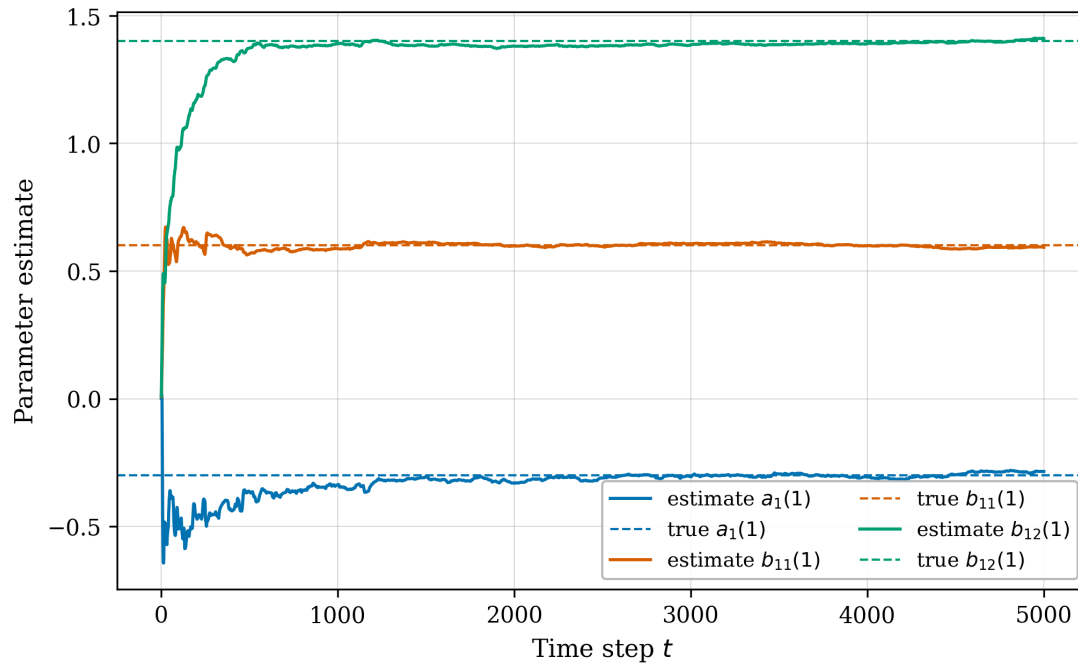


Figure 3. RAMF-HMISG estimates of $a_1(1)$, $b_{11}(1)$, and $b_{12}(1)$ versus t .

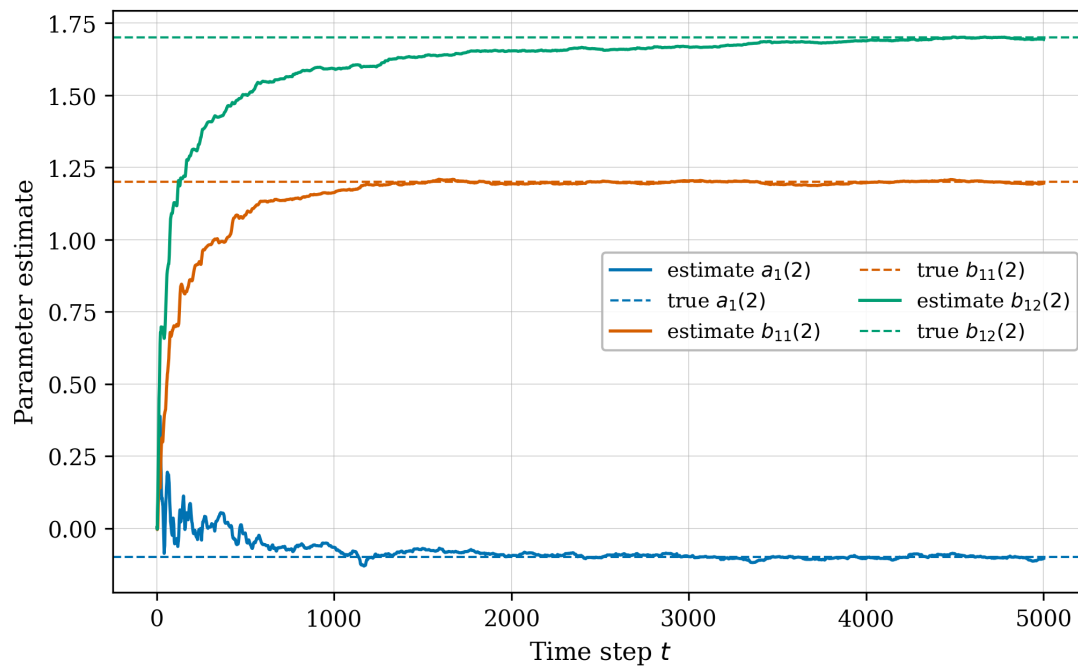


Figure 4. RAMF-HMISG estimates of $a_1(2)$, $b_{11}(2)$, and $b_{12}(2)$ versus t .

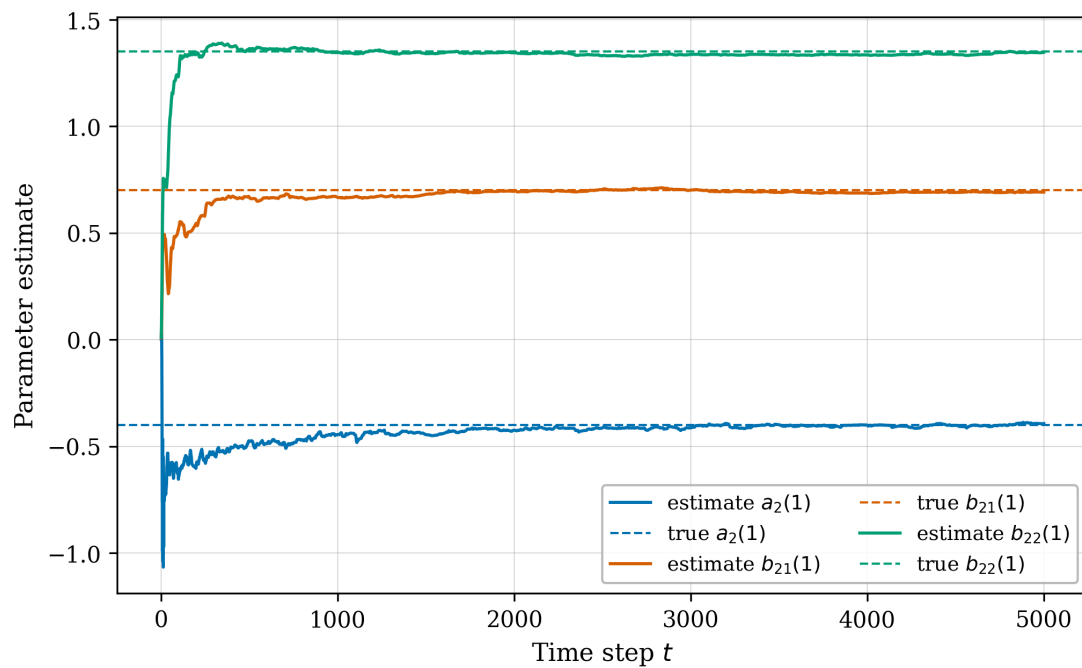


Figure 5. RAMF-HMISG estimates of $a_2(1)$, $b_{21}(1)$, and $b_{22}(1)$ versus t .

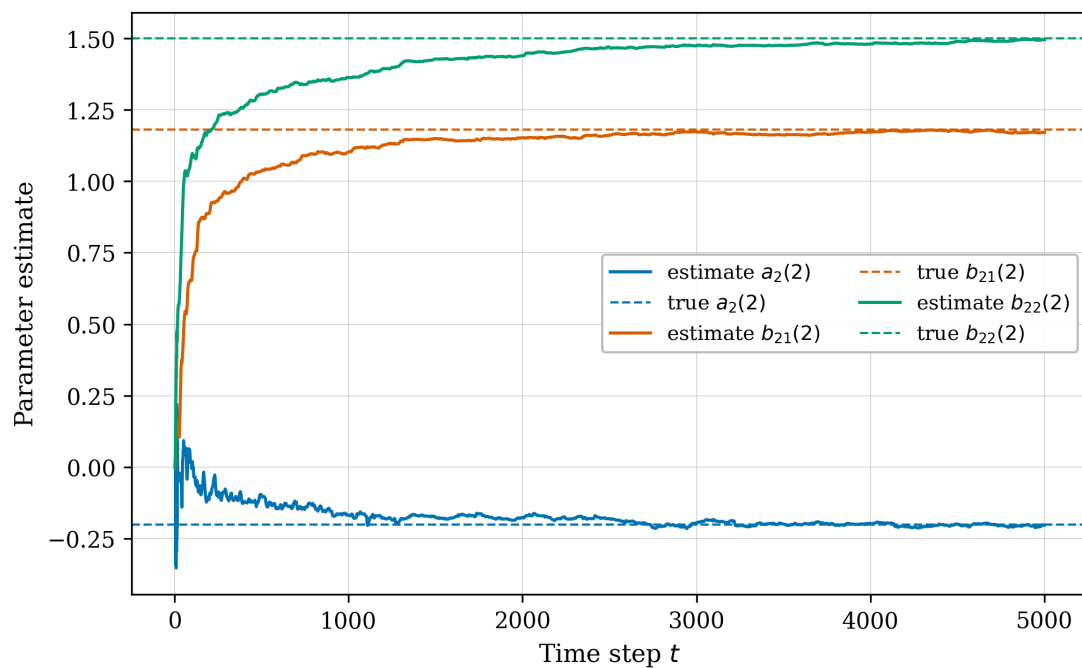


Figure 6. RAMF-HMISG estimates of $a_2(2)$, $b_{21}(2)$, and $b_{22}(2)$ versus t .

6.3. Validation and adaptive variables

Figures 7 and 8 show that the identified model reproduces unseen noiseless validation outputs. This confirms that the estimated coefficients are meaningful for input–output prediction, not only close in

the Euclidean norm.

