AIMS Mathematics, 11(9): 28202–28253.
DOI: 10.3934/math.20261124
Received: 23 May 2026
Revised: 16 July 2026
Accepted: 31 July 2026
Published: 03 September 2026

http://www.aimspress.com/journal/Math

Research article

Robust recursive identification of interconnected fractional Hammerstein systems under impulsive disturbances

Slim Dhahri[1,*], Mourad Elloumi[2,3], Hend Aljahani[4], Salem Albalawi[5], Sahar Almashaan[5,*], Hatem Alwardi[5] and Foued Mtiri[6]

[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: Interconnected fractional Hammerstein models are useful for nonlinear processes with hereditary local dynamics and cross-subsystem interactions, but ordinary fractional recursive extended least squares (FRELS) estimators remain fragile under impulsive measurements. This paper develops Huber and maximum correntropy criterion (MCC) robust stochastic FRELS (RS-FRELS) recursions for finite-memory fractional interconnected Hammerstein regressions. The key point is not a new robust score but a careful frozen-weight interpretation: The online update exactly solves a weighted quadratic surrogate whose weights are evaluated at prior innovations, rather than the full nonlinear robust empirical risk. This framing yields explicitly conditional guarantees. On the algebraic side, these cover surrogate optimality, covariance positivity under either a deterministic or an in-expectation weighted-excitation condition, bounded single-sample influence, and an estimator-side saturation variant for leverage control. On the statistical side, they cover Huber population consistency under score orthogonality, a local orthogonality-defect bias bound, MCC stationarity without global uniqueness, local stochastic approximation convergence in general, global convergence for projected Huber RS-FRELS on a compact convex projection set under uniform Huber curvature, and conservative tracking under slow drift. The assumptions are connected to checkable diagnostics for weighted information, score defects, saturation activation, and regressor norms, together with

a step-by-step verification workflow for practitioners. Simulations include 30-seed Monte Carlo comparisons, paired significance tests, heavy-tailed and leverage stress tests, MCC kernel and impulse severity sweeps, same-regressor algorithm baselines including a redescending recursive least M-estimate and adaptively tuned robust variants, model structure ablations, and scalability timing up to 50 subsystems. The synthetic study is complemented by a real-data validation on the DaISy (Database for the Identification of Systems) liquid–saturated steam heat exchanger benchmark, in which the robust recursions match ordinary FRELS on the raw record and improve held-out prediction accuracy under injected impulsive sensor faults. The results support the bounded score mechanism while clearly separating what the experiments validate from unresolved colored output error consistency questions.

Keywords: fractional-order systems; Hammerstein systems; interconnected systems; recursive identification; robust RLS; Huber M-estimation; maximum correntropy; impulsive noise; frozen-weight IRLS; scalability diagnostics
Mathematics Subject Classification: 26A33, 62F35, 62L20, 93E12, 93E24

1. Introduction

Block-oriented nonlinear models are widely used in system identification because they combine interpretable static nonlinearities with dynamic elements of moderate dimension. The Hammerstein architecture, consisting of a memoryless nonlinear block followed by a linear dynamic block, is among the most common such structures [1–3]. Fractional-order Hammerstein models further enrich this architecture by replacing integer-order difference operators with noninteger operators capable of representing hereditary and long-memory effects; the mathematical foundations of such operators are classical [4, 5]. The identification of discrete fractional-order Hammerstein systems was considered in [6], and evolutionary recursive estimators subsequently made this model class practically identifiable [7, 8]. Recent fractional Hammerstein identification methods have considered auxiliary-model recursive least squares (RLS) for interconnected fractional regressions [9], separable and orthogonal least squares fractional parameterizations [10, 11], and staged or two-stage estimation schemes for fractional and parameter-varying Hammerstein systems [12, 13].

A remaining limitation is robustness to impulsive disturbances. Ordinary least squares and RLS algorithms give quadratic emphasis to large innovations, so a small number of outliers can dominate the update. This problem has motivated classical robust M-estimation [14, 15], recursive least-M-estimate and Huber-type robust RLS filters designed for impulsive noise environments [16, 17], correntropy criteria for non-Gaussian signal processing [18, 19], and recursive maximum correntropy algorithms with generalized or proportionate extensions [20, 21]. In nonlinear system identification, recent work continues to address impulsive noise in recursive Hammerstein and errors-in-variables estimation [22, 23], metaheuristic robust identification of fractional Hammerstein models [24, 25], and multi-innovation recursions for sandwich nonlinear systems [26]. Nevertheless, the combination of fractional memory, interconnection, Hammerstein nonlinearities, scalability diagnostics, and robust online recursive estimation remains insufficiently developed.

To state the research gap precisely, it is useful to identify the mechanism by which each existing algorithm family fails in the scenario studied here. (i) Quadratic-score fractional recursions (fractional

recursive extended least squares (FRELS)/RLS applied to Grünwald–Letnikov (GL) regressors [7, 9]) have an unbounded influence function: A single innovation of size $|e|$ moves the estimate proportionally to $|e|$, so one impulsive measurement can undo hundreds of clean updates. The damage is amplified in the fractional setting, because the impulse also enters the finite-memory GL filters and persists in the own-output regressors for L subsequent samples, converting a single vertical outlier into a sequence of leverage points. (ii) Robust adaptive filters (recursive least M-estimate and Huber-RLS [16, 17], MCC and its recursive variants [20, 21]) bound the score, but they were developed for integer-order finite impulse response (FIR) or channel models with exogenous regressors; they provide neither fractional memory nor an interconnection structure, and their analyses do not address the frozen-weight surrogate interpretation or the score orthogonality issue raised by colored output regressors. (iii) Fractional Hammerstein identification methods (the evolutionary recursive estimator of [7], separable and orthogonal least squares [10, 11], staged Levenberg–Marquardt (LM) estimation [12], and metaheuristic estimators [24, 25]) assume finite variance and mostly Gaussian disturbances and use quadratic criteria or batch optimization, so they inherit the impulse fragility of (i) or give up the online recursive structure altogether. The gap addressed in this paper is therefore specific: An online estimator with RLS-level cost for finite-memory fractional interconnected Hammerstein regressions that provably bounds the influence of impulsive innovations, makes the required excitation and orthogonality conditions explicit and checkable, and quantifies what happens when they fail.

The method proposed here starts from the finite-memory fractional interconnected Hammerstein regression

$$y_i(k) = \theta_i^{\star\top}\psi_i(k) + \xi_i(k), \qquad \xi_i(k) = C_i(\Delta_L^{\nu_i})e_i(k) + \eta_i(k), \tag{1.1}$$

where the regressor $\psi_i(k)$ contains fractional own-output terms, nonlinear fractional input terms, and fractional interconnection terms, while $\eta_i(k)$ represents impulsive contamination. The symbol $\xi_i(k)$ always denotes this colored plus impulsive disturbance; the expression above fixes its form once and for all, and Section 3 only restates it with the shorthand $\chi_i(k) = C_i(\Delta_L^{\nu_i})e_i(k)$ for the colored component. The ordinary FRELS innovation is replaced by either a Huber score or a maximum correntropy score. The resulting recursion has the same computational structure as RLS, but its gain is multiplied by a residual-dependent weight.

A central mathematical point is that the algorithm is not claimed to be the exact recursive minimizer of the full nonlinear robust empirical risk. At time k, the robust weight is computed from the prior innovation and then frozen. The recursion therefore exactly minimizes a weighted quadratic surrogate and should be interpreted as an online one-step iteratively reweighted least squares (IRLS)/RLS method. This distinction is essential in systems with colored output regressors, because the frozen weights are path-dependent, and the score orthogonality need not follow automatically from the noise model. The convergence and tracking statements below are therefore conditional results under explicit boundedness, excitation, and local mean field assumptions.

The contributions are summarized as follows; established ingredients (robust scores, weighted RLS algebra, GL regressors) are explicitly credited to the cited literature, and only the items below are claimed as new for this model class.

(C1) Algorithm. A frozen-weight Huber/MCC robust stochastic FRELS (RS-FRELS) recursion for finite-memory fractional interconnected Hammerstein regressions under colored plus impulsive disturbances, with optional estimator-side regressor saturation and projection.

(C2) Exact surrogate and influence theory. The recursion is characterized as the exact minimizer of a frozen-weight quadratic surrogate with arbitrary initialization (Theorem 5.1), and the single-sample influence is proved to be bounded, with a saturation corollary that makes the bounded regressor requirement an algorithmic design choice rather than a data assumption.

(C3) Conditional convergence theory. Weighted persistent excitation is supported by two checkable sufficient conditions, namely a deterministic one and an in-expectation one based on a matrix Azuma–Hoeffding bound that covers the MCC case; Huber–Fisher consistency; a local orthogonality-defect bias bound (Proposition 5.2 in Section 5); local stochastic approximation convergence; a global convergence theorem for projected Huber RS-FRELS under uniform Huber curvature, together with an explanation of why MCC cannot be globalized the same way; and a conservative tracking lemma are all stated under explicit conditional assumptions.

(C4) Verification-oriented evidence. Assumption-linked diagnostics (weighted information, score defect, saturation activation), 30-seed Monte Carlo comparisons with paired tests and additional robust baselines, stress and sensitivity studies, scalability timing with per-method runtimes, and a real-data validation on the DaISy heat exchanger benchmark under injected impulsive sensor faults.

A detailed mathematical comparison with closely related recent work is provided in Appendix A (Table 13). In short, established ingredients such as robust RLS, Huber M-estimation, MCC adaptive filtering, and finite-memory fractional regressors are combined here into a finite-memory GL interconnected Hammerstein regression, together with the arbitrary initialization frozen surrogate identity, the estimator-side leverage–saturation influence corollary, and diagnostics that expose when the conditional assumptions are plausible. The paper therefore presents a disciplined synthesis and analysis for this model class rather than claiming to invent robust recursive estimation itself.

2. Preliminaries

This section contains only definitions and standard facts used later. Proofs are deliberately omitted for well-known results and appropriate references are given. Elementary results that are specific to the proposed robust recursion are moved to Section 5.

For a vector $x \in \mathbb{R}^n$, $\|x\|$ denotes the Euclidean norm. For a symmetric matrix A, $A \succ 0$ and $A \succeq 0$ denote positive definiteness and positive semidefiniteness, respectively. The smallest and largest eigenvalues of a symmetric matrix are denoted $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$. All signals are indexed by $k \in \mathbb{N}$. For a stochastic sequence, $\mathcal{F}_k$ denotes the natural filtration generated by the data up to time k.

Definition 2.1 (Finite-memory GL fractional difference [4, 5, 27])**.** *Let $\mu > 0$ and let $L \in \mathbb{N}$ be a prescribed finite memory length. The causal truncated GL fractional difference of a scalar sequence $x(k)$ is*

$$\Delta_L^{\mu} x(k) = \sum_{\ell=0}^{\min\{L,k\}} g_\ell^{(\mu)} x(k-\ell), \qquad g_\ell^{(\mu)} = (-1)^\ell \binom{\mu}{\ell} = (-1)^\ell \frac{\Gamma(\mu+1)}{\Gamma(\ell+1)\Gamma(\mu-\ell+1)}. \tag{2.1}$$

The finite-memory gain is denoted by

$$G_{\mu,L} = \sum_{\ell=0}^{L} \left| g_\ell^{(\mu)} \right|. \tag{2.2}$$

Proposition 2.1 (Known boundedness of finite-memory GL filters [5,7,9]). *If $|x(k)| \le M_x$ for all k and if $L < \infty$, then $\left|\Delta_L^{\mu} x(k)\right| \le M_x G_{\mu,L}$ for all k. This follows directly from the triangle inequality applied to the finite convolution in* (2.1).

Remark 2.1 (Finite-memory approximation). The analysis in this paper is for the finite-memory model defined by (2.1). If the true physical plant has infinite fractional memory, replacing it by (2.1) introduces truncation bias. Increasing L reduces this modeling approximation only when the omitted tail is sufficiently small; it does not follow from the recursive estimator itself. Therefore, the consistency statements below refer to the parameter of the finite-memory regression model.

Proposition 2.2 (Elementary GL truncation-tail bound [5]). *Let the infinite-memory GL difference be formally written as*

$$\Delta_{\infty}^{\mu} x(k) = \sum_{\ell=0}^{k} g_{\ell}^{(\mu)} x(k-\ell), \tag{2.3}$$

and suppose $|x(k)| \le M_x$. Then the finite-memory truncation error satisfies

$$\left|\Delta_{\infty}^{\mu} x(k) - \Delta_L^{\mu} x(k)\right| \le M_x \sum_{\ell=L+1}^{k} |g_{\ell}^{(\mu)}|. \tag{2.4}$$

Consequently, if a regression uses n_x such filtered signals with bounded coefficient magnitudes, the induced one-step modeling error is bounded by the corresponding coefficient-weighted sum of the tails in (2.4). *This proposition quantifies the finite-memory approximation error but does not remove the need to choose L large enough for the plant being modeled.*

Definition 2.2 (Huber [14, 15] and maximum correntropy [18, 19] scores). *The Huber loss with threshold $\delta > 0$ is*

$$\rho_{\delta}^{H}(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le \delta, \\ \delta|e| - \frac{1}{2}\delta^2, & |e| > \delta, \end{cases} \qquad s_{\delta}^{H}(e) = \frac{\mathrm{d}}{\mathrm{d}e}\rho_{\delta}^{H}(e) = \begin{cases} e, & |e| \le \delta, \\ \delta\,\mathrm{sgn}(e), & |e| > \delta. \end{cases} \tag{2.5}$$

The maximum correntropy-type loss used here is

$$\rho_{\sigma}^{C}(e) = \sigma^2\left(1 - \exp\left[-\frac{e^2}{2\sigma^2}\right]\right), \qquad s_{\sigma}^{C}(e) = e\exp\left[-\frac{e^2}{2\sigma^2}\right], \tag{2.6}$$

where $\sigma > 0$ is the kernel width. In both cases, the recursive algorithm uses

$$\omega(e) = \begin{cases} s(e)/e, & e \ne 0, \\ 1, & e = 0, \end{cases} \tag{2.7}$$

so that $\omega(e)e = s(e)$. Explicitly, we have

$$\omega_{\delta}^{H}(e) = \begin{cases} 1, & |e| \le \delta, \\ \delta/|e|, & |e| > \delta, \end{cases} \qquad \omega_{\sigma}^{C}(e) = \exp\left[-\frac{e^2}{2\sigma^2}\right]. \tag{2.8}$$

Proposition 2.3 (Weighted RLS identity [17, 28, 29]). *Let P^{-1} be nonsingular, $u \in \mathbb{R}^d$, $\lambda > 0$, and $\omega \ge 0$. The inverse of $\lambda P^{-1} + \omega uu^{\top}$ is*

$$\left(\lambda P^{-1} + \omega uu^{\top}\right)^{-1} = \lambda^{-1}P - \lambda^{-1}\frac{\omega Puu^{\top}P}{\lambda + \omega u^{\top}Pu}. \tag{2.9}$$

This is the Sherman–Morrison–Woodbury identity and is the standard algebra behind RLS and weighted RLS recursions.

3. Fractional interconnected Hammerstein regression model

Consider a process composed of N_s interconnected subsystems. For subsystem $i \in \{1, \dots, N_s\}$, the measured output is modeled as

$$y_i(k) = \theta_i^{\star\top}\psi_i(k) + \xi_i(k), \tag{3.1}$$

where $\theta_i^\star \in \mathbb{R}^{d_i}$ is a finite-dimensional lumped parameter vector, $\psi_i(k) \in \mathbb{R}^{d_i}$ is a fractional Hammerstein regressor, and $\xi_i(k)$ is a colored plus impulsive disturbance.

A representative regressor has the form

$$\begin{aligned}\psi_i(k) = \big[&- \Delta_L^{\alpha_{i,1}} y_i(k-1), \dots, -\Delta_L^{\alpha_{i,n_i}} y_i(k-n_i),\\ &\Delta_L^{\beta_{i,1}} u_i(k), \Delta_L^{\beta_{i,2}} u_i^2(k), \dots, \Delta_L^{\beta_{i,r_i}} u_i^{r_i}(k),\\ &\{\Delta_L^{\gamma_{ij,m}} y_j(k-m) : j \neq i,\ m \in \mathcal{D}_{ij}\}\big]^\top.\end{aligned} \tag{3.2}$$

The fractional orders $\alpha_{i,m}, \beta_{i,r}, \gamma_{ij,m} > 0$ are assumed known in the recursive estimation developed here. Unknown fractional orders can be estimated with evolutionary identification methods [7, 8] or with separable, orthogonal, and staged estimation schemes for fractional Hammerstein models [10–12], but order estimation is not the focus of the present work.

The disturbance is decomposed as

$$\xi_i(k) = \chi_i(k) + \eta_i(k), \qquad \chi_i(k) = C_i(\Delta_L^{\nu_i})e_i(k), \tag{3.3}$$

where $C_i(\Delta_L^{\nu_i})$ is a finite-memory fractional coloring filter, $e_i(k)$ is a nominal zero-mean innovation, and $\eta_i(k)$ is an impulsive component. This is the same disturbance already introduced in (1.1); the only new notation is the shorthand $\chi_i(k)$ for its colored part. Typical simulations use a contaminated distribution

$$\eta_i(k) = b_i(k)a_i(k), \qquad b_i(k) \sim \text{Bernoulli}(p_i), \tag{3.4}$$

where $a_i(k)$ is large-amplitude Gaussian or heavy-tailed noise. The robust recursion does not require a finite variance for $a_i(k)$ in the bounded single-sample influence theorem; however, convergence theorems require separate regularity assumptions, which are stated later.

Remark 3.1 (Lumped Hammerstein identifiability, scaling ambiguity, and normalization options). Equation (3.1) identifies the lumped regression parameter $\theta_i^\star$. It does not, by itself, separately identify the static nonlinearity coefficients and the dynamic-block coefficients of a classical Hammerstein cascade. Indeed, multiplying the static block by a nonzero scalar c and the dynamic block by $1/c$ gives the same input–output map. Separate block identification therefore requires an additional normalization, such as fixing one static coefficient or one dynamic gain. This paper estimates the lumped parameter vector unless such a normalization is imposed. The practical impact of this inherent scaling ambiguity should be stated explicitly. The lumped vector $\theta_i^\star$, the one-step predictor $\hat{\theta}_i^\top\psi_i$, all prediction-error metrics, and the normalized mean square deviation (NMSD) used in Section 6 are invariant to the ambiguity because they never require the blocks to be separated; in particular, every algorithm comparison in this paper is unaffected, since all estimators target the same lumped parameterization. The ambiguity matters only when the user wants physically interpretable block coefficients. In that case, two standard normalizations can be applied after (or during) estimation without changing the recursion: (i) Fix the leading static nonlinearity coefficient to one, $c_{i,1} = 1$, and

absorb the scale into the dynamic block, which is the most common convention and is well conditioned whenever the linear input term is non-negligible; or (ii) constrain the static coefficient vector to a unit Euclidean norm with a fixed sign convention, which distributes the scale more evenly when the leading coefficient may be small. Both normalizations amount to a deterministic rescaling of the subvectors of $\hat{\theta}_i$ and therefore inherit the convergence properties of the lumped estimate; they differ only in numerical conditioning.

4. Robust stochastic FRELS algorithm

For each subsystem, the recursion may be run either with the raw fractional Hammerstein regressor $\psi_i(k)$ or with an estimator-side saturated regressor. For a user-selected level $B_i > 0$, define

$$\mathcal{S}_{B_i}(x) = \begin{cases} x, & \|x\| \le B_i, \\ B_i x / \|x\|, & \|x\| > B_i. \end{cases} \tag{4.1}$$

The actual regressor supplied to the update is

$$\varphi_i(k) = \begin{cases} \psi_i^{\mathrm{raw}}(k), & \text{raw RS-FRELS}, \\ \mathcal{S}_{B_i}(\psi_i^{\mathrm{raw}}(k)), & \text{saturated RS-FRELS}. \end{cases} \tag{4.2}$$

For notational simplicity, the equations and theorems below write $\psi_i(k)$ for the regressor actually supplied to the recursion; thus $\psi_i(k)$ should be read as $\varphi_i(k)$ when saturation is enabled. The same convention applies to the optional projection Π_{Θ_i} of the parameter vector onto a convex set Θ_i: Projection is not needed for the algebraic surrogate theorem but it is useful in the stochastic approximation and bounded regressor variants.

For each subsystem, define the prior innovation

$$\varepsilon_i(k) = y_i(k) - \hat{\theta}_i^{\top}(k-1)\psi_i(k). \tag{4.3}$$

Given a robust score s_i and weight $\omega_i(e) = s_i(e)/e$, the proposed recursion is

$$K_i(k) = \frac{\omega_i(\varepsilon_i(k))P_i(k-1)\psi_i(k)}{\lambda_i + \omega_i(\varepsilon_i(k))\psi_i^{\top}(k)P_i(k-1)\psi_i(k)}, \tag{4.4}$$

$$\hat{\theta}_i(k) = \hat{\theta}_i(k-1) + K_i(k)\varepsilon_i(k), \tag{4.5}$$

$$P_i(k) = \lambda_i^{-1}P_i(k-1) - \lambda_i^{-1}K_i(k)\psi_i^{\top}(k)P_i(k-1), \tag{4.6}$$

where $0 < \lambda_i \le 1$ is the forgetting factor. In (4.4)–(4.6), and in all information-form recursions derived from them, the weight $\omega_i(k) = \omega_i(\varepsilon_i(k))$ is computed from the explicit formulas in (2.8), according to whether the Huber or the MCC score has been selected. If $\omega_i \equiv 1$, (4.4)–(4.6) reduce to the ordinary FRELS/RLS update applied to the fractional regressor.

Remark 4.1 (Exact surrogate, not the exact nonlinear robust minimizer). The nonlinear robust empirical criterion

$$\begin{aligned} J_{i,k}(\theta) = &\sum_{\ell=1}^{k} \lambda_i^{k-\ell}\rho_i\big(y_i(\ell) - \theta^{\top}\psi_i(\ell)\big) \\ &+ \frac{1}{2}\lambda_i^{k}(\theta - \theta_{i,0})^{\top}P_i(0)^{-1}(\theta - \theta_{i,0}), \end{aligned} \tag{4.7}$$

will have score equations in which every weight depends on the candidate vector θ. The implementable recursion instead evaluates $\omega_i(\varepsilon_i(k))$ using the previous estimate and then freezes that weight. Thus RS-FRELS is an online one-step IRLS/RLS approximation of (4.7); it exactly solves a frozen-weight quadratic surrogate, as proved below. This distinction is essential for the convergence statements.

