



BANK ĊENTRALI TA' MALTA  
EUROSISTEMA  
CENTRAL BANK OF MALTA

# Causality-Preserving Information-Theoretic Smoothing for Financial Time Series

A Rigorous RKHS Framework with Heteroskedastic Calibration  
and Cross-Regime Sensitivity Analysis<sup>1</sup>

Mohamed Yagob<sup>2</sup>

---

<sup>1</sup>I would like to thank Prof. David Suda for reviewing an earlier version of this paper and suggesting areas for improvement. I would also like to thank Wendy Zammit, Andrew Spiteri, Karen Caruana, and Jesmond Pule for their continued support during this work. I am also grateful for the opportunity to present this paper at the ISI RSC 2026 in Malta and would like to thank the participants for their valuable feedback and constructive comments. I also wish to thank the members of the Machine Learning and Big Data Network of the Central Bank of Malta for their comments and feedback.

The views expressed in this paper are the author's own and do not necessarily reflect the views of the Central Bank of Malta. Correspondence: [mohamedy@centralbankmalta.org](mailto:mohamedy@centralbankmalta.org)

<sup>2</sup>Senior Economist Analyst – Financial Stability Surveillance and Risk Assessment Department, Central Bank of Malta

## Abstract

This paper develops a self-contained mathematical framework for smoothing heteroskedastic financial time series within a reproducing kernel Hilbert space (RKHS) setting. The central object is the *Information Flow Kernel*, a positive definite kernel constructed as the pointwise (Hadamard) product of the Brownian motion covariance and an Ornstein–Uhlenbeck correlation function, whose positive definiteness is established rigorously via the Schur product theorem. Building on the Kimeldorf–Wahba correspondence between Bayesian estimation and regularisation, we derive an entropy-regularised objective that jointly optimises data fidelity, temporal smoothness, and divergence from a linear trend prior. We provide complete, self-contained proofs of existence and uniqueness via the direct method in the calculus of variations, a representer theorem for a direct-sum RKHS, a closed-form matrix solution, and minimax-optimal convergence rates derived from the spectral analysis of the associated integral operator.

A methodological point governs how the empirical results should be read: neither the Information Flow Smoother (IFS) nor Wahba’s cubic smoothing spline is a forecasting method. Both are *retrospective* smoothers whose hat matrix is defined over the full observed window simultaneously; there is no causal fitting structure. Empirical comparisons are therefore conducted on three distinct criteria: in-sample volatility-weighted residual calibration, which is the native operating criterion of the IFS; regime-stratified fit quality, which assesses performance under heteroskedastic stress; and an *anchoring stability test*, which asks whether better heteroskedastic calibration produces a more stable terminal fitted value when the smoother is extrapolated linearly into a hold-out window.

We evaluate all three criteria across three train/test splits spanning qualitatively different market regimes: a full bull-bear-bull cycle (2015–2022 train), the post-COVID recovery period (2015–2020 train), and the Federal Reserve rate-hiking stress episode (2015–mid-2022 train), and against two Wahba benchmarks: the classical uniformly-weighted cubic smoothing spline, and a heteroskedastic variant fitted with the same precision weights as the IFS, which isolates the marginal contribution of the Information Flow kernel from that of reweighting alone. The IFS achieves lower volatility-weighted error than both benchmarks across all three regimes, at its own GCV-optimal flexibility, at matched effective degrees of freedom, and with the advantage confirmed as statistically robust to the serial dependence of daily returns via block-bootstrap confidence intervals. The sensitivity heatmap confirms that the IFS advantage on vol-weighted criteria is a structural feature of the heteroskedastic weight design rather than an artefact of any particular train/test partition; this evidence is drawn from a single equity index and its generalisation to other assets remains to be established.

# 1 Introduction

Financial time series present a well-known cluster of challenges that sit uncomfortably within the classical nonparametric smoothing framework pioneered by Wahba [1990]. The fundamental difficulty is not merely statistical but structural: the data-generating process is non-stationary, exhibits pronounced heteroskedasticity driven by volatility clustering Engle [1982], and embeds a long-run upward trend that is economically meaningful and should not be indiscriminately smoothed away. A classical cubic smoothing spline, whose penalty functional treats all time points symmetrically with equal squared second derivative cost, is therefore a poor model for equity price trajectories. It has no mechanism to know that a large residual in March 2020 is mostly noise while a comparably large residual in a tranquil 2017 session is mostly signal.