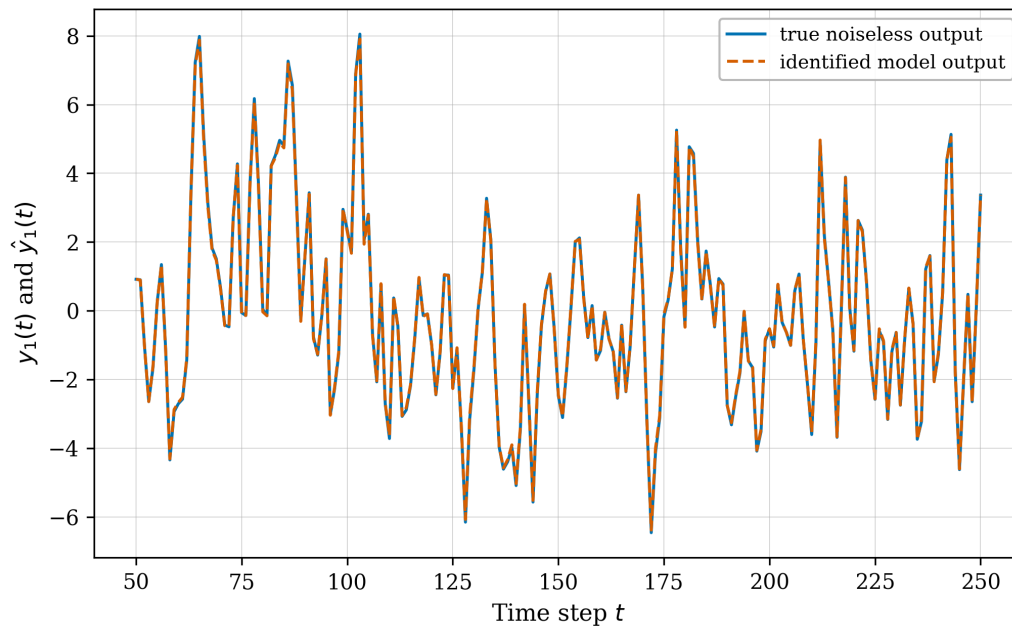


Figure 7. True validation output $y_1(t)$ and identified model output $\hat{y}_1(t)$.

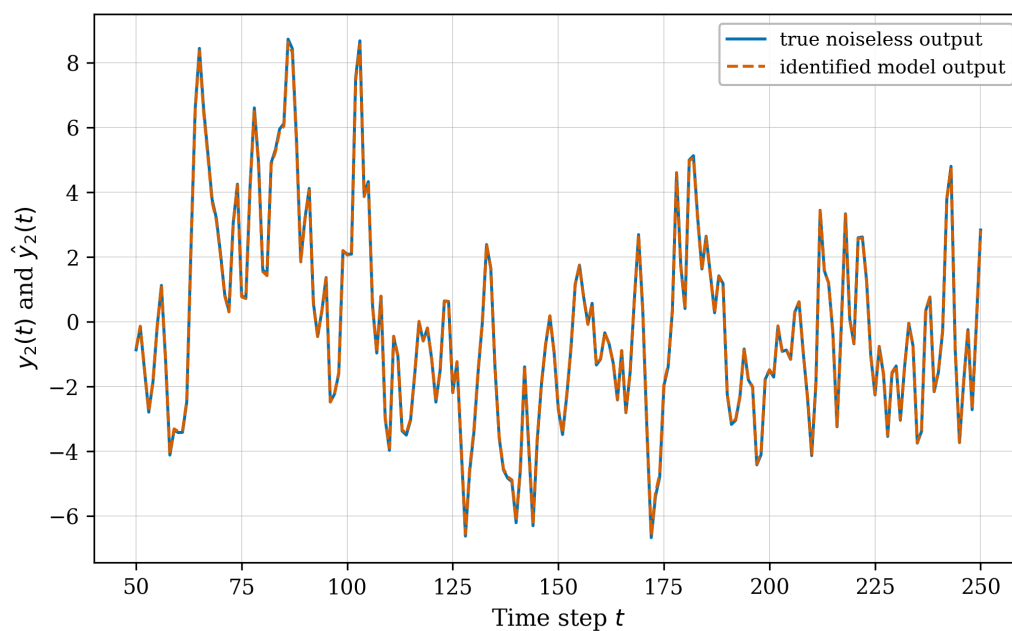


Figure 8. True validation output $y_2(t)$ and identified model output $\hat{y}_2(t)$.

Figures 9–11 validate the adaptive mechanisms. The index $\varepsilon(t)$ is smaller during the transient phase and larger when the innovation energy decreases. The innovation length changes according to the same energy and, after the initial transient, the rule reduces the window toward $p_{\min} = 5$ rather than using an unnecessarily long window. The residual-based filter estimate approaches the true common coloring

coefficient despite missing and impulsive data.

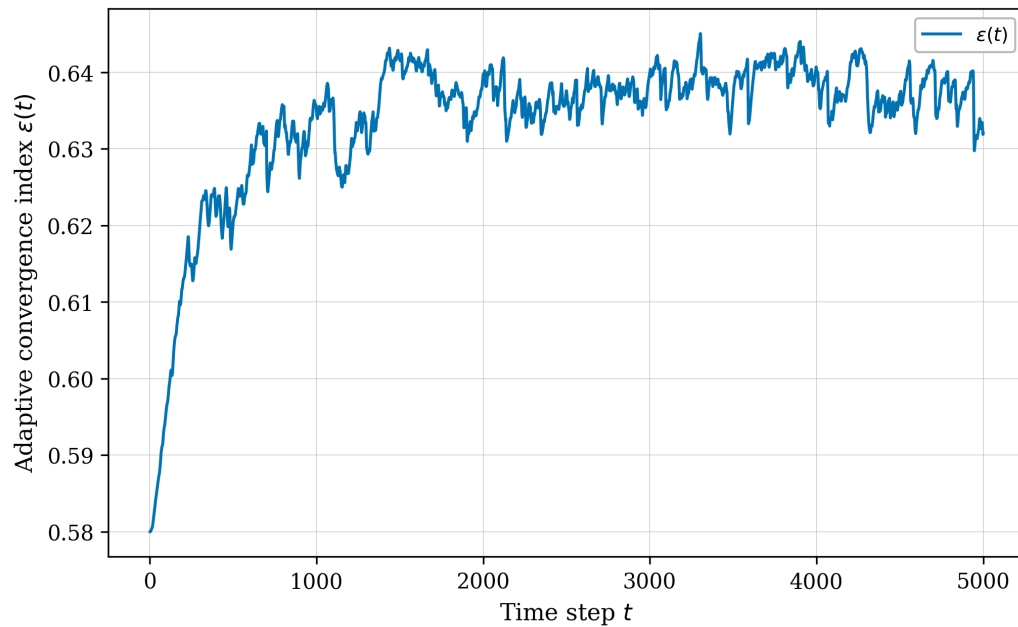


Figure 9. Adaptive convergence index $\varepsilon(t)$ of the proposed algorithm.

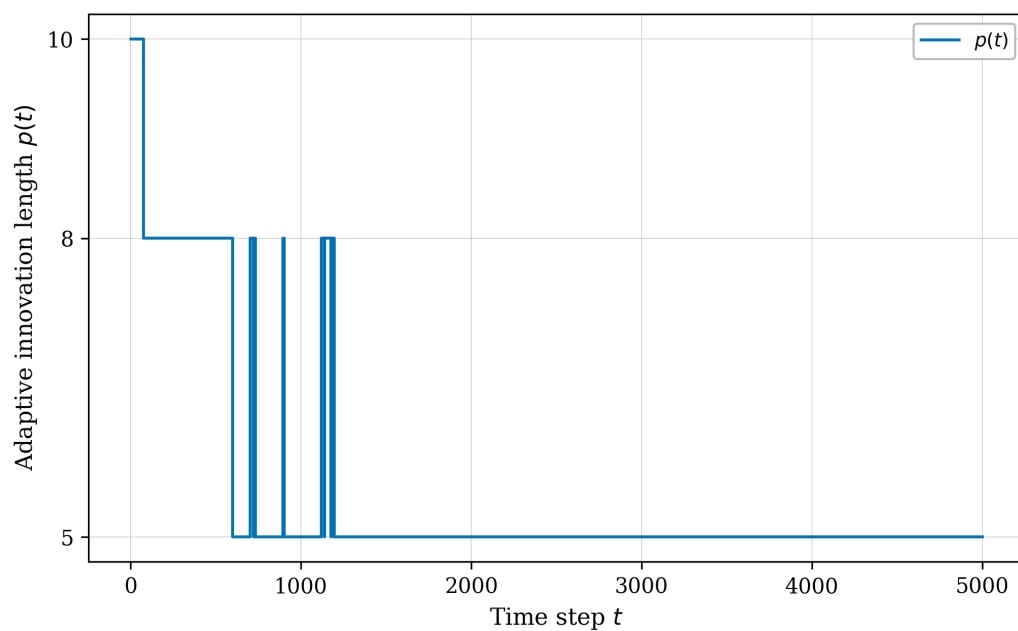


Figure 10. Adaptive innovation length $p(t)$ of the proposed algorithm.

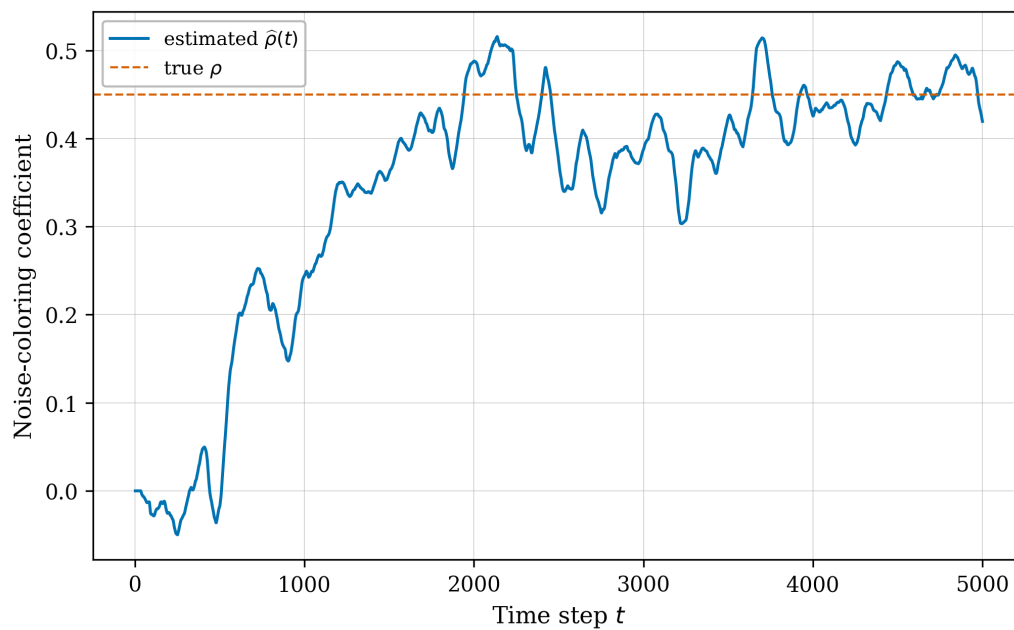


Figure 11. Robust residual-based estimate $\hat{\rho}(t)$ of the common colored-noise coefficient.