Algorithm 1 Huber/MCC RS-FRELS with optional regressor saturation and projection

Require: Forgetting factors λ_i, initial estimates $\hat{\theta}_i(0)$, $P_i(0) \succ 0$, robust score s_i, optional saturation bounds B_i, optional projection sets Θ_i.

for $k = 1, 2, \ldots$ **do**
  for $i = 1, \ldots, N_s$ **do**
    Form the raw fractional Hammerstein regressor $\psi_i^{\mathrm{raw}}(k)$ from (3.2).
    Set $\psi_i(k) \leftarrow \psi_i^{\mathrm{raw}}(k)$, or $\psi_i(k) \leftarrow \mathcal{S}_{B_i}(\psi_i^{\mathrm{raw}}(k))$ if saturation is enabled.
    Compute $\varepsilon_i(k) = y_i(k) - \hat{\theta}_i^{\top}(k-1)\psi_i(k)$.
    Set $\omega_i(k) = s_i(\varepsilon_i(k))/\varepsilon_i(k)$ if $\varepsilon_i(k) \neq 0$ and $\omega_i(k) = 1$ otherwise.
    Compute $K_i(k)$ from (4.4) and update $P_i(k)$ from (4.6).
    Compute $\theta_i^{+}(k) = \hat{\theta}_i(k-1) + K_i(k)\varepsilon_i(k)$.
    Set $\hat{\theta}_i(k) = \theta_i^{+}(k)$, or $\hat{\theta}_i(k) = \Pi_{\Theta_i}(\theta_i^{+}(k))$ if projection is enabled.
  end for
end for

Remark 4.2 (Projection versus exact surrogate optimality). When projection is disabled, Algorithm 1 is exactly the weighted RLS recursion characterized in Theorem 5.1. When projection is enabled, Theorem 5.1 applies to the unprojected intermediate vector $\theta^{+}(k)$. The projected iterate $\Pi_{\Theta}(\theta^{+}(k))$ should therefore be interpreted as a stabilized projected stochastic approximation variant, not as the unconstrained minimizer of the frozen-weight quadratic surrogate, unless $\theta^{+}(k) \in \Theta$. A constrained-surrogate interpretation would require replacing the minimization domain in (5.3) by Θ and solving the constrained quadratic problem.

Remark 4.3 (Per-sample computational cost and the path to a scalable implementation). Let d_i be the regressor dimension of subsystem i. With a full covariance matrix, the dominant update cost is the matrix-vector product $P_i(k-1)\psi_i(k)$ and the rank-one covariance correction, both $O(d_i^2)$ per subsystem and $O(d_i^2)$ for the memory. Thus the centralized per-sample update cost is $O(\sum_{i=1}^{N_s} d_i^2)$ and the memory requirement is $O(\sum_{i=1}^{N_s} d_i^2)$. Forming finite-memory GL regressors adds $O(Ld_i)$ arithmetic if implemented by direct finite convolutions, or reduces this cost when cached convolution states are used. For dense all-to-all interconnections of the representative form (3.2), $d_i = 5 + (N_s - 1)$; for sparse coupling with the maximum degree q_i, $d_i = 5 + q_i$. The scalability experiment in Section 6 reports the empirical runtime associated with this full-covariance implementation. To avoid overstatement, the scope of the word "scalable" used in this paper is made precise here, together with the route to genuinely large systems. For dense all-to-all coupling, d_i grows linearly in N_s and the full-covariance cost grows like $O(N_s^3)$ in total, so the dense full-covariance implementation is not claimed to be large-scale; the timing study only verifies that the measured cost follows the predicted $O(\sum_i d_i^2)$ law up to $N_s = 50$. Three structural reductions give a practical path to larger systems. First, physical interconnections are typically sparse: With bounded coupling degree $q_i \leq q$, the total cost is $O(N_s(5+q)^2)$, i.e., linear in the number of subsystems. Second, the update decouples across subsystems

given the measured interconnection signals, so the N_s recursions can be executed in parallel with no communication beyond the shared measurements; the per-processor cost is then $O(d_i^2)$ regardless of N_s. Third, when even $O(d_i^2)$ is too expensive, the same robust score can drive an $O(d_i)$ normalized-gradient update. The Huber-NLMS baseline in Section 6 is exactly this variant, and its per-method runtime is reported in Table 10 so that the accuracy–cost trade-off is documented rather than hidden.

5. Main results

The results are stated for one subsystem in order to avoid notational repetition. All statements apply componentwise to the interconnected model by replacing $(y, \psi, \theta, P, \lambda)$ with $(y_i, \psi_i, \theta_i, P_i, \lambda_i)$. The first result is the algebraic core of the method: Once the robust weights have been computed from prior innovations and are treated as fixed data, the recursion is an exact weighted RLS solver for a well-defined quadratic surrogate.

Theorem 5.1 (Exact solution of the frozen-weight surrogate)**.** *Let $P(0) \succ 0$, $\theta_0 \in \mathbb{R}^d$, $R(0) = P(0)^{-1}$, and $r(0) = P(0)^{-1}\theta_0$. Assume that the weights $\omega(\ell) \geq 0$ are frozen after they are computed from the prior innovations. Define*

$$R(k) = \lambda R(k-1) + \omega(k)\psi(k)\psi^\top(k), \tag{5.1}$$

$$r(k) = \lambda r(k-1) + \omega(k)\psi(k)y(k). \tag{5.2}$$

Then $\hat{\theta}(k) = R(k)^{-1}r(k)$ is the unique minimizer of

$$\widetilde{J}_k(\theta) = \frac{1}{2}\sum_{\ell=1}^{k} \lambda^{k-\ell}\omega(\ell)[y(\ell) - \theta^\top\psi(\ell)]^2 + \frac{1}{2}\lambda^k(\theta-\theta_0)^\top P(0)^{-1}(\theta-\theta_0). \tag{5.3}$$

Moreover, if $P(k) = R(k)^{-1}$, this minimizer is generated by the gain and covariance recursions (4.4)–(4.6).

Proof. The proof has three steps. First, the objective in (5.3) is a strictly convex quadratic function. Expanding the terms that depend on θ gives

$$\begin{aligned}\widetilde{J}_k(\theta) = \frac{1}{2}\theta^\top\left[\lambda^k P(0)^{-1} + \sum_{\ell=1}^{k}\lambda^{k-\ell}\omega(\ell)\psi(\ell)\psi^\top(\ell)\right]\theta \\ - \theta^\top\left[\lambda^k P(0)^{-1}\theta_0 + \sum_{\ell=1}^{k}\lambda^{k-\ell}\omega(\ell)\psi(\ell)y(\ell)\right] + c_k,\end{aligned} \tag{5.4}$$

where c_k is independent of θ. Expanding (5.1) and (5.2) yields exactly the two bracketed quantities in (5.4). Since $P(0) \succ 0$, $\lambda > 0$, and each $\omega(\ell)\psi(\ell)\psi^\top(\ell)$ is positive semidefinite, $R(k) \succ 0$. Hence $\widetilde{J}_k$ is strictly convex and its only stationary point

$$\nabla\widetilde{J}_k(\theta) = R(k)\theta - r(k) = 0, \tag{5.5}$$

is the unique minimizer $\hat{\theta}(k) = R(k)^{-1}r(k)$.

Second, the covariance recursion follows from the matrix inversion lemma. From $R(k) = \lambda R(k-1) + \omega(k)\psi(k)\psi^\top(k)$ and $P(k-1) = R(k-1)^{-1}$, we have

$$P(k) = \lambda^{-1}P(k-1) - \lambda^{-1}\frac{\omega(k)P(k-1)\psi(k)\psi^\top(k)P(k-1)}{\lambda + \omega(k)\psi^\top(k)P(k-1)\psi(k)}. \tag{5.6}$$

Defining

$$K(k) = \frac{\omega(k)P(k-1)\psi(k)}{\lambda + \omega(k)\psi^\top(k)P(k-1)\psi(k)} \tag{5.7}$$

turns (5.6) into (4.6).

Third, the parameter update follows by multiplying the information vector by the updated covariance

$$\hat{\theta}(k) = P(k)r(k) = P(k)\{\lambda r(k-1) + \omega(k)\psi(k)y(k)\}. \tag{5.8}$$

Using $r(k-1) = P(k-1)^{-1}\hat{\theta}(k-1)$ and the identity $P(k)\omega(k)\psi(k) = K(k)$, which is obtained by multiplying (5.6) by $\omega(k)\psi(k)$, gives

$$\begin{aligned}\hat{\theta}(k) &= [I - K(k)\psi^\top(k)]\hat{\theta}(k-1) + K(k)y(k)\\ &= \hat{\theta}(k-1) + K(k)[y(k) - \psi^\top(k)\hat{\theta}(k-1)].\end{aligned} \tag{5.9}$$

This is precisely (4.5). The factor λ^k in the prior term is not cosmetic: Without it, the recursion $R(k) = \lambda R(k-1) + \cdots$ would not be the normal equation of the stated objective. Similarly, if $r(0) = 0$, the exact algebra corresponds only to $\theta_0 = 0$. Thus the theorem establishes the precise objective solved by the implementable frozen-weight recursion. □

The preceding theorem justifies the algorithmic recursion but does not by itself guarantee that the weighted data remain informative. The following assumption and result record the information-matrix condition used later. The proof is routine RLS algebra and is placed in the appendix to keep the main text focused on the contribution-specific arguments.

Assumption 5.1 (Weighted persistent excitation)**.** *An integer $k_0 \geq 1$, a window length $T \in \mathbb{N}$, and the constants $0 < m \leq M < \infty$ exist such that, for every window starting at $k \geq k_0$, we have*

$$mI_d \leq \frac{1}{T}\sum_{\ell=k}^{k+T-1}\omega(\ell)\psi(\ell)\psi^\top(\ell) \leq MI_d. \tag{5.10}$$

Proposition 5.1 (Two checkable sufficient conditions for weighted excitation)**.** *Suppose that the raw or saturated regressor sequence satisfies ordinary finite-window excitation*

$$m_0 I_d \leq \frac{1}{T}\sum_{\ell=k}^{k+T-1}\psi(\ell)\psi^\top(\ell) \leq M_0 I_d, \qquad k \geq k_0. \tag{5.11}$$

Either of the following two conditions implies Assumption 5.1.

(i) Pathwise lower bound. If the robust weights along the estimator trajectory obey a deterministic lower bound

$$\omega(\ell) \geq \omega_{\min} > 0, \qquad \ell \geq k_0, \tag{5.12}$$

then Assumption 5.1 holds with $m = \omega_{\min}m_0$ and $M = M_0$, since $0 \leq \omega(\ell) \leq 1$ for the Huber and MCC weights used here.

(ii) Conditional expectation lower bound with bounded martingale difference deviation: Suppose instead that there is a constant $\omega_{\mathrm{E}} > 0$ such that the conditional expectation of the weight given the past and the current regressor satisfies

$$\mathbb{E}\Big[\omega(\ell) \,\Big|\, \mathcal{F}_{\ell-1}, \psi(\ell)\Big] \geq \omega_{\mathrm{E}}, \qquad \ell \geq k_0, \tag{5.13}$$

that the regressors and weights are uniformly bounded, $\|\psi(\ell)\| \leq \bar{\psi}$ and $0 \leq \omega(\ell) \leq 1$, and that the matrix sequence

$$D(\ell) = \omega(\ell)\psi(\ell)\psi^{\top}(\ell) - \mathbb{E}\Big[\omega(\ell)\psi(\ell)\psi^{\top}(\ell) \,\Big|\, \mathcal{F}_{\ell-1}\Big] \tag{5.14}$$

is a matrix martingale difference sequence with $\|D(\ell)\| \leq 2\bar{\psi}^2$ almost surely. Then, for every $\epsilon \in (0, \omega_{\mathrm{E}} m_0)$, the following two conclusions hold.

(ii.a) Fixed window, explicit probability: For every window start $k \geq k_0$, we have

$$\mathbb{P}\left((\omega_{\mathrm{E}} m_0 - \epsilon)\, I_d \leq \frac{1}{T} \sum_{\ell=k}^{k+T-1} \omega(\ell)\psi(\ell)\psi^{\top}(\ell) \leq M_0 I_d\right) \geq 1 - 2d \exp\left(-\frac{T\epsilon^2}{32\bar{\psi}^4}\right). \tag{5.15}$$

(ii.b) Slowly growing windows, eventually almost surely. If the window lengths are allowed to grow logarithmically, $T_k \geq (32\bar{\psi}^4/\epsilon^2)\,(2\ln k + \ln 2d)$ for the window starting at k, then there exists a finite random time k_1 such that

$$(\omega_{\mathrm{E}} m_0 - \epsilon)\, I_d \leq \frac{1}{T_k} \sum_{\ell=k}^{k+T_k-1} \omega(\ell)\psi(\ell)\psi^{\top}(\ell) \leq M_0 I_d, \qquad k \geq k_1, \tag{5.16}$$

so Assumption 5.1 holds eventually almost surely along such window sequences with $m = (\omega_{\mathrm{E}}\, m_0)/2$ (taking $\epsilon \leq \omega_{\mathrm{E}} m_0/2$) and $M = M_0$. For a fixed window length T, Conclusion (ii.a) is the operative statement: It is the quantitative, checkable form that the empirical weighted information diagnostics of Section 6 monitor.

Proof. See Appendix C. □

Remark 5.1 (Why two cases are needed: The MCC failure of the pathwise bound). Part (i) is the simplest sufficient condition and is realistic for Huber weights when residuals are bounded on the local trajectory, in which case $\omega_{\delta}^{H}(e) \geq \min\{1, \delta/E_{\max}\}$. It is, however, not realistic for MCC. For MCC, residual control would give $\omega_{\sigma}^{C}(e) \geq \exp[-E_{\max}^2/(2\sigma^2)]$, and this lower bound collapses to zero either when too small a value of σ is taken (a known practical risk) or when occasional large residuals occur during initial transients or impulsive events. A deterministic pathwise lower bound on $\omega(\ell)$ therefore generally fails for MCC under the very impulsive disturbances that motivate using MCC in the first place. Part (ii) is included precisely to give a checkable replacement: It only asks that the conditional mean of $\omega(\ell)$ given the past and current regressor be bounded below, which for MCC is a statement about the residual distribution rather than about pathwise residual extrema. For example, if the prediction error $\varepsilon(\ell)$ has a conditional density $f_{\varepsilon|\mathcal{F}_{\ell-1},\psi(\ell)}$ that is bounded below by $\underline{f}$ on an interval $[-A, A]$, then

$$\mathbb{E}\Big[\omega_{\sigma}^{C}(\varepsilon(\ell)) \,\Big|\, \mathcal{F}_{\ell-1}, \psi(\ell)\Big] \geq \underline{f} \int_{-A}^{A} \exp\left[-\frac{u^2}{2\sigma^2}\right] \mathrm{d}u = \underline{f}\, \sigma \sqrt{2\pi}\, [\Phi(A/\sigma) - \Phi(-A/\sigma)], \tag{5.17}$$

where the integration variable is written as u to avoid any confusion with the residual symbol e. The right-hand side stays bounded away from zero even when σ is moderately small. Part (ii) consequently captures the MCC case that Part (i) does not. The numerical section continues to report empirical weighted information eigenvalues, so that Assumption 5.1 is monitored on the actual trajectory rather than declared automatic.

Remark 5.2 (Trajectory dependence is reduced but not eliminated). Both parts of Proposition 5.1 remain trajectory-dependent because the residuals enter $\omega(\ell)$. Part (ii) reduces this dependence to a conditional moment of the residual distribution, which is checkable as long as the residual conditional law has a density bounded below on a nonvanishing window; it does not require pathwise residual control. The matrix martingale difference condition in (5.14) is automatic once ψ is bounded by saturation (4.1) or by an a priori bound on the data, and the deviation ϵ can be made arbitrarily small at the price of a longer window: Quantitatively through the explicit probability in (5.15) for a fixed window, and eventually almost surely through the slowly growing windows of Part (ii.b).

Theorem 5.2 (Covariance positivity and information-size bounds)**.** *Let $P(0) \succ 0$, $0 < \lambda \leq 1$, and $\omega(k) \geq 0$. Then $P(k) \succ 0$ for all k. Suppose Assumption 5.1 holds with the start time k_0, and for $k \geq k_0$, define*

$$B_k = \left\lfloor \frac{k - k_0 + 1}{T} \right\rfloor. \tag{5.18}$$

If $\lambda = 1$, then for $k \geq k_0$, we have

$$\|P(k)\| \leq \frac{1}{\lambda_{\min}(P(0)^{-1}) + mTB_k}. \tag{5.19}$$

If $0 < \lambda < 1$, then for $k \geq k_0$, we have

$$\|P(k)\| \leq \left[\lambda^k \lambda_{\min}(P(0)^{-1}) + mT\lambda^{T-1} \frac{1 - \lambda^{TB_k}}{1 - \lambda^T} \right]^{-1}. \tag{5.20}$$

Consequently, after the exponentially weighted excitation window has settled, $\|P(k)\| \leq C_P(1 - \lambda)$ for λ sufficiently close to one, with C_P independent of k.

Proof. See Appendix C. □

The next lemma contains only elementary score bounds. Because it is a standard calculus verification, its proof is deferred to Appendix C. These constants are nevertheless important because they are the quantities that enter the influence and tracking estimates.

Lemma 5.1 (Huber and maximum correntropy score bounds)**.** *For the Huber score in* (2.5)*,*

$$\left| s_\delta^H(e) \right| \leq \delta \qquad \forall e \in \mathbb{R}. \tag{5.21}$$

For the MCC score in (2.6)*,*

$$\left| s_\sigma^C(e) \right| \leq \sigma \exp(-1/2) \qquad \forall e \in \mathbb{R}. \tag{5.22}$$

Proof. See Appendix C. □

The first robustness statement quantifies the effect of one arbitrarily large vertical innovation while the regressor and covariance are fixed. It is intentionally local in one sample; it is not a claim of robustness against all possible leverage contamination.

Theorem 5.3 (Bounded single-sample influence)**.** *Assume* $0 < \lambda \le 1$, $\|P(k-1)\| \le \bar{P}$, *and* $\|\psi(k)\| \le \bar{\psi}$. *Let* $\Delta\hat{\theta}(k) = \hat{\theta}(k) - \hat{\theta}(k-1)$. *Then the robust update satisfies*

$$\left\|\Delta\hat{\theta}(k)\right\| \le \frac{\bar{P}\bar{\psi}}{\lambda} S, \tag{5.23}$$

where $S = \delta$ *for Huber RS-FRELS and* $S = \sigma\exp(-1/2)$ *for MCC RS-FRELS. In ordinary FRELS, for a fixed nonzero* $P(k-1)\psi(k)$*, the corresponding update grows linearly with* $|\varepsilon(k)|$.

Proof. See Appendix C; the full proof is located there, unchanged, to shorten the main text. □

Corollary 5.1 (Algorithm-enforced bounded influence with regressor saturation)**.** *Run Algorithm 1 with the saturated regressor* $\psi(k) = \mathcal{S}_B(\psi^{\mathrm{raw}}(k))$. *If* $\|P(k-1)\| \le \bar{P}$, *then for every raw regressor magnitude, including leverage values generated by impulsive past outputs,*

$$\left\|\Delta\hat{\theta}(k)\right\| \le \frac{\bar{P}B}{\lambda} S, \tag{5.24}$$

with $S = \delta$ *for Huber and* $S = \sigma\exp(-1/2)$ *for MCC. If the optional projection* Π_Θ *is used and* $\hat{\theta}(k-1) \in \Theta$*, the same bound holds for the projected increment.*

Proof. The saturation map satisfies $\|\mathcal{S}_B(x)\| \le B$ for every raw vector x. Applying Theorem 5.3 with $\bar{\psi} = B$ gives (5.24). For the projected update, Euclidean projection onto a closed convex set is nonexpansive. Since $\hat{\theta}(k-1) \in \Theta$

$$\left\|\Pi_\Theta(\hat{\theta}(k-1)+u) - \hat{\theta}(k-1)\right\| = \left\|\Pi_\Theta(\hat{\theta}(k-1)+u) - \Pi_\Theta(\hat{\theta}(k-1))\right\| \le \|u\|, \tag{5.25}$$

where u is the unprojected increment. Thus projection cannot increase the pathwise increment bound. □

Remark 5.3 (Vertical and leverage outliers). The bounded influence theorem controls a large innovation at a bounded regressor. In autoregressive or output error regressions, an impulsive output can also enter future regressors and become a leverage outlier. Such effects are outside (5.23) unless an additional bounded regressor mechanism is used. Corollary 5.1 gives one direct way to make the bounded regressor condition algorithmic: The pathwise influence bound then depends on the chosen saturation level B, not on the raw regressor norm. This does not make the model immune to all consequences of saturation; it trades leverage protection against possible saturation bias. The unsaturated estimator still requires bounded or moment-controlled regressors. The numerical section therefore reports both bounded output simulations and an unclipped leverage stress test, while the real plant guarantees require verifying the chosen preprocessing, saturation, instrumental variable, or whitening mechanism.

The next result identifies the population target for the Huber version. The statement is formulated under score orthogonality, not under a blanket independence assumption, because colored disturbances and output regressors may be statistically dependent.

