entropy

MDPI

Article

Pairwise-Agreement Symbolic Entropy Saturates by Construction: Implications for Delay Selection

Zhenmin Zhao [1], Jianghong Li [2,3,*] and Jiahang Cui [2]

[1] National Elite Institute of Engineering, Northwestern Polytechnical University, Xi'an 710072, China; 2022160894@mail.nwpu.edu.cn
[2] School of Power and Energy, Northwestern Polytechnical University, Xi'an 710072, China; cui275228896@mail.nwpu.edu.cn
[3] Science and Technology on Altitude Simulation Laboratory, Northwestern Polytechnical University, Xi'an 710072, China
* Correspondence: jhli@nwpu.edu.cn

Abstract

A family of methods selects the phase-space reconstruction delay τ as the point where a pairwise-agreement symbolic entropy $H(m, \tau)$ stops growing, on the premise that this onset marks the decorrelation delay. We show that the saturation is built into the construction rather than discovered in the data, so that the premise fails for the pairwise-agreement symbol. The agreement symbol is invariant under complementation of the thresholded sign vector, so the symbol space has exactly 2^{m-1} elements, the map from complement orbits to attainable agreement patterns is a bijection, and $H(m, \tau) \leq m - 1$ for every signal. The plateau is fixed by the symbolization and the threshold-crossing probability alone: for independent Bernoulli(p) signs it equals $H_\infty(m, p) \leq m - 1$, with equality if and only if $p = 1/2$, and an independent control confirms this to 0.005 bits for $m = 2, \ldots, 8$. Saturation therefore occurs for any input, including i.i.d. noise, and the plateau carries no dynamical information; what remains is the rate of approach, which depends on the marginal and not only on the dependence structure. Translating a record by half its standard deviation leaves the mutual-information curve and its first minimum unchanged to machine precision while moving the criterion's delay by 27% of its own value. We test the consequences on five chaotic systems at six noise levels, each verified to have a positive largest Lyapunov exponent under two integration steps and two accumulation lengths, and on two machinery datasets. On clean data our low-pass-filtered variant returns delays 1.08–4.69 times the average-mutual-information reference and, under the as-used protocol, carries the largest bias of the six; the filter raises the delay bias by 71% (31.6 $\rightarrow$ 54.1 samples) rather than reducing it; and the threshold that reproduces the reference is absent for one of five systems on clean data and for two at 10 dB. Under a harmonized protocol that gives all six estimators one search window, threshold and dimension range, the ordering changes and the variant is no longer the worst, so the ranking is protocol-dependent. On MFPT bearing data the comparison inverts: the most biased estimator under the as-used protocol classifies best (0.917 from repeated grouped cross-validation, 95% cluster interval over records $[0.798, 0.987]$) and the one closest to the reference worst (0.552, $[0.408, 0.697]$), while replacing the per-record (τ, m) by a single pooled pair drops the winner to 0.743, so about a fifth of its score is the record-specific calibration acting as a record fingerprint rather than any property of the delay. All conclusions are bounded to the five systems, six noise levels and two datasets examined. We advise against reading a decorrelation delay off the saturation onset of a pairwise-agreement symbolic entropy; where only classification accuracy is sought it may serve as a black-box feature heuristic, but its output should not be reported as a delay.

Check for updates

Academic Editor: Yong Deng

Received: 29 September 2026
Revised: 10 November 2026
Accepted: 11 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: phase space reconstruction; time delay; symbolic entropy; saturation; mutual information; recurrence quantification analysis; bearing fault diagnosis; negative results

1. Introduction

Phase space reconstruction is the standard route from a scalar measurement to the dynamics that produced it [1–3]. The method requires two parameters: a delay τ and an embedding dimension m. Both are chosen by the analyst, and both materially affect everything computed downstream: recurrence quantification, Lyapunov spectra, entropy measures, and the classifier features built from them [4]. Selecting them badly degrades all of these. Selecting them well is an old problem with a settled core: delay from the first minimum of the average mutual information (AMI) [5] or from the autocorrelation, dimension from the false nearest neighbors (FNN) criterion [6]. This general challenge, that of inferring unknown model parameters purely from measured time-series observations, also appears broadly across the mechanical-engineering community, for instance, within Bayesian probabilistic model-updating frameworks that infer model posteriors from experimental response data [7,8]. However, for nonlinear dynamical time-series, phase-space reconstruction has developed its own set of dedicated heuristic criteria.

A more recent family of methods abandons these criteria and reads both parameters off a symbolic entropy curve $H(m, \tau)$ [9]. The curve is obtained by thresholding the signal at ε, encoding the embedding vector by which of its components exceed ε, and computing the Shannon entropy of the resulting symbol distribution as a function of the delay. The proposal is that H grows while the components are still dependent, stops growing once they decorrelate, and that the transition is a robust and noise-tolerant delay estimator [9]. The C-C method [10] is the well-known precursor of this idea; multiscale and fuzzy-entropy variants followed [11–13]. We also construct a filtered variant of that criterion. The Low-Pass Filtered Symbolic Entropy (LPF-SEM) keeps the symbolization and the selection logic of the published criterion and adds a signal-processing front end: the entropy curve is low-pass filtered with a zero-phase Butterworth filter whose cutoff is chosen from the curve's own spectrum, the delay is placed where the normalized growth rate of the filtered curve falls below a threshold, and the dimension is inferred from the spacing of the filtered curve's minima. The intent is the natural one, to suppress the spurious local maxima that trap the unfiltered rule on clean data while leaving the noise tolerance of the criterion intact. This variant is our own construction rather than a published method, and Section 3 together with Appendix D specify it completely, so that it can be reproduced or replaced.

The variant does not work. Its delays are no closer to the reference than those of the unfiltered criterion it was built to repair, and asking why produced the exact result below: the quantity every criterion in this family thresholds is not a function of the dependence structure alone, so no choice of filter recovers the decorrelation delay.

1.1. What We Establish, and What We Do Not

The pairwise-agreement symbolization is invariant under complementation of the thresholded sign vector $\mathbf{s} \mapsto \mathbf{1} - \mathbf{s}$: the agreement indicators record which components share a side of the threshold, not which side. The map from sign vector to symbol is two-to-one, the symbol space has exactly 2^{m-1} elements rather than 2^m, the map from complement orbits to attainable agreement patterns is a bijection, and the entropy cannot exceed $m - 1$ bits (Lemma 1, Theorem 1). The plateau value is stronger than a ceiling: for independent components with threshold-crossing probability p it equals $H_\infty(m, p)$, a function of the symbolization and the marginal alone, which equals $m - 1$ only at $p = 1/2$

(Proposition 1). Saturation is thus guaranteed a priori for every signal, and the event "H has stopped growing" carries no information about the dynamics. Our argument is therefore in three layers, and we keep them apart throughout. The theorem is unconditional and covers the pairwise-agreement symbol only. The plateau result is a statement about symbolization and marginal. The claim that a delay read off the approach to that plateau is not the decorrelation delay is empirical, established for the five systems and six noise levels below; it is not a corollary of the theorem, and we do not present it as one. A normalized functional of H could in principle still carry dependence information even though the level of H is fixed a priori, which is why Section 2.5 states explicitly what the theorem does and does not imply and adds a structural argument that does not rest on the experiments: the mutual-information curve is invariant under translation of the record while $H(m, \tau)$ is not. What is closed off is the reading of a decorrelation delay off the saturation of a pairwise-agreement entropy; the theorem covers the two methods that actually compute that symbol, namely SEM and LPF-SEM, and Section 3.2 repositions the correlation-integral precursor and the multiscale and fuzzy variants as benchmarks rather than as members of the covered class. Nothing here bears on symbolic entropies of other design: permutation entropy [14], whose alphabet is the set of ordinal patterns, has no complement degeneracy, and no result below applies to it.

1.2. What We Measure

To test the family rather than a straw man, we rebuilt the pipeline from scratch on genuine dynamics: three autonomous ODEs, one periodically forced oscillator and one delay differential equation, each verified chaotic through a positive largest Lyapunov exponent computed with renormalization, under two integration steps and two accumulation lengths. All six estimators were run over five systems $\times$ six signal-to-noise ratios $\times$ twenty trials, with a component ablation, a filter-parameter sweep, a threshold-transferability study, a harmonized-protocol control and a translation-invariance check added. On clean data LPF-SEM returns delays 1.08–4.69$\times$ the reference, so the discrepancy is not a constant offset that could be calibrated away, and under the as-used protocol, in which each published estimator keeps the configuration of its source and our filtered variant keeps that of Section 3, it is the largest bias of the six at every noise level. That ranking is protocol-dependent; however, under a harmonized protocol it is no longer the worst, and Section 4.11 reports the ordering. The low-pass filter increases the mean delay bias from 31.6 to 54.1 samples, and adding the curvature criterion leaves it at 44.7, worse than the unfiltered criterion it was meant to improve. Finally, a negative result is only useful if it bounds the damage, so we carried the study through to the downstream task: RQA-feature fault classification on the MFPT bearing set and a rotor set, with repeated grouped cross-validation so that windows from one record cannot straddle the train/test split. On MFPT, the accuracy ordering is the reverse of the delay ordering: LPF-SEM classifies best (0.917) while C-C, the closest to the reference, is worst (0.552). A control that replaces the per-record calibration by a single pooled pair removes most of that advantage, so the downstream evidence sustaining this family was never evidence about the delay. On the rotor set the effect vanishes.

1.3. Structure

Section 2 proves the plateau result and derives a noise-inflation bound. Section 3 defines the estimators, the reference criteria and the scope of each. Section 4 reports the numerical study on chaotic systems. Section 5 applies the estimators to real machinery data. Section 6 discusses why the criteria fail, and Section 7 collects the advice with its scope. Appendices A–E give the Lyapunov protocol, the harmonized-protocol control, the

classification statistics, the implementation details and the data description. Tables A1–A11 collect the corresponding tabulated material.

2. Symbolization, Ceiling, and Saturation

2.1. Pairwise-Agreement Symbolization

Let $x = \{x_t\}_{t=0}^{N-1}$ be a scalar time series, fixed $m \geq 2$ and $\tau \in \mathbb{N}^+$, and take a threshold $\varepsilon > 0$. Write

$$\sigma(u) = \mathbf{1}[u > \varepsilon] \in \{0,1\}, \qquad u \in \mathbb{R}. \tag{1}$$

The embedding vector at time t is $\mathbf{x}_t^{(m)} = (x_t, x_{t+\tau}, \ldots, x_{t+(m-1)\tau})$, and its sign vector is

$$\mathbf{s}_t = \big(\sigma(x_{t+j\tau})\big)_{j=0}^{m-1} \in \{0,1\}^m. \tag{2}$$

The symbolization used by the methods studied here does not record $\mathbf{s}_t$ itself. It records, for every pair of components, whether the two agree:

$$A_t = \big([\sigma(x_{t+j\tau}) = \sigma(x_{t+k\tau})]\big)_{0 \leq j < k \leq m-1}, \tag{3}$$

so that $A_t \in \{0,1\}^{\binom{m}{2}}$. The symbolic entropy is the Shannon entropy of the empirical distribution of A_t over the $T = N - (m-1)\tau$ admissible positions, with $\hat{p}(A) = \frac{1}{T}\#\{t : A_t = A\}$:

$$H(m,\tau) = -\sum_{A} \hat{p}(A) \log_2 \hat{p}(A). \tag{4}$$

This is the quantity plotted as the entropy curve and, in the saturation family, used to select τ. The threshold is conventionally tied to the data through $\varepsilon = \alpha\, \mathrm{sd}(x)$; the methods below fix α and treat it as a design constant rather than a parameter to be estimated.

2.2. Complement Invariance and the Effective Alphabet

Lemma 1 (Complement invariance and the orbit–pattern bijection)**.** *Write $b_j = \sigma(x_{t+j\tau})$ for the j-th component of $\mathbf{s}_t$. For every t and every $j < k$,*

$$[b_j = b_k] = [(1 - b_j) = (1 - b_k)], \tag{5}$$

so A_t is invariant under the complementation $\mathbf{s}_t \mapsto \mathbf{1} - \mathbf{s}_t$, and the map $\mathbf{s} \mapsto A(\mathbf{s})$ is constant on the two-element orbits of that action on $\{0,1\}^m$. The induced map is a bijection between the orbits and the attainable agreement patterns, hence the image of A has exactly 2^{m-1} elements.

Proof. Both sides of the display assert that two bits agree; replacing both by their complements preserves agreement, which gives the identity and the invariance. The complement action is a fixed-point-free involution on $\{0,1\}^m$ (no $\mathbf{s}$ equals its complement), so it partitions the 2^m sign vectors into $2^m/2 = 2^{m-1}$ orbits and A factors through the orbit index.

For injectivity, read the agreement pattern as an equivalence relation on the coordinate set $\{1, \ldots, m\}$: the bit for the pair (j,k) says whether j and k lie on the same side of ε. That relation has at most two non-empty classes, the coordinates above ε and those below, one of them empty when every component falls on the same side, and the pair of classes is unordered because exchanging them complements $\mathbf{s}$ and changes no agreement bit. Conversely, any bipartition $\{S, S^c\}$ of $\{1, \ldots, m\}$ produces the pattern whose agreeing pairs are the pairs inside S and inside S^c, so every bipartition is attained and distinct bipartitions give distinct patterns. There are 2^{m-1} bipartitions, which matches the number of orbits; the correspondence is therefore a bijection for every $m \geq 2$. □

For $m = 3$ the eight sign vectors fall into four orbits and the four attainable patterns can be read off directly: 000 and 111 give the all-agree pattern, and 001/110, 010/101 and 100/011 give the three one-against-two patterns, one for each choice of the isolated coordinate. Nothing is lost in passing to the symbol: the alphabet is exactly 2^{m-1}, including at $m = 2$, where the two orbits are the all-agree and the all-disagree patterns.

Geometrically, replacing $\sigma(u) = \mathbf{1}[u > \varepsilon]$ by $\mathbf{1}[u < \varepsilon]$, equivalently $\varepsilon \to -\varepsilon$, flips every component bit without changing A: reflecting the signal about the threshold leaves the symbol sequence, and hence H, identically unchanged, so the symbolization is blind to which side of the threshold a component falls on.

Theorem 1 (Entropy ceiling). *For all $m \geq 2$, $\tau \in \mathbb{N}^+$ and any x and $\varepsilon > 0$,*

$$0 \leq H(m, \tau) \leq \log_2 2^{m-1} = m - 1. \tag{6}$$

Proof. A_t is a function of $\mathbf{s}_t$, which takes at most 2^{m-1} distinct values on orbits by Lemma 1. A random variable supported on k points has entropy at most $\log_2 k$, attained only when the distribution is uniform on the support. Hence $H(m, \tau) \leq \log_2 2^{m-1} = m - 1$. Non-negativity is immediate. □

The bound is tight, and the two things that a finite record can lose must not be confused. The symbol space has exactly 2^{m-1} elements; the empirical support on a record of length N may be smaller, because some of the 2^{m-1} symbols need never occur. A record therefore shows a measured entropy below $\log_2$ of the alphabet size for two separate reasons, unoccupied symbols and residual dependence between the components, and Section 4.2 separates them quantitatively. Attainment of the bound is a statement about the independent case and is treated in Proposition 1 and verified numerically there.

Remark 1 (Scope of the ceiling). *Theorem 1 and Proposition 1 are statements about the pairwise-agreement symbol A_t of (3), and only about it. Their engine is the complement degeneracy of that symbol; any symbolic construction whose alphabet is not invariant under complementation is untouched by the argument, and the plateau $H_\infty(m, p)$ does not apply to it. Permutation or ordinal-pattern entropy [14] is the clearest example of such a construction, and the class is wider than that. What is excluded is the use of the saturation of a pairwise-agreement entropy as a delay estimator, not symbolic entropy as a tool. Section 3.2 states which of the estimators compared below fall inside this class and which do not.*

2.3. Saturation for Mixing Processes

Proposition 1 (Plateau value). *Suppose that for some τ_0 the m components $(x_{t+j\tau})_{j=0}^{m-1}$ are mutually independent and that $\mathbb{P}(x_{t+j\tau} > \varepsilon) = p$ for every j, with $0 < p < 1$. Then, each complement orbit O carries probability*

$$P(O) = \sum_{\mathbf{s} \in O} p^{|\mathbf{s}|}(1-p)^{m-|\mathbf{s}|}, \tag{7}$$

where $|\mathbf{s}|$ is the number of ones in $\mathbf{s}$, the agreement symbol is distributed over the orbits with those probabilities, and

$$H(m, \tau) = H_\infty(m, p) := -\sum_{O} P(O) \log_2 P(O) \;\leq\; m - 1 \tag{8}$$

with equality if and only if $p = 1/2$. Consequently, if the process is mixing and ε is the median of the marginal so that $p = 1/2$, then $H(m, \tau) \to m - 1$ as $\tau \to \infty$; for any other threshold-crossing probability the limit is $H_\infty(m, p) < m - 1$.

Proof. Independence makes the sign vector a product measure, so $\mathbb{P}(\mathbf{s}) = p^{|\mathbf{s}|}(1-p)^{m-|\mathbf{s}|}$; summing over the two sign vectors of each orbit gives (7), and by the bijection of Lemma 1 the symbol distribution is exactly this orbit distribution, so (8) is its Shannon entropy. At $p = 1/2$ every sign vector has probability 2^{-m}; hence, every orbit probability is $2^{-(m-1)}$ and $H = m - 1$. For $p \neq 1/2$ the orbit probabilities are not equal, since the all-agree orbit carries $p^m + (1-p)^m$, which exceeds $2^{-(m-1)}$ whenever $|p - 1/2| > 0$; the distribution is then not uniform on its support and the entropy is strictly smaller. Entropy is maximized by the uniform law on a support of that size, which proves the bound and the equality condition. For a mixing process, the components become asymptotically independent as $\tau \to \infty$ and H converges to the independent plateau computed at the marginal's p. □

Proposition 1 is the crux of our argument, and it is a statement about the plateau, not only about a ceiling. Every signal has an entropy curve that rises and then flattens at $H_\infty(m, p)$, a number fixed by the symbolization and by the marginal threshold-crossing probability and by nothing else. The family locates τ at the onset of that flattening and reads it as the decorrelation delay, but the asymptote being approached is a function of (m, p); information about the dynamics is carried entirely in the rate of approach. The recovered delay is therefore fixed by the shape of H near its plateau, a nonlinear functional of the joint threshold-crossing statistics and of the local expansion rates near the threshold level set, and not by the mutual-information reference it is taken to estimate.

The hypothesis $p = 1/2$ deserves emphasis because it is what the submitted form of this statement assumed, and because the experiments do not satisfy it. Every measurement below fixes $\varepsilon = \alpha\,\mathrm{sd}(x)$ with $\alpha = 0.1$, and the resulting crossing probability is a property of the marginal: on the five systems it ranges from 0.44 to 0.54 (Section 4.2), and for a standard normal marginal it is $1 - \Phi(0.1) = \Phi(-0.1) = 0.460$, with Φ the standard normal cumulative distribution function. So the plateau is strictly below $m - 1$ in every experiment and by an amount that depends on the system, which is exactly the reviewer's objection that the asymptote depends on the data. The objection is correct as stated against the $m - 1$ form of the claim; Proposition 1 absorbs it, because it replaces "$m - 1$" by a value that depends on (m, p) and therefore depends on the marginal rather than on the dynamics. The numerical check reported in Section 4.2 compares (8) with measurement in both regimes: on an independent control, where the hypothesis of Proposition 1 is met up to the marginal, and on the five chaotic systems, where it is not.