6.4. Ablation, Monte Carlo, and sensitivity tests

Table 10 specifies what each compared algorithm contains. Table 11 shows that the proposed RAMF-HMISG method achieves the smallest final estimation error among the compared methods and a validation mean-square error (MSE) close to the best fixed filtered variant. Table 12 reports mean and standard deviation over 30 independent short-horizon runs, confirming that the proposed combination of masking, Huber score, filtering, and adaptation gives the lowest average parameter error; the same statistics are displayed in Figure 12, where the standard-deviation bars show that the ordering of the methods is not an artifact of a particular realization.

Table 10. Ablation interpretation of the compared algorithms.

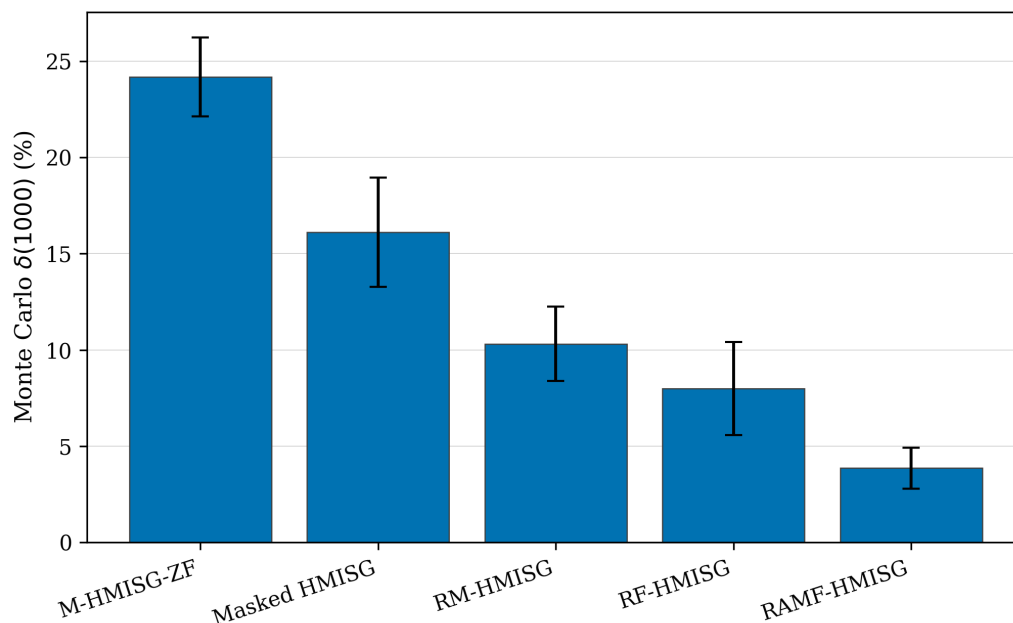
Algorithm	Missing mask	Reconstruction	Huber score	Filter estimate	Adaptive $p(t), \varepsilon(t)$
M-HMISG-ZF	No	No	No	No	No
Masked HMISG	Yes	Yes	No	No	No
RM-HMISG	Yes	Yes	Yes	No	No
RF-HMISG	Yes	Yes	Yes	Yes	No
RAMF-HMISG	Yes	Yes	Yes	Yes	Yes

Table 11. Comparison of final estimation errors, validation errors, and computational time.

Algorithm	$\delta(5000)$ (%)	Validation MSE	CPU time (s)
M-HMISG-ZF	16.9447	0.55115	3.406
Masked HMISG	7.5595	0.01588	3.244
RM-HMISG	2.5190	0.00330	2.938
RF-HMISG	0.8408	0.00256	8.090
RAMF-HMISG	0.7487	0.00265	7.908

Table 12. Monte Carlo statistics over 30 independent runs of length $L_{MC} = 1000$.

Algorithm	$\delta(1000)$ mean $\pm$ std (%)	Validation MSE mean $\pm$ std	CPU time mean $\pm$ std (s)
M-HMISG-ZF	24.180 ± 2.051	0.99831 ± 0.24981	0.611 ± 0.050
Masked HMISG	16.103 ± 2.840	0.08206 ± 0.02830	0.608 ± 0.048
RM-HMISG	10.309 ± 1.935	0.03879 ± 0.01748	0.615 ± 0.050
RF-HMISG	7.994 ± 2.419	0.03647 ± 0.01893	1.630 ± 0.104
RAMF-HMISG	3.858 ± 1.065	0.01617 ± 0.01112	1.529 ± 0.081

**Figure 12.** Monte Carlo mean relative errors with standard-deviation bars.

Tables 13 and 14 are stochastic short-horizon stress tests with $L_s = 1000$ and mean $\pm$ standard-deviation values. Therefore, their numerical values are not meant to coincide with the main $L_e = 5000$ final-error table. The missing-output sensitivity table shows that zero-filling degrades rapidly when the dropout probability increases. The proposed method remains more accurate because reconstructed values are used for regressor continuity but are not treated as observed innovations. The outlier sensitivity table is interpreted as a resampled stochastic stress test rather than a monotone deterministic

sweep: inputs, masks, and background-noise realizations are resampled, while the outlier probability is varied. It consistently shows that robust and filtered processing are important when impulsive contamination is present. Table 13 covers dropout probabilities up to 50% for all five algorithms, so that the boundary of the operating region can be read directly; this boundary is analyzed in Section 6.7.

Table 13. Robustness boundary in the missing-output probability. Final relative error $\delta(L_s)$ in percent at $L_s = 1000$ and mean $\pm$ standard deviation over 10 independent runs. The coloring coefficient is $\rho = 0.45$, the innovation standard deviation is 0.25, and the outlier probability is 2%.

Missing probability	M-HMISG-ZF	Masked HMISG	RM-HMISG	RF-HMISG	RAMF-HMISG
0%	20.273 $\pm$ 2.077	12.573 $\pm$ 1.773	6.529 $\pm$ 1.265	2.152 $\pm$ 0.629	1.473 $\pm$ 0.281
10%	18.191 $\pm$ 1.898	13.082 $\pm$ 3.342	8.323 $\pm$ 1.474	3.740 $\pm$ 1.341	2.224 $\pm$ 0.922
20%	23.506 $\pm$ 2.254	14.395 $\pm$ 3.676	10.368 $\pm$ 2.868	8.247 $\pm$ 3.433	3.862 $\pm$ 1.449
30%	31.651 $\pm$ 1.814	19.647 $\pm$ 4.429	13.932 $\pm$ 3.160	16.616 $\pm$ 5.303	8.862 $\pm$ 3.270
40%	39.928 $\pm$ 2.332	20.440 $\pm$ 2.951	15.853 $\pm$ 2.542	24.519 $\pm$ 3.679	13.347 $\pm$ 2.844
50%	48.064 $\pm$ 1.619	23.647 $\pm$ 3.911	20.880 $\pm$ 4.095	34.627 $\pm$ 5.416	20.983 $\pm$ 5.890

Table 14. Short-horizon stress test: Effect of impulsive contamination on the relative estimation error at $L_s = 1000$ over 10 independent runs with 20% missing outputs.

Outlier probability	Masked HMISG	RM-HMISG	RAMF-HMISG
0%	15.081 $\pm$ 2.668	14.526 $\pm$ 3.049	4.773 $\pm$ 2.044
2%	16.846 $\pm$ 2.921	15.257 $\pm$ 2.299	5.595 $\pm$ 1.949
5%	16.146 $\pm$ 3.392	16.460 $\pm$ 2.677	4.616 $\pm$ 1.916
8%	15.630 $\pm$ 1.600	17.122 $\pm$ 1.788	3.704 $\pm$ 0.927

6.5. Component-wise ablation: Isolating the contribution of each adaptive law

The ablation of Table 10 adds one mechanism at a time but treats the pair of adaptive laws as a single block, so it cannot attribute the final gain to $p(t)$ or to $\varepsilon(t)$ individually. Table 15 removes that limitation by inserting the two intermediate algorithms in which exactly one of the laws is active; the other is frozen at the fixed value used by RF-HMISG.

Two precautions make the comparison free of confounding factors. First, the Huber threshold $\kappa = 1.0$, the filter window $L_f = 80$, and the frozen values $p = 8$, $\varepsilon = 0.60$ are identical for every variant, so no difference can be attributed to a different robustness level or to a different filter memory. Second, the update gain is not shared. A single gain would be the more obvious way to equalize the comparison, but it would in fact bias it: the robust score bounds the innovation at κ , so robust variants require a larger gain to produce a correction of comparable size, and prewhitening reduces the regressor energy accumulated in $r(t)$, which shifts the useful gain range again. Imposing one value would therefore penalize whichever variant is furthest from its own optimum and would measure the choice of gain rather than the mechanism under study. Instead, each variant is given the gain that minimizes its own

error over the common grid $\mu \in \{0.4, 0.7, 1.0, 1.4, 2.0, 2.8\}$, the selection being performed on calibration records that are statistically independent of the records used for the reported evaluation. Every variant is thus compared at its own best operating point, which is the strongest form of fairness available here.

Table 15. Component-wise ablation isolating the individual contribution of each adaptive law. All variants use the *same* Huber threshold $\kappa = 1.0$, the same filter window $L_f = 80$, and the same fixed values $p = 8$, $\varepsilon = 0.60$ wherever the corresponding law is disabled; each variant is evaluated at the gain minimizing its own error over the common grid $\mu \in \{0.4, 0.7, 1.0, 1.4, 2.0, 2.8\}$, selected on calibration records independent of those used here. This table shows the mean $\pm$ standard deviation over 10 independent runs of length $L_s = 3000$.

Algorithm	Missing mask	Huber score	Pre-whitening	$p(t)$	$\varepsilon(t)$	$\delta(L_s)$ (%)	Validation MSE
Masked HMISG	Yes	No	No	fixed	fixed	2.051 ± 0.456	0.0242 ± 0.0214
RM-HMISG	Yes	Yes	No	fixed	fixed	1.809 ± 0.358	0.0180 ± 0.0090
RF-HMISG	Yes	Yes	Yes	fixed	fixed	1.109 ± 0.196	0.0076 ± 0.0071
RAF-HMISG (fixed p , adaptive ε)	Yes	Yes	Yes	fixed	adaptive	1.058 ± 0.182	0.0070 ± 0.0065
RAF-HMISG (adaptive p , fixed ε)	Yes	Yes	Yes	adaptive	fixed	1.085 ± 0.200	0.0067 ± 0.0062
RAMF-HMISG	Yes	Yes	Yes	adaptive	adaptive	1.007 ± 0.186	0.0062 ± 0.0057

The resulting ordering is strictly monotone in the number of active mechanisms. Adding the Huber score to the mask, then the prewhitening filter, then either adaptive law, and then both reduces the error at every step, and the same ordering holds for the validation error, which is an independent criterion computed on unseen data. The two intermediate algorithms are the informative entries: each adaptive law used alone recovers most of the gap between RF-HMISG and the full algorithm, and the two are close to each other, which indicates that they act on partly complementary mechanisms—the window length on the amount of information used per correction and the convergence index on the size of the correction—rather than on the same one. Enabling both together gives the smallest error and the smallest dispersion across runs, which is the quantitative justification for adapting the two variables jointly rather than selecting one of them.