Assumption 5.2 (Exogenous, instrumental, or whitened score condition)**.** *For the regression* $y(k) = \theta^{\star\top}\psi(k) + \xi(k)$*,*

$$\mathbb{E}[\psi(k)s(\xi(k))] = 0. \tag{5.26}$$

A sufficient condition is $\mathbb{E}[s(\xi(k)) \mid \psi(k)] = 0$ *with* $\mathbb{E}\,\|\psi(k)\|^2 < \infty$*. The condition is intended for exogenous regressors, instrumental regressors, whitened residual equations, or other settings in which the stated orthogonality can be verified.*

Assumption 5.3 (Huber positive weighted curvature)**.** *For some radius* $r > 0$ *and a constant* $c_H > 0$*, every unit vector* $v \in \mathbb{R}^d$ *and every* $t \in [-r, r]$ *satisfy*

$$\mathbb{E}\left[(\psi^\top v)^2 \mathbf{1}_{\{|\xi - t\psi^\top v| < \delta\}}\right] \geq c_H. \tag{5.27}$$

Theorem 5.4 (Huber–Fisher consistency and local quadratic growth)**.** *Let* $Q(\theta) = \mathbb{E}[\rho_\delta^H(y - \theta^\top\psi)]$ *and assume* $\mathbb{E}\,\|\psi\|^2 < \infty$*. Under Assumption 5.2,* $\nabla Q(\theta^\star) = 0$*; consequently, by convexity of the Huber loss,* $\theta^\star$ *is a global minimizer of* Q*. If Assumption 5.3 also holds, then* Q *has quadratic growth around* $\theta^\star$*. For every* $h \in \mathbb{R}^d$ *with* $\|h\| \leq r$*,*

$$Q(\theta^\star + h) \geq Q(\theta^\star) + \frac{c_H}{2}\|h\|^2. \tag{5.28}$$

In particular, $\theta^\star$ *is the unique minimizer of* Q *in the ball* $\{\theta : \left\|\theta - \theta^\star\right\| \leq r\}$*.*

Proof. See Appendix C; the full proof is located there, unchanged, to shorten the main text. □

Proposition 5.2 (Local target shift under a score orthogonality defect)**.** *Let*

$$b_\delta = \mathbb{E}[\psi s_\delta^H(\xi)], \tag{5.29}$$

so that $\nabla Q(\theta^\star) = -b_\delta$*. Assume* $\mathbb{E}\,\|\psi\|^2 < \infty$ *and Assumption 5.3. Let* $\bar\theta = \theta^\star + h$ *with* $0 < \|h\| \leq r$*. If* Q *is differentiable at* $\bar\theta$ *and* $\nabla Q(\bar\theta) = 0$*, then*

$$\|h\| \leq \frac{\|b_\delta\|}{c_H}. \tag{5.30}$$

More generally, if $\left\|\nabla Q(\bar\theta)\right\| \leq \eta$*, then*

$$\|h\| \leq \frac{\|b_\delta\| + \eta}{c_H}. \tag{5.31}$$

Proof. Set $t = \|h\|$ and $v = h/t$. With $q_v(\tau) = Q(\theta^\star + \tau v)$, the same one-dimensional argument used in Theorem 5.4 gives $q_v''(\tau) \geq c_H$ for almost every $\tau \in [0, t]$. Moreover

$$q_v'(0) = v^\top \nabla Q(\theta^\star) = -v^\top b_\delta. \tag{5.32}$$

Therefore

$$q_v'(t) = q_v'(0) + \int_0^t q_v''(\tau)\,\mathrm{d}\tau \geq -v^\top b_\delta + c_H t. \tag{5.33}$$

If $\bar\theta$ is stationary, then $q_v'(t) = v^\top \nabla Q(\bar\theta) = 0$; consequently $c_H t \leq v^\top b_\delta \leq \|b_\delta\|$. This proves (5.30). If only $\left\|\nabla Q(\bar\theta)\right\| \leq \eta$, then $q_v'(t) \leq \eta$, so $c_H t \leq \eta + v^\top b_\delta \leq \eta + \|b_\delta\|$, which proves (5.31). The same conclusion holds at the Huber threshold's nondifferentiability points by replacing the ordinary directional derivative with any element of the one-sided directional subdifferential. □

Remark 5.4 (Consequence of the orthogonality-defect bound). Exact Huber–Fisher consistency is recovered when $b_\delta = 0$. When colored output regressors make $b_\delta \neq 0$, Proposition 5.2 quantifies the local displacement of any nearby stationary population target. This identifies the quantity that must be controlled by whitening, instrumental variables, or dependence bounds in a fully endogenous output error analysis.

Remark 5.5 (MCC population target and locality). For the maximum correntropy loss, define the finite-memory population risk

$$Q_\sigma^C(\theta) = \mathbb{E}\left[\sigma^2\left(1 - \exp\left(-\frac{(y - \theta^\top \psi)^2}{2\sigma^2}\right)\right)\right]. \tag{5.34}$$

Under the score orthogonality condition in Assumption 5.2, with $s = s_\sigma^C$, and under the mild integrability condition $\mathbb{E}\,\|\psi\| < \infty$, the same directional differentiation argument used above gives

$$\nabla Q_\sigma^C(\theta^\star) = -\mathbb{E}\left[\psi s_\sigma^C(\xi)\right] = 0. \tag{5.35}$$

Thus the true finite-memory parameter is a stationary point of the MCC population criterion whenever the MCC score is orthogonal to the regressor. This statement should not be read as global MCC Fisher consistency. In contrast with the Huber loss, Q_σ^C is generally nonconvex and may have local minima away from $\theta^\star$; the redescending score may also suppress informative samples when σ is chosen too small. Consequently, MCC RS-FRELS is best interpreted as targeting a local population optimum of the finite-memory model, with practical performance depending on initialization, kernel width selection, and the persistence of the weighted information matrix. Operational safeguards are to start from Huber or ordinary FRELS estimates before switching to MCC, anneal σ from a broad to a narrower bandwidth, monitor the minimum eigenvalue of the weighted information matrix, and revert to Huber if the effective information collapses. This interpretation is consistent with the sensitivity results reported in Table 7. In particular, the smallest bandwidth in that table, $\sigma = 0.05$, already exhibits a visibly reduced weighted persistent excitation eigenvalue (3.386×10^{-2} versus at least 4.5×10^{-2} for every $\sigma \geq 0.10$), which is exactly the early signature of the information collapse that the warning above anticipates. The mechanism behind the local optima deserves a concrete description, because it dictates the practical rules. The MCC weight $\omega_\sigma^C(e) = \exp[-e^2/(2\sigma^2)]$ makes the effective information matrix residual-dependent: If the current estimate is far from $\theta^\star$, the residuals are large, most weights are nearly zero, the weighted information matrix loses rank in the directions that would correct the estimate, and the recursion can stall at a spurious stationary point in which only the (already fitted) small-residual samples retain influence. The basin of attraction of $\theta^\star$ shrinks with σ, which is why small kernels are simultaneously the most outlier-selective and the most locality-prone. The following rules translate this into practice. Initialization: Always warm-start MCC from an ordinary FRELS or Huber estimate obtained on an initial data segment (a few hundred samples suffice in the experiments of Section 6); a cold start with a small σ is the configuration most at risk of stalling, and the reduced weighted information eigenvalue at $\sigma = 0.05$ in Table 7 is the measurable precursor of exactly that risk. Kernel width schedule: Initialize with a broad kernel, $\sigma_0 \approx 2\hat{s}_r$ or larger, where $\hat{s}_r$ is the robust residual scale, so that the initial basin is wide; next, anneal σ downward geometrically or linearly toward the target bandwidth while the weighted information eigenvalue stays above a preset floor, and freeze the annealing the moment the diagnostic degrades. Fallback: If the minimum weighted information eigenvalue drops below the floor (in our experiments, half of its Huber value is a workable floor),

revert to the Huber score, which restores a convex landscape at the cost of weaker outlier suppression. The annealed-σ variant evaluated in Table 2 implements exactly this schedule and confirms that the safeguards are not merely defensive: It attains the best clean mean square error (MSE) in the comparison while keeping the information diagnostic healthy.

Remark 5.6 (Failure of the curvature condition). Assumption 5.3 is a sufficient local identifiability condition. If it fails, the score orthogonality still makes $\theta^\star$ a Huber population minimizer, but flat directions or nonunique minimizers may remain. This can occur, for example, when the Huber threshold is too small relative to quantized or degenerate noise, or when the regressor lacks excitation in a direction that is visible inside the quadratic region of the Huber loss.

Remark 5.7 (Colored disturbances and output regressors). If $\psi(k)$ contains past measured outputs while $\xi(k)$ is colored, then $\xi(k)$ and $\psi(k)$ are generally dependent. Assumption 5.2 may then fail unless the estimating equation is whitened, an instrumental variable construction is used, or an explicit dependence control argument is supplied. When the defect is not zero, Proposition 5.2 bounds the local Huber population target displacement by $\left\|\mathbb{E}[\psi s_\delta^H(\xi)]\right\| / c_H$. The theory is therefore presented as conditional; the simulations demonstrate the practical robust mechanism but do not prove a small orthogonality defect for every colored output error realization.

The convergence result below is written as a stochastic approximation theorem for $\lambda = 1$. The statement includes the bounded regressor condition needed to justify the denominator expansion in the exact update; alternatively, one may replace this by a stronger high-moment and summability condition that gives the same perturbation control.

Assumption 5.4 (Stochastic approximation regularity)**.** *For $\lambda = 1$, suppose that the following conditions hold for one subsystem.*

(SA1) $\hat{\theta}(k)$ *remains in a compact convex set* Θ*, either naturally or by projection.*

(SA2) $k^{-1}R(k) \to \bar{R} \succ 0$ *almost surely; equivalently along the considered trajectory,* $kP(k) \to \bar{R}^{-1}$ *almost surely.*

(SA3) $\sup_k \|\psi(k)\| < \infty$ *almost surely. An equivalent high-moment/summability condition may be used if it implies* $\psi^\top(k)P(k-1)\psi(k) = O(k^{-1})$ *and the gain perturbation bound in (SA6).*

(SA4) The mean field

$$h(\theta) = \mathbb{E}[\psi s(y - \theta^\top \psi)] \tag{5.36}$$

is locally Lipschitz on Θ*, has a unique zero* $\theta^\star \in \Theta$*, and satisfies the local monotonicity inequality*

$$(\theta - \theta^\star)^\top \bar{R}^{-1} h(\theta) \le -a \left\|\theta - \theta^\star\right\|^2 \tag{5.37}$$

for some $a > 0$ *in a neighborhood of* $\theta^\star$*.*

(SA5) With

$$M_k(\theta) = \psi(k)s(y(k) - \theta^\top \psi(k)) - h(\theta), \tag{5.38}$$

one has $\mathbb{E}[M_k(\theta) \mid \mathcal{F}_{k-1}] = 0$ *and* $\mathbb{E}[\|M_k(\theta)\|^2 \mid \mathcal{F}_{k-1}] \le C_M$ *uniformly for* θ *in the local region.*

(SA6) The gain and denominator perturbation produced by replacing the exact recursion with its limiting stochastic approximation form is locally negligible: the remainder ζ_k in (5.43) *satisfies* (5.44) *for every sufficiently small $\epsilon > 0$ after a finite random time. A sufficient, but not necessary, primitive condition is a summable rate for $kP(k-1) - \bar{R}^{-1}$, together with bounded regressors and bounded scores.*

Remark 5.8 (Primitive routes to Assumption 5.4). Assumption 5.4 is a local stochastic approximation condition, not a global theorem about every colored output error Hammerstein plant. A common route to these hypotheses is to use projection or estimator-side saturation to keep the trajectory and regressors bounded; assume that the finite-memory data process is stationary, ergodic, and geometrically mixing; verify the ordinary persistent excitation (PE) of the regressor and a positive lower bound on robust weights as in Proposition 5.1; and impose either an exogenous, instrumental, or whitened score equation so that the stochastic score has a martingale difference or a sufficiently mixing dependence decomposition. These conditions are stronger than needed but we should make it clear which plant-level ingredients must be checked. Without such ingredients, Theorem 5.5 should be read as a conditional local stability result for the recursion rather than as unconditional consistency for colored output regressor systems.

Remark 5.9 (Engineering feasibility of the assumptions and compensation strategies). The assumptions used in this section are idealizations, and it is legitimate to ask how a practitioner can approach them in an engineering setting. The following mapping lists, for each assumption family, its typical status in practice and the preprocessing or compensation strategy that brings a real plant closer to it. Bounded iterates and regressors (SA1, SA3) are rarely guaranteed a priori but are directly enforceable by the algorithm itself through projection onto physically motivated parameter bounds (e.g., known gain ranges and stability margins) and through the estimator-side saturation (4.1); no plant knowledge beyond rough magnitude bounds is required. Excitation (ordinary PE in Proposition 5.1) is a property of the experiment rather than of the algorithm; it is promoted by the persistently exciting input design (multisine or pseudorandom-binary-sequence-type inputs covering the bandwidth of interest) and is directly measurable from data, as described in Remark 5.10. Weight lower bounds (Proposition 5.1): Favored by moderate robust tuning (δ not too small, σ not too small) and monitored through the weighted information diagnostic; if the diagnostic degrades, enlarging δ or σ is the immediate compensation. Score orthogonality (Assumption 5.2): The most demanding item, because colored disturbances correlate with the output regressors. Practical mitigation follows classical system identification: Detrend and deseasonalize the signals, estimate the noise coloring and prewhiten the residual equation, or replace the output regressors by instrumental variables built from past inputs. Proposition 5.2 quantifies the residual bias when the compensation is imperfect. Stationarity and mixing: Approximately valid for plants operated around a fixed operating point; when the operating point drifts, the tracking lemma with $\lambda < 1$ is the relevant statement instead of the convergence theorem. None of these strategies converts a conditional theorem into an unconditional one, but, together, they give a concrete engineering route to a regime in which the stated assumptions are plausible and monitored rather than merely postulated.

Remark 5.10 (A practitioner's checklist for verifying the assumptions on a given plant). The key assumptions can be checked, or at least falsified, from the same data used for identification. The following workflow makes the verification concrete. Steps (V1), (V2), (V5), and (V6) are computed by the released code and reported for the simulated plant (Tables 4 and 6); Steps (V3) and (V4) are

standard offline inspections of logged residuals and weights that the user performs with any statistics package.

(V1) Excitation. Compute the windowed Gram matrix $T^{-1}\sum_{\ell=k}^{k+T-1}\psi(\ell)\psi^{\top}(\ell)$ over sliding windows (a window of $T \approx 10\,d$ samples is a reasonable default) and track its minimum eigenvalue. Values persistently close to zero falsify ordinary PE, and no robust reweighting can repair this; redesign the input first.

(V2) Weighted excitation. Repeat the computation with the robust weights included, i.e., the empirical version of (5.10). A large gap between the ordinary and the weighted minimum eigenvalue indicates that the score is suppressing informative samples; enlarge δ or σ until the gap closes to an acceptable factor (a factor of two is a practical alarm threshold, cf. Table 4).

(V3) Weight health. Plot the histogram of the realized weights $\omega(\ell)$ and the effective sample fraction $\sum_{\ell}\omega(\ell)/K$. For Huber with a well-chosen threshold, the bulk of weights should equal one with a contaminated tail below one; a weight mass accumulating near zero reproduces the MCC information collapse signature of Figure 7.

(V4) Residual scale and density. Estimate the robust residual scale $\hat{s}_r$ (median absolute deviation of recent residuals) and inspect a kernel density estimate of the residuals on the window $[-A, A]$ relevant to Remark 5.12. The estimated density being visibly bounded away from zero on the window supports the curvature conditions; a density with holes or hard truncation falsifies them; cf. Remark 5.13.

(V5) Score orthogonality. Compute the empirical score defect $\|K_e^{-1}\sum_k \psi(k)s(\varepsilon(k))\|$ over the evaluation window, as in Table 4. This is a finite-sample diagnostic, not a proof; a defect that does not shrink as the window grows indicates a colored noise orthogonality violation, and the instrumental variable or prewhitening compensations of Remark 5.9 are then required, with Proposition 5.2 bounding the induced bias.

(V6) Leverage and saturation. Record the raw regressor norm distribution and the saturation activation rate as in Table 6. A nonzero activation rate documents that leverage protection is actually being exercised; a high rate warns that the bound B is biasing the estimate.

Steps (V1)–(V3) are cheap online monitors; Steps (V4)–(V6) are offline checks on logged data. Together, they operationalize the assumption set: Each theoretical condition is paired with an observable quantity and a corrective action.

Theorem 5.5 (Conditional local convergence of projected RS-FRELS)**.** *Under Assumption 5.4, the projected RS-FRELS recursion with $\lambda = 1$ converges almost surely to $\theta^{\star}$ whenever it enters and subsequently remains in the neighborhood where* (5.37) *holds. If no projection is used, the same conclusion holds provided the unprojected trajectory is almost surely bounded and eventually remains in that neighborhood.*

Proof. Let $\mathcal{F}_k$ be the data filtration and write the exact recursion in score form as follows:

$$\hat{\theta}(k) = \hat{\theta}(k-1) + \frac{P(k-1)\psi(k)s(\varepsilon(k))}{1+\omega(\varepsilon(k))\psi^{\top}(k)P(k-1)\psi(k)}. \tag{5.39}$$

By Assumption 5.4, $P(k-1) = R(k-1)^{-1}$ satisfies

$$P(k-1) = \frac{1}{k}\bar{R}^{-1} + \frac{1}{k}A_k, \qquad \|A_k\| \to 0 \tag{5.40}$$

almost surely along the considered trajectory. Because $\psi(k)$ is almost surely bounded and the Huber/MCC weights satisfy $0 \le \omega \le 1$, we have

$$0 \le \omega(\varepsilon(k))\psi^\top(k)P(k-1)\psi(k) \le \frac{C}{k} \tag{5.41}$$

for all sufficiently large values of k. Hence

$$\frac{1}{1+\omega(\varepsilon(k))\psi^\top(k)P(k-1)\psi(k)} = 1+\rho_k, \qquad |\rho_k| \le \frac{C}{k}. \tag{5.42}$$

Substituting (5.40) and (5.42) into (5.39) gives

$$\hat{\theta}(k) = \hat{\theta}(k-1) + \frac{1}{k}\bar{R}^{-1}\psi(k)s\big(y(k) - \hat{\theta}^\top(k-1)\psi(k)\big) + \zeta_k, \tag{5.43}$$

where ζ_k collects the terms containing A_k, the denominator factor, and the difference between the exact gain and the limiting gain. By Assumption 5.4, for every sufficiently small $\epsilon > 0$, after a finite random time, we have

$$\|\zeta_k\| \le \frac{\epsilon}{k}\left\|\hat{\theta}(k-1) - \theta^\star\right\| + \frac{C_\epsilon}{k^2}. \tag{5.44}$$

This explicit perturbation condition is included to avoid hiding a rate requirement inside the denominator expansion.

Decompose the stochastic score as

$$\psi(k)s\big(y(k) - \hat{\theta}^\top(k-1)\psi(k)\big) = h(\hat{\theta}(k-1)) + M_k(\hat{\theta}(k-1)), \tag{5.45}$$

where $\mathbb{E}[M_k(\hat{\theta}(k-1)) \mid \mathcal{F}_{k-1}] = 0$ and $\mathbb{E}[\left\|M_k(\hat{\theta}(k-1))\right\|^2 \mid \mathcal{F}_{k-1}] \le C_M$ by Assumption 5.4. Let $V_k = \left\|\hat{\theta}(k) - \theta^\star\right\|^2$. Conditional on $\mathcal{F}_{k-1}$, the first-order drift term in V_k is

$$\frac{2}{k}(\hat{\theta}(k-1) - \theta^\star)^\top \bar{R}^{-1} h(\hat{\theta}(k-1)), \tag{5.46}$$

which is at most $-2aV_{k-1}/k$ in the local region. The martingale term has zero conditional mean, and its second-order contribution is $O(k^{-2})$. The perturbation bound (5.44) can be absorbed into the negative drift by choosing $\epsilon < a$. Therefore, for all sufficiently large values of k,

$$\mathbb{E}[V_k \mid \mathcal{F}_{k-1}] \le \left(1 - \frac{a}{k}\right)V_{k-1} + \frac{C}{k^2}. \tag{5.47}$$

The Robbins–Siegmund supermartingale theorem implies that V_k converges almost surely and that $\sum_k k^{-1}V_{k-1} < \infty$. Since $\sum_k k^{-1} = \infty$, the only possible non-negative limit compatible with this summability is zero. Hence $\hat{\theta}(k) \to \theta^\star$ almost surely.

If projection onto Θ is used, the Euclidean projection is nonexpansive and $\theta^\star \in \Theta$, so projection cannot increase $\left\|\hat{\theta}(k) - \theta^\star\right\|$. The same Lyapunov recursion therefore holds for the projected iteration. Without projection, the boundedness and eventual locality assumptions ensure that the same local argument applies after the finite entrance time. □

The locality in Theorem 5.5 is necessary in full generality because the mean field contraction (5.37) is only assumed in a neighborhood of $\theta^\star$. For the Huber score this restriction can be removed, because the Huber population risk is convex and admits a uniform curvature lower bound on any compact set. The next theorem records that strengthening.

Assumption 5.5 (Global Huber score orthogonality and uniform curvature on Θ)**.** *Let $\Theta \subset \mathbb{R}^d$ be a compact convex set with $\theta^\star$ in its interior, and let the iterates of Algorithm 1 be projected onto Θ. The Huber score $s = s_\delta^H$ is used. A filtration $\{\mathcal{F}_k\}$ exists such that $\psi(k)$ is $\mathcal{F}_{k-1}$-measurable, $y(k)$ is $\mathcal{F}_k$-measurable,* $\sup_k \|\psi(k)\| < \infty$ *almost surely, and*

(GH1) Global score orthogonality. For every $k \geq 1$

$$\mathbb{E}\Big[\psi(k) s_\delta^H(\xi(k)) \,\Big|\, \mathcal{F}_{k-1}\Big] = 0. \tag{5.48}$$

(GH2) Uniform Huber curvature on Θ. Here, $c_H > 0$ exists such that for every $\theta \in \Theta$, every unit vector $v \in \mathbb{R}^d$, and every $k \geq 1$,

$$\mathbb{E}\Big[(\psi(k)^\top v)^2 \, \mathbf{1}_{\{|y(k)-\theta^\top\psi(k)|<\delta\}} \,\Big|\, \mathcal{F}_{k-1}\Big] \geq c_H. \tag{5.49}$$

(GH3) Weighted information limit. The empirical information matrix satisfies $k^{-1}R(k) \to \bar{R} \succ 0$ almost surely (equivalently, $kP(k) \to \bar{R}^{-1}$ almost surely).

(GH4) Martingale and perturbation regularity. With $M_k(\theta) = \psi(k)s_\delta^H(y(k) - \theta^\top\psi(k)) - h(\theta)$ and $h(\theta) = \mathbb{E}[\psi s_\delta^H(y - \theta^\top\psi)]$, the conditional second moment of $M_k(\theta)$ is uniformly bounded by some C_M for $\theta \in \Theta$, and the gain perturbation remainder ζ_k in the stochastic approximation (SA) expansion satisfies (5.44) *after a finite random time.*

Theorem 5.6 (Global convergence of the projected Huber RS-FRELS on Θ)**.** *Let Algorithm 1 be run with the Huber score, $\lambda = 1$, and Euclidean projection onto Θ. Under Assumption 5.5, the projected iterate satisfies*

$$\hat{\theta}(k) \longrightarrow \theta^\star \qquad \text{almost surely as } k \to \infty, \tag{5.50}$$

for any initial estimate $\hat{\theta}(0) \in \Theta$. The convergence is global in the sense that no a priori localization to a neighborhood of $\theta^\star$ is required.