The purpose of this paper is to address this inconsistency systematically while maintaining mathematical rigour. We work throughout in the RKHS setting, which provides both a principled variational basis for penalised estimation and a direct connection to Bayesian Gaussian process regression Kimeldorf and Wahba [1970], Rasmussen and Williams [2006]. The central methodological contribution is a new covariance kernel for this setting, which we call the *Information Flow Kernel*. Informally, it is built from two ingredients acting together: a component that accumulates market information monotonically from a fixed origin, so that two time points sharing a longer common history are more strongly correlated, and a second component that discounts this shared history over a finite memory horizon, so that observations far apart in time become nearly independent regardless of how much history they share. The combination is deliberately asymmetric in time and encodes, directly in the geometry of the function space, the idea that recent information is more informative about the present level of the series than distant information of the same age. Section 3 gives the precise construction and establishes its positive definiteness rigorously via the Schur product theorem.

The information-theoretic component enters through an entropy-regularised objective that incorporates heteroskedastic observation precision weights—motivated directly by the local Gaussian information distance—and a Kullback–Leibler penalty measuring departure from a linear trend prior. The combined functional is strictly convex, coercive, and weakly lower semi-continuous on the RKHS, so a unique minimiser exists. Existence, uniqueness, the representer theorem, and the minimax convergence rate are then consequences of standard variational, RKHS, and empirical-process arguments once the Information Flow Kernel and the heteroskedastic objective are in place; we state these results precisely and give short proofs adapted to the present direct-sum setting, but we make no claim of novelty for the underlying machinery itself, only for its assembly around the new kernel and its application to heteroskedastic financial smoothing.

The paper’s empirical contributions, established across three distinct market regimes, are therefore the following. First, the IFS achieves lower *volatility-weighted* in-sample error than both Wahba benchmarks, the criterion economically appropriate for evaluation when observation noise varies substantially across the sample. Second, this advantage is most pronounced during high-volatility stress periods, exactly the episodes where credible trend estimation matters most for risk management and financial stability surveillance. Third, the advantage is structural rather than incidental: a sensitivity analysis across three train/test splits spanning qualitatively different market regimes confirms that the IFS’s heteroskedastic weight design consistently outperforms on vol-weighted criteria regardless of the particular split chosen.

The paper is organised as follows. Section 2 reviews the relevant literature. Section 3 establishes the mathematical building blocks. Section 4 introduces the entropy-regularised objective. Sections 5–6 contain the representer theorem and closed-form solution. Section 7 provides the complete theoretical analysis. Section 8 analyses the causal structure of the kernel. Section 9 describes the computational algorithm, and Section 10 reports empirical results on S&P 500 data across three market regimes.

## 2 Literature Review

The modern theory of smoothing splines traces its origins to the regularisation framework of Wahba [1990], which placed the classical polynomial smoothing spline in the setting of reproducing kernel Hilbert spaces. A foundational account of RKHS theory in probability and statistics is given by Berlinet and Thomas-Agnan [2004], which we follow for all definitions and standard results. The pivotal *Kimeldorf–Wahba correspondence* Kimeldorf and Wahba [1970] establishes that the penalised least-squares estimator in an RKHS is simultaneously the posterior mean of a Gaussian process with the corresponding kernel as covariance function, a duality that grounds much of our development. The definitive modern treatment of Gaussian process regression is Rasmussen and Williams [2006], which provides the spectral and approximation-theoretic results we draw upon in the convergence analysis.

The representer theorem, which reduces infinite-dimensional optimisation problems to finite-dimensional linear systems, was first proved in the smoothing spline context by Wahba [1990] and later generalised to arbitrary RKHS and convex loss functions by Schölkopf et al. [2001]. Our version requires a modest extension to handle the direct-sum space  $\mathcal{H}_{K_{\text{flow}}} \oplus \mathcal{H}_{K_{\text{trend}}}$ , for which we adapt the argument of Berlinet and Thomas-Agnan [2004].

Minimax convergence rates for nonparametric estimators in Sobolev-type spaces are classical; the definitive reference for the minimax theory we employ is Tsybakov [2009]. The connection between eigenvalue decay of the integral operator associated with a kernel and the corresponding optimal convergence rate is discussed in Rasmussen and Williams [2006] and rigorously developed in Steinwart and Scovel [2012]. Fischer and Steinwart [2020] establish optimal learning rates for regularised least-squares in RKHS under the interpolation space framework, characterising precisely how the effective smoothness of the target function and the eigenvalue decay of the kernel jointly determine the minimax rate; their results subsume the classical Wahba rates and clarify conditions under which the  $O(n^{-2/3})$  rate derived in Theorem 6 is tight.