6.6. Reliability mask: Hard certification versus softened weighting

Remark 2.2 shows that the hard mask $\gamma_{f,i}(t) = \gamma_i(t)\gamma_i(t-1)$ discards a fraction p_m of the samples that survive the raw mask, and introduces the softened family (2.35). Table 16 and Figure 13 quantify the resulting trade-off across dropout probabilities from 10% to 50%. As in Section 6.5, each combination of dropout probability and softening weight is given its own gain, selected on independent calibration records over the common grid $\mu \in \{0.8, 1.2, 1.6, 2.2, 3.0\}$. Without this precaution, a larger ω would simply admit more samples into the loss and thereby act as a larger effective step size, and the comparison would measure the step size instead of the mask.

Table 16. Effect of softening the filtered reliability mask (2.35). The columns $\omega = 0$ and $\omega = 1$ correspond, respectively, to the hard mask $\gamma_i(t)\gamma_i(t-1)$ retained in the paper and to the current-sample mask $\gamma_i(t)$. Entries are the final relative error $\delta(L_s)$ in percent, with mean $\pm$ standard deviation over 10 runs of length $L_s = 1000$. The second and third columns report the empirical availability fractions, which follow the prediction (2.34).

p_m	$\mathbb{E}\{\gamma_i\}$	$\mathbb{E}\{\gamma_{f,i}\}$	$\omega = 0.00$	$\omega = 0.50$	$\omega = 1.00$
10%	0.894	0.800	1.691 ± 0.220	1.650 ± 0.266	1.517 ± 0.260
20%	0.801	0.642	1.562 ± 0.372	1.469 ± 0.307	1.619 ± 0.324
30%	0.694	0.482	2.478 ± 0.748	1.763 ± 0.254	1.919 ± 0.458
40%	0.597	0.356	5.242 ± 1.575	2.561 ± 0.931	2.269 ± 0.648
50%	0.503	0.251	11.626 ± 4.065	5.142 ± 2.192	2.827 ± 0.732

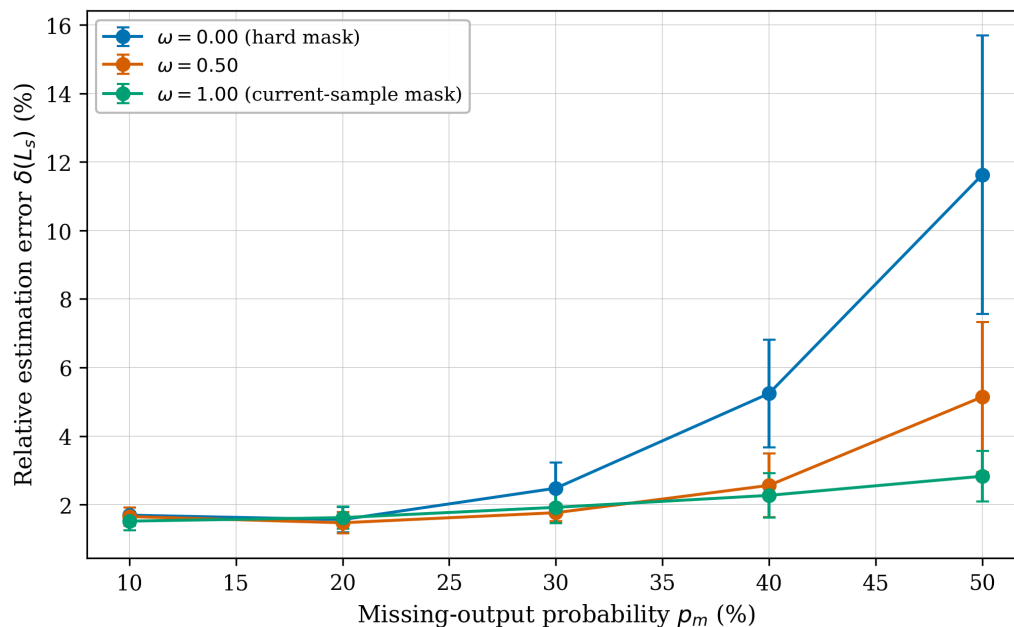


Figure 13. Final relative estimation error of the proposed algorithm as a function of the missing-output probability, for the hard reliability mask ($\omega = 0$), an intermediate weighting ($\omega = 0.5$), and the current-sample mask ($\omega = 1$) of the softened family (2.35). Error bars are one standard deviation over the replicates.

The measured availability fractions reproduce the prediction (2.34) closely, which confirms the information accounting of Remark 2.2. The performance picture separates into two regimes. Up to a dropout probability of about 20%, the three masks are statistically indistinguishable: few samples are affected and the reconstruction error injected by (2.36) is of the same order as the innovation of the sample it makes available, so nothing is gained by relaxing the certification. Beyond that point, the balance shifts decisively. The quadratic loss of certified samples predicted by (2.34) becomes the dominant effect: admitting the partially reconstructed samples with a positive weight recovers a large

part of the degradation, and at 50% dropout, the softened mask is several times more accurate than the hard one. This is precisely the regime in which Remark 2.2 predicts the hard constraint to be most costly, so the analysis and the measurement agree.

The practical conclusion is a recommendation rather than a replacement. The hard mask remains the default of the algorithm for two reasons: in the moderate-dropout regime for which it was designed, it loses nothing and it is the only choice for which the filtered sample is certified to be a function of genuine measurements, so that the mismatch term $\bar{\Delta}^2$ of Assumption 4.2 contains no contribution from imputed data and the guarantee of Theorem 4.1 is obtained under the weakest hypotheses. All other results of this paper use it. When the dropout probability exceeds roughly 30%, however, the certification is bought at a price that the present experiment measures explicitly, and the softened mask with ω close to one becomes the better engineering choice. The trade-off identified analytically in Remark 2.2 is therefore real, quantified, and resolved differently in the two regimes, and the designer is given the information needed to make the choice for a particular application. It should be noted that the gains used here are calibrated for each configuration, so the absolute levels of Table 16 are not directly comparable with those of Table 13, which uses the nominal gain of Table 8 throughout; only the comparison between columns within Table 16 is meaningful.

6.7. Robustness boundary: High data loss and low signal-to-noise ratio

The experiments reported so far use a dropout probability of 20% and an innovation standard deviation of 0.25. Two further studies locate the boundary of the region in which the proposed algorithm retains its advantage.

Table 13 extends the dropout probability to 50%. The proposed algorithm has the smallest error at every level up to 40%, and its margin over the masked and zero-filling baselines widens as the dropout increases. At 50%, however, the proposed algorithm and the unfiltered robust variant RM-HMISG become statistically indistinguishable, and the filtered fixed-parameter variant RF-HMISG is clearly worse than both. This is the boundary of the operating region, and its location is predicted by the analysis: by (4.42), the excitation available to a filtered method scales as $(1 - p_m)^2$ and has fallen to a quarter of its nominal value at $p_m = 0.5$, while the mismatch injected by the prewhitening step does not decrease. Beyond roughly 40% dropout, prewhitening therefore ceases to pay for itself, and a designer facing such severe data loss should either use the unfiltered robust variant or soften the reliability mask as described in Section 6.6.

Table 17 varies the innovation standard deviation from 0.25 to 0.75, the impulsive amplitude being scaled in proportion so that the contamination ratio is preserved, and crosses this with dropout probabilities of 20% and 40%. The proposed algorithm has the smallest error in every one of the six configurations, and its degradation with decreasing signal-to-noise ratio is modest: at 20% dropout, tripling the innovation standard deviation increases its error by about a quarter, whereas the unfiltered robust variant nearly doubles. The Huber score bounds the influence of the enlarged innovations while the prewhitening filter continues to remove the coloring. The zero-filling variant behaves differently and instructively: it barely degrades at all, but only because it is already limited by the imputation bias analyzed in Section 1.2, which the noise level does not affect. Its insensitivity to the noise is therefore a symptom of its floor, not a sign of robustness. Combining a low signal-to-noise ratio with 40% data loss is the most demanding configuration tested, and the proposed method still separates clearly from every baseline.

Table 17. Behavior under low signal-to-noise ratio and high data loss. The innovation standard deviation σ_ξ and the impulsive amplitude are increased together so that the contamination ratio is preserved. This table shows the final relative error $\delta(L_s)$ in percent at $L_s = 1000$ and mean $\pm$ standard deviation over 10 runs.

σ_ξ	SNR (dB)	Missing	M-HMISG-ZF	Masked HMISG	RM-HMISG	RF-HMISG	RAMF-HMISG
0.25	15.7	20%	24.106 $\pm$ 1.806	16.566 $\pm$ 1.766	10.584 $\pm$ 1.311	9.829 $\pm$ 2.146	4.757 $\pm$ 1.407
0.25	15.8	40%	39.401 $\pm$ 2.196	16.595 $\pm$ 3.248	13.193 $\pm$ 3.037	21.079 $\pm$ 5.831	11.363 $\pm$ 4.470
0.50	9.6	20%	26.002 $\pm$ 1.853	16.944 $\pm$ 3.528	13.692 $\pm$ 2.743	10.209 $\pm$ 3.560	3.667 $\pm$ 0.983
0.50	9.8	40%	42.005 $\pm$ 2.148	21.051 $\pm$ 6.185	19.888 $\pm$ 4.526	27.515 $\pm$ 4.400	13.409 $\pm$ 4.383
0.75	6.4	20%	27.009 $\pm$ 1.609	20.016 $\pm$ 3.312	20.131 $\pm$ 2.588	14.808 $\pm$ 3.007	5.903 $\pm$ 1.320
0.75	6.0	40%	40.171 $\pm$ 2.484	21.993 $\pm$ 5.149	23.120 $\pm$ 5.031	30.491 $\pm$ 6.757	15.201 $\pm$ 5.227

6.8. Coloring coefficient and impulsive intensity

The nominal experiment uses a single coloring coefficient $\rho = 0.45$ and a single impulse amplitude. Tables 18 and 19 vary each of them over a wide range.

Table 18. Effect of the coloring coefficient of the common disturbance $v_i(t) = \rho v_i(t-1) + \xi_i(t)$. The second column reports the coefficient identified by the robust estimator (2.25) in the proposed algorithm. The table shows the final relative error $\delta(L_s)$ in percent at $L_s = 1000$ and mean $\pm$ standard deviation over 10 runs with 20% missing outputs.