Proof. See Appendix C; the full proof is located there, unchanged, to shorten the main text. □

Remark 5.11 (Why the same globalization is not available for MCC RS-FRELS). Theorem 5.6 relies on two structural features that are specific to Huber RS-FRELS. The convexity of the population risk and a lower bound on Θ with uniform curvature both fail for the MCC context. First, the MCC population risk Q_σ^C is generically nonconvex: The score $s_\sigma^C(e) = e\exp(-e^2/(2\sigma^2))$ is increasing only on $|e| \leq \sigma$ and redescending for $|e| > \sigma$, so $\nabla^2 Q_\sigma^C$ takes both signs depending on the conditional residual distribution. Second, even where Q_σ^C is locally convex, the curvature analog

$$\mathbb{E}\Big[(\psi^\top v)^2(1 - (e_\theta/\sigma)^2)\exp(-e_\theta^2/(2\sigma^2)) \,\Big|\, \mathcal{F}_{k-1}\Big]$$

cannot be bounded below uniformly in $\theta \in \Theta$: Shifting θ moves $e_\theta = y - \theta^\top\psi$ into the redescending region, where the integrand becomes negative for $|e_\theta| > \sigma$. Consequently, the uniform-on-Θ analog

of (5.49) need not hold, and spurious stationary points of Q_σ^C can exist arbitrarily far from $\theta^\star$. MCC RS-FRELS therefore inherits the locality of Theorem 5.5 and is best operated with the safeguards listed in the MCC remark above (Huber initialization, bandwidth annealing, weighted information monitoring, and fallback to Huber when information collapses).

Remark 5.12 (Primitive sufficient conditions for uniform Huber curvature). Assumption 5.5 (GH2) is implied by the following pair of conditions, which are checkable from the data-generating process:

- Regressor nondegeneracy. Here, $\mu_\psi > 0$ exists such that $\mathbb{E}[\psi(k)\psi(k)^\top \mid \mathcal{F}_{k-1}] \geq \mu_\psi I_d$ uniformly in k.
- Noise density is bounded below on a neighborhood of zero with a width matching the Huber threshold and the projection diameter. Here, $A \geq \delta + \bar{\psi}\,\mathrm{diam}(\Theta)$ and $\underline{f} > 0$ exist such that the conditional density of $\xi(k)$ given $\mathcal{F}_{k-1}$ and $\psi(k)$ is at least $\underline{f}$ on $[-A, A]$.

Under these two conditions, for any $\theta \in \Theta$ and any unit vector v, we have

$$\mathbb{E}[(\psi^\top v)^2 \mathbf{1}_{\{|\xi-(\theta-\theta^\star)^\top\psi|<\delta\}} \mid \mathcal{F}_{k-1}] \geq 2\delta \underline{f}\,\mu_\psi,$$

because the indicator integrates against the noise density over a subinterval of $[-A, A]$ of length 2δ. Thus $c_H = 2\delta \underline{f}\,\mu_\psi$ is a valid uniform curvature constant. The role of the projection set Θ is therefore not only to keep iterates bounded but also to size the noise density neighborhood needed for global Huber curvature: A wider Θ requires the noise density to be bounded below on a wider window. For contaminated Gaussian noise of the form (3.4), the unimodal nominal component already satisfies the density lower bound on any compact interval, so this condition is realistic in the simulation setting of Section 6.

Remark 5.13 (How realistic is the uniform-curvature condition (GH2)?). It is a fair question whether (GH2) is ever verifiable for a real plant, or whether it is a theoretical construct that merely permits a clean proof. The answer is honest but not dismissive, and it has three parts. (a) When it holds: The reduction in Remark 5.12 needs the conditional noise density to be bounded below on a window of width $A \geq \delta + \bar{\psi}\,\mathrm{diam}(\Theta)$. For any noise model whose density is continuous and strictly positive on compact sets—Gaussian, Student's t, Laplace, and every contaminated mixture with such a nominal component—the bound $\underline{f} > 0$ exists on every compact window, so (GH2) is satisfied in a mathematically rigorous sense for this large and practically common class. In this precise sense, the condition is not vacuous. (b) Where the honesty lies: The constant degrades with the window width. For a Gaussian nominal component, $\underline{f}$ decays like the density tail at A, so a large projection set drives $c_H = 2\delta \underline{f} \mu_\psi$ toward zero and the convergence guaranteed by Theorem 5.6 becomes arbitrarily slow even though it remains global. Moreover, for noise densities that vanish on part of the window—quantized measurements, bounded support noise, or multimodal densities with gaps—(GH2) genuinely fails for a large Θ, and no proof-side repair is claimed here. In those cases, Theorem 5.6 should be read as global only on the largest Θ whose induced window keeps the density positive, which may reduce it to a semilocal statement. (c) What the practitioner should do: Two consequences follow. First, Θ should be chosen as small as prior engineering knowledge allows because $\mathrm{diam}(\Theta)$ enters the required window linearly; crude parameter bounds from physical gain and stability considerations are usually enough to make the window moderate. Second, the density lower bound is checkable from data: A kernel density estimate of the identification residuals on $[-A, A]$, as in Step (V4) of Remark 5.10,

either visibly supports a positive floor or exposes the gaps that falsify the condition. The practical value of Theorem 5.6 is therefore best summarized as a structural robustness statement—the convex Huber landscape admits globalization while the redescending MCC landscape does not—together with an explicitly computable certificate whose conservatism the user can measure, rather than as a claim that global convergence is free in practice.

The last theoretical statement concerns slowly varying parameters. Since the key contraction property is assumed rather than derived from weighted persistent excitation alone, the result is stated as a conditional tracking lemma.

Assumption 5.6 (Slow variation and local mean-square contraction)**.** *Let $\theta^\star(k)$ vary with*

$$\left\|\theta^\star(k) - \theta^\star(k-1)\right\| \le q. \tag{5.51}$$

Assume bounded regressors, bounded robust score $|s(e)| \le S$, and weighted persistent excitation so that $\|P(k)\| \le c_P(1-\lambda)$ for all sufficiently large values of k. In a local tracking region, suppose the no-drift part of the update is mean-square contractive; thus, $a, b > 0$ exist such that

$$\mathbb{E}[\left\|\tilde{\theta}^+(k)\right\|^2 \mid \mathcal{F}_{k-1}] \le (1 - a(1-\lambda))\left\|\tilde{\theta}(k-1)\right\|^2 + b(1-\lambda)^2 S^2, \tag{5.52}$$

where $\tilde{\theta}(k) = \hat{\theta}(k) - \theta^\star(k)$.

Lemma 5.2 (Conditional mean-square tracking estimate)**.** *Under Assumption 5.6, for any λ that is sufficiently close to one,*

$$\limsup_{k\to\infty} \mathbb{E}\left\|\hat{\theta}(k) - \theta^\star(k)\right\|^2 \le C_1(1-\lambda)S^2 + C_2\frac{q^2}{(1-\lambda)^2}, \tag{5.53}$$

where C_1, C_2 are finite constants independent of k, q, and S.

Proof. See Appendix C; the full proof is located there, unchanged, to shorten the main text. □

The term $C_1(1-\lambda)S^2$ is a variance-like contribution controlled by the bounded score, whereas $C_2q^2/(1-\lambda)^2$ is a drift contribution. The $q^2/(1-\lambda)^2$ scaling is conservative: It arises from Young's inequality applied to the squared norm in (C.32). A direct cross-term bound based on the Cauchy–Schwarz inequality can recover the classical $q/(1-\lambda)$-type scaling under additional drift regularity, but a sharper analysis is outside the scope of this conditional lemma. Thus a forgetting factor close to one improves noise averaging but worsens the tracking of time-varying parameters, exactly as in classical RLS theory.

6. Numerical experiments and benchmark validation

6.1. Simulation model and algorithms

The simulation uses a three-subsystem fractional interconnected Hammerstein process. Each subsystem uses a finite GL memory $L = 45$ and polynomial nonlinear input terms up to degree three. The disturbance contains a colored finite-memory component and impulsive contamination. The primary algorithm comparison keeps the same full fractional interconnected regressor dimension for

ordinary FRELS, Huber RS-FRELS, MCC RS-FRELS, a same-regressor Huber-NLMS baseline, and an optional regressor-saturated Huber RS-FRELS variant, which is labeled "+ sat" (short for saturation) in the tables and figures that follow. Three further same-regressor robust baselines are included to broaden the comparison with the state of the art in robust recursive estimation: A recursive least M-estimate variant with the redescending three-part Hampel score used in the RLM family of robust adaptive filters [16], with the knots $(a, b, c) = (\delta, 2\delta, 4\delta)$; a Huber variant whose threshold is retuned online by the rolling median absolute deviation (MAD) rule of Section 6.4, $\delta(k) = 1.5 \times 1.4826\,\mathrm{MAD}$ over the last 150 residuals; and an MCC variant that implements the broad-to-narrow bandwidth annealing safeguard, with $\sigma(k)$ decreasing linearly from 0.50 to 0.10 over the first 300 samples. All eight algorithms consume the identical regressor sequence, so the comparison isolates the scoring and gain mechanisms. The baseline pool is chosen for same-regressor comparability: The RLM family remains the reference redescending recursive M-estimator, while more recent proportionate and sparsity-penalized correntropy recursions [20, 21] encode sparse-channel priors that do not apply to the dense fractional Hammerstein regressor; the two adaptive variants represent the current variable-parameter practice (adaptive threshold selection and kernel annealing) in a form that is directly compatible with the protocol. The synthetic study is complemented by a real-data benchmark validation in Section 6.5, which addresses the synthetic-data-only limitation of the previous version. Model structure ablations that remove fractional memory or interconnections are reported separately, because they use different parameter spaces, and their NMSD values are not directly comparable with the full model. Unless otherwise stated, the generated outputs are clipped only in the data generator to avoid unbounded autoregressive leverage; the estimator itself sees the measured outputs and, when enabled, applies only the explicit regressor saturation map in (4.1). An additional unclipped leverage stress scenario is included below to expose the difference between output-bounded data generation and estimator-side leverage control. Table 1 collects the numerical values of all quantities that define the representative simulation, together with the settings of the stress, impulse grid, and evaluation protocols used throughout this section.

Table 1. Parameters used in the representative simulation.

Quantity	Value
Number of subsystems	$N_s = 3$
Samples per run	$K = 650$
GL memory	$L = 45$
Forgetting factor	$\lambda = 0.995$ unless varied
Huber threshold	$\delta = 0.35$ unless varied
MCC kernel width	$\sigma = 0.50$ unless varied
Nominal colored noise scale	0.045
Impulse probability, standard case	$p = 0.04$
Impulse amplitude scale, standard case	1.10
Estimator-side saturation bound	$B = 50$ when enabled
Monte Carlo seeds	30 independent runs
Stress test	Student's $t_{2.5}$ impulses, $p = 0.08$, scale = 2.0, five seeds per scenario
Impulse grid	$p \in \{0, 0.04, 0.08\}$ and scale $\in \{0.35, 1.10, 2.00\}$, three seeds per cell
Evaluation metric	Clean MSE, measured MSE, final NMSD averaged over $k = \lfloor 0.65K \rfloor, \ldots, K$

The NMSD, i.e., the normalized misalignment for subsystem i, is

$$\mathrm{NMSD}_i(k) = \frac{\left\|\hat{\theta}_i(k) - \theta_i^\star\right\|^2}{\left\|\theta_i^\star\right\|^2}, \qquad \mathrm{NMSD}(k) = \frac{1}{N_s}\sum_{i=1}^{N_s} \mathrm{NMSD}_i(k). \tag{6.1}$$

The clean prediction MSE evaluates the identified noiseless regression output $z_i(k) = \theta_i^{\star\top}\psi_i(k)$, while the measured output MSE includes unpredictable impulses and is therefore less directly connected to parameter accuracy.

6.2. Monte Carlo comparison

Table 2 reports mean and standard deviation over independent Monte Carlo runs. The robust RLS-type methods reduce clean model error and final misalignment relative to ordinary FRELS while using the same fractional interconnected regressor. The Huber-NLMS baseline uses the same robust score and regressor but lacks the recursive covariance normalization, which separates the gain-scheduling contribution from the robust score contribution. The saturated Huber variant is numerically identical to the unsaturated Huber run in the standard case because most raw regressors remain below $B = 50$; its role is to make the bounded influence condition enforceable in leverage-stress settings. The MCC method is typically the most aggressive outlier suppressor, whereas the Huber method is less sensitive to kernel width collapse. The three additional robust baselines refine this picture. The redescending RLM–Hampel score (1.265×10^{-4} clean MSE) sits between the Huber (1.703×10^{-4}) and MCC (1.112×10^{-4}) methods, which is coherent with its design: It clips moderate outliers like those of the Huber

method but fully rejects gross ones like MCC, without MCC's nonconvex risk. More notably, the two adaptively tuned variants outperform their fixed-constant counterparts: The rolling-MAD Huber threshold reaches 8.494×10^{-5} and the annealed bandwidth MCC reaches 7.195×10^{-5}, the best value in the comparison. This is not an accident of tuning generosity—both variants use only nonoracle statistics of the residual stream—but is a consequence of the fixed constants ($\delta = 0.35$, $\sigma = 0.50$) being deliberately conservative defaults, while the adaptive rules converge toward the sharper values that the sensitivity study of Table 7 identifies a posteriori. The annealed MCC also keeps its weighted information eigenvalue at 4.574×10^{-2}, comfortably above the collapse regime, confirming that the safeguard schedule of Section 5 is compatible with aggressive final bandwidths.

Table 2. Monte Carlo performance summary. Mean ± standard deviation over 30 independent seeds.

Method	Clean MSE	Measured MSE	Final NMSD	PE min eigenvalue
FRELS	$9.962e-04 \pm 4.5e-04$	$4.709e-02 \pm 1.8e-02$	$6.372e-03 \pm 2.6e-03$	$5.404e-02 \pm 3.4e-03$
Huber RS-FRELS	$1.703e-04 \pm 4.7e-05$	$4.862e-02 \pm 1.9e-02$	$1.135e-03 \pm 3.8e-04$	$5.314e-02 \pm 3.4e-03$
MCC RS-FRELS	$1.112e-04 \pm 2.8e-05$	$4.918e-02 \pm 1.9e-02$	$7.263e-04 \pm 2.2e-04$	$5.251e-02 \pm 3.3e-03$
Huber NLMS	$1.187e-03 \pm 2.4e-04$	$4.799e-02 \pm 1.9e-02$	$1.266e-02 \pm 3.4e-03$	$5.314e-02 \pm 3.4e-03$
Huber RS-FRELS + sat	$1.703e-04 \pm 4.7e-05$	$4.862e-02 \pm 1.9e-02$	$1.135e-03 \pm 3.8e-04$	$5.314e-02 \pm 3.4e-03$
RLM–Hampel RS-FRELS	$1.265e-04 \pm 3.4e-05$	$4.913e-02 \pm 1.9e-02$	$8.283e-04 \pm 2.5e-04$	$5.296e-02 \pm 3.4e-03$
Huber RS-FRELS (adaptive δ)	$8.494e-05 \pm 2.1e-05$	$4.918e-02 \pm 1.9e-02$	$5.484e-04 \pm 1.7e-04$	$5.116e-02 \pm 3.3e-03$
MCC RS-FRELS (annealed σ)	$7.195e-05 \pm 1.9e-05$	$4.942e-02 \pm 1.9e-02$	$4.644e-04 \pm 1.5e-04$	$4.574e-02 \pm 2.9e-03$

Table 3 reports paired tests on clean MSE over the same 30 seeds. These tests are not a substitute for real plant validation, but they address the finite-sample question of whether the reported Monte Carlo differences are driven by a few favorable seeds. The Huber-NLMS row also explains why the same robust score alone is insufficient: Without the covariance-normalized RLS gain, NLMS converges more slowly and can perform worse than ordinary FRELS in clean prediction error. The three new rows show that the improvements of the added baselines are themselves statistically resolved on the paired seeds: RLM–Hampel improves on fixed-threshold Huber ($t = -9.24$), the rolling-MAD threshold improves on the fixed threshold ($t = -13.18$), and the annealed bandwidth improves on the fixed bandwidth ($t = -9.48$), all with p-values below 10^{-9}.

Table 3. Paired clean MSE comparisons over the same 30 Monte Carlo seeds. Positive differences mean that the first method has a larger error.

Comparison	Mean paired clean MSE difference	t	p
FRELS – Huber RS-FRELS	$8.259e-04$	10.66	$1.51e-11$
FRELS – MCC RS-FRELS	$8.850e-04$	10.83	$1.06e-11$
Huber RS-FRELS – MCC RS-FRELS	$5.909e-05$	11.57	$2.20e-12$
Huber NLMS – Huber RS-FRELS	$1.017e-03$	24.15	$9.39e-21$
RLM–Hampel RS-FRELS – Huber RS-FRELS	$-4.380e-05$	−9.24	$3.81e-10$
Huber RS-FRELS (adaptive δ) – Huber RS-FRELS	$-8.534e-05$	−13.18	$8.90e-14$
MCC RS-FRELS (annealed σ) – MCC RS-FRELS	$-3.924e-05$	−9.48	$2.19e-10$

Table 4 reports two assumption-oriented diagnostics. The empirical score defect is the norm of the evaluation window average $K_e^{-1} \sum_k \psi(k)s(\varepsilon(k))$, averaged over subsystems and seeds. It is only a finite-sample diagnostic based on the fitted residuals, not a proof that $\mathbb{E}[\psi s(\xi)] = 0$ in the population.

The weighted information eigenvalue is likewise a trajectory diagnostic for Assumption 5.1.

Figures 1–3 display the same comparison along a single representative run, so that the aggregate numbers of Tables 2–4 can be read against the underlying trajectories. Figure 1 shows the normalized parameter misalignment NMSD(k), Figure 2 shows the moving clean prediction error, and Figure 3 shows the robust weights that the two scores assign to the realized innovations.

Table 4. Empirical diagnostics for score orthogonality, weighted excitation, and raw regressor size.

Method	Empirical score defect	PE min eigenvalue	Max raw $\|\|\psi\|\|$
FRELS	$1.720e-02 \pm 4.7e-03$	$5.404e-02 \pm 3.4e-03$	$4.921e+00 \pm 6.8e-01$
Huber RS-FRELS	$7.027e-03 \pm 1.8e-03$	$5.314e-02 \pm 3.4e-03$	$4.921e+00 \pm 6.8e-01$
MCC RS-FRELS	$5.661e-03 \pm 1.4e-03$	$5.251e-02 \pm 3.3e-03$	$4.921e+00 \pm 6.8e-01$
Huber NLMS	$1.307e-02 \pm 3.2e-03$	$5.314e-02 \pm 3.4e-03$	$4.921e+00 \pm 6.8e-01$
Huber RS-FRELS + sat	$7.027e-03 \pm 1.8e-03$	$5.314e-02 \pm 3.4e-03$	$4.921e+00 \pm 6.8e-01$

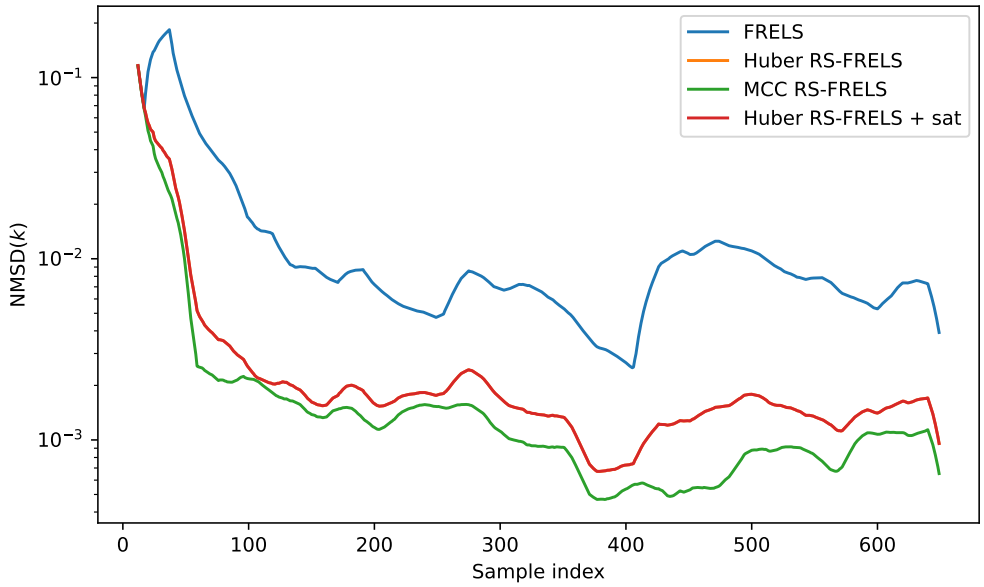

Figure 1. Representative normalized parameter misalignment NMSD(k) as defined in (6.1). Robust methods suppress the large excursions induced by impulsive innovations.

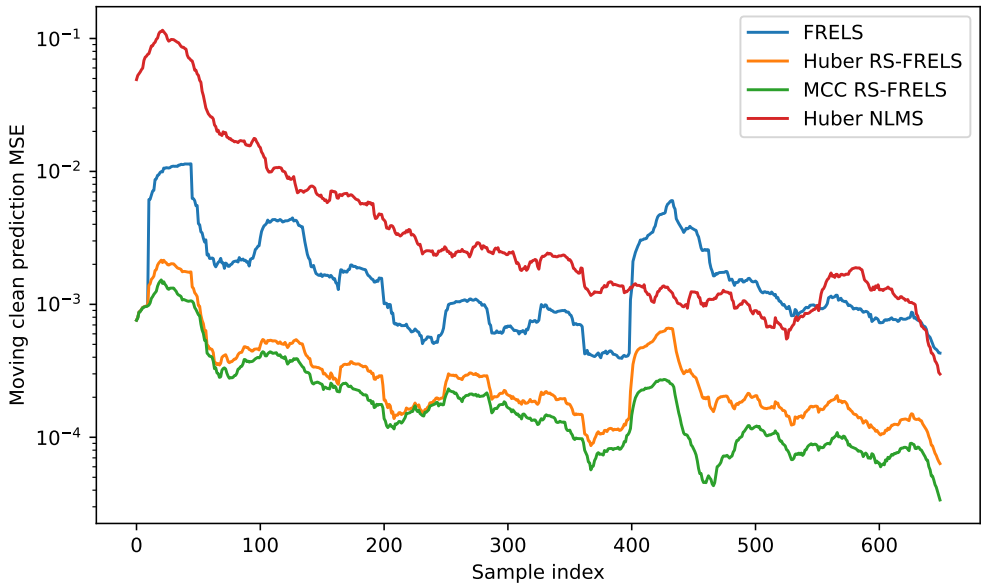

Figure 2. Representative moving clean prediction error. The clean prediction error is the relevant metric for validating parameter recovery because measured outputs include unpredictable impulses.