The econometric literature on heteroskedastic time series began in earnest with Engle [1982], whose ARCH model demonstrated that volatility clustering is a fundamental and persistent feature of financial returns. The subsequent GARCH literature Bollerslev [1986] formalised this into a complete framework for time-varying conditional variance, which provides the statistical motivation for our heteroskedastic precision weights  $\sigma_i^{-2}$ . The realised volatility literature, initiated by Andersen et al. [2003], provides model-free estimates of integrated variance that can serve as direct inputs to those weights without requiring parametric assumptions about the variance process. Gatheral et al. [2018] provide compelling empirical evidence that log-volatility behaves like a fractional Brownian motion with Hurst exponent  $H \approx 0.1$ , substantially rougher than the standard Brownian case implicitly assumed by GARCH-type models, which has direct implications for how

any smoothness prior over the variance process should be calibrated. Heteroskedastic nonparametric smoothing has been studied since [Eubank \[1984\]](#) and [Fan and Gijbels \[1996\]](#); our use of a non-stationary kernel addresses heteroskedasticity at the level of the function space rather than through post-hoc reweighting. More recently, [Wong-Toi et al. \[2025\]](#) show that mean-variance nonparametric regression is subject to a sharp signal-to-noise phase transition governed by the level of regularisation, with under-regularised models explaining all variation as signal and over-regularised models explaining it all as noise—the same tension we document empirically in [Section 10](#) between the near-interpolating GCV-optimal Wahba spline and its more heavily regularised variants.

Non-stationary covariance functions for Gaussian process regression were systematically introduced by [Paciorek and Schervish \[2004\]](#), who showed that a broad class of non-stationary kernels can be constructed from stationary correlation functions by allowing the characteristic length-scale to vary spatially. Their framework motivates our use of the exponential decay factor  $\exp(-|s - t|/\tau)$  as the correlation function of an Ornstein–Uhlenbeck process, which has a known and well-studied RKHS structure [Rasmussen and Williams \[2006\]](#). The Schur product theorem [Horn and Johnson \[2012\]](#) is the natural algebraic underpinning for our product-kernel construction. The connection to state-space representations is of considerable computational importance: [Hartikainen and Särkkä \[2010\]](#) establish that for kernels whose spectral density is a rational function of frequency, GP regression reduces exactly to a linear-Gaussian state-space model solvable in  $O(n)$  time via the Rauch–Tung–Striebel smoother. Since  $K_{\text{flow}} = K_{\text{BM}} \cdot K_{\text{OU}}$  and both factors correspond to solutions of first-order autonomous SDEs, a complete  $O(n)$  Kalman representation is in principle derivable, and we flag this as a natural direction for future work.

The computational challenge of GP regression with large  $n$  has motivated a substantial literature on sparse and structured approximations. [Titsias \[2009\]](#) introduces the variational inducing-point framework, which approximates the full GP posterior with a lower-rank surrogate, reducing the  $O(n^3)$  computation to  $O(nm^2)$ . [Hensman et al. \[2013\]](#) extend this to stochastic variational inference for datasets with millions of observations. [Burt et al. \[2020\]](#) provide a rigorous convergence analysis for sparse variational GP inference, establishing conditions under which the inducing-point approximation achieves the same generalisation bounds as the exact GP; their results are directly relevant to the Nyström low-rank approximation deployed in [Section 9](#). For structured kernels admitting Kronecker or Toeplitz decompositions, [Wilson and Nickisch \[2015\]](#) show that kernel matrix-vector products can be computed in  $O(n \log n)$  via the Fast Fourier Transform. [Wainwright \[2019\]](#) provides a non-asymptotic treatment of regularisation in high-dimensional and function spaces that complements the classical minimax approach. [Wilson and Adams \[2013\]](#) introduce the Spectral Mixture kernel, which provides a data-driven alternative to the GCV grid search for bandwidth selection. [Liu et al. \[2020\]](#) provide a recent systematic survey of scalable GP methods and their theoretical guarantees.

The information-geometric perspective, developed in [Amari \[2016\]](#), provides deeper grounding for the KL penalty in our objective: the KL divergence is the canonical squared distance between distributions on the Riemannian manifold of a statistical model, so the KL regulariser is interpretable as a constraint that the fitted signal should not deviate too far from the trend prior in the information-geometric sense. [Alquier \[2024\]](#) develops the PAC-Bayes framework for estimators with KL-type regularisers, providing finite-sample risk bounds that could in principle complement our minimax rates with non-asymptotic guarantees. The problem of recovering a smooth latent trajectory from discrete noisy

observations also falls within the scope of functional data analysis [Ramsay and Silverman \[2005\]](#), and [Müller et al. \[2011\]](#) provides a directly comparable use case in the form of yield curve dynamics.