ρ	$\hat{\rho}$	M-HMISG-ZF	Masked HMISG	RM-HMISG	RF-HMISG	RAMF-HMISG
0.00	-0.180	24.825 $\pm$ 1.557	15.995 $\pm$ 3.338	9.696 $\pm$ 1.462	12.736 $\pm$ 3.513	8.029 $\pm$ 2.461
0.20	0.110	24.479 $\pm$ 2.080	15.053 $\pm$ 2.286	8.993 $\pm$ 2.185	9.345 $\pm$ 3.909	5.090 $\pm$ 2.104
0.45	0.311	24.894 $\pm$ 1.881	16.296 $\pm$ 3.098	10.625 $\pm$ 2.742	9.684 $\pm$ 4.157	4.742 $\pm$ 2.104
0.70	0.599	25.496 $\pm$ 1.865	17.857 $\pm$ 3.645	12.413 $\pm$ 3.124	8.976 $\pm$ 5.050	4.135 $\pm$ 2.255
0.90	0.799	26.289 $\pm$ 2.268	19.614 $\pm$ 4.643	17.305 $\pm$ 3.499	5.936 $\pm$ 2.422	2.655 $\pm$ 0.906

Table 18 sweeps the coloring coefficient from 0, that is, white disturbance, to 0.9, that is, strongly correlated disturbance. The two families behave in opposite ways, and both behave as the theory predicts. The unfiltered variants degrade monotonically as the coloring strengthens because the correlated disturbance biases a gradient computed from raw innovations; the robust masked variant nearly doubles its error across the range. The filtered variants improve because the quantity the prewhitening step removes grows with ρ . The consequence is that the benefit of filtering is smallest for a white disturbance, where the proposed algorithm is only marginally ahead of the robust masked variant, and grows steadily with the coloring, until at $\rho = 0.9$, it is better by a factor of more than six. The experiment therefore delimits the useful domain of the filtering mechanism as clearly as it demonstrates its benefit: prewhitening is worth its cost exactly to the extent that the disturbance is actually colored.

The second column of Table 18 records the coefficient identified by the robust estimator (2.25). It

follows the true value across the whole range with an approximately unit slope and a small systematic underestimate of about one-tenth. That offset is not a defect of the design but an illustration of Remark 4.2: no consistency of $\widehat{\rho}$ is claimed anywhere in the analysis, its error is absorbed into the bounded mismatch term, and the accuracy of the parameter estimates is evidently unaffected by an offset of this size.

Table 19 fixes the outlier probability at 5% and multiplies the impulse amplitude up to 27 times the innovation standard deviation. The proposed algorithm has the smallest error at every level, and the shape of the degradation is as informative as its size. Its error is essentially flat up to fourteen standard deviations and rises only in the most extreme configuration, whereas the robust masked variant, which lacks the prewhitening filter, more than doubles over the same range, and the non-robust masked variant is worse than the proposed algorithm by a factor of three throughout. The flat region is the signature of the bounded score: once a residual exceeds κ , its contribution is capped at κ regardless of how large it becomes, so multiplying the amplitude of the impulses changes nothing until the truncated samples become numerous enough to erode the effective excitation, in the manner quantified by (4.43).

Table 19. Effect of the intensity of the impulsive contamination at a fixed outlier probability of 5% and 20% missing outputs. The table shows the final relative error $\delta(L_s)$ in percent at $L_s = 1000$ and mean $\pm$ standard deviation over 10 runs.

Impulse amplitude	Relative to σ_ξ	Masked HMISG	RM-HMISG	RF-HMISG	RAMF-HMISG
0.0	$0\sigma_\xi$	15.489 ± 2.159	9.234 ± 1.824	8.238 ± 2.635	4.434 ± 1.351
1.7	$7\sigma_\xi$	15.893 ± 3.338	10.687 ± 2.179	8.165 ± 3.516	3.735 ± 1.587
3.4	$14\sigma_\xi$	19.451 ± 4.894	14.589 ± 3.636	10.706 ± 3.877	4.650 ± 1.257
6.8	$27\sigma_\xi$	25.223 ± 2.881	22.138 ± 3.863	16.147 ± 4.976	7.389 ± 2.969

6.9. Channel-dependent coloring and the boundary of the common-filter model

Equation (2.22) assumes that all output channels share one coloring coefficient. Remark 2.1 explained why this assumption is structurally necessary for the coupled update and predicted, in (2.24), that a violation costs an additional mismatch growing quadratically with the spread of the true coefficients. Table 20 tests that prediction by coloring the two channels with different coefficients of increasing spread.

Table 20. Applicable boundary of the common-filter assumption of Remark 2.1. The two output channels are colored by different coefficients (ρ_1, ρ_2) with increasing spread $|\rho_1 - \rho_2|$ at constant mean 0.45. RM-HMISG applies no prewhitening and is therefore unaffected by the spread; it serves as the reference against which the benefit of a single common filter is measured. The table shows the final relative error $\delta(L_s)$ in percent at $L_s = 1000$ and mean $\pm$ standard deviation over 10 runs.

(ρ_1, ρ_2)	Spread	$\widehat{\rho}$	RM-HMISG	RF-HMISG	RAMF-HMISG
(0.45, 0.45)	0.00	0.363	9.385 ± 2.260	6.650 ± 2.790	3.535 ± 1.184
(0.30, 0.60)	0.30	0.368	10.429 ± 2.081	9.325 ± 3.448	4.762 ± 1.916
(0.15, 0.75)	0.60	0.464	10.049 ± 2.094	6.665 ± 1.967	3.291 ± 0.715
(0.00, 0.90)	0.90	0.545	16.950 ± 3.328	10.894 ± 3.899	4.875 ± 1.813
(-0.45, 0.45)	0.90	-0.044	9.358 ± 1.939	10.957 ± 3.405	6.129 ± 2.084

Two regimes appear. As long as the coefficients keep the same sign and are clustered around a common mean, the single estimated coefficient settles between them, both channels are partially whitened, and the proposed algorithm retains essentially the whole of its advantage over the unfiltered robust variant. The degradation predicted by (2.24) is present but remains small compared with the other error sources because the common part of the coloring, which is what the filter removes, is unchanged. The situation changes in the last configuration, where the two coefficients have opposite signs and no common part exists. The estimator then correctly returns a value close to zero—there is no common coloring to identify—so prewhitening removes almost nothing from either channel, and the error of the proposed algorithm rises to roughly 1.7 times its value at zero spread. The advantage over the unfiltered robust variant is not eliminated, because the mask, the reconstruction, and the adaptive laws continue to operate, but it is reduced by about half. This is the genuine applicable boundary of the model: the common-filter formulation is appropriate when the channel-coloring coefficients share a dominant common component, whatever their spread around it, and it should be replaced by a channel-dependent filter—at the cost of the m -fold duplication analyzed in Remark 2.1—when that common component is absent. Extending the hierarchical structure to channel-dependent filters without paying that cost is left as future work.

6.10. Higher-order multivariable example

All results reported so far concern one two-input two-output system with twelve unknown parameters. To show that the conclusions are not specific to that example, the whole comparison is repeated on a three-input three-output ARX system with $n_i = n_j = 3$, whose true parameters are listed in Table 21. The system has thirty-six unknown parameters, three times as many as in the two-input two-output example, and its regressor dimension per channel is $d_i = n_i + n = 12$. The operating conditions are unchanged: 20% missing outputs, coloring coefficient 0.45, and 2% impulsive contamination. As in the previous studies, each algorithm is given the gain that minimizes its own error over a common grid, selected on independent calibration records. Table 22 reports the resulting accuracy, prediction error, and execution time, and Figure 14 shows the corresponding convergence curves.

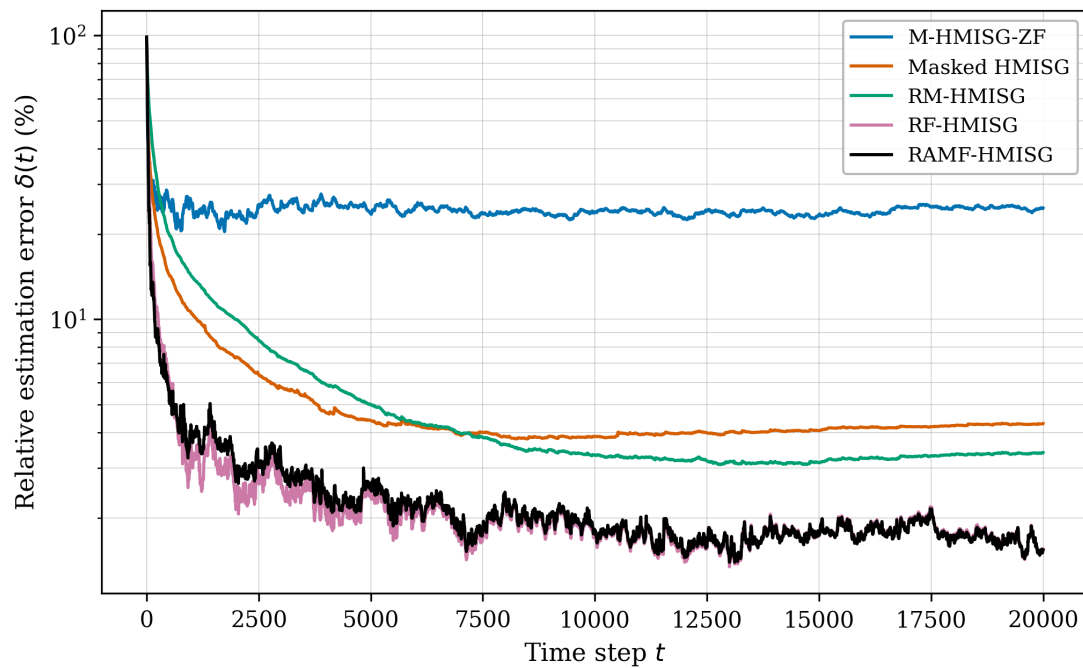


Figure 14. Relative parameter-estimation errors on the higher-order three-input three-output ARX system with 36 unknown parameters, 20% missing outputs, coloring coefficient 0.45, and 2% impulsive contamination.

Table 21. True parameters of the higher-order three-input three-output ARX system: autoregressive coefficients (top) and input coefficients (bottom), where j is the input index and ℓ the lag.

$a_1(1)$	$a_1(2)$	$a_1(3)$	$a_2(1)$	$a_2(2)$	$a_2(3)$	$a_3(1)$	$a_3(2)$	$a_3(3)$
-0.250	-0.150	-0.100	-0.300	-0.120	-0.080	-0.200	-0.180	-0.090

(j, ℓ)	$\theta_{1j}(\ell)$	$\theta_{2j}(\ell)$	$\theta_{3j}(\ell)$
(1, 1)	0.800	-0.550	0.650
(1, 2)	1.150	0.900	-1.050
(1, 3)	-0.700	0.620	0.480
(2, 1)	1.350	1.100	-0.950
(2, 2)	-0.950	1.280	1.120
(2, 3)	0.600	-0.850	0.720
(3, 1)	1.050	0.750	1.300
(3, 2)	-0.880	1.200	-0.900
(3, 3)	0.520	-0.660	1.080

Table 22. Higher-order example: three-input three-output ARX system with $n_i = n_j = 3$ (36 unknown parameters), 20% missing outputs, coloring coefficient 0.45, and 2% impulsive contamination. Each algorithm uses the gain minimizing its own error over the common grid $\mu \in \{1.6, 3.2, 6.4, 12.8, 25.6\}$, selected on independent calibration records. The table shows the mean $\pm$ standard deviation over 10 independent runs of length $L = 20,000$.