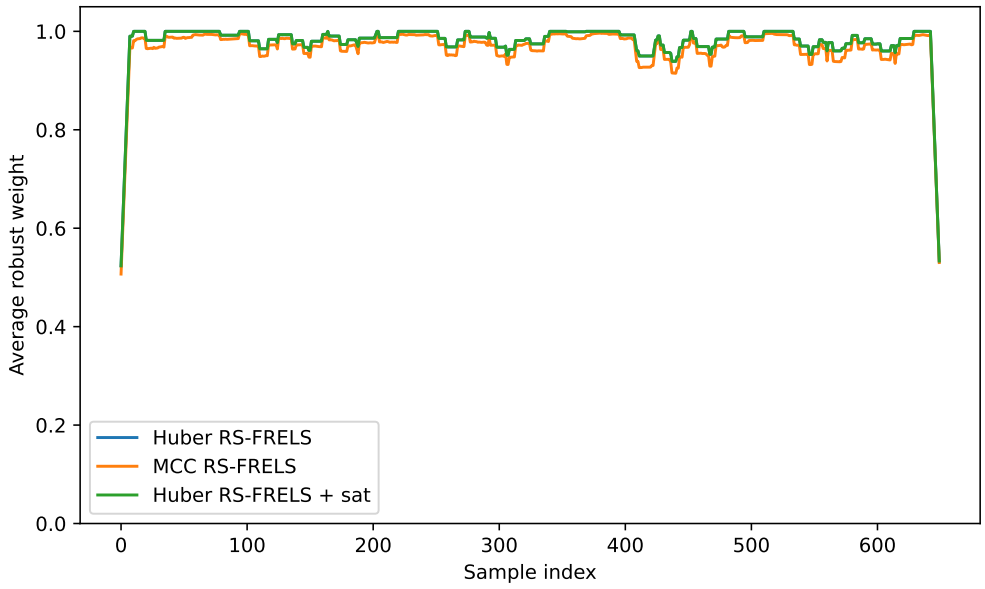

Figure 3. Representative robust weights. Huber weights clip large residuals, whereas MCC weights decay rapidly for large innovations.

6.3. Stress test and sensitivity analysis

Table 5 compares the standard contaminated Gaussian setting with a heavier-tailed Student's t stress case and an unclipped leverage stress case. The purpose is not to prove convergence under these stress

distributions but to check whether the bounded score mechanism and optional regressor saturation remain practically effective when the impulses are more severe and when large outputs are allowed to propagate into future autoregressive regressors. Figure 4 shows the corresponding clean prediction MSE for the standard and heavy-tailed scenarios.

Table 5. Stress test summary using the included simulation code.

Scenario	Method	Clean MSE	Final NMSD
Standard	FRELS	$8.649e-04 \pm 2.6e-04$	$6.387e-03 \pm 3.2e-03$
Standard	Huber RS-FRELS	$1.926e-04 \pm 4.5e-05$	$1.202e-03 \pm 3.7e-04$
Standard	MCC RS-FRELS	$1.262e-04 \pm 2.5e-05$	$6.822e-04 \pm 1.9e-04$
Standard	Huber RS-FRELS + sat	$1.926e-04 \pm 4.5e-05$	$1.202e-03 \pm 3.7e-04$
Heavy-tailed stress	FRELS	$1.700e-02 \pm 8.9e-03$	$5.009e-02 \pm 3.5e-02$
Heavy-tailed stress	Huber RS-FRELS	$3.005e-04 \pm 7.8e-05$	$7.910e-04 \pm 2.9e-04$
Heavy-tailed stress	MCC RS-FRELS	$1.261e-04 \pm 3.0e-05$	$3.770e-04 \pm 1.1e-04$
Heavy-tailed stress	Huber RS-FRELS + sat	$3.005e-04 \pm 7.8e-05$	$7.910e-04 \pm 2.9e-04$
Unclipped leverage stress	FRELS	$2.068e-02 \pm 1.1e-02$	$8.178e-02 \pm 7.1e-02$
Unclipped leverage stress	Huber RS-FRELS	$3.131e-04 \pm 6.8e-05$	$7.845e-04 \pm 2.8e-04$
Unclipped leverage stress	MCC RS-FRELS	$1.374e-04 \pm 4.3e-05$	$3.700e-04 \pm 1.1e-04$
Unclipped leverage stress	Huber RS-FRELS + sat	$3.131e-04 \pm 6.8e-05$	$7.845e-04 \pm 2.8e-04$

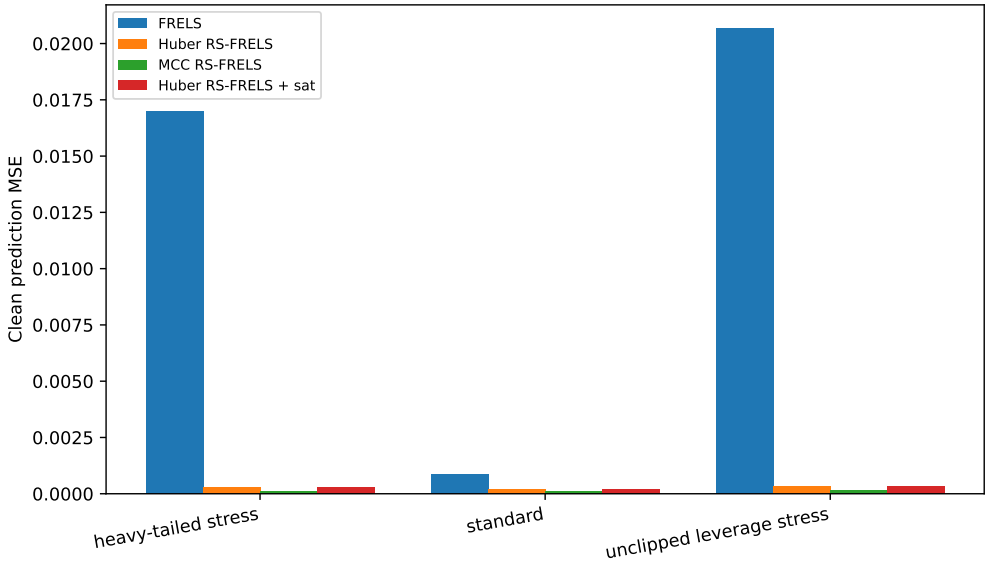

Figure 4. Clean prediction MSE under standard and heavy-tailed stress conditions. Robust variants remain substantially more stable than ordinary FRELS.

A separate leverage experiment in Table 6 deliberately removes output clipping, increases the impulse severity, and varies the saturation bound. This experiment makes the activation of (4.1) visible. It also shows the expected bias–protection trade-off: Hard bounds such as $B = 10$ cap the used regressor norm but can oversaturate informative samples, whereas mild bounds such as $B = 50$

activate rarely and mainly document that the algorithm can enforce a chosen regressor limit. The table should therefore be read as a diagnostic of leverage control, not as a claim that tighter saturation always improves prediction error. One entry deserves an explicit reading: For $B = 50$, the activation rate is only 0.21%, i.e., a handful of samples per run, and those few samples are clipped exactly onto the sphere of radius B. The reported maximum used norm $(4.119 \times 10^1 \pm 1.0 \times 10^1)$ nevertheless sits below $B = 50$ on average because, in some seeds, no raw regressor exceeds the bound at all; in those runs, the maximum used norm equals the maximum raw norm, which is below 50, and the mixture of activated and nonactivated seeds produces both the sub-B mean and the large dispersion. Figure 5 plots the resulting clean prediction MSE against the saturation bound.

Table 6. Dedicated saturation/leverage diagnostic. Here, "sat" abbreviates saturation, and activation is the fraction of samples for which $\|\psi^{\text{raw}}\| > B$.

Method	Clean MSE	Final NMSD	Activation	Max raw $\|\psi\|$	Max used $\|\psi\|$	Max jump
Huber no sat	$1.914e-03 \pm 3.4e-03$	$2.227e-03 \pm 6.6e-04$	0.0000 ± 0.0000	$5.024e+01 \pm 2.1e+01$	$5.024e+01 \pm 2.1e+01$	$6.517e+00 \pm 2.6e+00$
Huber sat $B = 10$	$2.082e+00 \pm 4.8e+00$	$5.644e-03 \pm 2.2e-03$	0.0921 ± 0.0310	$5.024e+01 \pm 2.1e+01$	$1.000e+01 \pm 1.6e-15$	$6.517e+00 \pm 2.6e+00$
Huber sat $B = 25$	$6.972e-01 \pm 1.7e+00$	$2.726e-03 \pm 9.4e-04$	0.0130 ± 0.0166	$5.024e+01 \pm 2.1e+01$	$2.500e+01 \pm 3.6e-15$	$6.517e+00 \pm 2.6e+00$
Huber sat $B = 50$	$8.834e-02 \pm 2.2e-01$	$2.224e-03 \pm 6.8e-04$	0.0021 ± 0.0030	$5.024e+01 \pm 2.1e+01$	$4.119e+01 \pm 1.0e+01$	$6.517e+00 \pm 2.6e+00$

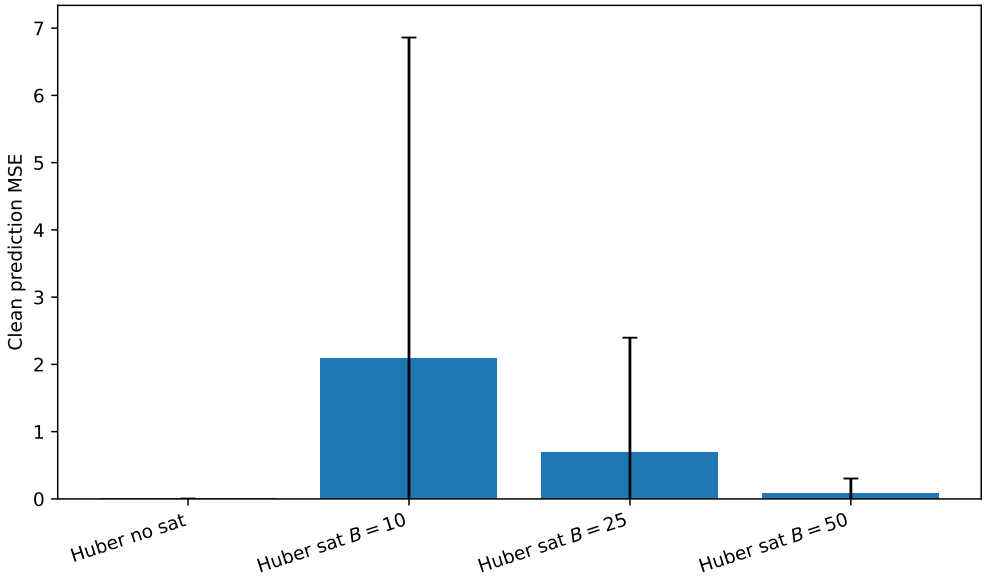

Figure 5. Clean prediction MSE in the saturation diagnostic. Tight saturation can introduce bias; the purpose of saturation is to enforce a leverage bound when such a bound is needed for stability or theory.

The sensitivity study checks the dependence on robust tuning constants. Table 7 and Figures 6–8 include MCC widths down to $\sigma = 0.05$ to bracket the weighted information collapse regime suggested by the theory. In this generator, $\sigma = 0.10$ remains the best clean MSE choice, while $\sigma = 0.05$ starts to reduce the weighted persistent excitation eigenvalue without causing a catastrophic collapse. Thus the MCC locality warning should be read as a finite-sample and model-dependent risk rather than as a universal U-shaped performance law. Excessively large Huber thresholds approach ordinary least

squares. The Huber and forgetting factor rows admit an equally mechanistic reading. Clean MSE increases monotonically in δ over the tested range because the standard impulse amplitude (scale 1.10) is far above the nominal residual scale (≈ 0.05), so smaller thresholds reject contamination earlier while barely touching nominal samples; the weighted information eigenvalue is nearly flat in δ, confirming that this rejection costs almost no excitation. For the forgetting factor, moving λ from 0.990 to 0.998 improves the clean MSE by roughly a factor 2.6 in this time-invariant generator, which is the variance-averaging effect predicted by the tracking analysis; the same choice would slow adaptation under drift, which is why Section 6.4 recommends selecting λ from the expected drift time scale rather than defaulting to the largest stable value.

Table 7. Sensitivity to δ, σ, and λ.

Setting	Clean MSE	Final NMSD	PE min eigenvalue
Huber $\delta = 0.25$	$1.476e-04 \pm 1.9e-05$	$8.497e-04 \pm 1.7e-04$	$5.214e-02 \pm 2.9e-03$
Huber $\delta = 0.35$	$2.032e-04 \pm 1.4e-05$	$1.187e-03 \pm 1.9e-04$	$5.235e-02 \pm 2.8e-03$
Huber $\delta = 0.50$	$3.077e-04 \pm 1.0e-05$	$1.831e-03 \pm 3.3e-04$	$5.259e-02 \pm 2.8e-03$
MCC $\sigma = 0.05$	$1.361e-04 \pm 4.8e-05$	$7.490e-04 \pm 2.9e-04$	$3.386e-02 \pm 2.0e-03$
MCC $\sigma = 0.10$	$7.824e-05 \pm 2.0e-05$	$4.285e-04 \pm 1.2e-04$	$4.502e-02 \pm 2.9e-03$
MCC $\sigma = 0.20$	$8.461e-05 \pm 1.9e-05$	$4.374e-04 \pm 1.4e-04$	$4.970e-02 \pm 3.0e-03$
MCC $\sigma = 0.35$	$9.839e-05 \pm 2.4e-05$	$5.026e-04 \pm 1.7e-04$	$5.116e-02 \pm 2.9e-03$
MCC $\sigma = 0.50$	$1.245e-04 \pm 2.6e-05$	$6.555e-04 \pm 1.7e-04$	$5.171e-02 \pm 2.8e-03$
MCC $\sigma = 0.70$	$1.812e-04 \pm 1.4e-05$	$1.009e-03 \pm 1.9e-04$	$5.213e-02 \pm 2.8e-03$
Huber $\lambda = 0.990$	$3.666e-04 \pm 5.0e-05$	$1.883e-03 \pm 5.4e-05$	$5.235e-02 \pm 2.8e-03$
Huber $\lambda = 0.995$	$2.032e-04 \pm 1.4e-05$	$1.187e-03 \pm 1.9e-04$	$5.235e-02 \pm 2.8e-03$
Huber $\lambda = 0.998$	$1.419e-04 \pm 3.4e-05$	$9.907e-04 \pm 3.4e-04$	$5.235e-02 \pm 2.8e-03$

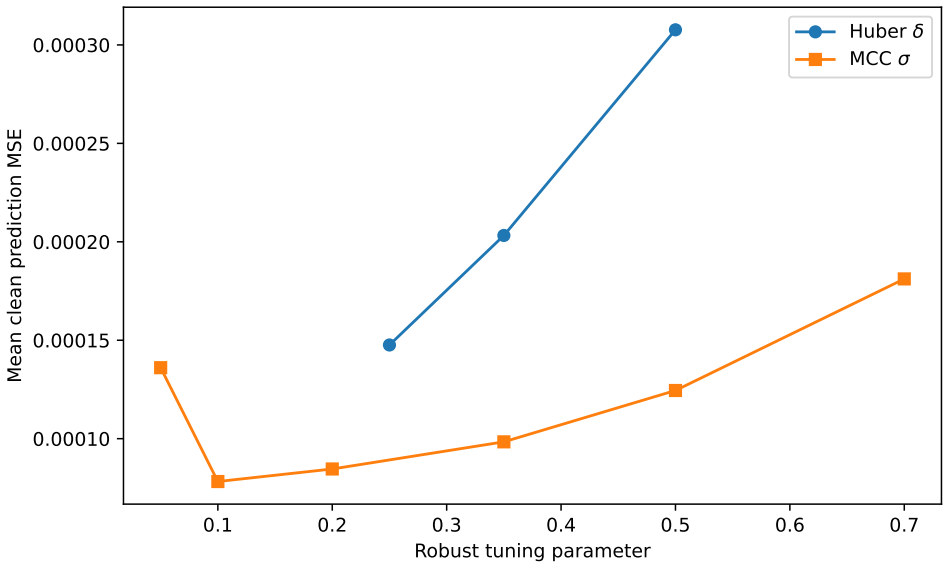

Figure 6. Sensitivity of clean prediction MSE to the Huber threshold and the MCC kernel width.

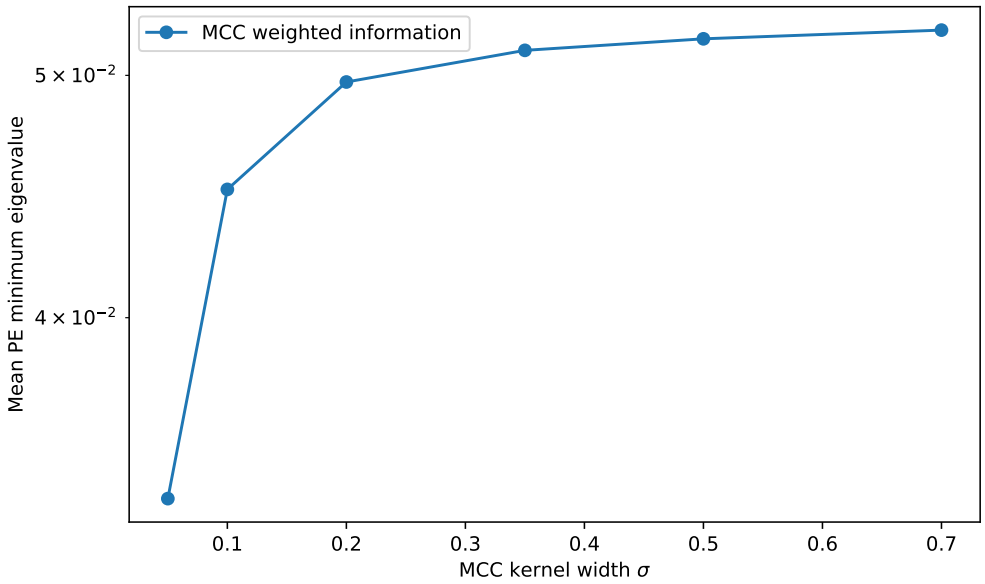

Figure 7. Weighted information retained by the MCC method as the kernel width changes. Very small kernels can suppress informative samples as well as outliers.

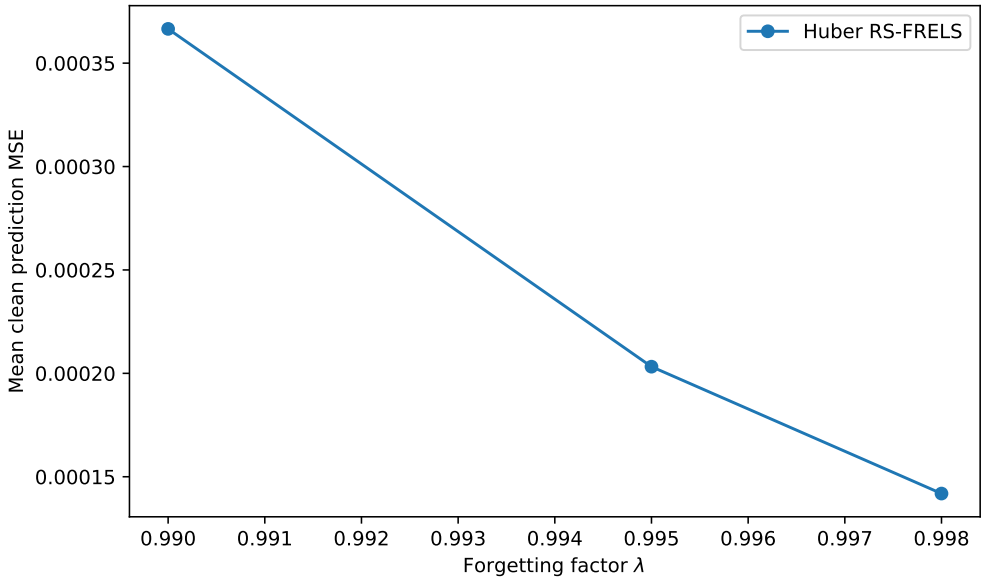

Figure 8. Sensitivity of Huber RS-FRELS to the forgetting factor. Higher λ improves steady-state averaging but can slow adaptation.

Table 8 and Figure 9 complete the stress picture by varying the impulse probability and amplitude scale. The gains over ordinary FRELS increase as the impulses become more frequent or larger, confirming that the large relative improvements in the standard case are tied to the severity of impulsive contamination rather than to a universal domination claim; at $p = 0$, the three methods are within a few percent of one another, quantifying the insurance premium paid for robustness on clean data.