Among very recent contributions most directly relevant to our framework, [Ober and Rasmussen \[2021\]](#) provide a systematic study of GP uncertainty quantification in interpolation regimes that clarifies when the Kimeldorf–Wahba posterior is well-calibrated and when it is not; their findings suggest that calibration improves precisely when the observation weights are informative about the local signal-to-noise ratio, which is the mechanism underlying our heteroskedastic design. [Wild et al. \[2022\]](#) develop a kernel method for quantile regression under heteroskedasticity that shares our motivation of encoding variance non-stationarity directly in the RKHS norm rather than through external reweighting. The stylised facts most relevant to the present work are catalogued comprehensively by [Cont \[2001\]](#), whose empirical benchmarks provide a natural reference for the residual diagnostics of Section 10.

**Scope of the present contribution.** Given how much of the machinery above this paper draws on, it is worth stating plainly, before the technical development begins, exactly what is being claimed as new and what is not. The Information Flow Kernel of Section 3 and its use as the fidelity-and-penalty-consistent basis for heteroskedastic smoothing (Section 4) are, to our knowledge, new. The positive-definiteness argument via the Schur product theorem, the representer theorem for the resulting direct-sum RKHS, the existence-uniqueness argument, and the minimax convergence rate (Sections 5–7) are all instances of standard RKHS, calculus-of-variations, and empirical-process results—respectively due to [Horn and Johnson \[2012\]](#), [Berlinet and Thomas-Agnan \[2004\]](#), [Schölkopf et al. \[2001\]](#), standard convex-analytic arguments, and [Tsybakov \[2009\]](#), [Fischer and Steinwart \[2020\]](#)—specialised to the new kernel and the heteroskedastic direct-sum objective. We prove these results explicitly, both for completeness and because the direct-sum structure and the non-stationary kernel require a modest but non-vacuous adaptation of each standard argument, but we do not present any of them as a theoretical contribution in its own right. The empirical contribution (Section 10) is the demonstration, across three market regimes and against a volatility-aware benchmark, that the resulting estimator achieves a real and statistically supported improvement on the criterion it is designed for.

## 3 Mathematical Foundations

### 3.1 Time Domain and Observation Model

Throughout the paper, the time domain is the compact interval  $\mathcal{T} = [0, T]$  for some  $T > 0$ . We observe  $n$  pairs  $\{(t_i, y_i)\}_{i=1}^n$  at distinct design points  $0 \leq t_1 < t_2 < \dots < t_n \leq T$ , where

$$y_i = f_0(t_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad \text{independently.}$$

The variances  $\sigma_i^2 > 0$  are assumed known (or estimated in a preliminary step, for instance from an EWMA or GARCH model) and are allowed to differ across observations, capturing heteroskedasticity. The function  $f_0 \in \mathcal{H}$  is the unknown signal we wish to recover.

### 3.2 Information Distance

A signal value  $f(t_i)$  is scored against the observation  $y_i$  through a precision-weighted discrepancy, motivated directly by the Gaussian observation model just introduced: weighting each squared residual by  $\sigma_i^{-2}$  makes the observation noise itself the natural unit of error. Formalised, this weighting is the pointwise information distance.

**Definition 1** (Pointwise Information Distance). For an observation  $y \in \mathbb{R}$ , a signal value  $f \in \mathbb{R}$ , and a precision parameter  $\sigma^{-2} > 0$ , the *pointwise information distance* is

$$I(y, f; \sigma) = \frac{1}{2} \log(2\pi\sigma^2) + \frac{(y - f)^2}{2\sigma^2}. \quad (1)$$

This is the negative log-likelihood of the Gaussian observation model  $y \mid f \sim \mathcal{N}(f, \sigma^2)$ . Its relevance as a proper notion of “distance” in the information-geometric sense follows from the fact that it equals  $-\log p(y \mid f, \sigma) + \text{const}$ , where the constant does not depend on  $f$ .