Two structural consequences follow from the plateau being a function of (m, p).

Remark 2 (τ–m coupling)**.** *The plateau $H_\infty(m, p)$ depends on m, so the curve whose saturation onset is being located changes with m. Because saturation is located through a threshold on the normalized growth rate $|dH/d\tau|/\max|dH/d\tau|$, changing m both raises the plateau and displaces the whole curve, and $\hat{\tau}$ therefore depends on the trial value of m. The two parameters are not separately identifiable within this family. In the LPF-SEM pipeline the dependence is circular: the dimension is derived from $\hat{\tau}$ through the pseudo-period test, so an m-dependent $\hat{\tau}$ feeds an $\hat{m}$ that justifies the m used to obtain $\hat{\tau}$. We quantify the drift in Section 4.8.*

Remark 3 (Thresholds are not identifiable either)**.** *The criterion interprets "saturation" as $|dH/d\tau| \leq \gamma \max|dH/d\tau|$ after low-pass filtering at cutoff f_c. Neither constant is determined by the data. Since H approaches its plateau monotonically and smoothly for every signal, the first crossing of any such threshold traces out a wide range of delays as the constants vary, with no reason for any particular choice to be the decorrelation delay. Sections 4.6 and 4.7 measure this.*

2.4. Noise Inflation and the Looseness of the Standard Bound

Symbolic entropies are often argued to be noise-robust on the grounds that additive noise only perturbs which side of the threshold a sample falls on. The following bound makes this precise, and the numerical study below shows it is valid but too loose to certify anything at practical noise levels.

Proposition 2 (Noise-inflation bound)**.** *Let x have symbolization entropy H_x and let $y = x + \eta$ with $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$ independent of x. Write q for the probability that a single threshold comparison of x is flipped by η, so that a comparison is preserved with probability $1 - q$. A pair of comparisons is then preserved when both flip or neither does, that is, with probability*

$$\beta \;=\; q^2 + (1-q)^2 \;=\; 1 - 2q(1-q), \tag{9}$$

and it is the complementary quantity $2q(1-q)$ that is the probability of exactly one flip, i.e., of the agreement indicator being inverted. With $h_2(\beta) = -\beta \log_2 \beta - (1-\beta)\log_2(1-\beta)$ the binary entropy function becomes

$$H_y - H_x \;\le\; \binom{m}{2} h_2(\beta). \tag{10}$$

Proof. Noise acts on the pair (j, k) through its two comparisons. An agreement bit is preserved when an even number of the two flips, and because $h_2(\beta) = h_2(1-\beta)$ the numerical value of the per-bit entropy penalty is the same whichever of β and $1-\beta$ is used; the proof below needs only that the bit be inverted with probability $2q(1-q)$, so it is unaffected by which reading of β is written down. Each agreement bit therefore behaves as the output of a binary symmetric channel with crossover probability $2q(1-q)$ applied to the clean bit: a binary symmetric channel cannot decrease conditional entropy, $H(\text{output}) \le H(\text{input}) + h_2(\cdot)$ per bit. Summing over the $\binom{m}{2}$ agreement coordinates uses only subadditivity of entropy, $H(A_1, \ldots, A_K) \le \sum_k H(A_k)$, and never the joint law, so no independence between agreement bits is assumed; the bound is correspondingly loose, as Section 4.10 quantifies. □

Remark 4 (Pair-specific flip probabilities)**.** *A single q presupposes that the m components sit at equal distance from the threshold, which they do not: the distance of $x_{t+j\tau}$ from ε decides how likely noise is to push it across, and those distances differ. Writing q_j, q_k for the flip probabilities of the j-th and k-th comparison, the agreement bit for that pair is inverted with probability*

$$\beta_{jk} = q_j(1-q_k) + q_k(1-q_j), \tag{11}$$

and with $\bar{q} = \frac{1}{2}(q_j + q_k)$ one has $2\bar{q}(1-\bar{q}) - \beta_{jk} = (q_j - q_k)^2/2 \ge 0$, so replacing every β_{jk} by the common $2\bar{q}(1-\bar{q})$ only enlarges the right-hand side of (10). Writing a single q is therefore an implicit assumption that costs validity nothing; the revision states it explicitly rather than leaving it silent.

Remark 5 (Correlation between agreement bits, and why the bound is loose)**.** *The $\binom{m}{2}$ agreement bits are not independent, and (10) does not need them to be. A single sign flip in one component inverts every agreement bit involving that component simultaneously, so the bits are strongly positively correlated and the entropy of the flip pattern is far below $\binom{m}{2} h_2$. Because the proof uses subadditivity, that correlation enters the bound only through the constant $\binom{m}{2}$, which overcounts: this is the structural reason the bound is loose, not an artifact of the argument. Accounting for the correlation exactly would require the joint flip distribution of the m comparisons, which is signal-dependent, and we do not attempt it.*

Remark 6 (A heuristic sharpening, not a bound). *Replacing* $\binom{m}{2}$ *in (10) by* $m-1$*, the number of independent binary degrees of freedom that Theorem 1 identifies, produces a much smaller quantity that we evaluate in Section 4.10. We do not present it as a bound. Passing from* $\binom{m}{2}$ *agreement bits to* $m-1$ *degrees of freedom would require noise to act as an independent channel on each of those degrees of freedom, and Remark 5 says it does not; the constant* $(m-1)$ *is a heuristic observed to track the measurements better, and no result in this paper depends on it.*

2.5. What Theorem 1 Does and Does Not Imply

Theorem 1 establishes a ceiling imposed by the pairwise-agreement symbolization. That ceiling alone does not logically exclude every saturation-based estimator: a normalized functional of $H(m,\tau)$ could in principle carry information about dependence even though the level of H is fixed a priori, and one such functional is discussed in Section 6.2. What the theorem does establish, together with Proposition 1, is that the plateau value is a function of (m,p) alone and is reached by any mixing process, so that the event "H has stopped growing" carries no information about the dynamics. What survives as dynamical content is the rate of approach, and every criterion in the family thresholds that rate; whether any such threshold recovers the decorrelation delay is a question we answer empirically in Sections 4 and 5, not a corollary of Theorem 1.

Two further points sharpen the distinction, and neither rests on the experiments. First, the ceiling is not what the family claims to use: it claims to use the onset, and an onset defined by a threshold on a normalized growth rate is a functional of the whole curve rather than of its limit. Second, and more decisively, the quantity the criterion depends on is not a function of the dependence structure alone. Let $c \in \mathbb{R}$ and $y_t = x_t + c$. Mutual information is invariant under any bijection of the marginals, so

Proposition 3 (Translation non-invariance). *For every* $c \in \mathbb{R}$ *and every* τ*,*

(i) $I(y_t; y_{t+\tau}) = I(x_t; x_{t+\tau})$*, so the AMI curve, its first minimum, and any quantile-based estimator of it are unchanged;*

(ii) $H_y(m,\tau) \neq H_x(m,\tau)$ *in general, because* $\varepsilon = \alpha\,\mathrm{sd}(y) = \alpha\,\mathrm{sd}(x)$ *is unchanged while* $p = \mathbb{P}(y_t > \varepsilon) = F_x(\varepsilon - c)$ *changes, where* $F_x(u) = \mathbb{P}(x_t > u)$ *is the upper-tail function of the marginal, so that* $1 - F_x$ *is its cumulative distribution function, and both the plateau* $H_\infty(m,p)$ *of (8) and the approach to it depend on* p*.*

Consequently $H(m,\tau)$ *is not a function of the dependence structure of the record alone, whereas the reference criterion is, and a delay read off the saturation of* H *cannot be an estimator of a quantity that is.*

Proof. (i) Translation is a bijection of $\mathbb{R}$ and mutual information is invariant under bijective transformations of either marginal applied jointly, so the joint distribution of $(y_t, y_{t+\tau})$ is the image of that of $(x_t, x_{t+\tau})$ under a bijection and both marginal entropies are unchanged.

(ii) $\mathrm{sd}(x+c) = \mathrm{sd}(x)$, so the threshold in absolute units is the same; the mass above it is $F_x(\varepsilon - c)$, which depends on c unless the marginal is degenerate. By Proposition 1 the plateau is then $H_\infty(m, p(c))$, and the orbit probabilities (7) show that the whole curve, not only its limit, is a function of p. Section 4.12 verifies both statements numerically. □

Remark 7 (What the non-invariance depends on). *The result turns on the threshold rule rather than on translation alone. Under the conventional scale rule used throughout this paper,* $\varepsilon = \alpha\,\mathrm{sd}(x)$*, the spread is translation-invariant, so the threshold in absolute units is the same before and after the shift while the mass above it changes; an absolute constant threshold behaves the same way, its position being fixed while the record moves. The non-invariance disappears only under a rule that follows the marginal, for instance a fixed quantile* $\varepsilon = F_x^{-1}(\alpha)$*, for which the tail probability*

is $p = \alpha$ by construction and H is translation-invariant. What the family's convention fixes is the scale of the threshold and not its location, and that asymmetry is what the proposition exploits.

The proposition is the structural statement the reviewer's first comment asks for. It does not say that no normalized functional of H can ever be informative; it says that the quantity every criterion in this family thresholds is sensitive to a transformation that changes no dependence property whatsoever, while the reference it is scored against is exactly invariant to that transformation. The two cannot be functionals of the same object.

3. Estimators Under Test

3.1. Reference Criteria

We take as the reference the two criteria the reconstruction literature has relied on longest, and we are explicit that they are conventions rather than optima.

The first minimum of the average mutual information [5], estimated on 64 equiprobable bins with x_t and $x_{t+\tau}$ cross-tabulated for lags up to 100. We use the first local minimum, and denote it τ^*.

The first embedding dimension at which the fraction of false nearest neighbors [6] drops below 1%, with tolerance parameters $r_{\text{tol}} = 15$ and $a_{\text{tol}} = 2$ on the normalized distances, and candidate dimensions up to 12. We denote it m^*.

Neither criterion is a ground truth, and we do not treat it as one. The first-minimum rule is the prevailing convention of the field and the criterion the methods under test claim to replace, which is what makes it the appropriate reference for this comparison; it is a heuristic, not an exact optimum for reconstruction. Mutual information is a functional of the full joint distribution of $(x_t, x_{t+\tau})$, hence sensitive to nonlinear dependence and invariant under any invertible transformation of either marginal. The autocorrelation, by contrast, is a second-order statistic and is not invariant under monotone transformations. The AMI curve therefore depends on the binning, and it is not guaranteed to have an interior minimum at all; both defects are theirs, not ours, and we quantify the first in Section 4.7 and list both in Section 6.4. Because the reference is a convention, Section 6.4 reports the delay ratios against three further references that involve no mutual information at all.

The FNN tolerances are the conventional defaults of the field rather than parameters tuned to this study: $r_{\text{tol}} = 15$ and $a_{\text{tol}} = 2$ are the values in the source criterion [6], and the only quantity we chose is the 1% stopping fraction.

3.2. The Saturation Family and Its Scope

We implement six estimators. SEM follows the published criterion [9] with its own internal constant. LPF-SEM is our filtered extension of SEM, built as described below and documented rule by rule in Appendix D. The remaining four are benchmarks, positioned in Section 3.2.

1. SEM [9]. The delay is the first local maximum of the unfiltered entropy curve $H(m, \tau)$ evaluated at $m = 6$, with $\varepsilon = 0.1\,\mathrm{sd}(x)$. The dimension is fixed at 6.
2. LPF-SEM. The entropy curve is low-pass filtered by a fourth-order zero-phase Butterworth filter with cutoff f_c selected adaptively as $f_c = \mathrm{clip}(2.5 f_p, 0.02, 0.3)$, where f_p is the dominant frequency of the curve's own power spectrum. The normalized filtered curve $\tilde{H}$ is differentiated, and the delay is the first index beyond $\tau_{\text{start}} = 3$ where $|\mathrm{d}\tilde{H}/\mathrm{d}\tau| < \gamma \max|\mathrm{d}\tilde{H}/\mathrm{d}\tau|$ with $\gamma = 0.05$; if no index qualifies, the maximizer of the normalized curvature $\kappa = |\tilde{H}''|/(1 + \tilde{H}'^2)^{3/2}$ is used instead. The dimension comes from the pseudo-period consistency test: the mean spacing of the filtered curve's minima is accepted as a pseudo-period T_p only if its coefficient of variation

is below 0.5, and then $m = \mathrm{clip}(\mathrm{round}(\tau_{\max}/T_p) + 1, 2, 30)$; otherwise m falls back to the design default 6.

3. MSE [11,13]. Coarse-grained multiscale sample entropy over scales $1 \ldots 20$ with $m = 2$, $r = 0.15\,\mathrm{sd}$; the delay is the scale minimizing the entropy.
4. SFE [12]. Symbolic fuzzy entropy over the delay axis for lags $1 \ldots 20$; the delay is the first local minimum.
5. MS-SFE. Multiscale fuzzy entropy over scales $1 \ldots 20$; the delay is the scale minimizing the entropy.
6. C-C [10]. The correlation-integral statistic $S(t)$ and $\Delta\bar{S}(t)$ over $t = 1 \ldots 60$, evaluated on 500 reference points drawn from the embedding [15]; the delay is the first local minimum of $\bar{S}(t)$, the window the minimum of $\Delta\bar{S}$, and the dimension is derived from the window. We take the delay at the first local minimum of $\bar{S}(t)$ rather than at its first zero crossing, which is the rule in the original paper, because in the statistic evaluated here $\bar{S} = C - C^2$ is built from one predicate: it is non-negative by construction and never reaches zero, so there is no crossing to locate on any of the hundred clean records analysed in Section 4. That choice is not doing the work of the comparison. With the alternative statistic of Kim et al., which replaces the second term by the m-th power of the scalar correlation integral; that term exceeds the m-dimensional integral at large m, so the statistic does take both signs, the delay error of C-C on the same hundred records is 16.5 samples against the 17.1 we report, and that statistic still crosses zero on only 27 of them.

The six do not belong to the same symbolic class, and the theorem covers only two of them. This matters for how the results should be read, so we state the positioning explicitly.

SEM and LPF-SEM are the covered class. They are the only two of the six that threshold the signal, encode agreement between the components of the embedding vector, and select the delay from the saturation of the resulting entropy curve. Theorem 1, Proposition 1 and Proposition 2 apply to them and to any future variant built on the same symbol.

C-C is the correlation-integral precursor of the idea, not a member of the class. It computes the statistics $S(t)$ and $\Delta\bar{S}(t)$ of the proportion of embedded vectors closer than a radius, and there is no pairwise-agreement symbol anywhere in it. We include it because the family's literature positions it as the ancestor of the saturation idea and because it is the natural accuracy benchmark, its delay being closest to the reference among the six, but no theorem here constrains its internal statistics or its selection rule, and we draw no structural conclusion about it.

MSE, SFE and MS-SFE are further benchmarks, and their selection rules differ from the family's: MSE and MS-SFE take the scale that minimizes entropy, SFE the first local minimum along the delay axis, none of them the point at which entropy stops growing. They are included because the applied literature compares them with SEM side by side, not because they share its symbolization. For these three the mapping from their internal scale to a delay is loose, which is one reason the harmonized control of Section 4.11 constrains their range without replacing their rule.

Where the source method returns a dimension through a rule of its own, we use that rule rather than imposing a common one, so that each estimator is judged as it would be used. All entropy curves are computed on at most 800 delay points per curve, and the search range is $\tau \leq 200$ throughout. The complete pipeline is implemented in Python 3.13.5 with NumPy [16], SciPy [17] and scikit-learn [18]; the full software environment is listed in Appendix D.

3.3. Reference Values Used

Applied to the five systems, the reference criteria give $\tau^* = 16, 30, 40, 43, 21$ and $m^* = 3, 2, 3, 3, 3$ for Lorenz, Rössler, Duffing, Mackey–Glass and Chen respectively. Those are the nominal values, computed once from one canonical record per system. They are the reference for the five noise rows of the noise study (Section 4.4), for the ablation of Section 4.5, for the threshold sweep of Section 4.7 and for every row of the harmonized control (Section 4.11 and Appendix B); all six estimators see identical records throughout. The clean results are the deliberate exception: in the clean-data table of Section 4, in the clean row of the noise study and in the clean column of the ablation study the reference is recomputed on each of the twenty independent records, so the reference column there reports the mean of those twenty per-record values (16.1, 38.5, 36.0, 46.2, 20.4), whose record-to-record spread is tabulated in Appendix C. That is why that reference column differs from the nominal list just given.

Two properties of these values should be stated at the outset, because they bound how much a comparison against them can establish. First, the reference is itself realization-dependent: regenerating each system from perturbed initial conditions and a shifted analysis window moves τ^* by 0.3 (Lorenz), 9.5 (Rössler), 8.0 (Duffing), 9.1 (Mackey–Glass) and 0.9 (Chen), one standard deviation over twenty realizations (Appendix C). Second, it is binning-dependent: over 32, 64 and 128 equiprobable bins the first minimum moves for every one of the five systems, by as much as $80 \to 43 \to 30$ for Mackey–Glass. Neither observation rescues the saturation criteria, since Section 4.7 shows that their ratios to the reference spread by a factor of four or more whichever reference is used, but both belong in the limitations and both are stated there.

4. Numerical Study on Chaotic Systems

4.1. Systems and Their Verification

We use three autonomous ODEs, one periodically forced oscillator and one delay differential equation, all integrated from scratch with fixed-step fourth-order Runge–Kutta. No surrogate, closed-form process or synthetic time series is used in place of the systems; i.i.d. Gaussian noise appears only as a control series in Sections 4.2 and 4.10.

Lorenz $\dot{x} = 10(y - x)$, $\dot{y} = x(28 - z) - y$, $\dot{z} = xy - \frac{8}{3}z$, at $\mathrm{d}t = 0.01$ [19]; Rössler $\dot{x} = -y - z$, $\dot{y} = x + 0.2y$, $\dot{z} = 0.2 + z(x - 5.7)$, at $\mathrm{d}t = 0.02$ [20]; Chen $\dot{x} = 35(y - x)$, $\dot{y} = -7x - xz + 28y$, $\dot{z} = xy - 3z$, at $\mathrm{d}t = 0.005$ [21]; these three are autonomous. Duffing, $\ddot{x} = -0.15\dot{x} + x - x^3 + 0.3\cos(1.2t)$ at $\mathrm{d}t = 0.02$, is a periodically forced oscillator and is therefore non-autonomous: its forcing period is $2\pi/1.2 = 5.236$ time units, i.e., 262 steps at $\mathrm{d}t = 0.02$, and the exponent we report for it is a time average. Mackey–Glass is the delay equation $\dot{x} = 0.2\,x(t - 17)/(1 + x(t - 17)^{10}) - 0.1\,x(t)$ at $\mathrm{d}t = 0.1$ [22,23], whose internal delay of 17 time units is 170 steps at that step size. The observed component is x in every case. Appendix A tabulates the initial condition, the discarded transient in both steps and time units, the integration step, the renormalization interval and the initial perturbation norm for each system.

Each system is verified chaotic rather than asserted to be, using a Benettin two-trajectory estimator with renormalization [24,25], and we report the convergence evidence that the verification requires. For the ODEs the perturbation is rescaled every one time unit; for the delay equation a scalar perturbation never reaches the delayed term, so the perturbation is applied to the whole history segment over the delay window, the separation is read at the current point, and that segment rather than the whole history is rescaled, every one time unit, over a long run. Every exponent is computed twice, at the step size of Table 1 and at half of it, and accumulated over both 200 and 2000 time units, and over 200 and 100,000 for Mackey–Glass, whose exponent is two orders of magnitude smaller, so that

the relative standard error is of the same order for every system; the standard error is taken over twenty accumulation blocks. Table 1 reports the long-accumulation values.