Table 8. Clean MSE impulse severity grid. Entries are the means over three independent seeds per cell.

p	Impulse scale	FRELS	Huber	MCC
0.00	0.35	$6.595e-05$	$6.596e-05$	$6.532e-05$
0.00	1.10	$6.595e-05$	$6.596e-05$	$6.532e-05$
0.00	2.00	$6.595e-05$	$6.596e-05$	$6.532e-05$
0.04	0.35	$1.902e-04$	$1.317e-04$	$1.156e-04$
0.04	1.10	$1.078e-03$	$1.640e-04$	$9.854e-05$
0.04	2.00	$3.048e-03$	$1.717e-04$	$9.379e-05$
0.08	0.35	$2.710e-04$	$1.809e-04$	$1.543e-04$
0.08	1.10	$1.899e-03$	$2.579e-04$	$1.459e-04$
0.08	2.00	$5.934e-03$	$2.758e-04$	$1.152e-04$

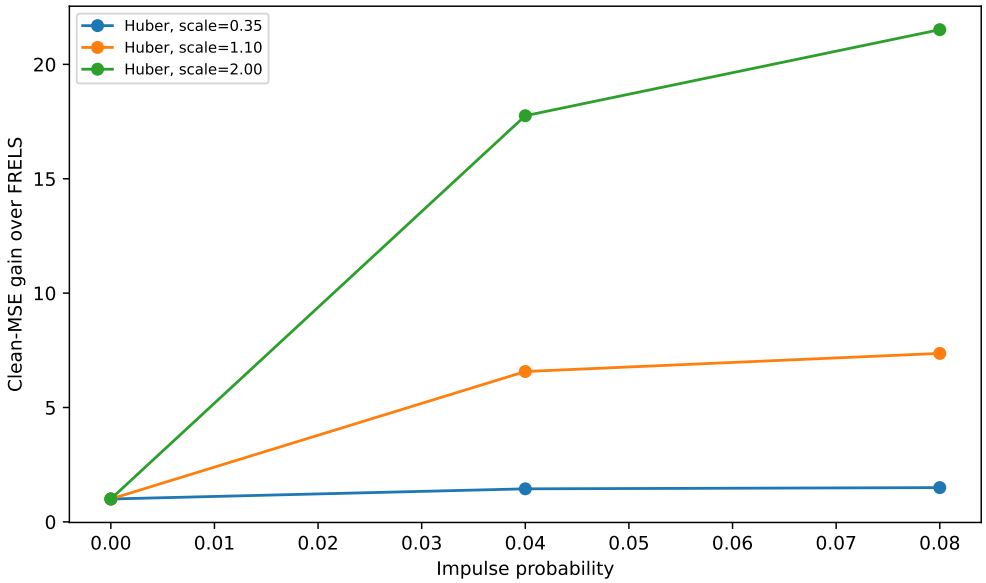

Figure 9. Clean prediction MSE over impulse probability and amplitude scale. Bounded score methods become most valuable under the high-contamination regime.

6.4. Practical tuning and nonoracle validation metrics

Nonoracle tuning rules for δ, σ, λ, and B. The clean MSE and NMSD used above are simulation metrics because the noiseless output and true parameters are known. In applications, tuning should instead use nonoracle quantities. A practical rule is to estimate a robust residual scale $\hat{s}_r$ from recent one-step residuals, for example by a median absolute deviation estimate, and set the Huber threshold near $\delta = c_\delta \hat{s}_r$ with $c_\delta \in [1.3, 2.0]$, depending on the desired outlier rejection. For the MCC context, a safe operational approach is to start with a broad bandwidth $\sigma = c_\sigma \hat{s}_r$ with $c_\sigma \geq 2$, monitor the weighted information minimum eigenvalue, and anneal σ only while the information diagnostic remains above a chosen threshold. The forgetting factor should be selected from the validation

prediction loss or expected drift time scale: A larger λ reduces variance, while a smaller λ adapts faster. The saturation bound B should be calibrated from the regressor stream rather than guessed: Record the norms $\|\psi^{\mathrm{raw}}(k)\|$ over a contamination-light calibration window and set B to a small multiple (two is used in the benchmark of Section 6.5) of a high quantile, e.g., the 99th percentile or the maximum. This keeps the activation rate near zero in normal operation, so no saturation bias is paid, while still capping the leverage of impulse-driven regressors. Table 6 quantifies what happens when B is instead pushed into the informative range. These four rules are not merely advisory: The rolling-MAD Huber threshold and the annealed MCC bandwidth evaluated in Table 2 are direct implementations, and both outperform their hand-fixed counterparts, so following the rules is empirically at least as good as expert defaults in this generator.

Practical guidelines for reporting without an oracle. Nonoracle reporting should include robust validation loss, trimmed measured output MSE, median absolute prediction error, residual quantiles, and parameter trajectory stability, in addition to the oracle clean MSE available only in simulation. (This paragraph is intentionally advice rather than a numerical result; it is kept in this dedicated subsection, next to the tuning rules it complements, and the benchmark study below applies these guidelines verbatim.)

6.5. Benchmark validation on a real heat exchanger record

To complement the synthetic study with real data, the algorithms are applied to the liquid–saturated steam heat exchanger record of the DaISy identification database (dataset 97-002) [30, 31]: 4000 samples at 1 s of the liquid flow rate (input) and outlet liquid temperature (output), a nonlinear nonminimum-phase process that is a standard benchmark for nonlinear identification and is exactly the plant class named in the conclusion of the previous version of this paper. The protocol is fully nonoracle and follows Section 6.4 verbatim. Signals are normalized by the training-half statistics. A single-subsystem ($N_s = 1$) fractional Hammerstein regressor with polynomial input terms up to degree three and GL memory $L = 45$ is used; the own-output fractional order pair is selected by the ordinary FRELS one-step validation error on the raw record over a coarse grid and then frozen for all methods and scenarios, so no method receives a tuning advantage. The selected pair $(0.30, 0.60)$ outperforms the integer-order profile $(1.00, 1.00)$ by a factor of about three in validation MSE (2.83×10^{-1} versus 8.38×10^{-1} in normalized units), which is direct real-data support for the fractional memory structure. The Huber threshold and MCC bandwidth follow the rolling median absolute deviation rules $\delta(k) = 1.5\hat{s}_r(k)$ and $\sigma(k) = 2.0\hat{s}_r(k)$, and the saturation bound is set from the warm-up regressor norms (B equal to twice their maximum). Estimation is performed online on the first 2000 samples; the identified parameters are then frozen and validated by one-step-ahead prediction on the clean second half, so that the reported errors isolate parameter quality exactly in the same way as the clean MSE does in the synthetic study. Impulsive sensor faults (Bernoulli occurrences with Student's $t_{2.5}$ amplitudes, three normalized standard deviations in scale) are injected into the estimation half only, corrupting both the innovations and the autoregressive regressors seen by the estimators; five contamination replicates are averaged per scenario.

Table 9 and Figure 10 summarize the outcome using only nonoracle metrics (validation root mean square error (RMSE), mean absolute error (MAE), and the 95% absolute error quantile, all in °C). Three observations match the theory. First, on the raw record, the three robust variants are within

about 2% of ordinary FRELS (RMSE 0.983–0.988 versus 0.968; the worst case, MCC, is 2.1% above), so the bounded score insurance is nearly free when no gross contamination is present. Second, under 5% contamination, the robust recursions clearly improve validation accuracy. The MCC method reduces RMSE from 1.710 to 1.325 (−23%) and the Huber method to 1.585 (−7%), with the same ordering in MAE and in the 95% quantile. Third, at 10% contamination the robust advantage persists but narrows (MCC 1.664 versus FRELS 1.797), because the injected impulses also propagate through the fractional own-output regressors and act as leverage points, a failure channel that residual-based weighting alone cannot fully repair; this is precisely the vertical versus leverage distinction drawn after Corollary 5.1, observed here for real data. The saturated variant pays a visible bias price at high contamination because the impulse-corrupted regressor directions are preserved by norm clipping; leverage-robust constructions such as instrumental variables remain the appropriate remedy and are left as future work. The benchmark is a single-subsystem record, so it validates the fractional Hammerstein regression and the robust scoring mechanism but not the interconnection structure; a multi-subsystem plant record remains the outstanding empirical target.

Table 9. Heat exchanger benchmark (DaISy 97-002): Frozen parameter one-step validation errors on the clean second half after estimation on the (possibly contaminated) first half. Mean ± standard deviation over five contamination replicates; the raw record is deterministic.

Scenario	Method	RMSE (°C)	MAE (°C)	$q_{0.95}$ (°C)
Raw record	FRELS	0.968	0.824	1.700
Raw record	Huber RS-FRELS	0.983	0.833	1.728
Raw record	MCC RS-FRELS	0.988	0.832	1.757
Raw record	Huber RS-FRELS + sat	0.985	0.831	1.744
5% impulses	FRELS	1.710 ± 0.084	1.417 ± 0.053	3.144 ± 0.231
5% impulses	Huber RS-FRELS	1.585 ± 0.097	1.325 ± 0.070	2.901 ± 0.203
5% impulses	MCC RS-FRELS	1.325 ± 0.212	1.118 ± 0.171	2.393 ± 0.425
5% impulses	Huber RS-FRELS + sat	1.716 ± 0.113	1.373 ± 0.071	3.504 ± 0.336
10% impulses	FRELS	1.797 ± 0.105	1.483 ± 0.070	3.368 ± 0.301
10% impulses	Huber RS-FRELS	1.726 ± 0.051	1.431 ± 0.033	3.184 ± 0.114
10% impulses	MCC RS-FRELS	1.664 ± 0.055	1.386 ± 0.040	3.072 ± 0.114
10% impulses	Huber RS-FRELS + sat	1.980 ± 0.088	1.546 ± 0.048	4.212 ± 0.312

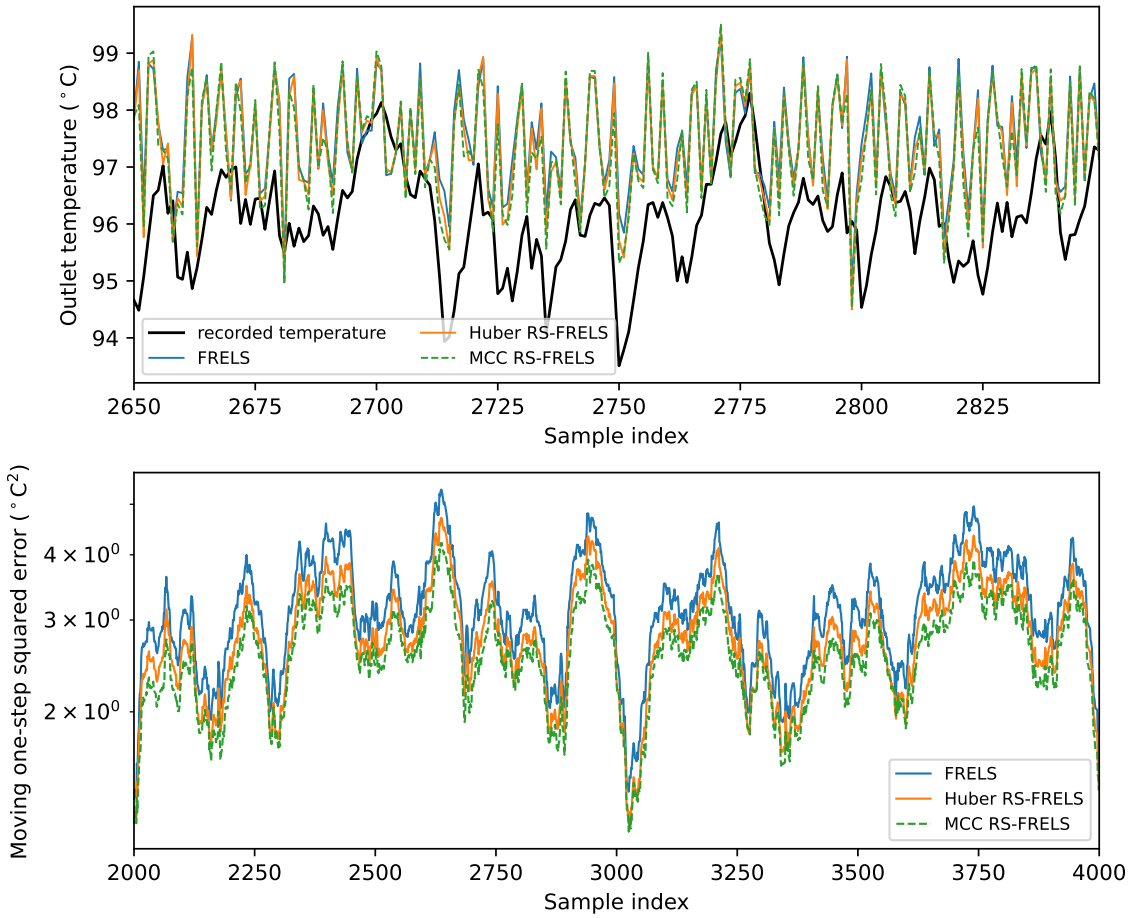

Figure 10. Heat exchanger benchmark under 10% injected impulses (first replicate). Top: Recorded outlet temperature and frozen parameter one-step predictions on a validation segment. Bottom: Moving one-step squared prediction error over the validation half.

6.6. Scalability timing

Table 11 and Figure 11 report a synthetic scalability diagnostic for the full-covariance implementation. The larger systems use deterministic stable parameter extensions with dense interconnection regressors and shorter runs for timing. The purpose is deliberately limited: The table verifies that the measured runtime grows no faster than the $O(\sum_i d_i^2)$ cost bound of Remark 4.3 (empirically, it is well described by a per-subsystem overhead plus an $O(d_i^2)$ covariance term, with the interpreter overhead dominating at small d_i), and that remark states precisely in what regime the method scales (sparse coupling, per-subsystem parallelism, or the $O(d_i)$ NLMS variant) and in what regime it does not (dense all-to-all coupling with full covariance). No claim of large-scale plant validation is attached to this timing study. For dense all-to-all coupling, $d_i = 5 + (N_s - 1)$, and the measured runtime stays within the $O(\sum_i d_i^2)$ cost envelope of Remark 4.3. The $N_s = 50$ row also illustrates that weighted information can weaken as dimension grows, so high-dimensional applications should use sparse coupling, regularization, or distributed covariance approximations. To prevent a misreading of that row, the degradation of clean MSE, final NMSD, and the persistent excitation eigenvalue (down

to 1.107×10^{-2}) at $N_s = 50$ is most plausibly an excitation effect, not an algorithmic failure. The dense $N_s = 50$ configuration has 2700 unknown parameters identified from short timing runs, and the same input class must excite a 54-dimensional regressor per subsystem; the weighted information matrix therefore thins in its worst direction as the dimension grows, which is exactly what the reported eigenvalue documents. Longer records or sparser coupling restore the excitation margin.

Table 10 adds the per-method runtimes that were missing from the previous version: All structurally distinct contrast algorithms are timed on the identical data and regressor streams (the adaptive threshold and annealed bandwidth variants share the per-sample cost of their base recursions to within measurement noise, and the saturated Huber reference is the implementation timed in Table 11). The RLS-type methods (FRELS, Huber, MCC, RLM–Hampel) are indistinguishable in terms of cost, as expected since the robust scoring adds only a scalar weight evaluation to the $O(d_i^2)$ covariance update; robustness is computationally free at this scale. The Huber-NLMS variant is the only structurally cheaper algorithm ($O(d_i)$ per update), and its advantage grows with dimension, reaching a factor of about 2.3 at $N_s = 50$. Together with its weaker accuracy in Table 2, this quantifies the accuracy–cost trade-off available when covariance updates become the bottleneck.

Table 10. Per-method runtime per sample (ms) in the scalability test. All methods run on identical data and regressors; means ± standard deviation over two seeds.

N_s	FRELS	Huber RS-FRELS	MCC RS-FRELS	RLM–Hampel	Huber NLMS
3	0.125 ± 0.002	0.126 ± 0.006	0.134 ± 0.000	0.130 ± 0.003	0.095 ± 0.003
5	0.211 ± 0.009	0.206 ± 0.006	0.202 ± 0.004	0.200 ± 0.006	0.150 ± 0.005
10	0.382 ± 0.001	0.386 ± 0.029	0.403 ± 0.025	0.414 ± 0.033	0.280 ± 0.013
20	0.820 ± 0.011	0.825 ± 0.015	0.849 ± 0.013	0.844 ± 0.021	0.544 ± 0.012
50	3.298 ± 0.019	3.376 ± 0.102	3.327 ± 0.005	3.561 ± 0.023	1.443 ± 0.014

Table 11. Scalability timing for Huber RS-FRELS with estimator-side saturation. Entries are the means over two seeds.

N_s	d_i	Total parameters	Runtime/sample (ms)	Clean MSE	Final NMSD	PE min eigenvalue
3	7	21	0.1377 ± 0.0112	$3.338e-04 \pm 1.3e-05$	$2.616e-03 \pm 4.5e-04$	$4.606e-02 \pm 6.3e-03$
5	9	45	0.2436 ± 0.0118	$3.338e-04 \pm 1.1e-05$	$2.254e-03 \pm 2.5e-04$	$4.816e-02 \pm 3.6e-05$
10	14	140	0.4190 ± 0.0028	$8.255e-04 \pm 2.1e-04$	$3.098e-03 \pm 8.3e-04$	$5.152e-02 \pm 2.6e-03$
20	24	480	1.0076 ± 0.2055	$1.251e-03 \pm 1.8e-04$	$6.432e-03 \pm 3.4e-04$	$4.990e-02 \pm 1.5e-03$
50	54	2700	3.6099 ± 0.2363	$8.280e-03 \pm 4.7e-04$	$4.766e-02 \pm 1.1e-02$	$1.107e-02 \pm 1.4e-04$

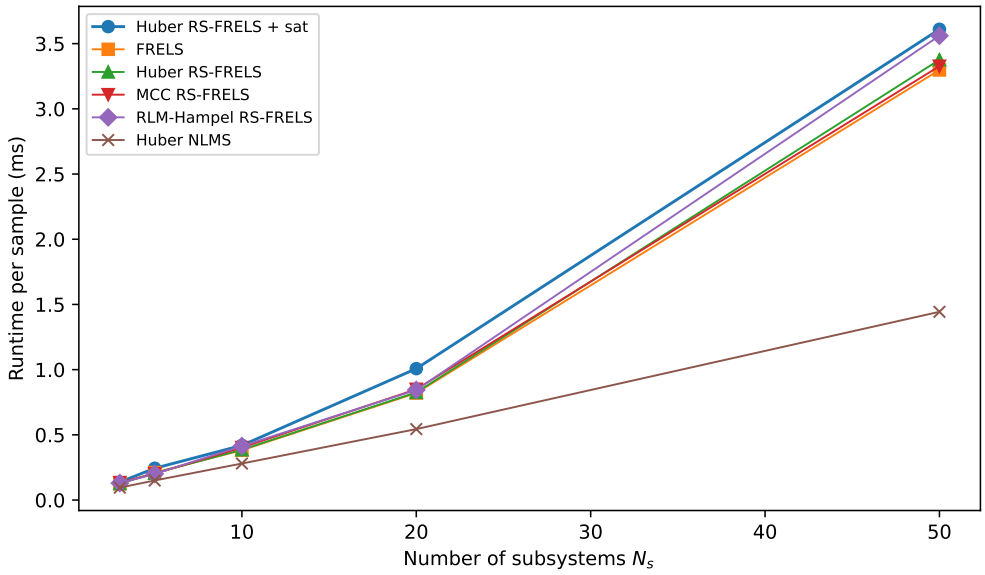

Figure 11. Measured runtime per sample as the number of subsystems increases for the saturated Huber reference implementation and all contrast methods. The $O(d_i)$ Huber-NLMS curve separates from the $O(d_i^2)$ RLS-type cluster as the dimension grows.

6.7. Ablation study

The model structure ablation removes one modeling component at a time. The short-memory and integer-order variants test whether fractional memory is useful, while the no-coupling variant tests the contribution of interconnection regressors. These are model mismatch diagnostics, not same-parameter-space algorithm comparisons. Accordingly, Table 12 reports the clean prediction MSE only; the final NMSD is intentionally omitted because the parameter vectors differ from the full fractional interconnected $\theta^\star$. Figure 12 reports the clean prediction MSE for a representative run.

Table 12. Model structure ablation. These variants use different regressor dimensions, so NMSD is not reported.

Model variant	Clean MSE
Huber RS-FRELS	$1.646e-04 \pm 4.2e-05$
Huber short-memory	$6.913e-04 \pm 1.2e-04$
Huber integer-order	$1.346e-02 \pm 6.3e-04$
Huber no coupling	$4.871e-03 \pm 5.4e-04$

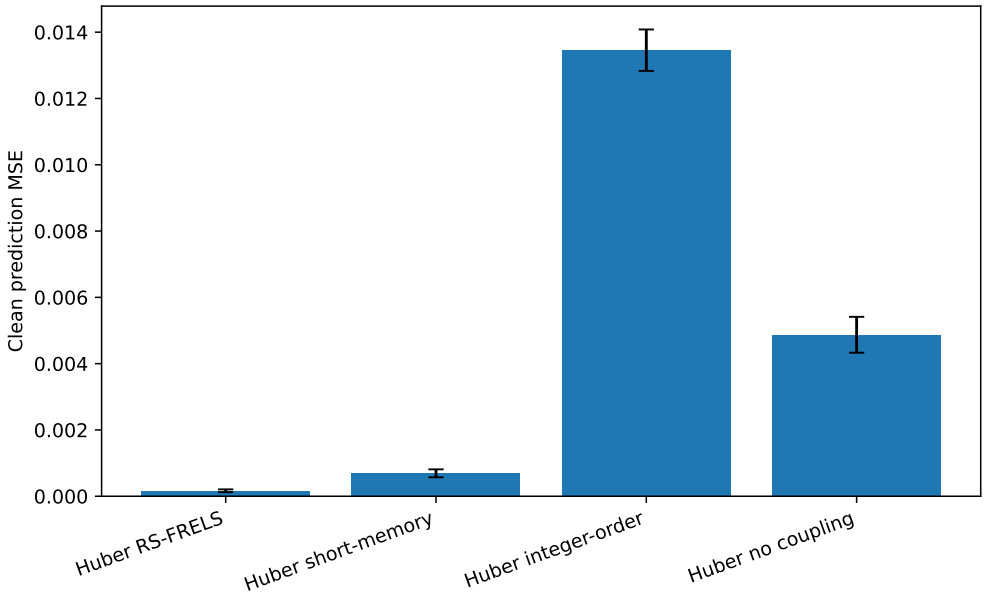

Figure 12. Ablation study. Removing fractional memory or interconnection terms increases clean prediction error in this simulated interconnected fractional plant.

Remark 6.1 (What the experiments validate). The simulations validate the algorithmic mechanism behind Theorem 5.3 and Corollary 5.1: Large innovations receive reduced score magnitudes, and optional saturation prevents raw leverage from entering the update unboundedly. The simulations are also consistent with the covariance and tracking results when weighted excitation remains adequate. They do not, by themselves, prove the stochastic approximation assumptions of Theorem 5.5 or verify a small score orthogonality defect; those assumptions are model-dependent and must be checked separately for a real plant. The heat exchanger benchmark of Section 6.5 extends this validation to real data for the single-subsystem case: It confirms the near-zero cost of the robust scores on clean records, their clear benefit under moderate impulsive contamination, and—honestly—the leverage limitation at high contamination, while leaving multi-subsystem real plant validation open.