Algorithm	Selected μ	$\delta(L)$ (%)	Validation MSE	CPU time (s)
M-HMISG-ZF	3.2	23.759 ± 0.849	0.3566 ± 0.0314	41.904 ± 1.657
Masked HMISG	1.6	3.435 ± 0.844	0.0046 ± 0.0015	41.623 ± 1.110
RM-HMISG	1.6	3.078 ± 0.409	0.0042 ± 0.0012	41.868 ± 1.551
RF-HMISG	6.4	1.603 ± 0.226	0.0041 ± 0.0018	89.415 ± 8.189
RAMF-HMISG	12.8	1.576 ± 0.217	0.0040 ± 0.0017	74.987 ± 11.319

The qualitative picture of Section 6 is reproduced without modification. Zero-filling is again limited by a bias that the longer record does not remove, and it stalls more than an order of magnitude above the other methods. Masking removes that bias, the robust score and the prewhitening filter each improve the result further, and the two filtered variants are the most accurate. Between those two, the difference is worth stating precisely rather than overstating: the proposed algorithm attains the smallest mean error, but its margin over the fixed-parameter filtered variant is smaller than one standard deviation across replicates, so on this system, the two should be regarded as equally accurate. The benefit of adaptation appears instead in the last column. Because the switching law spends the long window only during the transient and falls back to $p_{\min}$ afterward, the proposed algorithm reaches that accuracy in appreciably less computation than the fixed-window variant, which pays for a window of constant length over the whole record. This is the behavior predicted in Section 4.3, and it is the form the advantage takes once the record is long enough for a fixed window to converge as well.

Convergence is slower than in the two-input two-output case, as expected from the threefold increase in the number of unknowns at a comparable excitation level, which is why the record length is increased accordingly. It is worth noting what had to be re-tuned and what did not. The gain had to be re-selected, as anticipated in Section 4.6, because it is not dimensionless. Every other constant was carried over unchanged from the two-input two-output example, including the switching thresholds, which transfer by construction through the scale-free guideline (4.22), and the window candidate set, which was deliberately left at $\{5, 8, 10\}$ even though the band $[d_i/2, 2d_i]$ of Section 4.3 is $[6, 24]$ for this system. That the result is nevertheless competitive is consistent with the sensitivity study of Section 6.11, where accuracy was found to be flat with respect to the candidate set over a wide range, and it supports the claim that the adaptation rules are portable rather than tuned to one example.

6.11. Sensitivity to the tuning constants

Table 23 varies each tuning constant in turn, and all others are held at the nominal values of Table 8. The constants separate clearly into the three groups anticipated in Section 4.6.

Table 23. Sensitivity of the proposed algorithm to each tuning constant. One constant is varied at a time, with all others being held at the nominal values of Table 8. The last column reports the percentage of time spent at $p_{\min}$, p_{mid} , and $p_{\max}$, respectively, which diagnoses whether the switching law is exercised or degenerate. The table shows the mean $\pm$ standard deviation over 6 runs of length $L_e = 5000$ with 20% missing outputs.

Setting	$\delta(L_e)$ (%)	mean $p(t)$	time at $p_{\min}/p_{\text{mid}}/p_{\max}$ (%)
<i>Huber threshold κ</i>			
0.5	2.182 ± 1.107	5.48	84/15/1
1	0.669 ± 0.120	5.35	89/9/2
2	0.874 ± 0.202	6.02	67/31/2
4	0.979 ± 0.190	6.39	56/42/3
∞ (no truncation)	1.485 ± 0.750	6.79	44/49/6
<i>Update gain μ</i>			
0.8	4.217 ± 1.141	6.27	59/38/2
1.2	1.150 ± 0.233	5.55	83/16/2
1.6	0.669 ± 0.120	5.35	89/9/2
2	0.665 ± 0.119	5.28	92/7/1
2.4	0.707 ± 0.128	5.23	93/5/1
<i>Energy smoothing λ_ζ</i>			
0.9	0.726 ± 0.115	5.58	81/18/1
0.95	0.699 ± 0.123	5.43	86/13/1
0.98	0.669 ± 0.120	5.35	89/9/2
0.99	0.652 ± 0.119	5.35	90/8/2
0.995	0.623 ± 0.116	5.51	86/11/4
<i>Filter window L_f</i>			
20	0.672 ± 0.128	5.35	89/9/2
40	0.669 ± 0.124	5.35	89/9/2
80	0.669 ± 0.120	5.35	89/9/2
160	0.671 ± 0.129	5.35	89/9/2
<i>Window candidates $\{p_{\min}, p_{\text{mid}}, p_{\max}\}$</i>			
{3, 5, 8}	0.987 ± 0.204	3.37	84/14/2
{5, 8, 10}	0.669 ± 0.120	5.35	89/9/2
{8, 12, 16}	0.659 ± 0.114	8.36	92/6/1
{10, 20, 30}	0.667 ± 0.116	10.66	95/4/1
<i>Index range $[\varepsilon_{\min}, \varepsilon_{\max}]$</i>			
[0.51, 0.6]	0.828 ± 0.124	5.20	94/4/1
[0.56, 0.66]	0.669 ± 0.120	5.35	89/9/2
[0.6, 0.75]	5.684 ± 1.284	6.40	55/43/2
[0.7, 0.9]	31.655 ± 4.980	8.22	0/89/11

The Huber threshold and the update gain are the two constants that matter. Removing the truncation altogether, $\kappa = \infty$, restores the non-robust behavior and costs accuracy immediately, which confirms

that the bounded score, and not merely the mask, is responsible for the resistance to impulses. Very small thresholds are also harmful, for the reason identified in Remark 4.3: they truncate almost every residual and erode the effective excitation λ_G . The useful range is nevertheless broad and contains the recommended value $\kappa \approx 1.5\widehat{\sigma}_e$, so the threshold does not need to be found precisely. The gain behaves in the usual way for a stochastic-gradient recursion: the optimum is shallow and the degradation above it is mild, whereas reducing the gain below it costs transient speed rapidly.

Most of the remaining constants are second order. Varying the filter window over an eightfold range leaves the final error unchanged to three decimal places, which reflects the fact that $\widehat{\rho}$ needs to be approximately, not exactly, right: its error is absorbed by the mismatch term, as established in Remark 4.2. Window candidate sets are equivalent to within the run-to-run dispersion as soon as they lie from the middle of the band $[d_i/2, 2d_i]$ of Section 4.3 upward; only the shortest set tested, which sits at the very bottom of the band, loses accuracy, and pushing the candidates well beyond the band brings no further benefit while increasing the per-sample cost linearly, as Table 6 shows.

The convergence-index interval is the one second-order constant whose misspecification is expensive, and the penalty is strongly asymmetric. Shifting the interval down toward the admissible limit $1/2$ costs little, since the steps become slightly larger and the ultimate neighborhood slightly wider. Shifting it up is far more damaging: with $r(t)$ growing linearly, an index approaching one makes the step size decay like $1/t$, the transient stalls, and the final error degrades by more than an order of magnitude. The asymmetry is exactly what the stochastic-approximation argument of Section 4.3 predicts, and it is the reason the recommended interval is placed just above the lower admissible limit.

The forgetting factor of the innovation-energy estimator deserves a separate examination, because it is the constant that closes the feedback loop of the adaptive laws. Table 24 reports, for each value of λ_ζ , the effective memory of the estimator, the resulting accuracy, the extreme values reached by $\zeta(t)$, and the fraction of time spent above η_2 and below η_1 . Three observations follow. The final accuracy varies by less than a fifth over the whole range tested, so the algorithm is not sensitive to this constant, and the mild improvement obtained with a longer memory is consistent with the requirement, derived in Remark 4.1, that the estimator should average over several outlier inter-arrival times. The energy remains inside $[0, \kappa^2]$ in every run, as guaranteed by (4.21) with the initialization $\zeta(0) = \kappa^2$, which confirms experimentally that the feedback loop discussed in Remark 4.1 is bounded and cannot destabilize the recursion. Finally, the occupancy figures show that all three window levels are visited: the long window is used during the transient, which occupies a small fraction of a record of this length, and the short window for the remainder, which is the intended behavior. A degenerate tuning would be immediately visible as an occupancy of 100% in a single column, and this is the diagnostic recommended in Remark 4.1.

Table 24. Behavior of the smoothed innovation energy $\zeta(t)$ as a function of its forgetting factor and occupancy of the two switching thresholds $\eta_1 = 0.12$ and $\eta_2 = 0.55$. The energy remains inside $[0, \kappa^2] = [0, 1]$ in every run, in agreement with (4.21). The table shows the mean over 6 runs of length $L_e = 5000$.

λ_ζ	memory N_ζ	$\delta(L_e)$ (%)	$\max_t \zeta$	$\zeta(L_e)$	time above η_2 (%)	time below η_1 (%)
0.9	10	0.726 ± 0.115	0.972	0.0861	0.7	81.1
0.95	20	0.699 ± 0.123	0.986	0.0826	0.9	86.2
0.98	50	0.669 ± 0.120	0.994	0.0772	1.5	89.3
0.99	100	0.652 ± 0.119	0.997	0.0768	2.3	89.8
0.995	200	0.623 ± 0.116	0.999	0.0764	3.8	85.5

6.12. Numerical verification of the assumptions

The analysis of Section 4 rests on two quantitative hypotheses that are not directly observable: the excitation bounds of Assumption 4.3 and the mismatch bound of Assumption 4.2. Because the data are generated in simulation, both can be measured. Table 25 reports the smallest eigenvalues of the time-averaged windowed matrices appearing in (4.37) and (4.40), minimized over the output channels, together with the measured mismatch energy, for operating conditions ranging from ideal to severe.

Table 25. Numerical verification of Assumptions 4.2 and 4.3. λ_0 and λ_G are the smallest eigenvalues of the time-averaged windowed matrices of (4.37) and (4.40), minimized over the output channels. The “Predicted” column is the scaling $(1 - p_m)^2$ of (4.42). The last two columns report the measured mismatch energy $\mathbb{E}\{\Delta_i^2\}$ averaged over the early transient ($t \leq 400$) and over the steady state ($t \geq 1600$). The table shows the mean over 5 runs of length $L = 2000$.