Table 1. Largest Lyapunov exponents per unit time, from a renormalized Benettin estimator on the same trajectories used in the experiments, accumulated over 2000 time units, and over 100,000 for Mackey–Glass, whose exponent is two orders of magnitude smaller (±standard error over twenty accumulation blocks) and repeated at half the integration step. All five systems are genuinely chaotic. The last column is the relative difference between the two steps at the long accumulation, $|\lambda_{\max}(\mathrm{d}t/2) - \lambda_{\max}(\mathrm{d}t)|$ divided by the mean of the two.

System	$\lambda_{\max}$	$\lambda_{\max}$ (Half Step)	SE	Rel. Diff.
Lorenz	+0.918	+0.919	0.011	0.1%
Rössler	+0.078	+0.084	0.011	7.6%
Duffing †	+0.162	+0.157	0.013	2.8%
Chen	+2.049	+2.054	0.020	0.2%
Mackey–Glass	+0.0049	+0.0050	0.0001	1.8% ‡

† periodically forced, so this is a time-averaged exponent; the forcing period is 5.236 time units. ‡ Mackey–Glass is accumulated over 100,000 time units rather than 2000. Its exponent is two orders of magnitude smaller than the ODEs', and at 2000 time units its standard error (0.0027) is of the same order as the value itself, so the relative difference between the steps measures the noise of the estimate rather than any disagreement. With the longer accumulation the two entries are $\mathrm{d}t = 0.10$ and $\mathrm{d}t = 0.02$ with standard errors 0.0001 and 0.0002; they agree within 0.4 combined standard errors, a relative difference of 1.8% comparable to the 7.6% of Rössler. Table A2 of Appendix A lists both accumulations for every system. The column is reported to show that each system is chaotic, not to rank the systems: their time scales differ (the delay equation has an internal delay of 17 time units), so the values are not comparable across rows.

The four ODE exponents agree with the standard literature values to within a few percent (0.0714 for Rössler against our 0.078, 2.03 for Chen against our 2.049), and the two step sizes agree to 7.6% or better at the long accumulation, which is the convergence the verification is meant to establish. For Mackey–Glass the exponent is two orders of magnitude smaller, so the same 2000 time units leave a standard error of the same order as the value itself and the two steps would appear to differ by 50% for that reason alone; we therefore accumulate this one system over 100,000 time units, which gives 0.0049 ± 0.0001 at $\mathrm{d}t = 0.10$ and 0.0050 ± 0.0002 at $\mathrm{d}t = 0.02$, a relative difference of 1.8% and 0.4 combined standard errors. The same system at 200 time units of accumulation gives 0.017 ± 0.029, which is a finite-time fluctuation rather than a converged value; that is why Table 1 quotes the long accumulation and why Appendix A reports both. Throughout, $\tau^* = 43$ for Mackey–Glass is in steps, i.e., 4.3 time units, comfortably shorter than the system's own internal delay of 17 time units.

White Gaussian noise is added at the specified signal-to-noise ratio as $y = x + \sigma_\eta \xi$ with $\sigma_\eta^2 = \mathbb{E}[x^2]/10^{\mathrm{SNR}/10}$, so that the reported ratio is the in-band SNR of the record. Twenty realizations are used per cell, with fixed seeds so that every estimator sees identical data, and what varies between them is not the same at every noise level. Under added noise the clean segment is held fixed and twenty independent Gaussian noise realizations are superimposed, so the dispersion reported for the noise and ablation studies (Sections 4.4 and 4.5) isolates the noise sensitivity of the estimator. On clean data there is no noise to vary, and the twenty realizations are twenty independent records: the initial condition is perturbed uniformly in $[-1, 1]$ per coordinate and the analysis window is shifted by a random number of steps, so that both the trajectory and the analyzed segment differ, and the reference is recomputed on each record. The clean-data table of Section 4 reports that protocol; Appendix C gives the full twenty-realization table, including the realization-to-realization scatter of the reference itself.

4.2. Numerical Verification of the Ceiling

Proposition 1 predicts that $H(m, \tau)$ flattens at $H_\infty(m, p)$ and that the plateau equals $m - 1$ if and only if the threshold-crossing probability is $1/2$. Table 2 tests the prediction in the two regimes, evaluating $H(m, \tau)$ at a large delay ($\tau = 97$) for $m = 2, \ldots, 8$ on (i) i.i.d. standard normal samples, which meet the hypotheses of the proposition up to the marginal, and (ii) each of the five chaotic systems, which do not.

Three things are reported for every cell, because three effects have to be separated. $H_\infty(m, p)$ is the prediction of (8) at the measured crossing probability; the finite-sample column is the expected plug-in entropy of independent Bernoulli(p) signs at the same sample size, obtained by Monte Carlo over 400 replicates, and isolates the bias that an alphabet of 2^{m-1} symbols incurs on a finite record; and the total-variation distance between the measured orbit law and the independent prediction measures the residual dependence of the m components at that delay.

On the independent control the prediction is confirmed. For $m = 2, \ldots, 8$ the measured plateau is 1.0000, 1.9999, 2.9999, 3.9997, 4.9991, 5.9979, 6.9949 against predictions 1.0000, 1.9999, 2.9999, 3.9998, 4.9997, 5.9996, 6.9995: the largest discrepancy is 0.0045 bits, and the finite-sample null leaves a residual of only 0.0006 bits. This is the case in which the hypothesis of Proposition 1 holds, and there the formula is accurate to the fourth decimal.

Table 2. Verification of Proposition 1 at $\tau = 97$, N = 20,000, $\varepsilon = 0.1\,\mathrm{sd}(x)$. Upper block: the independent control (i.i.d. Gaussian, $p = 0.4637$), where the hypothesis of the proposition holds; the effective alphabet is exactly 2^{m-1} and the measured plateau matches $H_\infty(m, p)$ to 0.005 bits. Lower block: $m = 6$ on the five systems, which have $p \neq 1/2$, so their predicted plateau is strictly below $m - 1$; the residual shortfall and the total-variation distance are attributed to dependence between the components at $\tau = 97$.

i.i.d. Gaussian Control, $p = 0.4637$, Alphabet 2^{m-1}					
m	2^{m-1}	$H_\infty(m, p)$	**Finite-Sample**	H **Measured**	**Difference**
2	2	1.0000	0.9999	1.0000	+0.0000
3	4	1.9999	1.9998	2.0000	+0.0000
4	8	2.9999	2.9996	2.9999	+0.0000
5	16	3.9998	3.9993	3.9997	−0.0001
6	32	4.9997	4.9985	4.9991	−0.0006
7	64	5.9996	5.9973	5.9979	−0.0017
8	128	6.9995	6.9947	6.9949	−0.0045
Five Systems, $m = 6$, $m - 1 = 5$					
System	p	$H_\infty(6, p)$	**Finite-Sample**	H **Measured**	**TV**
Lorenz	0.4644	4.9997	4.9986	4.8976	0.136
Chen	0.4682	4.9998	4.9987	4.7187	0.269
Duffing	0.4758	4.9999	4.9988	3.2185	0.662
Mackey–Glass	0.5387	4.9996	4.9985	3.0457	0.686
Rössler	0.4422	4.9981	4.9970	2.8847	0.726

On the chaotic systems the measured plateau is lower, for two reasons that the table separates. The first is the marginal: at $\varepsilon = 0.1\,\mathrm{sd}(x)$ the crossing probability is 0.44–0.54 rather than $1/2$, so the correct prediction is $H_\infty(m, p) < m - 1$ and not $m - 1$. The second is residual dependence: at $m = 6$ the shortfall beyond the finite-sample expectation is 0.10 bits for Lorenz and 2.11 bits for Rössler, and the total-variation distance rises in the same order (0.136 for Lorenz, 0.726 for Rössler). Phase-coherent flows keep their thresholded coordinates correlated far beyond the delay at which their linear autocorrelation has decayed, which is why the flows that fall furthest short of the plateau are the two with the most regular rotation. Neither reason damages the argument: a plateau below $m - 1$

is what Proposition 1 predicts once $p \neq 1/2$, and an unattained plateau only makes the saturation criterion's premise weaker, not stronger. What the table rules out is the reading in which $m - 1$ is the number every curve approaches regardless of the data.

Two further observations belong here because they bound what the plateau can be used for. The plateau is a function of the record length: at $m = 6$, Lorenz reads 4.7554 at $N = 4000$ and 4.8976 at $N = 20{,}000$, so any table of plateau values must state N, as Table 2 does. And the plateau is a function of the marginal, hence of the threshold: it is not a dynamical quantity at all, which is the point of Proposition 1.

Figure 1a shows the full curves. Every system's $H(m, \tau)$ rises and flattens, and the flattening occurs at a system-dependent delay that has the visual signature the family reads as a decorrelation onset. Figure 1b shows what noise does to the same construction on Lorenz: the approach to the plateau becomes faster and the plateau becomes noisy, but the curve has the same qualitative shape, which is precisely why the criterion is described as noise-tolerant: the feature it detects is present in every curve, including one that is pure noise.

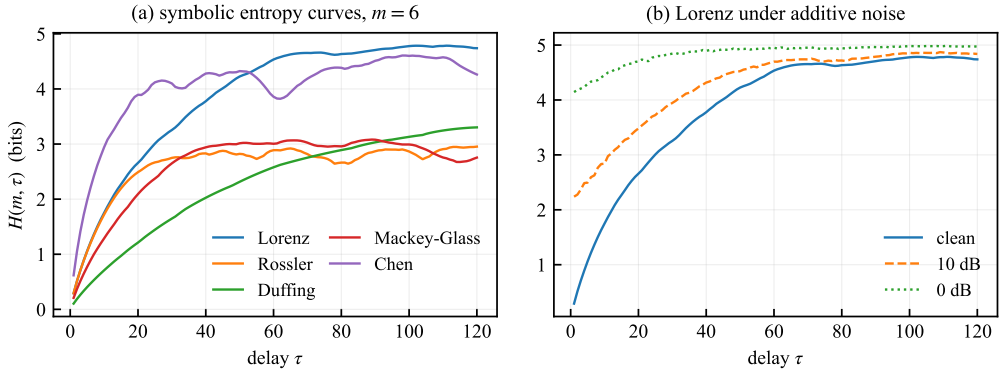


Figure 1. Symbolic entropy curves. (**a**) $H(6, \tau)$ for the five systems; all rise and flatten, and the onset is system-dependent. (**b**) Lorenz under additive white noise; noise changes the rate of approach to the plateau without changing the fact of saturation, because the plateau is imposed by the symbolization.

4.3. Delay Recovery on Clean Data

Table 3 is the first substantive result. It reports the delay each estimator returns on noiseless data, together with the ratio to that record's own AMI reference. The twenty realizations are twenty independent records, as described in Section 4: the initial condition is perturbed and the analysis window is shifted, and the reference is recomputed on each record, so the dispersion column is a genuine record-to-record scatter and the reference column shows the scatter of the reference itself.

Neither accuracy nor a constant offset holds. LPF-SEM's ratio ranges from 1.08 (Rössler) to 4.69 (Duffing), a four-fold spread with no common factor, and on three of the five systems it returns at least twice the reference (Lorenz, Duffing and Chen), while on the other two it is close to it (1.08 and 1.20). The unfiltered SEM is worse where it matters most, returning 131 for Duffing and 57 for Lorenz against references of 36 and 16. MSE and MS-SFE collapse onto small delays ($\hat{\tau} = 2.4$–7.0 across the five systems, ratios 0.08–0.43), the signature of an arg min rule landing on the smallest admissible scale, and SFE returns 3–7 samples throughout, effectively a constant floor. C-C is the most consistent of the six, with ratios 0.67–1.91, but it is consistent in the wrong way for Duffing, where it returns 24 samples against a reference of 36 while its own ratio for Lorenz is 1.91.

Two features of this table are worth separating, because they bound how much a delay comparison can establish. The reference itself scatters: its own 95% interval over the twenty records is ± 0.1 for Lorenz but ± 4.5 for Rössler, so part of every interval in this table is the reference's own instability rather than the estimator's. And with the record-to-record

variation restored, the deviation of a given estimator from the reference is comparable to, and for several systems larger than, the spread of the systems themselves, which is why the bias in Table 4 carries the wide intervals it does.

No member of the family returns a value proportional to the decorrelation delay, and the failure mode differs from member to member. This is not a noise problem.

Table 3. Delay returned on noiseless data, mean ± the half-width of the 95% t-interval over twenty independent realizations ($n = 20$, $t_{0.975,19} = 2.093$; this is a confidence interval, not a standard deviation), with the ratio to that record's own AMI reference in parentheses. A realization here is one independent record of the system: the initial condition is perturbed uniformly in $[-1, 1]$ per coordinate, the analysis window is shifted by a random number of steps, and the AMI reference is recomputed on that record. Each ratio is formed per record and then averaged, so the parenthesised value is the mean of the twenty per-record ratios and not the ratio of the two tabulated means, which differ whenever the reference itself scatters across records. A correct estimator returns τ^*. LPF-SEM spans a 1.08×–4.69× ratio across systems; MSE, SFE and MS-SFE collapse to a floor; C-C is closest on average but returns $0.67\tau^*$ for Duffing.

System	τ^*	C-C	SEM	MSE	SFE	MS-SFE	LPF-SEM
Lorenz	16.1 ± 0.1	30.9 ± 5.1 (×1.91)	56.5 ± 6.1 (×3.51)	7.0 ± 2.9 (×0.43)	3.5 ± 0.9 (×0.22)	5.4 ± 2.5 (×0.34)	67.6 ± 3.3 (×4.20)
Rössler	38.5 ± 4.5	48.9 ± 8.9 (×1.39)	31.3 ± 0.5 (×0.89)	4.0 ± 2.5 (×0.14)	3.6 ± 0.8 (×0.11)	3.4 ± 0.9 (×0.09)	36.8 ± 3.0 (×1.08)
Duffing	36.0 ± 3.7	23.7 ± 12.4 (×0.67)	131.1 ± 14.7 (×3.98)	4.2 ± 0.8 (×0.12)	7.2 ± 2.9 (×0.21)	5.0 ± 0.0 (×0.16)	157.4 ± 12.5 (×4.69)
Mackey–Glass	46.2 ± 4.3	53.5 ± 8.5 (×1.20)	46.0 ± 1.7 (×1.04)	3.6 ± 0.9 (×0.08)	5.3 ± 2.6 (×0.12)	5.0 ± 0.0 (×0.11)	53.5 ± 1.8 (×1.20)
Chen	20.4 ± 0.4	28.5 ± 1.9 (×1.40)	26.1 ± 1.5 (×1.28)	3.8 ± 0.9 (×0.19)	3.5 ± 0.8 (×0.17)	2.4 ± 0.9 (×0.12)	52.7 ± 0.4 (×2.59)

4.4. Behavior Under Noise

Figure 2 and Table 4 extend the comparison to six noise levels. Two quantities are separated: the bias $|\hat{\tau} - \tau^*|$ averaged over systems, and the run-to-run dispersion over the twenty realizations. Table 4 now reports an interval alongside every mean, taken across the five system means, because the system-to-system spread is comparable to the effect being reported, so a bare mean would overstate the precision.

Under the as-used protocol, in which every estimator is run with the configuration of its source and our filtered variant with the configuration of Section 3, which is what a practitioner would obtain by default, LPF-SEM carries the largest mean bias at every SNR, and the bias is not degraded by noise so much as uniformly large: 44.7 samples on clean data rising to 53.8 at 5 dB. Its interval is also the widest and nearly constant, between ±8.6 and ±9.7 samples, indicating a systematic offset rather than noise-driven scatter. On clean data C-C is the closest of the six, and once noise is added MSE has the smallest bias, falling from 27.0 on clean data to 13.9 at 0 dB: its floor-like behavior is less wrong than the plateau-based methods, for the wrong reason. C-C is competitive at high SNR but degrades monotonically (18.3 at 20 dB to 26.3 at 0 dB), and SEM falls from 31.6 on clean data to 20.9 at 20 dB before drifting back to 24.4 at 0 dB, an artifact of the noise flattening the spurious local maxima that trap it on clean data.

That ranking is protocol-dependent, and we report the dependency rather than the ranking alone. Section 4.11 repeats the comparison with one common search window, one common threshold and one common dimension range for all six estimators, and there the ordering changes: LPF-SEM moves from last to fifth and the five estimators other than C-C then lie within 1.8 samples of one another (25.0 to 26.8), with C-C clear at 17.1, so no ordering among the six is meaningful. MSE, whose low bias rests partly on being confined to scales $1 \ldots 20$, improves in absolute terms too ($27.0 \rightarrow 25.6$) but drops one place, because MS-SFE and SEM improve by more. The sentence "LPF-SEM carries the largest bias" is therefore true as stated only of the as-used protocol, and we keep that qualifier wherever the superlative appears.

Table 4. Mean absolute delay deviation $|\hat{\tau} - \tau^*|$ averaged over the five systems, at six noise levels, with the half-width of the 95% t-interval over the $5 \times 20 = 100$ realizations of each cell ($t_{0.975,99} = 1.984$). Because the five systems differ by more than the estimators do, that interval is dominated by the between-system spread, and the per-system intervals over their own twenty realizations are given in Table A8. Protocol: as-used, each published estimator with the configuration of its source and our filtered variant with that of Section 3; the harmonized protocol of Section 4.11 gives a different ordering. The clean row is computed on the twenty independent records of Table 3, each against its own AMI reference; the five noise rows instead hold one clean segment per system fixed and superimpose noise on it, which is what makes their dispersion a measure of noise sensitivity rather than of record-to-record variation, and compare against that segment's own reference.

SNR	C-C	SEM	MSE	SFE	MS-SFE	LPF-SEM
clean	17.1 ± 2.5	31.6 ± 7.4	27.0 ± 2.9	26.8 ± 2.7	27.2 ± 2.7	44.7 ± 8.6
20 dB	18.3 ± 2.2	20.9 ± 3.8	23.2 ± 2.6	26.2 ± 2.1	23.2 ± 2.6	50.8 ± 9.7
15 dB	18.4 ± 2.8	23.0 ± 3.4	20.4 ± 2.6	26.1 ± 2.1	22.8 ± 2.7	52.6 ± 9.5
10 dB	22.9 ± 2.8	22.3 ± 2.7	15.4 ± 2.1	26.4 ± 2.1	21.3 ± 2.8	53.8 ± 9.3
5 dB	24.6 ± 2.6	21.6 ± 2.9	14.0 ± 2.0	26.3 ± 2.1	17.6 ± 2.7	53.8 ± 8.8
0 dB	26.3 ± 2.1	24.4 ± 2.6	13.9 ± 2.0	26.4 ± 2.1	14.7 ± 2.3	53.0 ± 8.6

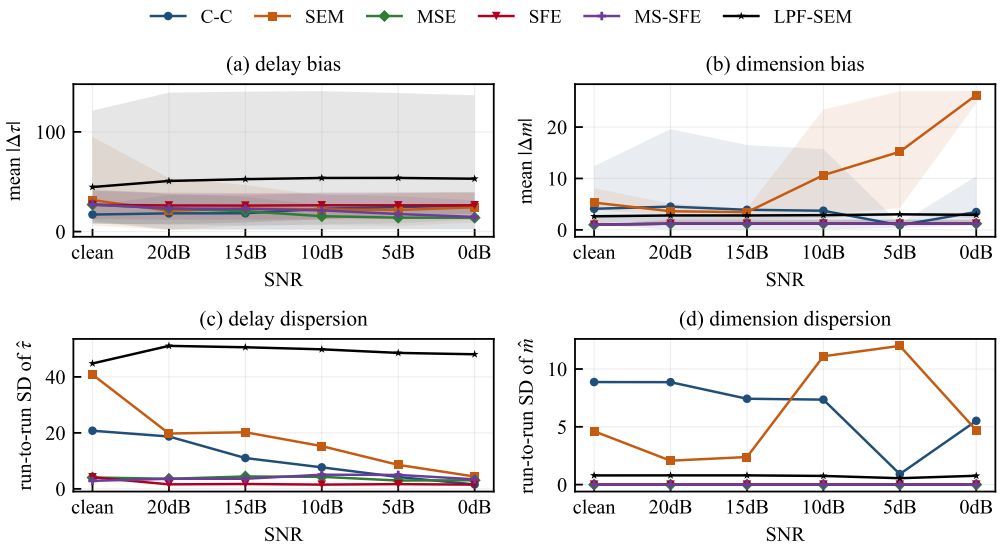


Figure 2. Bias and dispersion against noise level. Black stars are LPF-SEM: highest bias in every panel and a flat dispersion profile, the signature of a systematic offset rather than noise sensitivity. The dimension panels show SEM's breakdown at low SNR (orange, up to 26 dimensions of error against a reference of 3).