7. Conclusions

A robust stochastic FRELS method has been developed for finite-memory fractional interconnected Hammerstein regressions under impulsive disturbances. The analysis separates the standard cited tools from the contribution-specific results and establishes the precise interpretation of the recursion as an exact weighted RLS solution of a frozen-weight surrogate. Under explicit boundedness and excitation assumptions, the Huber and MCC scores provide bounded single-sample influence; an optional regressor saturation variant turns the bounded regressor condition into an estimator-side design choice. Population and recursive convergence results are given in conditional form: The Huber approach has Fisher consistency with local quadratic growth under a positive curvature condition, an explicit orthogonality-defect bound quantifies local population bias when colored output regressors break score orthogonality, and the MCC approach is stated only as a stationary local population target because of its nonconvexity. The weighted persistent excitation condition, which previously

rested on a deterministic pathwise lower bound on the robust weight that is unrealistic for MCC RS-FRELS under impulsive residuals, is now supported by an alternative conditional-expectation lower bound combined with a matrix-Azuma deviation control, so the assumption is checkable in terms of the residual conditional density rather than residual extrema. The local recursive convergence theorem is complemented by a global statement for the projected Huber RS-FRELS on any compact convex projection set under uniform Huber curvature, whose key curvature condition is accompanied by an explicit discussion of when it is verifiable and when it degenerates to a semilocal statement, while the MCC equivalent retains only the local guarantee because its nonconvex redescending score precludes the same globalization. For MCC RS-FRELS, the nonconvexity is handled operationally rather than dismissed: Warm starting, bandwidth annealing, and information monitoring are stated as concrete rules and are validated as an algorithm in the Monte Carlo comparison. Numerical evidence supports the robustness benefits through Monte Carlo statistics, paired tests, same-regressor algorithm baselines including a redescending RLM–Hampel score and adaptively tuned variants, stress tests, MCC information diagnostics, impulse severity grids, saturation activation diagnostics, scalability timing with per-method runtimes, and scoped model structure ablations. A first real-data validation on the DaISy liquid–saturated steam heat exchanger benchmark shows that the robust recursions cost almost nothing on the raw record, improve held-out prediction accuracy substantially under moderate injected impulsive sensor faults, and lose part of their advantage under heavy contamination through the leverage channel, in agreement with the vertical versus leverage analysis; the fractional-order profile selected on this record also outperforms its integer-order counterpart, supporting the fractional model class with real data. The claimed scalability is likewise scoped precisely: The full-covariance cost law is verified empirically, and sparse coupling, per-subsystem parallelism, or the $O(d)$ NLMS variant constitute the documented path to genuinely large systems. The most important remaining empirical step is validation on a multi-subsystem interconnected plant record, such as a Wiener–Hammerstein cascade or a multi-loop process. Future work should also develop instrumental or whitened robust fractional regressions for colored output error noise (the leverage channel exposed by the benchmark), sparse or distributed covariance approximations for very large systems, and joint estimation of fractional orders together with robust recursive parameters.

Author contributions

Slim Dhahri: Conceptualization, methodology, software, writing – original draft; Mourad Elloumi: Investigation, validation; Hend Aljahani: Resources, funding acquisition, writing – review and editing; Salem Albalawi: Investigation, visualization; Sahar Almashaan: Data curation, resources, writing – review and editing; Hatem Alwardi: Formal analysis, validation; Foued Mtiri: Supervision, project administration, validation. 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, Kingdom of Saudi Arabia, for funding this research work through the project number "NBU-FPEJ-2026-1257-02".

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.

Funding

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, Kingdom of Saudi Arabia, for funding this research work through the project number "NBU-FPEJ-2026-1257-02".

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

The authors declare no conflicts of interest.

References

1. K. S. Narendra, P. G. Gallman, An iterative method for the identification of nonlinear systems using a Hammerstein model, *IEEE Trans. Autom. Control*, **11** (1966), 546–550. https://doi.org/10.1109/TAC.1966.1098387
2. S. A. Billings, S. Y. Fakhouri, Identification of systems containing linear dynamic and static nonlinear elements, *Automatica*, **18** (1982), 15–26. https://doi.org/10.1016/0005-1098(82)90022-X
3. F. Giri, E. W. Bai, *Block-oriented nonlinear system identification*, London: Springer, 2010. https://doi.org/10.1007/978-1-84996-513-2
4. K. B. Oldham, J. Spanier, *The fractional calculus: Theory and applications of differentiation and integration to arbitrary order*, Elsevier Science, 1974.
5. I. Podlubny, *Fractional differential equations*, Academic Press, 1999.
6. D. V. Ivanov, Identification of discrete fractional order Hammerstein systems, In: *2015 International Siberian conference on control and communications (SIBCON)*, 2015, 1–4. https://doi.org/10.1109/SIBCON.2015.7147074
7. Q. Jin, Y. Ye, W. Cai, Z. Wang, Recursive identification for fractional order Hammerstein model based on ADELS, *Math. Probl. Eng.*, **2021** (2021), 6629820. https://doi.org/10.1155/2021/6629820

8. Q. Jin, B. Wang, Z. Wang, Recursive identification for MIMO fractional-order Hammerstein model based on AIAGS, *Mathematics*, **10** (2022), 212. https://doi.org/10.3390/math10020212
9. M. Elloumi, O. Naifar, I. Bouzida, Modeling and parameter estimation for fractional large-scale interconnected Hammerstein systems, *Asian J. Control*, 2015, 1–11. https://doi.org/10.1002/asjc.70009
10. J. Wang, W. Xiong, F. Ding, Y. Zhou, E. Yang, Parameter estimation method for separable fractional-order Hammerstein nonlinear systems based on the online measurements, *Appl. Math. Comput.*, **488** (2025), 129102. https://doi.org/10.1016/j.amc.2024.129102
11. X. Yin, Y. Liu, A new orthogonal least squares identification method for a class of fractional Hammerstein models, *Algorithms*, **18** (2025), 201. https://doi.org/10.3390/a18040201
12. C. Liu, H. Wang, Y. An, Staged parameter identification method for non-homogeneous fractional-order Hammerstein MISO systems using multi-innovation LM: Application to heat flow density modeling, *Fractal Fract.*, **9** (2025), 150. https://doi.org/10.3390/fractalfract9030150
13. S. Marzougui, S. Bedoui, K. Abderrahim, Two-stage identification approach for fractional parameter-varying Hammerstein system, *Eur. J. Control*, **89** (2026), 101504. https://doi.org/10.1016/j.ejcon.2026.101504
14. P. J. Huber, Robust estimation of a location parameter, *Ann. Math. Stat.*, **35** (1964), 73–101. https://doi.org/10.1214/aoms/1177703732
15. P. J. Huber, *Robust statistics*, John Wiley & Sons, 1981. https://doi.org/10.1002/0471725250
16. 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
17. M. Z. A. Bhotto, A. Antoniou, Robust recursive least squares adaptive-filtering algorithm for impulsive noise environments, *IEEE Signal Process. Lett.*, **18** (2011), 185–188. https://doi.org/10.1109/LSP.2011.2106119
18. 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
19. B. Chen, L. Xing, H. Zhao, S. Du, J. C. Principe, Effects of outliers on the maximum correntropy estimation: A robustness analysis, *IEEE Trans. Syst. Man Cybern. Syst.*, **51** (2021), 4007–4012. https://doi.org/10.1109/TSMC.2019.2931403
20. W. Ma, J. Duan, B. Chen, G. Gui, W. Man, Recursive generalized maximum correntropy criterion algorithm with sparse penalty constraints for system identification, *Asian J. Control*, **19** (2017), 1164–1172. https://doi.org/10.1002/asjc.1448
21. Z. Qin, J. Tao, L. Yang, M. Jiang, Proportionate recursive maximum correntropy criterion adaptive filtering algorithms and their performance analysis, *Digit. Signal Process.*, **139** (2023), 104073. https://doi.org/10.1016/j.dsp.2023.104073
22. J. Lin, S. C. Chan, Recursive identification of nonlinear Hammerstein systems with FIR linear parts under impulsive noise with model selection and order determination, *IEEE Access*, **12** (2024), 138702–138715. https://doi.org/10.1109/ACCESS.2024.3433583

23. X. Wang, F. Zhu, Robust recursive estimation for the errors-in-variables nonlinear systems with impulsive noise, *Sci. Rep.*, **15** (2025), 6031. https://doi.org/10.1038/s41598-025-89969-z

24. Z. A. Khan, T. A. Khan, M. Waqar, N. I. Chaudhary, M. A. Z. Raja, C. M. Shu, Nonlinear marine predator algorithm for robust identification of fractional Hammerstein nonlinear model under impulsive noise with application to heat exchanger system, *Commun. Nonlinear Sci. Numer. Simul.*, **146** (2025), 108809. https://doi.org/10.1016/j.cnsns.2025.108809

25. M. A. Ali, N. I. Chaudhary, T. A. Khan, W. L. Mao, C. C. Lin, Z. A. Khan, et al., Auxiliary model-based chameleon swarm optimization for robust parameter estimation of fractional order nonlinear Hammerstein systems, *J. Comput. Nonlinear Dyn.*, **20** (2025), 091005. https://doi.org/10.1115/1.4068718

26. Q. Liu, Q. Dong, H. Jiang, F. Chen, X. Wang, Multi-innovation recursive identification methods for nonlinear sandwich systems using the auxiliary model, *Appl. Math. Model.*, **150** (2026), 116398. https://doi.org/10.1016/j.apm.2025.116398

27. B. M. Vinagre, I. Podlubny, A. Hernandez, V. Feliu, Some approximations of fractional order operators used in control theory and applications, *Fract. Calc. Appl. Anal.*, **3** (2000), 231–248.

28. L. Ljung, *System identification: Theory for the user*, 2 Eds., Englewood Cliffs: Prentice Hall, 1999.

29. S. Haykin, *Adaptive filter theory*, 4 Eds., Prentice Hall, 2002.

30. B. L. R. De Moor, DaISy: Database for the identification of systems, 1997, Dataset 97-002: liquid–saturated steam heat exchanger. Available from: `https://homes.esat.kuleuven.be/~smc/daisy/`.

31. S. Bittanti, L. Piroddi, Nonlinear identification and control of a heat exchanger: A neural network approach, *J. Franklin Inst.*, **334** (1997), 135–153. https://doi.org/10.1016/S0016-0032(96)00059-2

32. J. A. Tropp, User-friendly tail bounds for sums of random matrices, *Found. Comput. Math.*, **12** (2012), 389–434. https://doi.org/10.1007/s10208-011-9099-z

A. Literature comparison table

Table 13 gives a mathematical comparison with closely related recent work. The notation "quadratic" denotes least squares type criteria, while "robust score" denotes a bounded score update of the form $P\psi s(\varepsilon)$. The table is intended to clarify the specific gap addressed in this paper rather than to rank the cited methods.

Table 13. Mathematical comparison with representative related identification methods.

Reference	Model class	Disturbance treatment	Recursive criterion/update	Difference from the present work
[9]	Fractional interconnected Hammerstein model with GL regressors	Colored finite variance stochastic disturbance	FRELS/extended least squares with quadratic innovation	Provides the fractional interconnected baseline; no bounded robust score for impulsive contamination.
[22]	Integer-order Hammerstein FIR model with model selection	Impulsive output noise	Robust recursive M-estimation with smoothly clipped absolute deviation (SCAD)/variable forgetting factor (VFF) mechanisms	Robust to impulses but not fractional and not an interconnected fractional GL formulation.
[23]	Errors-in-variables nonlinear systems	Impulsive noise in a nonlinear errors-in-variables (EIV) setting	Continuous logarithmic mixed p-norm recursive estimation	Treats noisy inputs and nonlinear EIV systems; does not address fractional interconnected Hammerstein regressors.
[10–12]	A fractional Hammerstein or fractional nonlinear model	Mainly finite variance/noisy measurements	Separable least squares, orthogonal least squares, or staged least squares (LS)/LM identification	Estimate fractional Hammerstein parameters, but the update is not a bounded score online robust FRELS recursion for interconnections.
[24, 25]	Fractional Hammerstein type nonlinear models	Gaussian and impulsive noise scenarios in metaheuristic studies	Population-based optimization with auxiliary models	Robust empirical performance is investigated, but not an RLS-type online frozen-weight surrogate with covariance and influence analysis.
[20, 21]	Adaptive filtering/sparse system identification	Non-Gaussian or impulsive noise	Recursive maximum correntropy type updates	Provides correntropy-based robustness, but without fractional Hammerstein interconnection structure.
This paper	Fractional interconnected Hammerstein regression	Colored plus impulsive/heavy-tailed disturbance	Frozen-weight Huber/MCC RS-FRELS: $\Delta\hat{\theta} = P\psi s(\varepsilon)/(\lambda + \omega\psi^{\top}P\psi)$	Combines GL memory, interconnection regressors, Hammerstein lumped parameters, bounded robust scores, and conditional recursive theory.

B. Reproducibility details

The representative three-subsystem experiment uses the finite-memory orders

$$\{0.18,\ 0.27,\ 0.31,\ 0.42,\ 0.55,\ 0.63,\ 0.76\}, \tag{B.1}$$

where 0.42 and 0.76 are own-output orders; 0.18, 0.31, and 0.55 are input-polynomial orders; 0.27 is the interconnection order; and 0.63 is the colored noise order. The three lumped parameter vectors are

$$\begin{aligned}\theta_1^\star &= (0.44, -0.16, 0.82, -0.20, 0.055, 0.095, -0.070)^\top,\\ \theta_2^\star &= (0.38, -0.12, 0.74, -0.16, 0.045, -0.085, 0.060)^\top,\\ \theta_3^\star &= (0.41, -0.14, 0.78, -0.18, 0.050, 0.070, -0.065)^\top.\end{aligned} \tag{B.2}$$

For each subsystem, the exogenous input is a first-order autoregressive (AR(1)) sequence with a coefficient 0.72 and an innovation standard deviation $0.65 + 0.05(i - 1)$, followed by a hyperbolic tangent nonlinearity and additive Gaussian jitter of standard deviation 0.15. The colored noise component is generated by applying the finite GL filter of order 0.63 to Gaussian innovations of standard deviation 0.045. The standard impulse process has a Bernoulli probability $p = 0.04$ and a Gaussian amplitude scale 1.10; stress cases use Student's $t_{2.5}$ amplitudes with the probabilities and scales reported in Table 1. Unless stated otherwise, the data generator clips simulated outputs to $[-8, 8]$ only to prevent unbounded autoregressive leverage during synthetic plant generation; this clipping is not part of the estimator.

All main algorithms start from $\hat{\theta}_i(0) = 0$ and $P_i(0) = 80I$. The default forgetting factor is $\lambda = 0.995$, the default Huber threshold is $\delta = 0.35$, and the default MCC width is $\sigma = 0.50$. The RLM–Hampel baseline uses the three-part redescending score with the knots $(a, b, c) = (0.35, 0.70, 1.40)$. The adaptive threshold Huber baseline recomputes $\delta(k) = 1.5 \times 1.4826 \times \mathrm{MAD}$ over the last 150 residuals (per subsystem, with a floor of 0.05, activated after 30 residuals). The annealed MCC baseline decreases $\sigma(k)$ linearly from 0.50 to 0.10 over the first 300 samples and holds it constant afterwards. The main Monte Carlo table uses Seeds $7, 8, \ldots, 36$; the stress tests use Seeds $30, \ldots, 34$; the sensitivity tests use Seeds $60, 61, 62$; the impulse grid uses Seeds $80, 81, 82$; the saturation diagnostic uses Seeds $110, \ldots, 115$; and the scalability timing uses Seeds 150 and 151. For $N_s > 3$ in the scalability diagnostic, the script creates deterministic stable parameter extensions with small distance-decaying interconnection coefficients. These extended systems are used only for timing and scalability diagnostics, not as additional evidence of real plant validity.

The heat exchanger benchmark uses the public DaISy dataset 97-002 [30, 31] (the liquid flow rate as input, the outlet liquid temperature as output, 4000 samples at 1 s), downloadable from the DaISy repository. Signals are normalized by the mean and standard deviation of the first 2000 samples. The regressor uses own-output fractional orders selected from the grid $\{(0.30, 0.60), (0.42, 0.76), (0.60, 0.90), (1.00, 1.00)\}$ by the FRELS validation error on the raw record, the input-polynomial orders $(0.18, 0.31, 0.55)$, the GL memory $L = 45$, the forgetting factor $\lambda = 0.998$, and $P(0) = 80I$. The rolling residual-scale estimate uses a 300-sample window with a 50-sample activation delay and a floor of 0.02; $\delta(k) = 1.5\hat{s}_r(k)$, $\sigma(k) = 2.0\hat{s}_r(k)$, and B is twice the maximum warm-up regressor norm over the first 200 samples. Contamination replicates use NumPy generators seeded as $200, \ldots, 204$; impulses are Bernoulli with a probability of 0.05 or 0.10 and Student's $t_{2.5}$ amplitudes scaled by 3.0 normalized units, injected into the first 2000 samples only.

C. Auxiliary proofs

Proof of Proposition 5.1. Define the ordinary and weighted finite-window information matrices

$$G_k = \frac{1}{T}\sum_{\ell=k}^{k+T-1} \psi(\ell)\psi^\top(\ell), \qquad G_k^\omega = \frac{1}{T}\sum_{\ell=k}^{k+T-1} \omega(\ell)\psi(\ell)\psi^\top(\ell).$$

Part (i): Pathwise lower bound. For every vector $x \in \mathbb{R}^d$, we have

$$x^\top G_k^\omega x = \frac{1}{T}\sum_{\ell=k}^{k+T-1} \omega(\ell)(\psi^\top(\ell)x)^2.$$

Using $\omega(\ell) \geq \omega_{\min}$ gives

$$x^\top G_k^\omega x \geq \omega_{\min}\frac{1}{T}\sum_{\ell=k}^{k+T-1} (\psi^\top(\ell)x)^2 = \omega_{\min} x^\top G_k x.$$

By the ordinary PE condition, we have

$$x^\top G_k x \geq m_0\|x\|^2,$$

and therefore

$$x^\top G_k^\omega x \geq \omega_{\min} m_0\|x\|^2.$$

Since this holds for every x, we obtain $G_k^\omega \succeq \omega_{\min} m_0 I_d$. For the upper bound, the Huber and MCC weights satisfy $0 \leq \omega(\ell) \leq 1$, so $x^\top G_k^\omega x \leq x^\top G_k x \leq M_0\|x\|^2$ and $G_k^\omega \preceq M_0 I_d$. Combining the two bounds yields Assumption 5.1 with $m = \omega_{\min} m_0$ and $M = M_0$.

Part (ii): Conditional expectation lower bound. Define the conditional and martingale difference components

$$\bar{G}_k^\omega = \frac{1}{T}\sum_{\ell=k}^{k+T-1} \mathbb{E}\Big[\omega(\ell)\psi(\ell)\psi^\top(\ell)\,\Big|\,\mathcal{F}_{\ell-1}\Big], \qquad \widehat{G}_k^\omega = \frac{1}{T}\sum_{\ell=k}^{k+T-1} D(\ell),$$

so that $G_k^\omega = \bar{G}_k^\omega + \widehat{G}_k^\omega$ with $D(\ell)$ defined in (5.14). Because $\psi(\ell)$ is $\mathcal{F}_{\ell-1}$-measurable (it is built from past data passing through the GL filter)

$$\mathbb{E}\Big[\omega(\ell)\psi(\ell)\psi^\top(\ell)\,\Big|\,\mathcal{F}_{\ell-1}\Big] = \mathbb{E}\Big[\omega(\ell)\,\Big|\,\mathcal{F}_{\ell-1}, \psi(\ell)\Big]\psi(\ell)\psi^\top(\ell) \geq \omega_{\mathrm{E}}\psi(\ell)\psi^\top(\ell)$$

by (5.13). Averaging over the window gives $\bar{G}_k^\omega \succeq \omega_{\mathrm{E}} G_k \succeq \omega_{\mathrm{E}} m_0 I_d$, and the upper bound argument from Part (i) shows $\bar{G}_k^\omega \preceq M_0 I_d$.

It remains to control the martingale difference part $\widehat{G}_k^\omega$. The increments $D(\ell)$ form a matrix martingale difference sequence with respect to $\{\mathcal{F}_\ell\}$ and satisfy $\|D(\ell)\| \leq 2\bar{\psi}^2$ almost surely, because $0 \leq \omega(\ell) \leq 1$ and $\|\psi(\ell)\| \leq \bar{\psi}$. The matrix Azuma–Hoeffding inequality for self-adjoint martingale sums (see, e.g., [32]) with the deterministic dominating sequence $A_\ell = 2\bar{\psi}^2 I_d$, so that the variance proxy of a length-T window is $\sigma^2 = \big\|\sum_\ell A_\ell^2\big\| = 4T\bar{\psi}^4$, gives the following for any $t > 0$ and any window start k:

$$\mathbb{P}\Big(\Big\|\widehat{G}_k^\omega\Big\| \geq t\Big) = \mathbb{P}\Big(\Big\|\sum_{\ell=k}^{k+T-1} D(\ell)\Big\| \geq Tt\Big) \leq 2d\exp\Big(-\frac{(Tt)^2}{8\sigma^2}\Big) = 2d\exp\Big(-\frac{Tt^2}{32\bar{\psi}^4}\Big).$$

Taking $t = \epsilon$ and combining with $\bar{G}_k^{\omega} \succeq \omega_{\mathrm{E}} m_0 I_d$ proves the high-probability statement (ii.a); the upper bound holds pathwise because $G_k^{\omega} \preceq G_k \preceq M_0 I_d$ deterministically.

For (ii.b), note that this bound holds for each window separately, and for a fixed window length the failure probability does not decrease with the window start k; an eventual almost-sure statement therefore genuinely requires the window length to grow. Let the window starting at k have a length $T_k \geq (32\bar{\psi}^4/\epsilon^2)(2\ln k + \ln 2d)$. Its failure probability is at most

$$2d \exp\left(-\frac{T_k \epsilon^2}{32\bar{\psi}^4}\right) \leq 2d \exp(-2\ln k - \ln 2d) = k^{-2},$$

which is summable in k. By the Borel–Cantelli lemma, almost surely only finitely many windows fail, i.e., there is a finite random time k_1 such that (5.16) holds for all $k \geq k_1$. Choosing $\epsilon \leq \omega_{\mathrm{E}} m_0/2$ gives the claimed eventual weighted-PE rate $m = (\omega_{\mathrm{E}}\, m_0)/2$. □

Proof of Theorem 5.2. The information matrix has the explicit representation

$$R(k) = \lambda^k P(0)^{-1} + \sum_{\ell=1}^{k} \lambda^{k-\ell} \omega(\ell)\psi(\ell)\psi^{\top}(\ell). \tag{C.1}$$

Every term in the summation is positive semidefinite, and the forgotten initial information term is positive definite. Thus $R(k) \succ 0$ and $P(k) = R(k)^{-1} \succ 0$.