Missing	Outlier	λ_0	$\lambda_0/\lambda_0(0)$	Predicted	λ_G	λ_G/λ_0	$\bar{\Delta}^2$ early	$\bar{\Delta}^2$ late
0%	0%	3.033	1.000	1.000	3.018	0.995	0.0123	0.0002
20%	0%	2.009	0.662	0.640	1.979	0.985	0.2563	0.0297
20%	2%	2.168	0.715	0.640	2.146	0.989	0.1577	0.0393
20%	8%	2.554	0.842	0.640	2.478	0.970	0.1287	0.1424
40%	2%	1.273	0.420	0.360	1.228	0.965	0.4581	0.0668
50%	2%	0.934	0.308	0.250	0.881	0.943	0.7557	0.1291

Three conclusions can be drawn. First, the excitation constant λ_0 decreases with the dropout probability in the manner predicted by (4.42): the measured ratio $\lambda_0(p_m)/\lambda_0(0)$ follows the quadratic law $(1 - p_m)^2$ closely at moderate dropout and stays slightly above it at 40% and 50%. The excess is expected and is informative: the mask suppresses samples in the loss, but the reconstructed outputs keep the regressor matrix populated, so the realized excitation is a little better than the purely combinatorial count predicts. This validates the information accounting of Remark 2.2 and confirms that missing data attenuates the excitation without ever destroying it. Second, the ratio λ_G/λ_0 , which

measures how much excitation the robust score sacrifices, is 0.99 at the nominal setting, in close agreement with the value $1 - p_o$ predicted by (4.43), and remains above 0.94 in every configuration tested, including under 8% impulsive contamination and at 50% dropout. Assumption 4.3 is therefore satisfied with a comfortable margin, and the bounded score costs at most a few percent of the available information. Third, the mismatch energy is finite everywhere, as Remark 4.2 argues it must be, and it falls by at least a factor of four between the early transient and the steady state in every configuration except the most heavily contaminated one, where it stays essentially constant because the impulsive component of the disturbance propagates into the reconstructed outputs and does not diminish as the parameter estimates improve.

Figure 15 shows this decay directly. It has a concrete consequence for the interpretation of Theorem 4.1, discussed in Remark 4.5: the constant $\bar{\Delta}^2$ entering the ultimate bound (4.46) is a uniform-in-time bound, hence it is necessarily governed by the transient, whereas the quantity that actually determines the steady-state accuracy is the much smaller late-time value. The bound is consequently conservative by roughly the ratio between the two columns, which is one of the reasons why the observed errors of Section 6 lie well inside the region the theorem delimits.

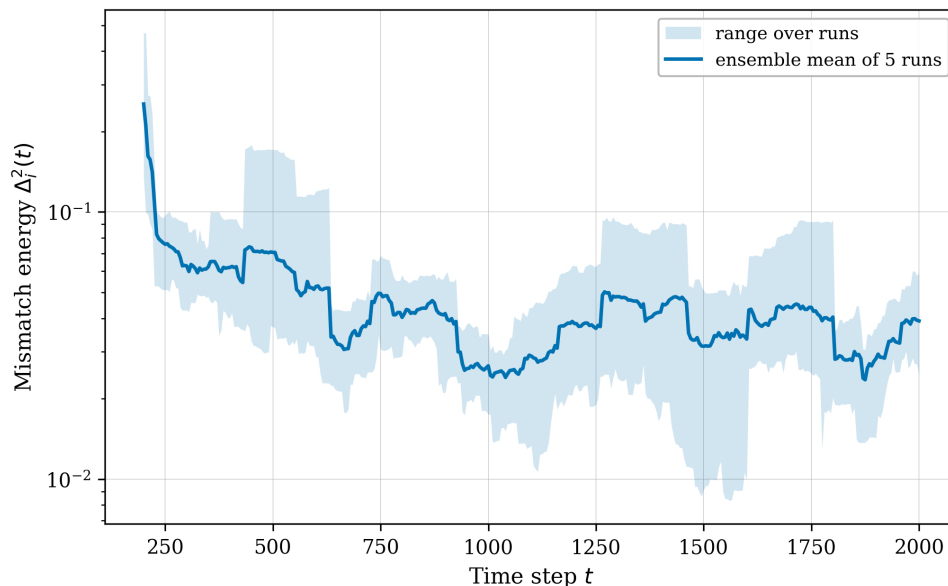


Figure 15. Evolution of the measured mismatch energy $\Delta_i^2(t)$ during adaptation, at 20% missing outputs and 2% impulsive contamination, with an ensemble mean over the independent runs, with the range across runs shaded. The energy falls sharply over the first few hundred samples, as the filter estimate settles and the reconstruction error shrinks, and then fluctuates about a level several times below its initial value. The uniform-in-time constant $\bar{\Delta}^2$ of Assumption 4.2 is necessarily governed by the initial part of this trajectory.

7. Conclusions

This paper designed a strong adaptive identification method for multivariable ARX systems based on the masked and filtered hierarchical multi-innovation stochastic gradient (MHISG) identification algorithm with missing outputs and common colored non-Gaussian disturbances. The missing outputs,

filtering, robust scoring, and adaptive normalization are captured explicitly in the formulation, which are the primary mathematical challenges arising from the presence of missing values. A reconstructed-regressor approach is used to address missing outputs, as well as a frozen-regressor stochastic-approximation interpretation. Robustness is based on the exact Huber loss and its bounded score. The output of the filtered sample is based on current and past outputs, thus a reliability mask is used. The adaptive laws for $p(t)$ and $\varepsilon(t)$ are non-circular. The boundedness theorem actually contains block excitation, strict limits on the step size, and the reconstruction/filter mismatch term, and must be viewed as a neighborhood boundedness result when noise/mismatch are nonzero, not a consistency theorem. The simulation results indicate that the proposed method is superior to zero-filling, masked-only, robust-only, and fixed filtered methods in the aspects of parameter estimation and validation.

The evaluation was carried out on a two-input two-output example and on a higher-order three-input three-output system with three times as many unknowns, over missing-output probabilities from zero to one-half, innovation standard deviations spanning a factor of three, coloring coefficients from zero to 0.9, and impulse amplitudes up to 27 standard deviations. A component-wise ablation, in which each variant is evaluated at its own best gain on a common grid, attributes the final accuracy to the individual mechanisms and shows that the two adaptive laws contribute separately and complementarily. The excitation constants and the mismatch level entering the analysis were measured directly rather than assumed: the excitation follows the quadratic law in the availability probability predicted by the theory, the robust score costs less than six percent of it in every configuration tested, and the mismatch energy falls by at least a factor of four during adaptation, which explains why the ultimate bound of Theorem 4.1 is conservative. A scalability study confirms that the cost of the adaptive machinery is linear in the number of channels while the core recursion is quadratic, so the relative overhead decreases as the system grows.

The same experiments delimit where the method should not be used. Prewhitening with a single common coefficient ceases to pay for itself beyond roughly forty percent data loss, where the certified filtered samples become too scarce, and it loses its benefit when the channel-coloring coefficients have no dominant common component, the extreme case being coefficients of opposite signs. In those regimes, the unfiltered robust variant, or the softened reliability mask, is the better choice, and the paper reports the measurements needed to make that decision.

Future research will include channel-dependent colored-noise filters, extensions to multivariable autoregressive moving average with exogenous input (ARMAX) and Hammerstein–Wiener systems, and missing-data likelihood methods. A further priority is validation on industrial records in which the missing-data pattern is generated by a real acquisition system rather than by an independent Bernoulli mechanism, since burst losses and losses correlated with the operating point are expected to affect the reliability mask differently from the independent dropout considered here.

Author contributions

Slim Dhahri: Conceptualization, methodology, software, writing – original draft; Mourad Elloumi: Conceptualization, formal analysis, methodology, supervision; Hend Aljahani: Formal analysis, funding acquisition, validation; Salem Albalawi: Data curation, investigation, validation; Sahar Almashaan: Project administration, visualization, writing – review and editing; Hatem Alwardi: Investigation, resources, software; Foued Mtiri: Formal analysis, funding acquisition, writing – review

and editing. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA, for funding this research work through the project number “NBU-FPEJ-2026-1257-03”. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through the Large Research Project under grant number RGP2/202/47.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. L. Ljung, System identification, In: *Wiley encyclopedia of electrical and electronics engineering*, John Wiley & Sons, 1999. <https://doi.org/10.1002/047134608X.W1046>
2. M. Deistler, System identification: T. Söderström, P. Stoica, Prentice Hall International, 1989, *Economet. Theor.*, **10** (1994), 813–815. <https://doi.org/10.1017/S026646660000880X>
3. R. Pintelon, J. Schoukens, *System identification: A frequency domain approach*, 2 Eds., Hoboken, NJ: IEEE Press/Wiley, 2012. <https://doi.org/10.1002/9781118287422>
4. H. F. Chen, L. Guo, *Identification and stochastic adaptive control*, Boston, MA: Birkhäuser, 1991. <https://doi.org/10.1007/978-1-4612-0429-9>
5. H. Robbins, S. Monro, A stochastic approximation method, *Ann. Math. Stat.*, **22** (1951), 400–407. <https://doi.org/10.1214/aoms/1177729586>
6. A. Benveniste, M. Métivier, P. Priouret, *Adaptive algorithms and stochastic approximations*, Berlin, Heidelberg: Springer, 1990. <https://doi.org/10.1007/978-3-642-75894-2>
7. A. H. Sayed, *Adaptive filters*, Hoboken, NJ: John Wiley & Sons, 2008. <https://doi.org/10.1002/9780470374122>
8. P. S. R. Diniz, *Adaptive filtering: Algorithms and practical implementation*, 4 Eds., New York: Springer, 2013. <https://doi.org/10.1007/978-1-4614-4106-9>
9. H. J. Kushner, G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, 2 Eds., New York: Springer, 2003. <https://doi.org/10.1007/b97441>
10. F. Ding, X. Liu, G. Liu, Several multi-innovation identification methods, *Digit. Signal Process.*, **20** (2010), 1027–1039. <https://doi.org/10.1016/j.dsp.2009.10.030>