The dimension results (right-hand panels) are more revealing still. On clean data LPF-SEM returns $m = 6$ for Lorenz, Duffing, Mackey–Glass and Chen in 100% of trials, which is exactly the design default used when the pseudo-period test is rejected, giving a dimension error of 3.0 on Lorenz, Duffing and Mackey–Glass and 2.9 on Chen, against a per-record reference that is 3 on almost every record. Rössler is the exception: its pseudo-period test invariably passes on clean data and returns $m = 4$ (error 2.0 against $m^* = 2$), though under added noise the fallback $m = 6$ reappears. The pseudo-period test does not in general produce a dimension; it returns the default. The reported dimension is inert, which explains the otherwise puzzling uniformity and fixes the τ–m coupling of Remark 2 at $m \approx 6$: LPF-SEM's delay output is in effect an unfiltered-curve criterion evaluated at a single dimension with two extra smoothing stages.

4.5. Component Ablation

The ablation isolates each ingredient by building the method up in stages, holding the data fixed. Results are in Table 5.

Table 5. Ablation , mean $|\hat{\tau} - \tau^*|$ / mean $|\hat{m} - m^*|$, with the half-width of the 95% t-interval over the $5 \times 20 = 100$ realizations pooled across the five systems ($t_{0.975,99} = 1.984$); the second entry of each pair is a mean only, because those estimators return an almost constant dimension and their own intervals are below 0.06 in every cell. Adding the low-pass filter raises the clean-data delay bias by about 70%; the pseudo-period test is the only component with a measurable benefit, and only for the dimension.

Configuration	Clean	10 dB	5 dB	0 dB
SEM	$31.6 \pm 7.4/3.0$	$22.2 \pm 2.6/3.2$	$22.0 \pm 2.8/3.2$	$24.2 \pm 2.6/3.2$
SEM + LPF	$54.1 \pm 11.1/3.0$	$65.1 \pm 9.9/3.2$	$69.6 \pm 9.4/3.2$	$70.7 \pm 9.4/3.2$
SEM + LPF + NCC	$44.7 \pm 8.6/3.0$	$53.7 \pm 9.3/3.2$	$54.1 \pm 8.9/3.2$	$51.9 \pm 8.7/3.2$
SEM + PCT	$31.6 \pm 7.4/2.6$	$22.2 \pm 2.6/2.8$	$22.0 \pm 2.8/3.0$	$24.2 \pm 2.6/2.9$
Full (LPF + NCC + PCT)	$44.7 \pm 8.6/2.6$	$53.7 \pm 9.3/2.8$	$54.1 \pm 8.9/3.0$	$51.9 \pm 8.7/2.9$

All three readings are unfavorable. Filtering raises the clean-data bias by 71% (31.6 → 54.1) and the penalty grows with noise, reaching 70.7 at 0 dB: the filter is not neutral but actively harmful. The curvature criterion recovers part of the loss (54.1 → 44.7) but never returns to the unfiltered level, so the two-stage NCC remains worse than the criterion it replaces. The pseudo-period test is the only component with a consistent benefit, and its gain is confined to the dimension (3.0 → 2.6 on clean data), is small, and, as Section 4.4 established, it is usually the design default rather than a value derived from the signal.

4.6. Filter Parameter Sensitivity

Figure 3 sweeps the two filter constants over the full grid $f_c \in \{0.02, 0.04, 0.06, 0.08, 0.10, 0.15, 0.20, 0.25\}$ and filter order $\in \{2, 4, 6\}$, twenty-four combinations in all, for each system and noise level, using the same crossing rule for the delay.

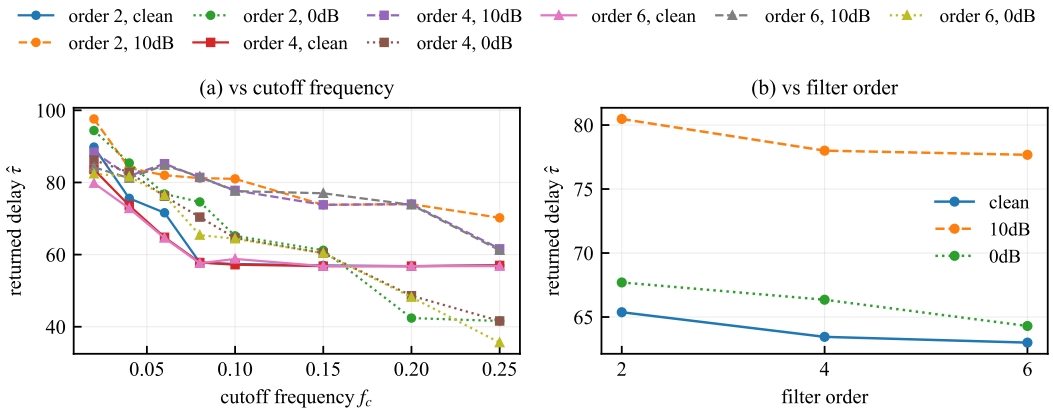


Figure 3. Filter sensitivity. (**a**) Returned delay against cutoff frequency for several filter orders and noise levels; the curves are non-monotone and system-dependent. (**b**) Returned delay against filter order. Across the full grid the estimate ranges from 5 to 197 samples.

Over this grid the returned delay spans 5 to 197 samples on clean data, essentially the entire search window $\tau \leq 200$, with a standard deviation of 42 across the grid, and the spread is larger under noise (52 at 10 dB). The reported delay is therefore dominated by the filter constants rather than by the signal: a plausible arbitrary choice changes the answer forty-fold, two orders of magnitude beyond the deviation the method is claimed to reduce.

The adaptive cutoff rule fixes one point in this space but chooses it from the dominant peak of the entropy curve's own spectrum, a property of the same flattening curve whose onset is being located, so it inherits the circularity rather than resolving it.

4.7. Non-Transferability of the Threshold

If the criterion measured something physical, the threshold γ that reproduces the reference delay would be roughly the same across systems. We sweep γ over $[0.005, 1]$ on a grid of 0.005 up to 0.1 and 0.05 above it, record the implied delay, and ask which γ, if any, returns τ^* to within one sample. Table 6 reports the set of admissible values and its size, not a single representative, because the width of that set is itself the quantity of interest: a wide admissible band means the "recovered" threshold is not identified by the data.

No single value works. On clean data the admissible set is $\{0.15, 0.20\}$ for Rössler, $\{0.25\}$ for Mackey–Glass, $\{0.75\}$ for Duffing and $\{0.75, 0.80\}$ for Chen, a five-fold spread between the lowest and highest admissible value, and for Lorenz no value in the entire swept range reproduces the reference at all. The admissible set then drifts with noise: Rössler moves to $\{0.20, 0.25\}$ and then $\{0.25\}$, Mackey–Glass to $\{0.50, 0.55\}$ and then $\{0.60, 0.65\}$, and Duffing loses feasibility entirely. At 10 dB two of the five systems admit no feasible threshold, and at 0 dB only four of five admit one.

Two readings follow. The first is the one Remark 3 predicts: a criterion whose free parameter must be re-tuned per system, drifts as the data degrades, and sometimes has no valid setting cannot serve as the basis of a general estimator. The second is narrower and more useful for the revision: even where a value does exist, the admissible band is usually one or two grid points wide, which means the requirement is not a property of the criterion but a coincidence at the resolution of the sweep. Reporting the set rather than its first element makes both readings visible.

Table 6. Set of γ values that reproduce the AMI reference delay to within one sample, and the size of that set (number of grid points), evaluated on the single clean record per system that defines the reference. No single value transfers: on clean data the admissible values span 0.15–0.80, Lorenz admits none, and the sets drift or vanish as noise is added.

System	Clean	10 dB	0 dB
Lorenz	none	none	$\{0.95\}$ (1)
Rössler	$\{0.15, 0.20\}$ (2)	$\{0.20, 0.25\}$ (2)	$\{0.25\}$ (1)
Duffing	$\{0.75\}$ (1)	none	none
Mackey–Glass	$\{0.25\}$ (1)	$\{0.50, 0.55\}$ (2)	$\{0.60, 0.65\}$ (2)
Chen	$\{0.75, 0.80\}$ (2)	$\{0.80\}$ (1)	$\{0.75\}$ (1)

4.8. Coupling Between the Delay and the Dimension

Remark 2 predicts that $\hat{\tau}$ depends on the trial embedding dimension, because the ceiling defining saturation is $m - 1$. Sweeping $m = 3, \ldots, 8$ at fixed data and fixed reference, we find the drift is real and system-dependent: over that range the LPF-SEM delay moves $73 \to 70$ for Lorenz, $45 \to 33$ for Rössler, $177 \to 173$ for Duffing, $69 \to 50$ for Mackey–Glass and, in the opposite direction, $31 \to 56$ for Chen. The drift reaches $1.8\times$ (Chen) and $1.4\times$ (Rössler, Mackey–Glass) even though τ^* is a fixed property of the signal.

Because the pipeline then derives the dimension from the delay, the two outputs are not independent estimates of two parameters; they are two views of one arbitrarily parameterized computation. Reporting them as separate calibrated quantities, as the source methodology does, misrepresents what has been computed.

4.9. Computational Cost

Figure 4a reports wall-clock time for every estimator on identical hardware, using the exact implementations compared everywhere else in this study, with median times over seven repeats at each of five series lengths ($N = 500$, 2000, 5000, 10,000, 20,000). The fitted exponents p in $T \sim N^p$ are 0.20 for C-C, 0.80 for LPF-SEM, 0.85 for SEM, 0.78 for MSE, 0.96 for MS-SFE and 0.30 for SFE. The measurement environment, as recorded by the timing script itself, is: Intel Core i7-6820HQ (Skylake, four physical cores, eight logical processors) at its nominal 2.70 GHz; 15.8 GB of DDR4-2133 memory; Windows 10 Pro, version 10.0.19045, 64-bit, on the balanced power plan; CPython 3.13.5 (64-bit) with NumPy 2.5.0, SciPy 1.18.1, scikit-learn 1.9.0 and matplotlib 3.11.2, linked against scipy-openblas 0.3.33.112.0 (OpenBLAS 0.3.33, Haswell kernel, `MAX_THREADS=24`); one process throughout, with no `OMP_NUM_THREADS`, `OPENBLAS_NUM_THREADS` or `MKL_NUM_THREADS` override, so the BLAS thread pool ran at its default width on those eight logical processors; each timing is the median of seven repeats taken with `time.perf_counter`. Because the machine is a laptop under a balanced power plan, its clock is not fixed, and we report below the spread we measured under varying load rather than quoting the exponents to more digits than they support.

The C-C exponent needs a caveat that the previous version omitted. The implementation used here caps its reference set at $n_{\mathrm{sub}} = 500$ points and forms the distance matrix among those points only, so beyond $N \approx 600$ the amount of work it performs stops growing: the number of pairwise distances evaluated rises by a factor of 1.37 between $N = 500$ and N = 20,000, while N itself rises by a factor of 40. The measured time nevertheless rises by a factor of 2.13 over the same range, and that is what the exponent is fitted to; the rest of the growth is the per-distance cost and the bookkeeping around it, so the fitted p is not the exponent of the distance count. The fitted slope therefore measures the subsampling cap, not the scaling of the C-C algorithm, and it is correspondingly unstable: repeated runs of the same measurement under different machine load move it by more than 0.1 with R^2 between 0.47 and 0.98. Removing the cap and evaluating the full N-point distance matrix on a tractable range ($N \leq 4000$) gives $p = 2.60$ with $R^2 = 0.995$, which is the scaling the published $O(N^2)$ formulation implies; scaled from that fit, the published form is some four orders of magnitude more expensive than the capped implementation at N = 20,000.

Accordingly we withdraw the earlier claim that LPF-SEM is cheaper than C-C by an order of magnitude. In the implementation actually compared, the ratio of LPF-SEM to C-C time is 0.030 at $N = 500$ but 0.30 at N = 20,000: LPF-SEM starts about thirty times cheaper and ends only about three times cheaper, the advantage shrinking by roughly a factor of ten across the tested range. Its ordering against the unfiltered SEM also reverses between $N = 5000$ and N = 10,000. SFE's fit is poor ($R^2 = 0.67$ against 0.96–0.998 for the others) because its timings are non-monotone in N, so its exponent is not a scaling law at all.

Nothing above depends on cost: an estimator can be cheap and wrong. The filtered variant is at once among the most biased and among the cheapest at short record lengths, and low cost is a plausible reason why filtering the curve looks attractive, without being a reason to trust it. We also stress a distinction that is easy to blur: the complexity of a correlation-integral algorithm is an operation count, whereas every number here is wall-clock time of one particular implementation, and the two need not agree in either direction. Section 4 gives an instance of exactly that, where a subsampling cap hides the scaling entirely.

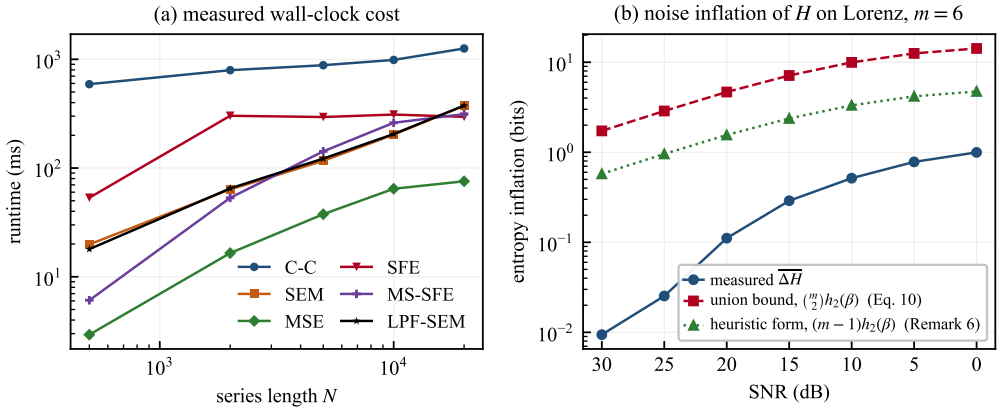


Figure 4. Secondary diagnostics. (**a**) Median wall-clock time against series length for the six estimators on identical hardware. (**b**) Noise inflation of H on Lorenz, $m = 6$, against the union bound of (10) and the heuristic $(m-1)$ form of Remark 6; the bound holds at every SNR and is loose by one to two orders of magnitude, while the heuristic form is closer but is not a bound.

4.10. The Noise-Inflation Bound

Proposition 2 bounds the inflation of the symbolic entropy by additive noise, and Remark 6 describes a heuristic sharpening of its constant from $\binom{m}{2}$ to $m-1$. Figure 4b evaluates both against the measured inflation on Lorenz at $\tau \leq 120$, $m = 6$.

Both hold at every SNR, as they must. Both are also loose by a wide margin. At 0 dB the measured inflation is 1.00 bit while (10) predicts 14.2 bits and the heuristic form predicts 4.7; at 30 dB the measured inflation is 0.009 bit against 1.73 and 0.58. The looseness is 14×–183× for the original bound and 5×–61× for the heuristic form.

The reason connects to the main theorem, and it is the reason the bound cannot be tightened without the joint flip law. The bound treats the $\binom{m}{2}$ agreement bits as if they were independently degraded, whereas by Theorem 1 they originate in only $m-1$ independent binary degrees of freedom, and noise flipping one sign bit inverts all agreement bits involving that component simultaneously, so the degradations are strongly positively correlated (Remark 5). Positive correlation lowers the entropy of the flip pattern, hence the information gained, which is why the subadditive bound is loose; accounting for it requires the joint flip distribution of the m comparisons, which is signal-dependent and which we do not attempt.

We report this as a secondary negative result: bounds of the form (10) are often offered as the warrant for calling symbolic methods noise robust, but a bound permitting fourteen times the observed inflation cannot certify robustness at 0 dB. It is worth adding that the heuristic $m-1$ form, though numerically closer, is not a bound at all, so it cannot carry that warrant either.

4.11. A Harmonized-Protocol Control

Each estimator above was run under the as-used protocol: the five published estimators with the configuration of their source, and our filtered variant with the configuration of Section 3. The argument for that baseline is that the question is what a practitioner obtains from these methods by default. The reviewer's objection is that this leaves the search ranges and selection rules different, so the comparison confounds the methods with their configurations. The objection is fair, and we answer it with a control rather than a defense.

The control uses the same records and the same noise realizations as the as-used run, the seed stream of the main experiment being replayed so that every cell is a paired comparison, and imposes one common configuration: delay and scale search window $1 \ldots 100$ for every estimator, symbolization threshold $\varepsilon = 0.10\,\mathrm{sd}(x)$ for the symbolic methods and tolerance $r = 0.20\,\mathrm{sd}(x)$ for the tolerance-based ones, reported dimension

restricted to $[2, 12]$, and the nominal dimension $m = 4$ that MSE, SFE and MS-SFE used to report replaced by their actual internal dimension $m = 2$. The rules are untouched, because replacing them would compare our rewrites rather than the published methods; only the ranges are put inside the common window. For MSE and MS-SFE, whose scale is not a delay, the control constrains the range only and this is stated in Appendix B.

The outcome changes the ranking, and we report it as it stands. Table 7 gives the mean absolute delay deviation under both protocols, and Table 8 the resulting order on clean data. C-C stays first, but LPF-SEM moves from last to fifth (44.7 $\rightarrow$ 26.2). The reason is visible in the per-system delays: giving MSE and MS-SFE a scale window out to 100 lets them reach $\hat{\tau} = 60$ on Rössler instead of four, so part of the low bias they show under the as-used protocol was an artifact of being confined to twenty scales. The spread across the six estimators narrows from 27.6 to 9.7 samples, and almost all of that remainder is the gap between C-C (17.1) and the other five, which lie within 1.8 samples of one another (25.0 to 26.8); SFE is nominally last, by 0.6 samples, and we do not read that difference as a ranking.

Three consequences follow for how the results of this paper should be quoted. The statement that LPF-SEM carries the largest delay bias is true of the as-used protocol, which is the protocol under which it would actually be used, and we qualify it that way from here on; it is not a statement about the estimator in isolation. The ordering of the middle of the table is protocol-dependent and we make no claim about it. And the two conclusions that the paper rests on are unaffected, because both survive in the harmonized column as well: no member of the family returns a value proportional to the reference, and LPF-SEM's ratio to the reference still spans a factor of four.