Assume now that Assumption 5.1 holds from the start time k_0. For $k \geq k_0$, let $B_k = \lfloor (k - k_0 + 1)/T \rfloor$, as in (5.18). This is the number of complete length-T windows that can be selected from the recent portion $\{k_0, \ldots, k\}$. When $\lambda = 1$, group the summation in (C.1) into these B_k complete blocks, plus a non-negative remainder. Assumption 5.1 gives at least mTI_d from each complete block, so

$$R(k) \succeq P(0)^{-1} + mTB_k I_d. \tag{C.2}$$

Taking the reciprocal eigenvalue bounds yields (5.19).

For $0 < \lambda < 1$, use complete blocks counted backward from the present time, while keeping only blocks whose starting index is at least k_0. The j-th most recent complete block is

$$\{k - (j+1)T + 1, \ldots, k - jT\}, \qquad j = 0, \ldots, B_k - 1. \tag{C.3}$$

For every sample in this block, the exponential factor $\lambda^{k-\ell}$ is at least $\lambda^{(j+1)T-1}$, because $0 < \lambda < 1$. Therefore

$$\begin{aligned} R(k) &\succeq \lambda^k P(0)^{-1} + mT \sum_{j=0}^{B_k-1} \lambda^{(j+1)T-1} I_d \\ &= \lambda^k P(0)^{-1} + mT\lambda^{T-1} \frac{1 - \lambda^{TB_k}}{1 - \lambda^T} I_d. \end{aligned} \tag{C.4}$$

The inverse eigenvalue bound gives (5.20). Finally, for any λ close to one, $1 - \lambda^T = (1-\lambda)(1 + \lambda + \cdots + \lambda^{T-1})$ is comparable with $T(1-\lambda)$, while λ^{T-1} remains bounded away from zero. Hence, after enough excited blocks have accumulated, the steady inverse information norm is of order $1 - \lambda$. □

Proof of Lemma 5.1. For Huber's score, $s_\delta^H(e) = e$ when $|e| \le \delta$ and $s_\delta^H(e) = \delta\,\mathrm{sgn}(e)$ when $|e| > \delta$. Thus $\left|s_\delta^H(e)\right| \le \delta$ for all e. For the MCC score, set $r = |e|/\sigma \ge 0$. Then

$$\left|s_\sigma^C(e)\right| = \sigma r \exp(-r^2/2). \tag{C.5}$$

The scalar function $f(r) = r\exp(-r^2/2)$ satisfies $f'(r) = \exp(-r^2/2)(1-r^2)$, so its maximum on $[0,\infty)$ occurs at $r = 1$, with the value $\exp(-1/2)$. Hence $\left|s_\sigma^C(e)\right| \le \sigma\exp(-1/2)$. □

Proof of Theorem 5.3 (relocated from Section 5). Fix the time instant k. The claim is pathwise: No distributional assumption on the innovation is required. Condition on the past information and on the current regressor, so that $P(k-1)$ and $\psi(k)$ are fixed objects satisfying the stated bounds. Write

$$e = \varepsilon(k), \qquad P = P(k-1), \qquad \psi = \psi(k), \qquad z = \psi^\top P\psi. \tag{C.6}$$

Since the covariance recursion preserves positive definiteness, $P \succ 0$ and therefore $z \ge 0$. For both robust choices used in the paper, $\omega(e) \ge 0$. Hence, the scalar denominator

$$d(e) = \lambda + \omega(e)z \tag{C.7}$$

is strictly positive and satisfies the uniform lower bound

$$d(e) \ge \lambda. \tag{C.8}$$

This elementary inequality is the place where the forgetting factor enters the proof. It also shows that the bound does not depend on any cancellation between the numerator and the denominator.

The parameter increment produced by (4.4)–(4.5) is

$$\begin{aligned} \Delta\hat{\theta}(k) &= K(k)e \\ &= \frac{\omega(e)P\psi}{\lambda + \omega(e)\psi^\top P\psi}\, e. \end{aligned} \tag{C.9}$$

By the definition of the IRLS weight, $\omega(e)e = s(e)$, with the convention $\omega(0) = 1$ and $s(0) = 0$. Thus the same increment can be written in the score form

$$\Delta\hat{\theta}(k) = \frac{P\psi s(e)}{\lambda + \omega(e)\psi^\top P\psi}. \tag{C.10}$$

Taking the Euclidean norms and using (C.8) gives

$$\begin{aligned} \left\|\Delta\hat{\theta}(k)\right\| &\le \frac{\|P\psi\|}{\lambda}\,|s(e)| \\ &\le \frac{\|P\|\,\|\psi\|}{\lambda}\,|s(e)| \\ &\le \frac{\bar{P}\bar{\psi}}{\lambda}\,|s(e)|\,. \end{aligned} \tag{C.11}$$

It remains only to insert the explicit score bound corresponding to the chosen loss.

For the Huber score, two cases exhaust all possibilities. If $|e| \le \delta$, then $s_\delta^H(e) = e$ and $|s_\delta^H(e)| \le \delta$. If $|e| > \delta$, then $s_\delta^H(e) = \delta \operatorname{sgn}(e)$ and again $|s_\delta^H(e)| = \delta$. Substitution in (C.11) yields

$$\left\|\Delta\hat{\theta}(k)\right\| \le \frac{\bar{P}\bar{\psi}}{\lambda}\,\delta. \tag{C.12}$$

Moreover, along a sequence of vertical innovations with $|e| \to \infty$ while P and ψ remain fixed, the Huber weight satisfies $\omega_\delta^H(e) = \delta/|e| \to 0$ and the denominator tends to λ. The update therefore remains bounded and approaches the one-sided limits

$$\Delta\hat{\theta}(k) \to \pm\frac{P\psi\delta}{\lambda}, \tag{C.13}$$

where the sign is the sign of e. Thus the Huber update clips the innovation contribution at a finite level.

For the MCC score, we have

$$|s_\sigma^C(e)| = |e| \exp\!\left(-\frac{e^2}{2\sigma^2}\right). \tag{C.14}$$

Set $r = |e|/\sigma$. The scalar function $r\exp(-r^2/2)$ is maximized at $r = 1$, so

$$|s_\sigma^C(e)| \le \sigma \exp(-1/2). \tag{C.15}$$

Equation (C.11) gives

$$\left\|\Delta\hat{\theta}(k)\right\| \le \frac{\bar{P}\bar{\psi}}{\lambda}\,\sigma\exp(-1/2). \tag{C.16}$$

The redescending nature of the MCC approach gives an even stronger limiting statement: As $|e| \to \infty$, both $s_\sigma^C(e)$ and $\omega_\sigma^C(e)$ tend to zero, while the denominator tends to λ. Consequently, $\Delta\hat{\theta}(k) \to 0$ for a fixed bounded regressor and covariance.

Combining the two cases gives (5.23) with $S = \delta$ for Huber RS-FRELS and $S = \sigma\exp(-1/2)$ for MCC RS-FRELS. The proof is uniform in the magnitude of the single innovation e; in particular, the right-hand side is not a function of $|e|$.

The final statement follows by repeating the same calculation for ordinary FRELS. In that case, $s(e) = e$ and $\omega(e) = 1$, so

$$\Delta\hat{\theta}_{\mathrm{LS}}(k) = \frac{P\psi}{\lambda + \psi^\top P\psi}\,e. \tag{C.17}$$

If $P\psi \neq 0$, the vector multiplying e is fixed and nonzero for the considered sample. Hence $\left\|\Delta\hat{\theta}_{\mathrm{LS}}(k)\right\|$ grows linearly as $|e| \to \infty$. This establishes the claimed contrast between bounded score RS-FRELS and ordinary quadratic FRELS for a vertical outlier. □

Proof of Theorem 5.4 (relocated from Section 5). The proof separates the Fisher consistency from local quadratic growth and uniqueness. The first part identifies the population estimating equation. For any candidate parameter θ, write

$$e_\theta = y - \theta^\top\psi = \xi - (\theta - \theta^\star)^\top\psi. \tag{C.18}$$

The Huber score is bounded by δ and is Lipschitz continuous with a Lipschitz constant of one. Since $\mathbb{E}\,\|\psi\|^2 < \infty$, particularly $\mathbb{E}\,\|\psi\| < \infty$, dominated convergence justifies differentiating the population risk in every direction. For any direction $a \in \mathbb{R}^d$, we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} Q(\theta + ta)\Big|_{t=0} &= \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}t} \rho_\delta^H(e_\theta - ta^\top \psi)\Big|_{t=0} \right] \\ &= -\mathbb{E}\left[(a^\top \psi) s_\delta^H(e_\theta) \right]. \end{aligned} \tag{C.19}$$

Thus the gradient is

$$\nabla Q(\theta) = -\mathbb{E}\left[\psi s_\delta^H(e_\theta) \right]. \tag{C.20}$$

At the true finite-memory regression parameter, $e_{\theta^\star} = \xi$. Assumption 5.2 gives

$$\nabla Q(\theta^\star) = -\mathbb{E}[\psi s_\delta^H(\xi)] = 0. \tag{C.21}$$

This proves Fisher consistency in the estimating equation sense.

Because ρ_δ^H is convex and $\theta \mapsto y - \theta^\top \psi$ is affine, the function $\theta \mapsto \rho_\delta^H(y - \theta^\top \psi)$ is convex for each data realization. Taking expectations preserves convexity. Therefore, for every θ, the first-order convexity inequality gives

$$Q(\theta) \geq Q(\theta^\star) + \nabla Q(\theta^\star)^\top (\theta - \theta^\star) = Q(\theta^\star), \tag{C.22}$$

so $\theta^\star$ is a global minimizer of the Huber population risk. This part uses only the score orthogonality assumption; it does not require the curvature condition.

We now prove local quadratic growth, which implies local uniqueness. Let $h \neq 0$ with $\|h\| \leq r$, set $v = h/\|h\|$, and define the one-dimensional restriction

$$q_v(t) = Q(\theta^\star + tv), \qquad |t| \leq r. \tag{C.23}$$

Using (C.20), we have

$$q_v'(t) = -\mathbb{E}\left[(\psi^\top v) s_\delta^H(\xi - t\psi^\top v) \right] \tag{C.24}$$

for all t at which the ordinary derivative is considered. The Huber score is absolutely continuous and its derivative with respect to its scalar argument is $\mathbf{1}_{\{|x|<\delta\}}$ except at the two threshold points $x = \pm\delta$. Hence, in the almost-everywhere sense in t, we have

$$q_v''(t) = \mathbb{E}\left[(\psi^\top v)^2 \mathbf{1}_{\{|\xi - t\psi^\top v| < \delta\}} \right]. \tag{C.25}$$

The threshold nondifferentiability causes no difficulty because the integral representation below uses the absolutely continuous derivative of the Huber score; equivalently, (C.25) may be interpreted through the generalized Hessian.

By Assumption 5.3, $q_v''(t) \geq c_H$ for almost every $t \in [-r, r]$. For $0 \leq t \leq r$, the fundamental theorem of calculus yields

$$\begin{aligned} q_v(t) - q_v(0) - t q_v'(0) &= \int_0^t (t - \tau) q_v''(\tau)\, \mathrm{d}\tau \\ &\geq c_H \int_0^t (t - \tau)\, \mathrm{d}\tau = \frac{c_H}{2} t^2. \end{aligned} \tag{C.26}$$

For $-r \le t \le 0$, the same identity integrated over $[t, 0]$, i.e., over the correctly ordered interval when $t < 0$, gives the same lower bound with t^2. Since $q_v'(0) = v^\top \nabla Q(\theta^\star) = 0$, we obtain

$$q_v(t) \ge q_v(0) + \frac{c_H}{2} t^2, \qquad |t| \le r. \tag{C.27}$$

Taking $t = \|h\|$ gives exactly (5.28).

The lower bound (5.28) is strict for every nonzero h with $\|h\| \le r$. Thus no parameter other than $\theta^\star$ can have the same value of Q inside the local ball, and $\theta^\star$ is locally unique. Notice that the conclusion established here is radial quadratic growth around $\theta^\star$. It is sufficient for local uniqueness and for the local mean field arguments used later; it is not stated as full pairwise strong convexity on the entire ball, which would require a stronger uniform Hessian condition between arbitrary pairs of points in the ball. □

Proof of Theorem 5.6 (relocated from Section 5). Define the Huber population risk $Q(\theta) = \mathbb{E}[\rho_\delta^H(y - \theta^\top \psi)]$, which is finite for $\theta \in \Theta$ because the almost-sure regressor bound $\sup_k \|\psi(k)\| < \infty$ in the preamble of Assumption 5.5 guarantees finite moments along the trajectory, and ρ_δ^H has at most linear growth at infinity. The Huber loss is convex and $\theta \mapsto y - \theta^\top \psi$ is affine, so Q is convex on $\mathbb{R}^d$. By the same dominated-convergence calculation that yielded (C.20), we have

$$\nabla Q(\theta) = -\mathbb{E}\Big[\psi s_\delta^H(y - \theta^\top \psi)\Big] = -h(\theta).$$

By (5.48) evaluated at $\theta = \theta^\star$, $\nabla Q(\theta^\star) = 0$, so $\theta^\star$ is a global minimizer of Q on $\mathbb{R}^d$.

Global radial monotonicity on Θ. Fix any $\theta \in \Theta \setminus \{\theta^\star\}$. Set $t = \left\|\theta - \theta^\star\right\|$, $v = (\theta - \theta^\star)/t$, and consider the one-dimensional restriction $q_v(\tau) = Q(\theta^\star + \tau v)$ for $\tau \in [0, t]$. The same generalized-Hessian calculation as in the proof of Theorem 5.4 (extended from a local ball to the whole segment $[0, t]$, which lies in Θ by convexity) gives

$$q_v''(\tau) = \mathbb{E}\Big[(\psi^\top v)^2 \mathbf{1}_{\{|\xi - \tau\psi^\top v| < \delta\}}\Big] \ge c_H, \qquad \text{for a.e. } \tau \in [0, t],$$

where (5.49) is applied at $\theta = \theta^\star + \tau v \in \Theta$ in unconditional form (taking total expectation over $\mathcal{F}_{k-1}$). Since $q_v'(0) = -v^\top h(\theta^\star) = 0$ by (5.48), integration from 0 to t yields

$$(\theta - \theta^\star)^\top \nabla Q(\theta) = t\, q_v'(t) = t \int_0^t q_v''(\tau)\, \mathrm{d}\tau \ge c_H t^2 = c_H \left\|\theta - \theta^\star\right\|^2.$$

Equivalently,

$$(\theta - \theta^\star)^\top h(\theta) \le -c_H \left\|\theta - \theta^\star\right\|^2, \qquad \theta \in \Theta. \tag{C.28}$$

This is the key strengthening of (5.37): It holds on all of Θ, not only on a neighborhood of $\theta^\star$.

Lyapunov contraction with the gain-adapted metric. Choose the Lyapunov function

$$V_k = (\hat{\theta}(k) - \theta^\star)^\top \bar{R}\, (\hat{\theta}(k) - \theta^\star).$$

Throughout this proof two norms are used, and their relationship is fixed once here. The Euclidean norm $\|\cdot\|$ is used for the mean field inequality (C.28) and for the nonexpansiveness of the projection; the $\bar{R}$-weighted quadratic V_k is used as the Lyapunov function because the limiting gain in (5.43) is

$\bar{R}^{-1}$, and pairing the update with the metric $\bar{R}$ cancels the gain matrix in the first-order drift term. The two are equivalent with the explicit constants

$$\lambda_{\min}(\bar{R}) \left\|\hat{\theta}(k) - \theta^\star\right\|^2 \;\le\; V_k \;\le\; \lambda_{\max}(\bar{R}) \left\|\hat{\theta}(k) - \theta^\star\right\|^2,$$

where $\lambda_{\min}(\bar{R}) > 0$ by (GH3); equivalently, V_k is also comparable with the $\bar{R}^{-1}$-weighted norm through the same eigenvalue bounds applied to $\bar{R}^{-1}$, whose extreme eigenvalues are the reciprocals $1/\lambda_{\max}(\bar{R})$ and $1/\lambda_{\min}(\bar{R})$. Every passage between the Euclidean and the weighted quantities below costs at most the condition number $\kappa(\bar{R}) = \lambda_{\max}(\bar{R})/\lambda_{\min}(\bar{R})$, which is finite and constant, and is absorbed into the drift constants. Substituting the SA expansion (5.43) and using $\bar{R}\bar{R}^{-1} = I$ gives, for the unprojected iterate $\hat{\theta}^+(k) = \hat{\theta}(k-1) + (1/k)\bar{R}^{-1}\psi(k)s(\varepsilon(k)) + \zeta_k$,

$$\begin{aligned} V_k^+ = V_{k-1} &+ \frac{2}{k}(\hat{\theta}(k-1) - \theta^\star)^\top \psi(k) s\big(y(k) - \hat{\theta}(k-1)^\top \psi(k)\big) \\ &+ N_k + \varrho_k, \end{aligned}$$

where N_k collects the quadratic and cross terms in the unbiased martingale increment and ϱ_k collects the contributions of ζ_k (the symbol ϱ_k is used to avoid a clash with the information matrix $R(k)$).

Taking the conditional expectations and using the decomposition $\psi s = h + M$ from Assumption 5.5 (GH4), the conditional mean of the first-order term becomes

$$\frac{2}{k}(\hat{\theta}(k-1) - \theta^\star)^\top h(\hat{\theta}(k-1)) \;\le\; -\frac{2c_H}{k} \left\|\hat{\theta}(k-1) - \theta^\star\right\|^2$$

by (C.28). The quadratic part of N_k contributes at most C/k^2 uniformly on Θ because ψ is bounded, $|s_\delta^H| \le \delta$, and the conditional second moment of M_k is bounded by C_M. The remainder term satisfies $\mathbb{E}[\|\varrho_k\| \mid \mathcal{F}_{k-1}] \le (\epsilon/k)\left\|\hat{\theta}(k-1) - \theta^\star\right\|^2 + C_\epsilon/k^2$ by (GH4). Choosing $\epsilon < c_H\lambda_{\min}(\bar{R})$ absorbs the perturbation into the negative drift. Therefore, for all values of k larger than a finite random time,

$$\mathbb{E}[V_k^+ \mid \mathcal{F}_{k-1}] \le \left(1 - \frac{c_H'}{k}\right) V_{k-1} + \frac{C'}{k^2}, \tag{C.29}$$

with $c_H' = c_H\lambda_{\min}(\bar{R})/\lambda_{\max}(\bar{R}) > 0$ and a deterministic constant C'.

Projection step and Robbins–Siegmund supermartingale. Euclidean projection onto Θ is nonexpansive in the standard inner product, and hence $\left\|\Pi_\Theta(\hat{\theta}^+(k)) - \theta^\star\right\| \le \left\|\hat{\theta}^+(k) - \theta^\star\right\|$ because $\theta^\star \in \Theta$. The $\bar{R}$-weighted norm differs from the Euclidean norm only by the constant condition number of $\bar{R}$, so the same inequality holds for V_k up to a multiplicative constant that is absorbed into c_H' (after recomputing the geometric coefficient). Substituting into (C.29) preserves the supermartingale-type inequality $\mathbb{E}[V_k \mid \mathcal{F}_{k-1}] \le (1 - c_H''/k)V_{k-1} + C''/k^2$.

The Robbins–Siegmund supermartingale theorem yields $V_k \to V_\infty$ almost surely and $\sum_k k^{-1}V_{k-1} < \infty$. Since $\sum_k k^{-1} = \infty$, the only possible non-negative finite limit consistent with this summability is $V_\infty = 0$. Equivalence of norms gives $\left\|\hat{\theta}(k) - \theta^\star\right\| \to 0$ almost surely.

Globality is automatic: (C.28) holds for every $\theta \in \Theta$, so no entrance time argument is needed and no a priori localization is assumed. The result depends only on the iterate staying in Θ, which is enforced by projection. □

Proof of Lemma 5.2 (relocated from Section 5). Set $\alpha = 1 - \lambda$. The no-drift update produces $\tilde{\theta}^+(k)$, and the parameter drift changes the reference parameter by

$$d(k) = \theta^\star(k) - \theta^\star(k-1), \qquad \|d(k)\| \le q. \tag{C.30}$$

Thus

$$\tilde{\theta}(k) = \tilde{\theta}^+(k) - d(k). \tag{C.31}$$

For any $\eta > 0$, Young's inequality gives

$$\|x - d\|^2 \le (1+\eta)\,\|x\|^2 + (1+\eta^{-1})\,\|d\|^2\,. \tag{C.32}$$

Apply (C.32) with $x = \tilde{\theta}^+(k)$, take the conditional expectations, and use (5.52). Choosing $\eta = a\alpha/2$ yields, for any α small enough,

$$\mathbb{E}[\left\|\tilde{\theta}(k)\right\|^2] \le \left(1 - \frac{a\alpha}{2}\right)\mathbb{E}[\left\|\tilde{\theta}(k-1)\right\|^2] + C_b\alpha^2 S^2 + C_q\frac{q^2}{\alpha}, \tag{C.33}$$

for the finite constants C_b and C_q. Iterating the scalar inequality (C.33) gives

$$\mathbb{E}[\left\|\tilde{\theta}(k)\right\|^2] \le \left(1 - \frac{a\alpha}{2}\right)^{k-k_0}\mathbb{E}[\left\|\tilde{\theta}(k_0)\right\|^2] + \sum_{j=k_0+1}^{k}\left(1 - \frac{a\alpha}{2}\right)^{k-j}\left(C_b\alpha^2 S^2 + C_q\frac{q^2}{\alpha}\right). \tag{C.34}$$

The geometric sum is bounded by $2/(a\alpha)$. Taking the upper limit gives

$$\limsup_{k\to\infty}\mathbb{E}\left\|\tilde{\theta}(k)\right\|^2 \le \frac{2C_b}{a}\alpha S^2 + \frac{2C_q}{a}\frac{q^2}{\alpha^2}, \tag{C.35}$$

which is (5.53). The result is conservative because it uses only the worst-case score bound and a local contraction inequality. □

AIMS Press