11. Y. Liu, L. Yu, F. Ding, Multi-innovation extended stochastic gradient algorithm and its performance analysis, *Circuits Syst. Signal Process.*, **29** (2010), 649–667. <https://doi.org/10.1007/s00034-010-9174-8>
12. Y. Liu, F. Ding, Y. Shi, An efficient hierarchical identification method for general dual-rate sampled-data systems, *Automatica*, **50** (2014), 962–970. <https://doi.org/10.1016/j.automatica.2013.12.025>
13. H. Xing, F. Ding, F. Pan, E. Yang, Hierarchical recursive least squares parameter estimation methods for multiple-input multiple-output systems by using the auxiliary models, *Int. J. Adapt. Control Signal Process.*, **37** (2023), 2983–3007. <https://doi.org/10.1002/acs.3669>
14. H. Xing, F. Ding, F. Pan, Highly-computational hierarchical iterative identification methods for multiple-input multiple-output systems by using the auxiliary models, *Int. J. Robust Nonlinear Control*, **33** (2023), 10845–10863. <https://doi.org/10.1002/rnc.6917>
15. H. Xing, F. Ding, F. Pan, Auxiliary model-based hierarchical stochastic gradient methods for multiple-input multiple-output systems, *J. Comput. Appl. Math.*, **442** (2024), 115687. <https://doi.org/10.1016/j.cam.2023.115687>
16. F. Ding, Y. Xiao, L. Xu, Z. Fang, Hierarchical stochastic gradient and hierarchical multi-innovation stochastic gradient identification for multivariable ARX models, *Int. J. Adapt. Control Signal Process.*, **40** (2026), 79–102. <https://doi.org/10.1002/acs.4081>
17. F. Ding, X. Luan, L. Xu, X. Zhang, Hierarchical recursive gradient parameter identification for multi-input ARX systems with partially-coupled information vectors, *Int. J. Adapt. Control Signal Process.*, **39** (2025), 1978–1995. <https://doi.org/10.1002/acs.4036>
18. F. Ding, L. Xu, X. Zhang, H. Xu, Y. Zhou, X. Luan, Hierarchical generalized extended parameter identification for multivariable equation-error ARMA-like systems by using the filtering identification idea, *Annu. Rev. Control*, **60** (2025), 100993. <https://doi.org/10.1016/j.arcontrol.2025.100993>
19. H. Wang, Q. Zhang, X. Liu, Auxiliary model-based maximum likelihood multi-innovation recursive least squares identification for multiple-input multiple-output systems, *J. Franklin Inst.*, **361** (2024), 107352. <https://doi.org/10.1016/j.jfranklin.2024.107352>
20. H. Wang, X. Liu, Highly efficient three-stage maximum likelihood recursive least squares identification method for multiple-input multiple-output systems, *Syst. Control Lett.*, **200** (2025), 106094. <https://doi.org/10.1016/j.sysconle.2025.106094>
21. H. Wang, X. Liu, Highly efficient two-stage filtering-based maximum likelihood stochastic gradient algorithm for multiple-input multiple-output systems, *Circuits Syst. Signal Process.*, **44** (2025), 5608–5639. <https://doi.org/10.1007/s00034-025-03068-y>
22. H. Wang, X. Liu, Auxiliary model-based maximum likelihood multi-innovation forgetting gradient identification for a class of multivariable systems, *Optim. Control Appl. Met.*, **46** (2025), 1249–1264. <https://doi.org/10.1002/oca.3257>
23. Q. Zhang, H. Wang, X. Liu, Auxiliary model maximum likelihood least squares-based iterative algorithm for multivariable autoregressive output-error autoregressive moving average systems, *P. I. Mech. Eng. I-J. Sys.*, **239** (2025), 441–463. <https://doi.org/10.1177/09596518241280323>

24. Q. Zhang, X. Liu, Auxiliary model maximum likelihood moving-data-window generalized extended gradient-based iterative algorithm for multivariable autoregressive output-error autoregressive moving-average systems, *Optim. Control Appl. Meth.*, **46** (2025), 2255–2275. <https://doi.org/10.1002/oca.3321>
25. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc. B*, **39** (1977), 1–22. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
26. R. J. A. Little, D. B. Rubin, *Statistical analysis with missing data*, 3 Eds., Hoboken, NJ: John Wiley & Sons, 2019. <https://doi.org/10.1002/9781119482260>
27. R. H. Shumway, D. S. Stoffer, An approach to time series smoothing and forecasting using the EM algorithm, *J. Time Ser. Anal.*, **3** (1982), 253–264. <https://doi.org/10.1111/j.1467-9892.1982.tb00349.x>
28. A. J. Isaksson, Identification of ARX-models subject to missing data, *IEEE Trans. Automat. Contr.*, **38** (1993), 813–819. <https://doi.org/10.1109/9.277253>
29. J. Chen, Q. Zhu, Y. Liu, Modified Kalman filtering based multi-step-length gradient iterative algorithm for ARX models with random missing outputs, *Automatica*, **118** (2020), 109034. <https://doi.org/10.1016/j.automatica.2020.109034>
30. Y. Gu, J. Chen, Y. Tan, A scheme of gradient-based iterative identification with separable synchronous variable-innovation for Exp-ARX systems with arbitrary data loss, *IEEE Trans. Ind. Inform.*, **21** (2025), 7499–7509. <https://doi.org/10.1109/TII.2025.3574425>
31. F. Ding, L. Xu, X. Zhang, Y. Zhou, X. Luan, Recursive identification methods for general stochastic systems with colored noises by using the hierarchical identification principle and the filtering identification idea, *Annu. Rev. Control*, **57** (2024), 100942. <https://doi.org/10.1016/j.arcontrol.2024.100942>
32. H. Xing, F. Ding, X. Zhang, X. Luan, E. Yang, Highly-efficient filtered hierarchical identification algorithms for multiple-input multiple-output systems with colored noises, *Syst. Control Lett.*, **186** (2024), 105762. <https://doi.org/10.1016/j.sysconle.2024.105762>
33. H. Xing, G. Miao, F. Ding, Three-stage filtered gradient identification methods for multivariable ARX systems with colored noise, *Optim. Control Appl. Meth.*, **47** (2026), 372–386. <https://doi.org/10.1002/oca.70063>
34. H. Xu, L. Xu, B. Shen, A novel three-stage filtering identification algorithm for the exponential autoregressive time-series model, *IEEE Signal Proc. Lett.*, **32** (2025), 2089–2093. <https://doi.org/10.1109/LSP.2025.3567804>
35. Y. Zou, L. Xu, Parameter identification and open-circuit voltage estimation of lithium-ion battery circuit model based on multi-innovation theory, *Int. J. Circ. Theor. Appl.*, **54** (2026), 2747–2764. <https://doi.org/10.1002/cta.70128>
36. Y. Liu, X. Yin, F. Ding, Q. Zhu, Greedy orthogonal least squares identification for multivariable Hammerstein models, *Circuits Syst. Signal Process.*, **45** (2026), 1865–1886. <https://doi.org/10.1007/s00034-025-03327-y>

37. Z. Li, S. Ding, L. Zhang, G. Song, Filter-based predefined-time optimal fault-tolerant consensus control for nonlinear multi-agent systems via reinforcement learning, *Neurocomputing*, **679** (2026), 133224. <https://doi.org/10.1016/j.neucom.2026.133224>
38. W. Liu, P. P. Pokharel, J. C. Principe, Correntropy: Properties and applications in non-Gaussian signal processing, *IEEE Trans. Signal Process.*, **55** (2007), 5286–5298. <https://doi.org/10.1109/TSP.2007.896065>
39. X. Wang, Y. Duan, Heterogeneous mixed kernel correntropy-based robust iterative estimation methods and convergence analysis for the nonlinear system with outliers, *Circuits Syst. Signal Process.*, **44** (2025), 6550–6581. <https://doi.org/10.1007/s00034-025-03156-z>
40. X. Wang, Y. Duan, Gaussian–Versoria mixed kernel correntropy-based robust parameter and state estimation for bilinear state–space systems with non-Gaussian process and measurement noises, *Axioms*, **14** (2025), 630. <https://doi.org/10.3390/axioms14080630>
41. F. Barani, A. Savadi, H. Sadoghi Yazdi, A distributed learning based on robust diffusion SGD over adaptive networks with noisy output data, *J. Parallel Distr. Com.*, **190** (2024), 104883. <https://doi.org/10.1016/j.jpdc.2024.104883>
42. W. Wang, S. Liu, Y. Jiang, J. Sun, W. Xu, X. Chen, A robust filter and smoother-based expectation-maximization algorithm for bilinear systems with heavy-tailed noise, *Mech. Syst. Signal Pr.*, **236** (2025), 112912. <https://doi.org/10.1016/j.ymssp.2025.112912>
43. G. Qian, L. Shen, Y. Guan, J. Qian, S. Wang, Robust recursive widely linear adaptive filtering algorithm for censored regression, *Signal Process.*, **230** (2025), 109854. <https://doi.org/10.1016/j.sigpro.2024.109854>
44. B. Chen, L. Xing, H. Zhao, N. Zheng, J. C. Príncipe, Generalized correntropy for robust adaptive filtering, *IEEE Trans. Signal Process.*, **64** (2016), 3376–3387. <https://doi.org/10.1109/TSP.2016.2539127>
45. B. Chen, X. Liu, H. Zhao, J. C. Príncipe, Maximum correntropy Kalman filter, *Automatica*, **76** (2017), 70–77. <https://doi.org/10.1016/j.automatica.2016.10.004>
46. Y. Zou, S. C. Chan, T. S. Ng, Least mean M-estimate algorithms for robust adaptive filtering in impulse noise, *IEEE Trans. Circuits Syst. II*, **47** (2000), 1564–1569. <https://doi.org/10.1109/82.899657>
47. S. C. Chan, Y. X. Zou, A recursive least M-estimate algorithm for robust adaptive filtering in impulsive noise: Fast algorithm and convergence performance analysis, *IEEE Trans. Signal Process.*, **52** (2004), 975–991. <https://doi.org/10.1109/TSP.2004.823496>
48. P. J. Huber, Robust estimation of a location parameter, *Ann. Math. Stat.*, **35** (1964), 73–101. <https://doi.org/10.1214/aoms/1177703732>
49. P. J. Huber, E. M. Ronchetti, *Robust statistics*, 2 Eds., Hoboken, NJ: John Wiley & Sons, 2009. <https://doi.org/10.1002/9780470434697>
50. F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel, *Robust statistics: The approach based on influence functions*, New York: John Wiley & Sons, 1986. <https://doi.org/10.1002/9781118186435>

51. H. Raghavan, A. K. Tangirala, R. B. Gopaluni, S. L. Shah, Identification of chemical processes with irregular output sampling, *Control Eng. Pract.*, **14** (2006), 467–480. <https://doi.org/10.1016/j.conengprac.2005.01.015>
52. S. Gibson, B. Ninness, Robust maximum-likelihood estimation of multivariable dynamic systems, *Automatica*, **41** (2005), 1667–1682. <https://doi.org/10.1016/j.automatica.2005.05.008>
53. T. B. Schön, A. Wills, B. Ninness, System identification of nonlinear state-space models, *Automatica*, **47** (2011), 39–49. <https://doi.org/10.1016/j.automatica.2010.10.013>
54. X. Chen, S. Zhao, F. Liu, Robust identification of linear ARX models with recursive EM algorithm based on Student's t-distribution, *J. Franklin Inst.*, **358** (2021), 1103–1121. <https://doi.org/10.1016/j.jfranklin.2020.06.003>
55. F. Ding, T. Chen, Performance analysis of multi-innovation gradient type identification methods, *Automatica*, **43** (2007), 1–14. <https://doi.org/10.1016/j.automatica.2006.07.024>



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)