Table 7. Harmonized -protocol control. Mean absolute delay deviation $|\hat{\tau} - \tau^*|$ averaged over the five systems, as-used versus harmonized, at six noise levels; each entry is 'as-used / harmonized'. The records, the noise realizations and the references are identical between the two columns, so the comparison is paired. Full protocol, per-system delays and reported dimensions are in Appendix B.

SNR	C-C	SEM	MSE	SFE	MS-SFE	LPF-SEM
clean	17.1/17.1	31.6/25.4	27.0/25.6	26.8/26.8	27.2/25.0	44.7/26.2
20 dB	18.3/18.3	20.9/20.9	23.2/27.4	26.2/26.2	23.2/27.8	50.8/29.9
15 dB	18.4/18.4	23.0/23.0	20.4/27.4	26.1/26.1	22.8/29.0	52.6/31.1
10 dB	22.9/22.9	22.3/22.3	15.4/28.1	26.4/26.4	21.3/36.1	53.8/31.1
5 dB	24.6/24.6	21.6/21.6	14.0/28.5	26.3/26.3	17.6/36.3	53.8/30.2
0 dB	26.3/26.3	24.4/24.4	13.9/31.0	26.4/26.4	14.7/34.8	53.0/29.1

Table 8. Estimator ranking by mean absolute delay deviation on clean data, under the two protocols, on the twenty-record basis. The top of the ranking is stable; the bottom is not, and under the harmonized protocol all but C-C lie within 1.8 samples of one another (25.0 to 26.8), so the order among them is not meaningful.

Rank	As-Used	Harmonized
1	C-C (17.1)	C-C (17.1)
2	SFE (26.8)	MS-SFE (25.0)
3	MSE (27.0)	SEM (25.4)
4	MS-SFE (27.2)	MSE (25.6)
5	SEM (31.6)	LPF-SEM (26.2)
6	LPF-SEM (44.7)	SFE (26.8)

4.12. Translation Non-Invariance of the Entropy Curve

Proposition 3 makes a claim that can be checked in a few lines of computation, and the check is worth reporting because it is the sharpest available statement of what is wrong with the criterion. We take one clean Lorenz record, translate it by

$c \in \{-0.5, -0.25, 0, 0.25, 0.5\}\,\mathrm{sd}(x)$, add no noise, and recompute everything: the crossing probability, the plateau, the delay LPF-SEM returns, the delay the unfiltered SEM rule returns, and the whole AMI curve.

Table 9 gives the result. The AMI curve is unchanged to the last bit, the maximum absolute deviation over all c and all τ being 0.0 in double precision, and its first minimum is 16 in all five cases, so the reference cannot see the translation at all. The criterion can: the LPF-SEM delay moves over 55–75 samples, a range of 20 samples or 27% of its own value, and the unfiltered SEM rule moves from 28 to 70, a factor of 2.5. The mechanism is the one the proposition describes: $\mathrm{sd}(x + c)$ is unchanged, so ε is unchanged in absolute units, but the mass above it moves from 0.254 to 0.632, and by Proposition 1 that moves the plateau ($H_\infty(6, p)$ rises from 4.571 at $p = 0.254$, passes through 5.000 near $p \approx 0.53$, and returns to 4.955 at $p = 0.632$) and with it the whole curve and the point at which it flattens.

This is not a sensitivity to a perturbation of the record; it is a sensitivity to a change of units. Translating a vibration record, or offsetting a signal before analysis, changes no dependence property whatsoever, and a delay estimator that responds to it by more than a quarter of its own output cannot be estimating a property of the dependence. The criterion is a functional of the marginal as well as of the joint law, while the reference it is scored against is a functional of the joint law alone. That is the structural version of the reviewer's first comment, and it does not depend on any of the experiments in this section.

Table 9. Translation non-invariance. A clean Lorenz record ($N = 4000$, $m = 6$, $\tau \le 200$, $\varepsilon = 0.1\,\mathrm{sd}(x)$) is translated by c standard deviations. The mutual-information curve and its first minimum are unchanged; the crossing probability, the plateau and both criterion delays are not. $c = 0$ is the untranslated record, so $\max|\Delta H|$ is the largest deviation of the entropy curve from the untranslated one over the whole delay window, $\max_{1\le\tau\le200}|H_c(\tau) - H_0(\tau)|$, in bits.

$c/\mathrm{sd}(x)$	p	$H_\infty(6, p)$	LPF-SEM	SEM	τ^*	$\max\|\Delta H\|$
−0.50	0.2542	4.5707	57	34	16	0.782
−0.25	0.3422	4.9123	74	70	16	0.323
0	0.4380	4.9975	75	70	16	0.000
+0.25	0.5255	4.9999	63	47	16	0.681
+0.50	0.6315	4.9554	55	28	16	1.226

5. Downstream Task: Fault Diagnosis

The measurements so far score the delay against a reference. A practitioner cares about it only through the task the reconstruction feeds. We therefore carry the comparison through to the application that motivates the family, RQA feature [26,27] fault classification on real machinery data [28,29], because the outcome reframes the practical claim, and it is more useful reported against the numbers than described.

5.1. Data and Protocol

Two sets are used, both described in detail in Appendix E.

MFPT bearing [30]: three baseline (healthy) records, seven outer-race and seven inner-race records under variable load, seventeen records in all (the archive's fourth group, three constant-load outer-race records, is not used; Appendix E gives the inventory and the bearing geometry). The acquisition metadata stored in each file gives the sampling rate, the shaft rate and the load for every record, and Appendix E tabulates them: the baseline records are stored at 97,656 Hz against 48,828 Hz for the fault records, and we therefore decimate the baseline records by two, which equalises bandwidth across classes rather than favoring one. Loads span 0–300 lb at a shaft rate of 25 Hz throughout. Each record is then

cut into windows of 1024 samples, without overlap (the stride equals the window length), at most forty per record, giving 680 windows over three classes, chance 1/3.

Rotor [31]: a gas-turbine scaled-model rotor test bench run at 3300 rpm and sampled at 10,240 Hz, giving 180 denoised records of 2048 samples (0.2 s each), forty-five per class over normal, rub-impact, unbalance and misalignment. The 180 records are independent acquisitions rather than slices of longer files, so no analysis stride or start offset is involved; each yields two non-overlapping 1024-sample windows (0.1 s, about 5.5 shaft revolutions), giving 360 windows over four classes, chance 1/4. A single vibration channel is used. The rig, the sensors, the fault-seeding method of each class, the `db4` denoising and its threshold rule are documented in Appendix E, together with the two items that the released archive does not record.

For each estimator the delay and dimension are calibrated on a held-out slice of each record, taken as the first 30,000 samples or the whole record where shorter, and then held fixed while that record's windows are embedded. Every classification result in this section uses the as-used protocol: each published estimator with the configuration of its source and our filtered variant with that of Section 3. The harmonized control of Section 4.11 is reported for the delay comparison only and is not used here. Each window is mapped to a nine-dimensional feature vector by recurrence quantification at the calibrated (τ, m) with a Theiler window [32] of 2τ: recurrence rate, determinism, mean and maximum diagonal line length, diagonal-line entropy, laminarity, trapping time, and the maximum and the entropy of the vertical line lengths. Four classifiers are applied (an RBF-kernel SVM, k-nearest neighbors with $k = 5$, a 200-tree random forest, and linear discriminant analysis) under repeated grouped cross-validation: five repeats of five-fold stratified grouped splitting, so that windows from one record never straddle the train/test split and the reported scores are not inflated by within-record similarity, and so that the split-to-split variance is itself measured.

Because the 680 MFPT windows come from 17 records and the 360 rotor windows from 180 segments, the independent unit is the record, not the window, and intervals are computed accordingly: we report the standard deviation across the five repeat means and a 95% cluster bootstrap interval over records (2000 resamples). Appendix C shows what this choice costs: the cluster standard error is 6.7–12.0 times the window-level one on MFPT, which is why a window-level interval would have been the wrong one to quote.

5.2. Results

Table 10 and Figures 5 and 6 give the accuracies, with the mean calibrated delay and dimension alongside.

On MFPT the estimator matters a great deal, and the ranking is the reverse of the one the delay study predicts. LPF-SEM is best on every classifier (0.917 with the random forest, 0.815 even with a linear discriminant, the weakest model in the set) for a mean over the four classifiers of 0.856. C-C is worst on every classifier, with a mean over the four of 0.552: its random forest score is only 0.22 above the 0.333 chance line, for a feature set that supports 0.917 when the delay is LPF-SEM's. The two cluster intervals over records, $[0.408, 0.697]$ and $[0.798, 0.987]$, do not overlap, so the difference survives the resampling unit that the benchmark's design imposes. The spread between best and worst, 0.37 of accuracy, exceeds any difference reported between the methods under test in the source literature.

The ordering does not follow delay accuracy. C-C carried the smallest mean delay deviation of the six on clean data (17.1 ± 2.5 samples) and at 20 dB (18.3), and it is worst here. LPF-SEM carried the largest deviation under the as-used protocol at every noise level, with a ratio to the reference spanning 1.08–4.69, and it is best here.

Table 10. RQA-feature classification accuracy under five repeats of grouped five-fold cross-validation (25 folds): mean ± standard deviation across the five repeat means, with the 95% cluster bootstrap interval over records in brackets for the random forest. $\bar{\tau}$ is the mean delay calibrated per record. The last column is the random-forest accuracy obtained from the calibrated pair (τ, m) alone, with the RQA features discarded. On MFPT, LPF-SEM (the most biased delay estimator under the as-used protocol) wins under every classifier and C-C (the least biased under either protocol) loses under every classifier; on the rotor set the estimator effect washes out and all the intervals overlap.

Method	$\bar{\tau}$	SVM	KNN	RF	LDA	RF 95% CI	τ, m
MFPT bearing (3 classes, chance 0.333**)**							
C-C	3.2	0.604 ± 0.060	0.456 ± 0.056	0.552 ± 0.052	0.594 ± 0.037	$[0.408, 0.697]$	0.433
SEM	3.1	0.796 ± 0.045	0.790 ± 0.037	0.783 ± 0.028	0.724 ± 0.035	$[0.606, 0.938]$	0.617
MSE	12.9	0.739 ± 0.019	0.696 ± 0.010	0.749 ± 0.015	0.777 ± 0.033	$[0.589, 0.908]$	0.450
SFE	4.1	0.744 ± 0.023	0.685 ± 0.005	0.733 ± 0.029	0.737 ± 0.013	$[0.588, 0.846]$	0.400
MS-SFE	19.6	0.751 ± 0.015	0.682 ± 0.012	0.751 ± 0.017	0.773 ± 0.012	$[0.651, 0.848]$	0.283
LPF-SEM	41.4	0.870 ± 0.015	0.822 ± 0.021	0.917 ± 0.051	0.815 ± 0.026	$[0.798, 0.987]$	0.817
Rotor (4 classes, chance 0.250**)**							
C-C	12.4	0.977 ± 0.008	0.963 ± 0.004	0.971 ± 0.005	0.962 ± 0.005	$[0.948, 0.989]$	0.844
SEM	8.9	0.962 ± 0.007	0.954 ± 0.002	0.962 ± 0.005	0.950 ± 0.006	$[0.939, 0.982]$	0.767
MSE	10.3	0.958 ± 0.008	0.919 ± 0.003	0.937 ± 0.004	0.884 ± 0.004	$[0.907, 0.964]$	0.606
SFE	4.5	0.957 ± 0.004	0.942 ± 0.007	0.953 ± 0.011	0.939 ± 0.007	$[0.930, 0.973]$	0.339
MS-SFE	7.1	0.936 ± 0.006	0.913 ± 0.005	0.941 ± 0.010	0.892 ± 0.006	$[0.914, 0.964]$	0.717
LPF-SEM	11.9	0.979 ± 0.004	0.972 ± 0.007	0.976 ± 0.002	0.969 ± 0.004	$[0.956, 0.992]$	0.606

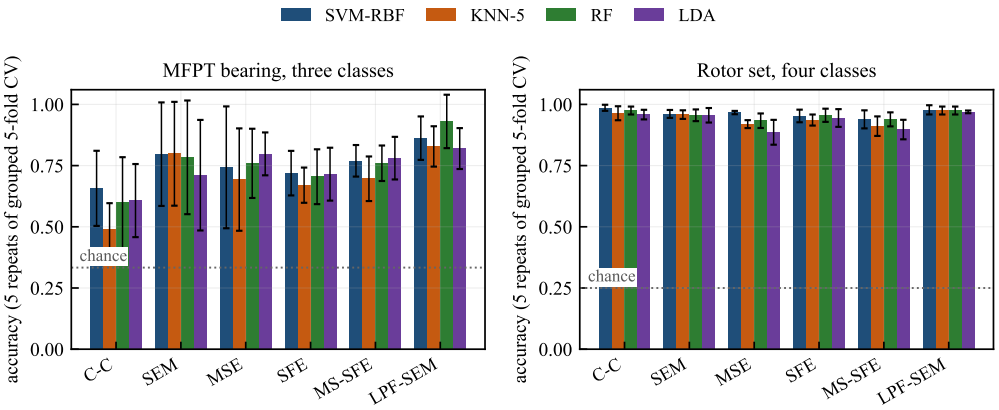


Figure 5. Fault classification on the two real datasets, grouped by record, with the standard deviation across five repeats of grouped five-fold cross-validation as error bars. Dotted lines mark chance. On MFPT the delay estimator changes the outcome by more than thirty accuracy points and the best and worst cluster intervals over records do not overlap; on the rotor set every estimator clears 0.93 and the choice is immaterial.

On the rotor set the picture inverts again, and less informatively: every estimator lands between 0.937 and 0.976 under the random forest (0.936 to 0.979 under the SVM), with LPF-SEM and C-C at the top and the rest within four points. All six cluster intervals overlap, so on this set the choice of estimator is not resolvable at the sample size available. The four classes are nearly linearly separable in these features, so the delay has little to do. This insensitivity is real, but it is a property of the rotor set rather than of the procedure, and it does not generalize to MFPT.

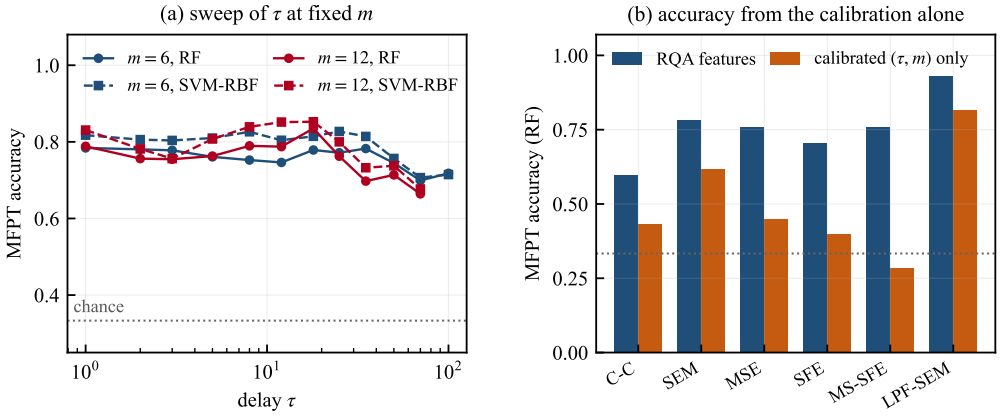


Figure 6. Why the MFPT ordering is what it is. (**a**) Accuracy against the delay with m fixed, at $m = 6$ and $m = 12$; the curves are flat or peaked, never monotone in τ, and never reach the table's best score. (**b**) Accuracy when each window is described by its calibrated (τ, m) alone, with the RQA features discarded. LPF-SEM's parameters carry most of its score; the ordering of the methods is largely reproduced without any signal features. The dotted rule in both panels is the chance level, 1/3, of the three-class task.

5.3. What Drives the MFPT Result

The obvious reading is that the delay is simply too short for the losers: LPF-SEM embeds at $\bar{\tau} = 41.4$ and wins, MS-SFE and MSE at 19.6 and 12.9 occupy the middle, and C-C and SFE at 3.2 and 4.1 occupy the bottom. A recurrence plot at $\tau = 3$ on a 1024-sample window is close to diagonal (the embedding barely separates successive coordinates), so its features should be correspondingly impoverished.

That reading cannot be the whole explanation, because SEM returns a delay as small as C-C's ($\bar{\tau} = 3.1$) yet places second at a mean of 0.773; what separates them is the dimension, 11.6 against 4.5. The task responds to the (τ, m) pair, and neither estimate being accurate is what selects the pair that works. We test the magnitude reading, and a less flattering one, directly.

5.4. Does the Delay Magnitude Explain It?

We sweep τ over 1–100 with m held fixed and everything else (windows, features, folds, classifier) unchanged. The sweep is run at the two dimensions the main experiment actually used: $m = 6$ (LPF-SEM's calibrated mean of 6.5) and $m = 12$ (SEM's 11.6); Figure 6a gives the result.

At $m = 6$ the accuracy is flat to within 0.04 for every delay from $\tau = 1$ to $\tau = 35$, and falls only past $\tau = 50$, well beyond the range any estimator returns. At $m = 12$ the curve is likewise shallow but peaked: it rises to 0.835 (RF) near $\tau = 18$ and declines by 0.17 to $\tau = 70$. Neither curve is monotone, and neither approaches the 0.917 that LPF-SEM reaches in Table 10. The delay's magnitude is not what separates the estimators; if it were, the sweep would climb toward LPF-SEM's $\bar{\tau} = 41.4$, yet it does the opposite.

5.5. Does the Calibration Explain It?

Since the delay itself does not, we remove the signal and keep the calibration. Each window's nine-dimensional feature vector is replaced by the two numbers (τ, m) used to compute it, and the same grouped cross-validation is re-run; the last column of Table 10 and Figure 6b give the part of the reported score attributable to the calibration alone. For LPF-SEM that control returns 0.817 under the random forest, 89% of the feature-based figure, from two numbers per record; for the other five it returns 0.28 to 0.62, and their ordering is close to the full pipeline's.

The mechanism is straightforward: every window of a record is embedded at that record's own (τ, m), and the classes are disjoint in records, so a parameter estimate that separates the classes is copied into every feature vector of that record, and the classifier can read the class partly off the calibration rather than off the signal. LPF-SEM's delays separate the 17 MFPT records well; C-C's range from 2 to 5 and do not, which is why the most accurate delay estimator scores 0.433 here and finishes last.

Three further analyses turn that mechanism from an observation into a measurement, all reported in full in Appendix C.

5.5.1. A Shuffling Test

If the calibration string carries class information, then describing each record by the calibrated (τ, m) of a randomly chosen other record should destroy the control's accuracy. It does, for the methods whose strings are informative: over 150 random reassignments, the LPF-SEM control falls from 0.817 to 0.368 ± 0.140, the observed value exceeding every one of the 150 permutations, so $p = 1/151$; the SEM control falls from 0.617 to 0.361 ± 0.140 with $p = 0.033$, while the four methods whose delays do not separate the records show no significant effect. On the rotor set the same test collapses every method's control to the 0.25 chance level with $p \leq 0.027$, which says the calibration is a record fingerprint on that set too, even though it buys no accuracy there because the features are already saturated.

5.5.2. A Pooled-Calibration Control

Replace the per-record calibration by a single pair per method, calibrated on the pooled calibration slices of every record, and the record-specific channel is switched off. LPF-SEM then falls to 0.743 from 0.917, and the six estimators bunch between 0.720 and 0.790: the ranking is gone. The record-specific calibration is worth 0.174 accuracy, 19% of the winner's score. On the rotor set the pooled control changes nothing (0.964–0.986), consistent with the features already saturating there.

5.5.3. A Label-Permutation Test

Permuting the class labels at the record level and re-running the same cross-validation puts every reported MFPT accuracy far outside its null distribution ($p \leq 0.010$ for five of the six methods, and $p = 0.069$ for C-C), so the scores are not an artifact of the grouped splitting procedure itself.

One caution limits how far the mechanism should be taken, and it is the reason the pooled control is the decisive one. The calibration is unsupervised and uses only the record's own data, and a deployment would calibrate a new record in exactly the same way, so this is not label leakage in the cross-validation sense. What it is instead is a per-record scalar that correlates with the class and therefore acts as a record fingerprint: the classifier can read part of the class off an identifier that happens to be computable from the record. That is a property of any benchmark that groups folds by record and calibrates per record, and it is a warning about the benchmark as much as about the method.

Two qualifications bound this result. The control does not exhaust the scores: the features add 0.10 over the parameters for LPF-SEM and much more for SEM and MS-SFE, so the signal still does work. And the effect is not universal: on the rotor set the control is high for several methods (0.844 for C-C, 0.767 for SEM) while every method's feature accuracy lies within four points of every other's, so where the features are saturated the calibration adds nothing either way. The MFPT ranking, however, is dominated by it.

5.6. What the Inversion Means

Saturation-based symbolic entropy is used in the applied literature because it improved fault-classification scores, and on this testbed it does improve them: the evidence

was real, the mechanism inferred from it was not. Our filtered variant reproduces that pattern, and it is the most biased estimator of the six by the delay measure under the as-used protocol while being the best classifier on MFPT. What the sweep and the controls together show is that this downstream evidence was never evidence about the delay: the delay's size does not drive the score, its accuracy does not drive it, and about a fifth of the variant's MFPT advantage is reproduced by its calibrated parameters alone once the per-record channel is removed. A negative result about delay accuracy measured against AMI therefore has no purchase on that literature, which is why the criterion's inability to recover the decorrelation delay can persist through several variants and a decade of use.

We also record what the revision changed in this section, because it affects how the numbers should be cited. The submitted version of Table 10 reported a single grouped five-fold pass with no interval; it now reports twenty-five folds with intervals taken over the unit the design actually resamples. The accuracies moved by 0.01–0.05 in the process (LPF-SEM 0.930 $\rightarrow$ 0.917, C-C 0.599 $\rightarrow$ 0.552) and no qualitative conclusion changed. The quantitative mechanism is what the revision adds.

6. Discussion

6.1. Why the Saturation Criterion Cannot Work

The failure is not an implementation defect and not a matter of tuning; it follows from what the entropy curve is. Saturation-based criteria rest on the premise that the delay at which $H(m, \tau)$ stops growing is the delay at which the embedding components become independent. Theorem 1 and Proposition 1 show that this happens for every signal whatsoever, because the symbolization guarantees a ceiling at $m - 1$ and every mixing process approaches the plateau $H_\infty(m, p)$ of Proposition 1, which equals that ceiling only when $p = 1/2$. The delay at which it happens is set by the rate at which the joint distribution of the m thresholded signs equilibrates: a nonlinear functional of the threshold geometry and the local expansion rates near that level set. The AMI minimum, by contrast, is a functional of the full joint distribution of $(x_t, x_{t+\tau})$, not of its second moments, and is therefore sensitive to nonlinear dependence and invariant under any invertible transformation of either marginal; the autocorrelation, which is a second-order statistic, is not invariant under monotone transformations, and the two must not be conflated. Nothing forces the AMI minimum and a threshold-crossing rate to coincide; the measured ratios of 1.08 to 4.69 are what one expects of two unrelated functionals. Section 4.12 sharpens this into a structural statement: translating the record leaves the AMI curve and its first minimum unchanged to machine precision while moving the criterion's delay by 27% of its own value, so the criterion depends on a quantity, the marginal, that the reference does not.

The extreme case illustrates this. Applied to i.i.d. noise, which has $\tau^* = 1$ trivially, since consecutive samples are independent, the curve still rises, still saturates smoothly at its plateau $H_\infty(m, p)$, which for this control is 4.9997 rather than the ceiling $m - 1 = 5$ (Table 2), and still shows a plausible saturation onset. The criterion assigns a delay to a signal with no dynamics at all: the definition of a detector responding to an artifact. Two problems compound this. The ceiling depends on m, so the delay and the dimension are not independent though they are reported as such; and the thresholds (γ, and f_c in the filtered version) are unconstrained by the data, so any threshold picks out some delay from a smooth monotone curve. The returned number is better described as a function of the analyst's constants than of the signal.

6.2. What a Usable Replacement Should Preserve

Nothing in our argument disturbs the standing criteria. First-minimum AMI and first-false-neighbor selection remain the sensible default, for the reasons they always had:

AMI is well defined for any process, has no free parameter beyond the binning, and its first minimum is the delay at which the reconstruction coordinates share the least mutual information; FNN has a threshold, but one calibrated against the noise floor and attractor size rather than against a shape heuristic.

If one wants a symbolic criterion, and there are good reasons to want cheap threshold-based indicators of dependence, it must compare H against its own plateau rather than a fixed threshold. Define the normalized occupancy

$$g(m, \tau) = \frac{m - 1 - H(m, \tau)}{m - 1} \in [0, 1], \tag{12}$$

which measures the distance from the uniform state of Proposition 1 and is free of the ceiling artifact by construction. Selecting the first τ with g below a level tied to the sampling noise of H would at least measure saturation relative to what saturation means for that m. We have not established that such a criterion estimates the decorrelation delay, and our results give no reason to expect it: by Proposition 3 the residual dependence on p survives the normalization, because $m - 1$ is not the plateau when $p \neq 1/2$. This is therefore not a repair of the estimator but a redefinition of what is being measured, and we offer it as a diagnostic to be validated rather than as a method. Evaluating whether $g(m, \tau)$ tracks the reference delay monotonically, and how much of its variation is still attributable to p, is left open.

6.3. Practice: A Decision Rule for the Two Goals

Section 5 answers the question reviewers will ask first: if the estimator is this wrong, how was it ever adopted? Because on the downstream task it is not wrong but better, for a reason that, as Sections 5.4 and 5.5 show, has nothing to do with the delay. The practical advice therefore has to be split by the practitioner's goal, and we state it as two branches.

If the goal is physically correct reconstruction parameters, such as a delay for a recurrence plot, a correlation or Lyapunov dimension, an attractor reconstruction, a Theiler window, or any analysis in which τ is itself the object of interest, then do not use a saturation-based entropy criterion of the kind tested here (SEM, LPF-SEM, MSE, SFE, MS-SFE, C-C). For the pairwise-agreement members the argument is structural (Section 2); for the others the recommendation rests on the empirical result that no configuration tracked the reference in Sections 4 and 5, and symbolic entropies of other design, such as permutation entropy, are outside the scope of both. Take τ from the first minimum of the AMI and m from FNN, as the field has done. This is not a matter of tuning: Sections 4.6 and 4.7 show that no setting of LPF-SEM's filter constants or threshold recovers the reference consistently, and its τ is wrong by a factor reaching 4.69, so the number carries no dynamical meaning even where it happens to be close.

If the goal is only classification accuracy on a task already solvable from RQA features, LPF-SEM may be used as a black-box heuristic feature-engineering step, and on MFPT it is the best of those compared here (0.917 under repeated grouped cross-validation, cluster interval over records $[0.798, 0.987]$). Three caveats carry over. First, its τ must not be reported or interpreted as the decorrelation delay: the parameter is a per-record index whose separating power is what helps the classifier, not the fact that it is a delay (Section 5.5), and a pooled-calibration control shows that about a fifth of its MFPT advantage is that index rather than the signal. Second, the accuracy figures are protocol-specific in the sense of Section 4.11: a harmonized search configuration changes the delay ranking, and there is no reason to expect the classification ranking to be immune to the same change. Third, the advantage is task-specific rather than a property of the method; it is not reproduced on the

rotor set, where every estimator lands within four points of every other, so the heuristic should be validated on the target data rather than assumed to transfer.

6.4. Limitations

Nine limitations should be stated, and we order them by how much they affect the reading of the paper.

First, and most important, the object of criticism is the saturation criterion, not the reference criteria. Our target throughout is the specific claim that the saturation onset of a pairwise-agreement entropy estimates the decorrelation delay, and specifically the methods that present themselves as replacements for AMI and FNN. Theorem 1, Proposition 1 and the empirical study bear on that claim. They do not show that AMI and FNN are optimal, and we do not claim it. The critique is of a substitution, not of the standard practice that substitution proposed to replace.

Second, the theorem covers two of the six estimators, not all of them. Theorem 1 and its consequences apply to SEM and LPF-SEM, which compute the pairwise-agreement symbol and read a delay off its saturation. C-C computes correlation integrals and has no such symbol; MSE, SFE and MS-SFE use different selection rules and, in the multiscale cases, a scale rather than a delay. The negative results about those four are empirical statements about the implementations compared here and carry no structural claim (Section 3.2). Where the submitted version described the family as covering "from the C-C method through LPF-SEM", that description was too wide and is corrected.

Third, AMI and FNN have defects of their own, and we now quantify them. The first-minimum rule is a heuristic, not an exact optimum for reconstruction, and we use it as the field's convention and as the criterion under test claims to replace. Two of its defects are measured here. It depends on the binning: over 32, 64 and 128 equiprobable bins, the first minimum moves for every one of the five systems, by as much as $80 \to 43 \to 30$ steps for Mackey–Glass (Section 3.3). And it is realization-dependent: regenerating the systems from perturbed initial conditions and shifted windows moves τ^* by one standard deviation of 0.3 to 9.5 over twenty records (Section 3.3 and Appendix C); Table 3 reports the 95% interval of each reference instead. We showed that our conclusions do not rest on this reference by recomputing the delay ratios against three references that involve no mutual information at all, on the same twenty independent records as Table 3: the first zero crossing of the autocorrelation, its $1/e$ decay point, and its first local minimum. LPF-SEM's ratio spans 0.44–1.98 against the first, 0.61–2.79 against the second and 0.22–1.01 against the third, so the bias remains uncalibratable whichever reference is used. What does not survive a change of reference is the ordering of the estimators: LPF-SEM has the largest mean absolute ratio to one among the six under the AMI reference, but it is the closest of the six under the two autocorrelation references, which are themselves large (109–645 samples on a single record). A method that returns large delays therefore looks accurate against a large reference. We make no ordering claim that is not tied to the AMI convention.

Fourth, the ranking of the estimators is protocol-dependent. Under the as-used protocol LPF-SEM carries the largest delay bias at every noise level; under a harmonized protocol that gives all six estimators one search window, one threshold and one dimension range, it is fifth of six, and the five estimators other than C-C lie within 1.8 samples of one another, so no ordering among them is meaningful (Section 4.11). Both numbers are reported, and every superlative in this paper is qualified with the protocol it refers to. What survives both protocols is that no member of the family returns a value proportional to the reference, and that LPF-SEM's ratio to it spans a factor of four.

Fifth, the plateau is a function of the record and of the marginal, not a dynamical constant. Its value depends on N, since Lorenz reads 4.7554 at $N = 4000$ and 4.8976 at $N = 20{,}000$ for $m = 6$, and on the threshold through p; Table 2 therefore states N and reports the crossing probability. Readers should not carry the numbers of that table to other record lengths or other ε.

Sixth, the symbolization threshold was fixed at $\varepsilon = \alpha\,\mathrm{sd}(x)$ with $\alpha = 0.1$ throughout. Other choices move all curves, though Theorem 1 holds for every $\varepsilon > 0$ and the ceiling is unaffected, so the structural argument is independent of this choice while the specific numbers are not. Proposition 3 formalises exactly how much the numbers move.

Seventh, our systems are stationary, and we did not evaluate non-stationary or intermittently chaotic signals: records with drifting parameters, transient laminar phases, or intermittent bursting. The theorem is unconditional in x, so the ceiling and the plateau hold for such records too and the structural argument is unaffected; but the measured delay ratios, the ablation, and the downstream inversion are established only for the stationary benchmarks examined here.

Eighth, the classification evidence rests on two datasets and seventeen independent records on the informative one. On MFPT the unit of resampling is the record, there are seventeen of them, and the resulting cluster intervals are correspondingly wide ($[0.408, 0.697]$ and $[0.798, 0.987]$ for the extremes); on the rotor set every estimator's interval overlaps every other's, so no ranking is resolvable there. The mechanism identified in Section 5.5, a per-record calibration acting as a record fingerprint, is a property of any benchmark that groups folds by record and calibrates per record, and we regard it as a warning about such benchmarks generally rather than a fact about any one method.

Ninth, all timings describe one implementation on one machine. The cost results are wall-clock measurements of the Python implementations compared here, with the environment stated in Section 4.9; a wall-clock exponent can be dominated by an implementation choice, as it is for C-C, whose cost in this implementation is set by a subsampling cap rather than by the algorithm. The five systems are standard benchmarks plus one delay equation, and we do not claim the measured ratios transfer beyond them.

7. Conclusions

We set out to reproduce a reported improvement in phase-space parameter estimation and instead replaced the premise on which the family rests with a correct but narrower one. The result is a boundary statement as much as a theorem, and we now keep the boundary visible.

The theorem, and its three layers. The pairwise-agreement symbolization is invariant under complementation of the thresholded sign vector, so the symbol space has exactly 2^{m-1} elements, the map from complement orbits to attainable agreement patterns is a bijection, and $H(m, \tau) \in [0, m-1]$ for every signal (Lemma 1, Theorem 1). That is the first layer and it is unconditional. The second layer is stronger: the plateau equals $H_\infty(m, p)$, a function of the symbolization and of the threshold-crossing probability, equal to $m - 1$ if and only if $p = 1/2$ (Proposition 1); it is confirmed to 0.005 bits on an independent control and it sits strictly below $m - 1$ in every one of our experiments, where $p = 0.44$–0.54. The consequence is that saturation is built into the symbolization rather than discovered in the data, and that the event "H has stopped growing" carries no information about the dynamics; what remains is the rate of approach. The third layer is empirical and conditional: the criteria that thresholds that rate do not track the reference delay in the five systems and six noise levels examined, and we present that as an empirical finding rather than as a corollary.

What makes the criterion unfit for its stated purpose, structurally. It is not that the ceiling is reached; it is that the quantity the criterion thresholds is not a function of the dependence structure alone. Translating a record by half its standard deviation leaves the mutual-information curve and its first minimum bit-for-bit unchanged while moving the criterion's delay over 55–75 samples, a range of 27% of its own value (Proposition 3, Table 9). A change of units changes no dependence property, and an estimator that responds to it by a quarter of its output is not estimating one.

The evidence, with its qualifiers. The low-pass filter, the method's headline ingredient, increases the delay bias instead of reducing it (31.6 $\rightarrow$ 54.1 samples); the returned delay sweeps a forty-fold range across the filter's own constant grid; the set of thresholds that would reproduce the AMI reference spans 0.15–0.80 across systems and is empty for Lorenz, while at 10 dB it is empty for two systems; the dimension output is the method's own fallback constant in 100% of runs for four of five systems; and LPF-SEM's delay spans 1.08–4.69 times the reference with no common factor. Under the as-used protocol LPF-SEM carries the largest bias of the six at every noise level; under a harmonized protocol it is fifth of six, the five estimators other than C-C then lie within 1.8 samples of one another, and SFE is nominally last by 0.6 samples, so we state the ranking as protocol-dependent and treat no ordering among the six as meaningful. A secondary negative result is that the standard noise-inflation bound for symbolic entropies is valid but loose by one to two orders of magnitude at practical SNRs, and that the sharper $m - 1$ constant sometimes quoted is a heuristic rather than a bound.

The downstream reversal. In the downstream fault-classification study on MFPT bearing data the ordering inverts: the most biased estimator under the as-used protocol classifies best (0.917, 95% cluster interval over records $[0.798, 0.987]$) and the one closest to the reference worst (0.552, $[0.408, 0.697]$). Neither the delay's magnitude nor its accuracy explains it, a sweep of τ at fixed dimension being flat or non-monotone over the whole range, and replacing the per-record calibration by a single pooled pair drops the winner to 0.743, so about a fifth of its score is the record-specific calibration acting as a record fingerprint. A criterion can be useful and wrong at once, and the downstream evidence that sustained this family was never evidence about the delay.

The advice, with its scope. For a physically meaningful reconstruction, such as a delay for a recurrence plot, a correlation or Lyapunov dimension, an attractor reconstruction, a Theiler window, or any analysis in which τ is itself the object of interest, the delay should come from the first minimum of the average mutual information and the dimension from false nearest neighbors, as the field has done; the saturation onset of a pairwise-agreement entropy should not be used as a delay, and no tuning of its constants repairs this. Where only classification accuracy on an RQA feature set is wanted, LPF-SEM may be used as a black-box heuristic, and on MFPT it is the best of those compared here; but its output must be read as a tuned per-record index rather than as the decorrelation delay, and validated on the target data rather than assumed to transfer. All of these statements are bounded by what we tested: five systems, six noise levels, one symbolization threshold, two machinery datasets, and the implementations described in Section 3.

Three directions follow directly from the boundary just drawn. The first is to test the normalized occupancy $g(m, \tau)$ of Section 6.2 as a diagnostic: whether it varies monotonically with the reference delay, and how much of its residual variation is still attributable to the threshold-crossing probability, are open questions that a symbolic criterion would have to answer before being proposed again. The second is to extend the non-invariance argument beyond one-dimensional thresholds, to multivariate and multiscale symbolizations, where the analogous question is which transformations of a record leave the symbol law invariant and which do not. The third is methodological and applies well beyond this

paper: a benchmark that groups folds by record and calibrates parameters per record gives its classifiers a per-record identifier, and quantifying how much of a reported gain is that identifier, as the pooled-calibration and shuffling controls of Section 5.5 do here, should become routine in fault-diagnosis comparisons.

Author Contributions: Conceptualization, Z.Z. and J.L.; methodology, Z.Z.; software, Z.Z.; validation, Z.Z., J.C. and J.L.; formal analysis, Z.Z.; investigation, Z.Z.; resources, J.L.; data curation, Z.Z. and J.C.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z., J.L. and J.C.; visualization, Z.Z.; supervision, J.L.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Research and Development Program of Shaanxi, grant number 2025PT-ZCK-42.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MFPT bearing fault dataset is a public benchmark [30] and is described record by record in Appendix E; no new raw experimental data were generated for it in this work. The rotor set is our own: it was acquired on the test rig described in Appendix E and reported in our earlier conference paper, and it is not a public dataset. The processed segments used here, together with the calibration and window-extraction scripts, are available from the corresponding author on reasonable request; requests for the raw acquisitions should be directed to the same address. An expanded description of the processing chain is given in Appendix E.

Acknowledgments: The authors thank the anonymous reviewers for their constructive suggestions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Lyapunov Verification Protocol

Table A1. Generation and verification protocol for the five systems. Time is given in time units (t.u.) and in integration steps. The transient is removed before the analyzed record starts, and the discard is a further interval removed so that the analyzed segment is fully settled.

System	dt	Initial Condition	Transient	Discard	Autonomous
Lorenz	0.01	$(1, 1, 1)$	50.0 t.u. (5000 steps)	20.0 (2000 steps)	yes
Rössler	0.02	$(1, 1, 1)$	200.0 (10,000 steps)	60.0 (3000 steps)	yes
Duffing	0.02	$(0.1, 0.0)$	400.0 (20,000 steps)	100.0 (5000 steps)	no
Chen	0.005	$(-10, 0, 37)$	50.0 (10,000 steps)	15.0 (3000 steps)	yes
Mackey–Glass	0.1	$x_0 = 1.2$	300.0 (3000 steps)	300.0 (3000 steps)	yes

Table A1 gives, for each system, the initial condition, the discarded transient in both steps and time units, the integration step, and the renormalization parameters used for the Benettin estimate. Table A2 gives the measured exponent under two integration steps and two accumulation lengths, with the standard error over twenty accumulation blocks.

Three remarks complete the protocol. For Mackey–Glass the perturbation is applied to the whole history segment over the delay window rather than to a scalar, because a scalar perturbation does not propagate into the delayed term in the right proportion; the separation is read at the current point, and the segment rather than the whole history is rescaled, every one time unit, over the 100,000-time-unit accumulation, which is preceded by a 400-time-unit discard of the Lyapunov run itself; the 300-time-unit transient and 300-time-unit discard listed in Table A1 belong to the generation protocol of the series the

estimators see, not to this run. The internal delay of 17 time units corresponds to 170 steps at $dt = 0.1$, so the delay the estimators return for this system ($\tau^* = 43$ steps) is 4.3 time units, considerably shorter than the system's own delay. And the truncation of the history at 3000 steps corresponds to 300 time units, about 17.6 internal delays, which is what makes the initial condition irrelevant to the analyzed segment.

Table A2. Measured largest Lyapunov exponent per unit time under two integration steps and two accumulation lengths, renormalizing every one time unit with an initial perturbation norm of 10^{-8}; the standard error is taken over twenty accumulation blocks. 'Combined SE' is the difference between the two steps expressed in units of the combined standard error. The accumulation lengths are 200 and 2000 time units for the four systems integrated from ODEs and 200 and 100,000 for Mackey–Glass, whose exponent is two orders of magnitude smaller.

System	dt	Accumulation	λ_{max}	SE	Rel. Diff. to Half Step	Combined SE
Lorenz	0.01	2000	+0.91818	0.01082	0.1%	0.0
	0.005	2000	+0.91879	0.01099		
	0.01	200	+1.01075	0.10084	1.0%	0.1
Rössler	0.02	2000	+0.07762	0.01073	7.6%	0.4
	0.01	2000	+0.08372	0.01004		
	0.02	200	+0.16608	0.10242	6.5%	0.1
Duffing	0.02	2000	+0.16190	0.01289	2.8%	0.2
	0.01	2000	+0.15738	0.01376		
	0.02	200	+0.21800	0.09810	8.2%	0.1
Chen	0.005	2000	+2.04874	0.02038	0.2%	0.2
	0.0025	2000	+2.05378	0.02362		
	0.005	200	+2.12705	0.10905	2.3%	0.3
Mackey–Glass	0.1	100,000	+0.00489	0.00014	1.8%	0.4
	0.02	100,000	+0.00498	0.00016		
	0.1	200	+0.01693	0.02852	0.7%	0.0

At 200 time units of accumulation every estimate carries a standard error of order 0.1 and is a finite-time fluctuation rather than a converged value; the difference from the long-accumulation column ranges from 3.8% to 246% across the ten pairs, which is why the main text quotes the long accumulation only.

Appendix B. Harmonized-Protocol Control

Table A3. Mean delay per system on clean data under the two protocols (as-used/harmonized), with the mean of the twenty per-record AMI references alongside. The ranking change of Table 8 is produced by the Rössler and Duffing columns.

System	τ^*	C-C	SEM	MSE	SFE	MS-SFE	LPF-SEM
Lorenz	16.1	30.9/30.9	56.5/56.5	7.0/7.7	3.5/3.5	5.4/8.0	67.6/61.4
Rössler	38.5	48.9/48.9	31.3/31.3	4.0/60.0	3.6/3.6	3.4/60.5	36.8/45.1
Duffing	36.0	23.7/23.7	131.1/99.7	4.2/4.0	7.2/7.2	5.0/5.0	157.4/91.0
Mackey–Glass	46.2	53.5/53.5	46.0/46.0	3.6/3.2	5.3/5.5	5.0/5.0	53.5/52.4
Chen	20.4	28.5/28.5	26.1/26.1	3.8/2.8	3.5/3.5	2.4/2.6	52.7/34.1

The control behind Section 4.11 reuses the records, the noise realizations and the references of the main experiment, so every cell is a paired comparison, and changes one thing: the configuration.

1. Input: identical series, $N = 4000$, identical noise seeds replayed from the main experiment.
2. Delay and scale search window: $1 \leq \tau \leq 100$ for every estimator. The as-used windows were 200 (SEM, LPF-SEM), 60 (C-C) and 20 (MSE, SFE, MS-SFE).
3. Symbolization threshold: $\varepsilon = 0.10\,\mathrm{sd}(x)$ for SEM and LPF-SEM; tolerance $r = 0.20\,\mathrm{sd}(x)$ for MSE, SFE and MS-SFE, which raises MSE from its native 0.15 so that the three tolerance-based variants share one value.
4. Reported dimension: restricted to $[2, 12]$ for every method, and the nominal $m = 4$ that MSE, SFE and MS-SFE reported under the as-used protocol is replaced by their actual internal dimension $m = 2$.
5. Selection rules: unchanged. MSE and MS-SFE take the entropy-minimizing scale and SFE the first local minimum along the lag axis, as published; only their ranges are put inside the common window, because replacing the rules would compare rewritings rather than the published methods.

Table A3 gives the per-system delays that explain the change in ranking. Extending the scale window to 100 lets MSE and MS-SFE reach $\hat{\tau} = 60$ on Rössler instead of four, and allows SEM and LPF-SEM to shorten on Duffing; the reported dimensions move accordingly.

Figure A1 shows the same comparison as a bar chart, and makes the one asymmetry of the design visible: the harmonized protocol only ever relaxes the constraints that the as-used windows imposed on MSE, SFE and MS-SFE, whereas SEM and LPF-SEM lose half of their search range. The change in ranking therefore cannot be read as the harmonized protocol favoring LPF-SEM; if anything it works the other way, and LPF-SEM still moves up one place.

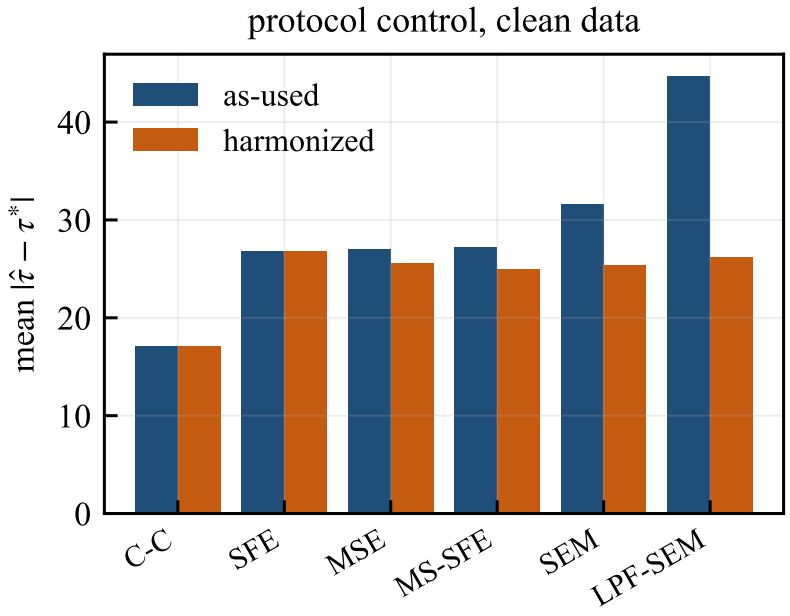


Figure A1. Mean absolute delay deviation on clean data under the as-used and harmonized protocols, for the six estimators. The top of the ranking is stable; the bottom is not, and LPF-SEM moves from last to fifth while all but C-C converge to within 1.8 samples of one another.

Appendix C. Classification Statistics

Table A4. Cluster bootstrap over records against a naive window bootstrap, random forest, 2000 resamples. The point estimate is the accuracy pooled over all out-of-fold windows, which is the quantity resampled; Table 10 reports instead the unweighted mean of the 25 fold accuracies, and the two differ by at most 0.011 because the folds are not of equal size. The cluster interval is the one quoted in Table 10.

Dataset	Method	Accuracy	Cluster CI	Cluster SE	Window CI	Window SE	SE Ratio
MFPT	C-C	0.554	[0.408, 0.697]	0.0734	[0.538, 0.571]	0.0085	8.60
MFPT	SEM	0.787	[0.606, 0.938]	0.0846	[0.773, 0.801]	0.0071	11.99
MFPT	MSE	0.760	[0.589, 0.908]	0.0830	[0.746, 0.775]	0.0073	11.41
MFPT	SFE	0.729	[0.588, 0.846]	0.0673	[0.714, 0.744]	0.0078	8.68
MFPT	MS-SFE	0.751	[0.651, 0.848]	0.0507	[0.736, 0.765]	0.0076	6.68
MFPT	LPF-SEM	0.915	[0.798, 0.987]	0.0488	[0.906, 0.924]	0.0048	10.15
Rotor	C-C	0.971	[0.948, 0.989]	0.0108	[0.962, 0.978]	0.0040	2.68
Rotor	SEM	0.962	[0.939, 0.982]	0.0111	[0.952, 0.971]	0.0047	2.36
Rotor	MSE	0.937	[0.907, 0.964]	0.0148	[0.926, 0.948]	0.0057	2.61
Rotor	SFE	0.953	[0.930, 0.973]	0.0112	[0.943, 0.963]	0.0050	2.23
Rotor	MS-SFE	0.941	[0.914, 0.964]	0.0133	[0.930, 0.951]	0.0056	2.39
Rotor	LPF-SEM	0.976	[0.956, 0.992]	0.0094	[0.968, 0.983]	0.0036	2.63

Table A5. Pooled out-of-fold confusion matrices over twenty-five folds, random forest, MFPT bearing data. Rows are the true class, columns the predicted class; each class contributes 5 times its window count.

Method	True\Pred	Healthy	Inner	Outer
C-C	healthy	318	100	182
	inner	123	893	384
	outer	41	686	673
LPF-SEM	healthy	595	0	5
	inner	5	1304	91
	outer	24	163	1213

Table A6. Pooled-calibration control. Each row replaces the per-record calibration by a single global (τ, m) for that method. On MFPT the six estimators collapse to a common band and LPF-SEM's advantage disappears; on the rotor set nothing changes.

Dataset	Method	(τ, m)	SVM	KNN	RF	LDA	Per-Record RF
MFPT	C-C	$(5, 2)$	0.752	0.717	0.720	0.748	0.552
MFPT	SEM	$(2, 6)$	0.806	0.723	0.780	0.783	0.783
MFPT	MSE	$(17, 4)$	0.813	0.762	0.784	0.809	0.749
MFPT	SFE	$(3, 4)$	0.805	0.718	0.790	0.789	0.733
MFPT	MS-SFE	$(17, 4)$	0.813	0.762	0.784	0.809	0.751
MFPT	LPF-SEM	$(49, 6)$	0.757	0.721	0.743	0.748	0.917
Rotor	C-C	$(14, 3)$	0.992	0.978	0.986	0.992	0.971
Rotor	SEM	$(13, 12)$	0.972	0.961	0.967	0.969	0.962
Rotor	MSE	$(20, 4)$	0.975	0.958	0.964	0.900	0.937
Rotor	SFE	$(9, 4)$	0.997	0.978	0.983	0.961	0.953
Rotor	MS-SFE	$(20, 4)$	0.975	0.958	0.964	0.900	0.941
Rotor	LPF-SEM	$(13, 9)$	0.978	0.958	0.975	0.975	0.976

Table A7. Permutation tests. Test 1: the (τ, m)-only control under shuffled record-to-calibration assignment. Test 2: full random-forest accuracy under record-level label permutation. The smallest attainable p at these permutation counts is $1/151$ and $1/101$ respectively.

Dataset	Method	Test 1 Observed	Test 1 Null	Test 1 p	Test 2 Null	Test 2 p
MFPT	C-C	0.433	0.364 ± 0.129	0.305	0.356 ± 0.118	0.069
MFPT	SEM	0.617	0.361 ± 0.140	0.033	0.398 ± 0.116	0.010
MFPT	MSE	0.450	0.376 ± 0.136	0.325	0.346 ± 0.119	0.010
MFPT	SFE	0.400	0.370 ± 0.144	0.391	0.341 ± 0.116	0.010
MFPT	MS-SFE	0.283	0.388 ± 0.150	0.801	0.364 ± 0.098	0.010
MFPT	LPF-SEM	0.817	0.368 ± 0.140	0.0066	0.357 ± 0.107	0.010
Rotor	C-C	0.844	0.252 ± 0.045	0.0066	0.251 ± 0.036	0.010
Rotor	SEM	0.767	0.255 ± 0.041	0.0066	0.251 ± 0.036	0.010
Rotor	MSE	0.606	0.254 ± 0.038	0.0066	0.248 ± 0.035	0.010
Rotor	SFE	0.339	0.245 ± 0.044	0.027	0.256 ± 0.035	0.010
Rotor	MS-SFE	0.717	0.249 ± 0.045	0.0066	0.248 ± 0.031	0.010
Rotor	LPF-SEM	0.606	0.255 ± 0.037	0.0066	0.247 ± 0.031	0.010

Table A8. Delay error per system and noise level, mean $\pm$ the half-width of the 95% t-interval over the twenty realizations of that cell ($n = 20$, $t_{0.975,19} = 2.093$). This is the interval behind every aggregate in Tables 4 and 5. On clean data the twenty realizations are twenty independent records of the system, each measured against its own AMI first minimum; at each noise level one clean segment is held fixed and twenty independent Gaussian realizations are superimposed, measured against that segment's reference.

System	SNR	C-C	SEM	MSE	SFE	MS-SFE	LPF-SEM
Lorenz	clean	14.8 ± 5.1	40.4 ± 6.1	9.1 ± 2.9	12.6 ± 0.9	10.7 ± 2.5	51.5 ± 3.3
	20 dB	16.4 ± 1.1	53.8 ± 0.4	2.0 ± 0.0	11.9 ± 1.1	2.0 ± 0.0	63.1 ± 0.3
	15 dB	12.6 ± 2.6	46.8 ± 6.7	2.0 ± 0.0	12.1 ± 0.8	1.9 ± 0.1	63.4 ± 0.4
	10 dB	7.2 ± 2.5	23.5 ± 8.2	2.0 ± 0.4	12.3 ± 0.7	1.8 ± 0.4	59.8 ± 0.7
	5 dB	8.3 ± 2.3	6.9 ± 2.1	2.5 ± 0.6	12.6 ± 0.6	2.5 ± 0.6	52.5 ± 0.7
	0 dB	12.0 ± 1.0	9.1 ± 2.0	2.5 ± 0.8	12.2 ± 0.8	2.2 ± 0.8	45.2 ± 0.5
Rössler	clean	19.9 ± 5.7	10.0 ± 3.0	35.0 ± 4.9	34.9 ± 4.8	35.1 ± 4.4	7.9 ± 4.1
	20 dB	13.7 ± 4.0	1.8 ± 0.5	25.0 ± 0.0	26.5 ± 0.6	25.0 ± 0.0	7.2 ± 0.2
	15 dB	7.0 ± 3.0	8.7 ± 1.9	17.7 ± 2.7	26.3 ± 0.8	25.0 ± 0.0	9.2 ± 0.3
	10 dB	24.1 ± 2.1	14.5 ± 2.0	12.0 ± 1.4	26.8 ± 0.6	24.8 ± 0.2	12.3 ± 2.1
	5 dB	26.6 ± 0.6	20.9 ± 2.4	12.2 ± 1.3	26.5 ± 0.8	15.7 ± 2.4	16.4 ± 3.8
	0 dB	26.1 ± 0.6	24.6 ± 2.2	14.2 ± 2.0	26.6 ± 0.7	13.8 ± 1.6	17.7 ± 4.4
Duffing	clean	26.4 ± 6.2	95.0 ± 14.1	31.9 ± 3.7	28.9 ± 4.7	31.1 ± 3.7	121.3 ± 11.7
	20 dB	36.8 ± 0.7	27.1 ± 4.0	35.0 ± 0.0	36.6 ± 0.5	35.0 ± 0.0	139.3 ± 0.3
	15 dB	36.4 ± 0.7	32.6 ± 3.7	34.9 ± 0.1	35.9 ± 0.9	35.0 ± 0.0	140.3 ± 0.5
	10 dB	36.9 ± 0.5	36.3 ± 0.7	26.7 ± 2.6	36.4 ± 0.6	33.4 ± 0.3	140.9 ± 0.4
	5 dB	36.6 ± 0.9	36.6 ± 0.8	24.1 ± 1.7	36.2 ± 0.8	29.9 ± 0.6	139.0 ± 0.6
	0 dB	36.6 ± 0.5	37.1 ± 0.5	22.3 ± 1.1	36.3 ± 0.8	23.4 ± 1.3	136.7 ± 0.8
Mackey–Glass	clean	16.2 ± 5.8	7.0 ± 2.8	42.6 ± 4.2	41.0 ± 5.2	41.2 ± 4.3	10.4 ± 2.7
	20 dB	17.2 ± 1.8	15.8 ± 3.4	38.0 ± 0.0	39.2 ± 0.8	38.0 ± 0.0	11.2 ± 0.3
	15 dB	29.9 ± 4.7	20.8 ± 3.8	34.3 ± 1.7	38.6 ± 0.9	38.0 ± 0.0	16.5 ± 1.7
	10 dB	39.0 ± 1.4	30.6 ± 3.6	26.8 ± 1.2	39.1 ± 0.9	37.8 ± 0.2	21.9 ± 1.1
	5 dB	39.3 ± 0.7	38.4 ± 1.0	26.2 ± 1.0	39.1 ± 0.9	35.5 ± 0.4	27.6 ± 0.9
	0 dB	39.4 ± 0.8	40.0 ± 0.5	27.1 ± 1.2	39.2 ± 0.7	31.4 ± 0.8	33.6 ± 2.0
Chen	clean	8.1 ± 1.9	5.8 ± 1.6	16.6 ± 0.9	16.9 ± 0.9	18.0 ± 0.9	32.3 ± 0.5
	20 dB	7.2 ± 0.6	6.0 ± 0.2	15.9 ± 0.1	16.9 ± 0.8	15.9 ± 0.1	33.0 ± 0.0
	15 dB	5.8 ± 1.6	6.2 ± 0.4	13.3 ± 0.6	17.5 ± 0.7	13.9 ± 0.6	33.5 ± 0.2
	10 dB	7.7 ± 2.2	6.5 ± 0.8	9.7 ± 1.9	17.4 ± 0.8	8.8 ± 2.3	34.1 ± 0.1
	5 dB	12.0 ± 1.9	5.5 ± 1.7	5.1 ± 1.7	17.2 ± 0.9	4.5 ± 1.8	33.6 ± 0.3
	0 dB	17.2 ± 0.8	11.1 ± 1.5	3.7 ± 1.4	17.5 ± 0.7	2.6 ± 1.0	31.9 ± 0.3

This appendix collects the statistics behind Section 5.

Appendix C.1. Why the Interval Must Be Taken over Records

Table A4 compares the cluster bootstrap over records with a naive bootstrap over windows for the same out-of-fold correctness vector. On MFPT the cluster standard error is 6.7–12.0 times the window-level one, so a window-level interval would have understated the uncertainty by an order of magnitude; on the rotor set, where there are 180 independent segments, the factor falls to 2.2–2.7. The unit that resamples is the unit the design actually holds out, and that is the record.

Appendix C.2. Confusion Matrices

Table A5 gives the pooled out-of-fold confusion matrices for the two methods at the extremes of the MFPT ranking, over all twenty-five folds. The failure of C-C is concentrated on the healthy-versus-outer boundary (318 of 600 healthy windows called outer) and on healthy-versus-inner (100); LPF-SEM's errors are almost entirely inner-versus-outer ($91 + 163$ of 2700). Per-class recall ranges across the six methods are 0.530–0.992 (healthy), 0.638–0.931 (inner) and 0.481–0.866 (outer) on MFPT, and 0.907–0.987, 0.969–1.000, 0.833–0.976 and 0.982–1.000 for misalignment, normal, rub-impact and unbalance on the rotor set.

Appendix C.3. Pooled-Calibration Control

Table A6 gives the full control: one global (τ, m) per method, calibrated on the pooled calibration slices of every record, with the features re-extracted at that pair and the same cross-validation re-run.

Appendix C.4. Permutation Tests

Table A7 reports both. Test 1 shuffles the record-to-calibration assignment and re-runs the window-level (τ, m)-only control, 150 permutations; the observed value is the same control under the true assignment. Test 2 permutes the class labels at the record level and re-runs the full grouped cross-validation, 100 permutations.

Appendix C.5. Twenty-Realization Reference Statistics

Regenerating the systems from perturbed initial conditions and shifted windows, the reference criteria give $\tau^* = 16.1 \pm 0.3$ (Lorenz), 38.5 ± 9.5 (Rössler), 36.0 ± 8.0 (Duffing), 46.2 ± 9.1 (Mackey–Glass) and 20.4 ± 0.9 (Chen), and $m^* = 3.0 \pm 0.0$, 2.9 ± 0.3, 3.0 ± 0.3, 3.0 ± 0.0 and 3.1 ± 0.3. The one-standard-deviation scatter of the delay reference is therefore non-negligible for three of the five systems, which is why Section 3.3 treats it as a limitation rather than a detail.

Appendix D. Implementation Details and Reproducibility

The complete pipeline is provided to this submission: a Python package with the module that implements every estimator, the scripts that produce every table and figure in this paper, the random seeds used for the noise realizations, and a configuration file recording every constant named in Section 3. A unit-test file accompanies it and checks the three identities on which the theory rests. The two machinery datasets are located through environment variables (`MFPT_DIR` and `ROTOR_MAT`) rather than by a hard-coded path, so the pipeline can be re-run on another machine without editing code.

https://doi.org/10.3390/e28111272

Table A9. Provenance of each element of LPF-SEM. The middle column is the SEM criterion as published [9]; the right column is our filtered extension. Rows marked 'added here' are not part of the published criterion.

Element	SEM Criterion (Published)	LPF-SEM (This Work)
filter	not part of the criterion	added here: zero-phase Butterworth, order 4, `scipy.signal.butter(4,` f_c`, 'low')` + `filtfilt`
cutoff f_c	not part of the criterion	added here: adaptive from the curve's own PSD, $f_c = \mathrm{clip}(2.5 f_p, 0.02, 0.3)$, f_p = argmax PSD in $(0, 0.5)$
clipping range	not part of the criterion	added here: $(0.02, 0.3)$
crossing rule	first local extremum of H	$\lvert \mathrm{d}\tilde{H}/\mathrm{d}\tau \rvert < \gamma \max \lvert \mathrm{d}\tilde{H}/\mathrm{d}\tau \rvert$ on the filtered curve, $\gamma = 0.05$, $\tau_{\mathrm{start}} = 3$
fallback for τ	none	added here: maximizer of $\kappa = \lvert \tilde{H}'' \rvert / (1 + \tilde{H}'^2)^{3/2}$
pseudo-period test	accept T_p if CoV < 0.5	identical, applied to the minima of the filtered curve
dimension	$m = \mathrm{clip}(\mathrm{round}(\tau_{\max}/T_p) + 1, 2, 30)$	identical, except that the clip is $[2, 30]$ here and $[2, 12]$ under the harmonized control
fallback for m	design default	$m = 6$ (Section 4.4)

Software Environment

Every table and figure in this paper was produced by the deposited package under one environment, which the timing script records in its own output file: Intel Core i7-6820HQ (four physical cores, eight logical processors) at 2.70 GHz, 15.8 GB DDR4-2133, Windows 10 Pro build 19,045 (64-bit, balanced power plan), CPython 3.13.5 with NumPy 2.5.0, SciPy 1.18.1, scikit-learn 1.9.0 and matplotlib 3.11.2, on scipy-openblas 0.3.33.112.0 (OpenBLAS 0.3.33, Haswell kernel) with no BLAS thread-count override. Section 4.9 gives the same list where the wall-clock results are reported, together with the timer and the number of repeats behind each median. Absolute times will differ on other hardware, so we draw conclusions only from ratios measured on this one machine.

Table A9 separates the part of LPF-SEM that comes from the published criterion from the part we added. The symbolization, the crossing rule, the pseudo-period test and the dimension formula are taken from SEM [9]; the low-pass stage, its adaptive cutoff and its clipping range, and the curvature fallback are ours, and are specified in the table so that a reader can rebuild the variant from the two columns. One design choice deserves emphasis: where the pseudo-period test is rejected we return the default $m = 6$, and Section 4.4 shows that this fallback is what the variant reports for four of the five systems.

Unit tests cover the three identities the theory rests on, namely complement invariance of A_t, the orbit-to-pattern bijection for $m = 2, \ldots, 8$, and (8) against a Monte Carlo of independent Bernoulli signs; the plateau routine is checked against a direct pairwise-agreement implementation, which must return the same entropy.

Appendix E. Data Description

Table A10. The seventeen MFPT records as used, read from the metadata inside each file. 'Duration' is after decimation; every record is truncated by the forty-window cap.

File	Class	Samples (Raw)	f_s (Hz)	Decim.	Load (lb)	Shaft (Hz)	Duration (s)	Windows
`baseline_1.mat`	healthy	585,936	97,656	2	270	25	6.00	40
`baseline_2.mat`	healthy	585,936	97,656	2	270	25	6.00	40
`baseline_3.mat`	healthy	585,936	97,656	2	270	25	6.00	40

Table A10. *Cont.*

File	Class	Samples (Raw)	f_s (Hz)	Decim.	Load (lb)	Shaft (Hz)	Duration (s)	Windows
`OuterRaceFault_vload_1.mat`	outer	146,484	48,828	1	25	25	3.00	40
`OuterRaceFault_vload_2.mat`	outer	146,484	48,828	1	50	25	3.00	40
`OuterRaceFault_vload_3.mat`	outer	146,484	48,828	1	100	25	3.00	40
`OuterRaceFault_vload_4.mat`	outer	146,484	48,828	1	150	25	3.00	40
`OuterRaceFault_vload_5.mat`	outer	146,484	48,828	1	200	25	3.00	40
`OuterRaceFault_vload_6.mat`	outer	146,484	48,828	1	250	25	3.00	40
`OuterRaceFault_vload_7.mat`	outer	146,484	48,828	1	300	25	3.00	40
`InnerRaceFault_vload_1.mat`	inner	146,484	48,828	1	0	25	3.00	40
`InnerRaceFault_vload_2.mat`	inner	146,484	48,828	1	50	25	3.00	40
`InnerRaceFault_vload_3.mat`	inner	146,484	48,828	1	100	25	3.00	40
`InnerRaceFault_vload_4.mat`	inner	146,484	48,828	1	150	25	3.00	40
`InnerRaceFault_vload_5.mat`	inner	146,484	48,828	1	200	25	3.00	40
`InnerRaceFault_vload_6.mat`	inner	146,484	48,828	1	250	25	3.00	40
`InnerRaceFault_vload_7.mat`	inner	146,484	48,828	1	300	25	3.00	40

Table A11. Rotor set: acquisition, fault seeding and processing. The upper block is the rig and the measurement chain; the lower block is the processing and record structure of the released matrix.

Item	Value
test bench	gas-turbine scaled-model rotor test bench: rotor bench, speed controller, multi-channel data acquisition instrument, host computer
rotor	diameter 10 mm, length 850 mm, supported by bearing brackets on both sides
material	high-strength alloy steel, chosen for high-temperature, high-load and high-speed operation
drive	motor through a coupling to a mass turntable of diameter 75 mm
transducer	eddy-current displacement sensors on the stationary bearing pedestal, measuring the displacement of the rotating shaft; two channels acquired simultaneously
operating speed	3300 rpm (55 Hz), constant throughout acquisition
sampling frequency	10,240 Hz
classes	normal; rub-impact; unbalance; misalignment (45 segments each)
rub-impact seeding	a 3 g friction screw screwed in until it contacts the rotating shaft
unbalance seeding	a 5.8 g mass block fixed on the edge of the mass turntable
misalignment seeding	the relative position of the two axes on the coupling is changed
segment length	2048 samples (0.2 s); a 1024-sample window is 0.1 s, about 5.5 shaft revolutions
denoising	`db4` (Daubechies order 4), level 3, soft thresholding
threshold	universal, $\lambda_j = \hat{\sigma}_j\sqrt{2\ln N}$ with $\hat{\sigma}_j = \mathrm{MAD}(d_j)/0.6745$
boundary handling	symmetric extension
robustness control	level 4 with hard thresholding; main conclusions unchanged
windows	two non-overlapping 1024-sample windows per record, 360 total
record structure	180 independent acquisitions, not cut from continuous files; no stride, no start offset
calibration slice	whole record ($2048 < 30{,}000$)

Appendix E.1. MFPT Bearing Set

The archive distributes twenty records in four condition groups. Three baseline (healthy) records were acquired at a 270 lb load, a 25 Hz shaft rate, 97,656 Hz sampling and 6 s duration. Three outer-race records were acquired at the same operating point (270 lb, 25 Hz, 97,656 Hz, 6 s). Seven further outer-race records were acquired at 25 Hz and 48,828 Hz for 3 s, at loads of 25, 50, 100, 150, 200, 250 and 300 lb. Seven inner-race records were acquired at 25 Hz and 48,828 Hz for 3 s, at loads of 0, 50, 100, 150, 200, 250 and 300 lb.

Seventeen of the twenty are used, namely the three baseline, the seven variable-load outer-race and the seven inner-race records, listed in Table A10. The three constant-load outer-race records are not used, so that the two fault classes are each represented by the same seven-load sweep at a common sampling rate and the healthy class by the three baseline records. Every column of the table is read from the acquisition metadata stored inside each `.mat` file (`sr`, `load`, `rate`), so the table is reproducible from the released data without any external source. The baseline records store 585,936 samples at 97,656 Hz and are decimated by two, which brings them to 48,828 Hz and equalizes the bandwidth of the classes; the fault records store 146,484 samples at 48,828 Hz. Each record is cut into windows of 1024 samples with a stride equal to the window length, so the windows do not overlap, and no more than forty windows are taken per record; every record is longer than 40×1024 samples, so the cap binds and the seventeen records give the 680 windows analyzed in Section 5.1.

Appendix E.2. Bearing Geometry and Fault Characteristic Frequencies

The test bearing is designated NICE; the data sheet gives a ball diameter of 0.235, a pitch diameter of 1.245, eight rolling elements and a contact angle of 0°. The archive does not record the unit of the two diameters, and this does not prevent the characteristic frequencies from being computed: each of them depends on the diameters only through the dimensionless ratio $d/D = 0.1888$, so the ratio, not the unit, fixes them. With $f_r = 25$ Hz and the standard relations, the outer-race ball-pass frequency is $\mathrm{BPFO} = \frac{N}{2} f_r (1 - \frac{d}{D}) = 81.12$ Hz, the inner-race frequency is $\mathrm{BPFI} = \frac{N}{2} f_r (1 + \frac{d}{D}) = 118.88$ Hz, the ball spin frequency is $\mathrm{BSF} = \frac{D}{2d} f_r [1 - (d/D)^2] = 63.86$ Hz and the cage frequency is $\mathrm{FTF} = \frac{1}{2} f_r (1 - \frac{d}{D}) = 10.14$ Hz, with $N = 8$, $d = 0.235$ and $D = 1.245$. No analysis in this paper depends on these numbers, which are recorded so that the class labels can be checked against the modulation frequencies of the spectra. The released archive contains no photograph or drawing of the test rig.

Appendix E.3. Rotor Set

The rotor set was acquired on a gas-turbine scaled-model test bench consisting of a rotor test bench, a speed controller, a multi-channel data acquisition instrument and a host computer. The rotor is 10 mm in diameter and 850 mm long, supported by bearing brackets on both sides, made of high-strength alloy steel to withstand high-temperature, high-load and high-speed operation; a coupling connects it to a mass turntable of diameter 75 mm and to the drive motor. Eddy-current displacement sensors were mounted on the stationary bearing pedestal and measured the displacement of the rotating shaft; their signals were transmitted to the computer for storage, with two channels acquired simultaneously. The operating speed was held at 3300 rpm (55 Hz) and the sampling frequency was 10,240 Hz throughout.

Four rotor states were recorded: normal, rub-impact, unbalance and misalignment, forty-five segments per class. The faults were seeded physically rather than simulated. Rub-impact was produced by screwing a 3 g friction screw in until it contacted the rotating shaft. Unbalance was produced by fixing a 5.8 g mass block on the edge of the mass turntable. Misalignment was produced by changing the relative position of the two axes on the coupling. The 180 records are 180 independent acquisitions, forty-five per class, not slices cut from longer continuous files: each is a separate 2048-point acquisition window, so the windows share no samples, no analysis stride and no start offset are involved, and the two 1024-sample windows taken from a record do not overlap. At 10,240 Hz a record spans 0.2 s and a window 0.1 s, about 5.5 shaft revolutions at 3300 rpm; all delays reported for this set are in samples, and one sample is 0.098 ms.

Denoising was applied to the single-channel vibration signal with a `db4` (Daubechies order 4) wavelet at decomposition level 3. The detail coefficients d_j at level j were soft-thresholded with the universal threshold

$$\lambda_j = \hat{\sigma}_j \sqrt{2 \ln N}, \qquad \hat{\sigma}_j = \frac{\mathrm{MAD}(d_j)}{0.6745}, \tag{A1}$$

where N is the record length and MAD is the median absolute deviation, and the boundary was handled by symmetric extension. A control with a four-level decomposition and hard thresholding is reported; the main conclusions are unchanged under it. Both acquired channels were recorded, but only one entered this study, and the released matrix holds 180×2048 values for that channel.

References

1. Packard, N.H.; Crutchfield, J.P.; Farmer, J.D.; Shaw, R.S. Geometry from a time series. *Phys. Rev. Lett.* **1980**, *45*, 712–716.
2. Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*; Lecture Notes in Mathematics 898; Springer: Berlin/Heidelberg, Germany, 1981; pp. 366–381.
3. Kantz, H.; Schreiber, T. *Nonlinear Time Series Analysis*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2004.
4. Marwan, N.; Romano, M.C.; Thiel, M.; Kurths, J. Recurrence plots for the analysis of complex systems. *Phys. Rep.* **2007**, *438*, 237–329.
5. Fraser, A.M.; Swinney, H.L. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A* **1986**, *33*, 1134–1140.
6. Kennel, M.B.; Brown, R.; Abarbanel, H.D.I. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A* **1992**, *45*, 3403–3411.
7. Hong, F.; Wei, P.; Bi, S.; Beer, M. Efficient variational Bayesian model updating by Bayesian active learning. *Mech. Syst. Signal Process.* **2025**, *224*, 112113. https://doi.org/10.1016/j.ymssp.2024.112113.
8. Song, J.; Liang, Z.; Wei, P.; Beer, M. Sampling-based adaptive Bayesian quadrature for probabilistic model updating. *Comput. Methods Appl. Mech. Eng.* **2025**, *433*, 117467. https://doi.org/10.1016/j.cma.2024.117467.
9. Matilla-García, M.; Morales, I.; Rodríguez, J.M.; Marín, M.R. Selection of embedding dimension and delay time in phase space reconstruction via symbolic dynamics. *Entropy* **2021**, *23*, 221. https://doi.org/10.3390/e23020221.
10. Kim, H.S.; Eykholt, R.; Salas, J.D. Nonlinear dynamics, delay times, and embedding windows. *Physica D* **1999**, *127*, 48–60.
11. Costa, M.; Goldberger, A.L.; Peng, C.-K. Multiscale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.* **2002**, *89*, 068102.
12. Chen, W.; Wang, Z.; Xie, H.; Yu, W. Characterization of surface EMG signal based on fuzzy entropy. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2007**, *15*, 266–272.
13. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **2000**, *278*, H2039–H2049.
14. Bandt, C.; Pompe, B. Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Lett.* **2002**, *88*, 174102.
15. Grassberger, P.; Procaccia, I. Characterization of strange attractors. *Phys. Rev. Lett.* **1983**, *50*, 346–349.
16. Harris, C.R.; Millman, K.J.; Van Der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362.
17. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberl, ; M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272.
18. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
19. Lorenz, E.N. Deterministic nonperiodic flow. *J. Atmos. Sci.* **1963**, *20*, 130–141.
20. Rössler, O.E. An equation for continuous chaos. *Phys. Lett. A* **1976**, *57*, 397–398.
21. Chen, G.; Ueta, T. Yet another chaotic attractor. *Int. J. Bifurc. Chaos* **1999**, *9*, 1465–1466.
22. Mackey, M.C.; Glass, L. Oscillation and chaos in physiological control systems. *Science* **1977**, *197*, 287–289.
23. Farmer, J.D. Chaotic attractors of an infinite-dimensional dynamical system. *Physica D* **1982**, *4*, 366–393.
24. Benettin, G.; Galgani, L.; Giorgilli, A.; Strelcyn, J.-M. Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; A method for computing all of them. Part 1: Theory. *Meccanica* **1980**, *15*, 9–20.

25. Rosenstein, M.T.; Collins, J.J.; Luca, C.J.D. A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D* **1993**, *65*, 117–134.
26. Zbilut, J.P.; Webber, C.L. Embeddings and delays as derived from quantification of recurrence plots. *Phys. Lett. A* **1992**, *171*, 199–203.
27. Webber, C.L.; Zbilut, J.P. Dynamical assessment of physiological systems and states using recurrence plot strategies. *J. Appl. Physiol.* **1994**, *76*, 965–973.
28. Smith, W.A.; Randall, R.B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **2015**, *64–65*, 100–131.
29. Randall, R.B.; Antoni, J. Rolling element bearing diagnostics: A tutorial. *Mech. Syst. Signal Process.* **2011**, *25*, 485–520.
30. Society for Machinery Failure Prevention Technology. MFPT Bearing Fault Data Sets. 2013. Available online: https://www.mfpt.org/fault-data-sets/ (accessed on 28 September 2026).
31. Zhao, Z.; Li, J.; Cui, J. Feature extraction for gas turbine rotor fault diagnosis in complex networks based on phase space method. In *Proceedings of the International Conference on Mechanical Design and Simulation (ICMDS), Xi'an, China*; Springer: Berlin/Heidelberg, Germany, 2025.
32. Theiler, J. Spurious dimension from correlation algorithms applied to limited time-series data. *Phys. Rev. A* **1986**, *34*, 2427–2432.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.