**Lemma 1** (Strict Convexity and Smoothness of the Information Distance). *The function  $f \mapsto I(y, f; \sigma)$  is strictly convex and  $C^\infty$  on  $\mathbb{R}$ . Its second derivative with respect to  $f$  is*

$$\frac{\partial^2 I}{\partial f^2}(y, f; \sigma) = \frac{1}{\sigma^2} > 0. \quad (2)$$

*Proof.* Differentiating (1) twice with respect to  $f$ :

$$\frac{\partial I}{\partial f} = -\frac{y - f}{\sigma^2}, \quad \frac{\partial^2 I}{\partial f^2} = \frac{1}{\sigma^2}.$$

Since  $\sigma^2 > 0$ , the Hessian is strictly positive everywhere, so  $I(\cdot, f; \sigma)$  is strictly convex and, being a quadratic function of  $f$ , is  $C^\infty$ .  $\square$

Strict convexity here is what ultimately makes the full estimator well-posed: combined with the RKHS penalties, it is this property that pins down a unique minimiser in Theorem 5. What remains open is the covariance structure that minimiser should live in.

### 3.3 The Information Flow Kernel

Before stating the construction formally, it is worth fixing the intuition it is meant to capture. We want a notion of covariance between two observation times  $s$  and  $t$  that reflects *causal information accumulation*: the amount of market information the two times have jointly accumulated since a common origin, discounted by how far apart in time they are. A natural measure of shared history is  $\min(s, t)$ , the length of the time interval both  $s$  and  $t$  have had to accumulate information from time 0; this is exactly the covariance kernel of standard Brownian motion. On its own, however, this factor has infinite memory: two very distant times with a long shared prefix would remain strongly correlated indefinitely, which is not a plausible description of a market that reprices continuously. We therefore discount the shared history by an exponentially decaying factor in the elapsed time  $|s - t|$ , governed by a memory-horizon parameter  $\tau > 0$ , so that closely spaced observations remain strongly correlated while distant ones become nearly independent regardless of how much history they share. Multiplying the two factors together, rather than adding them, is what makes the memory horizon act as a genuine discount on shared history instead of an independent additive correlation structure. This is the construction we now make precise.

**Definition 2** (Information Flow Kernel). Let  $\tau > 0$  be a memory-horizon parameter. The *Information Flow Kernel* is the function  $K_{\text{flow}} : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}$  defined by

$$K_{\text{flow}}(s, t) = \min(s, t) \cdot \exp\left(-\frac{|s - t|}{\tau}\right). \quad (3)$$

The kernel is the pointwise (Hadamard) product of two well-known kernels:

$$K_{\text{BM}}(s, t) = \min(s, t) \quad \text{and} \quad K_{\text{OU}}(s, t) = \exp\left(-\frac{|s - t|}{\tau}\right).$$

The first is the covariance kernel of standard Brownian motion; the second is the covariance kernel of the stationary Ornstein–Uhlenbeck (OU) process. Before proving positive definiteness, we recall the Schur product theorem in the form we need.

**Lemma 2** (Schur Product Theorem, [Horn and Johnson 2012](#)). Let  $K_1, K_2 : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}$  be positive definite kernels on a set  $\mathcal{T}$ . Then their pointwise product  $K_1 \cdot K_2$ , defined by  $(K_1 \cdot K_2)(s, t) = K_1(s, t)K_2(s, t)$ , is also a positive definite kernel on  $\mathcal{T}$ .

This is the classical Schur (Hadamard) product theorem for positive semi-definite matrices, applied pointwise to the Gram matrices of  $K_1$  and  $K_2$  at any finite set of points; we do not reproduce its proof and refer the reader to [Horn and Johnson \[2012\]](#) for the underlying matrix-analytic argument (Cholesky factorisation of the two Gram matrices followed by a sum-of-squares identity for the Hadamard product). What is specific to the present paper is not this classical fact but its application: taking  $K_1 = K_{\text{BM}}$  and  $K_2 = K_{\text{OU}}$  in Theorem 1 below.

**Theorem 1** (Positive Definiteness of the Information Flow Kernel).  $K_{\text{flow}}$  is a continuous, symmetric, and strictly positive definite kernel on  $\mathcal{T} = [0, T]$ .

*Proof. Symmetry.* Both  $K_{\text{BM}}(s, t) = \min(s, t)$  and  $K_{\text{OU}}(s, t) = e^{-|s-t|/\tau}$  are symmetric, so their product is symmetric.

**Continuity.** The function  $(s, t) \mapsto \min(s, t)$  is Lipschitz continuous on  $\mathcal{T} \times \mathcal{T}$ , and  $(s, t) \mapsto e^{-|s-t|/\tau}$  is  $C^\infty$ . Their product is therefore continuous.

**Positive semi-definiteness.** Both factor kernels are positive definite:

1. *Brownian motion kernel:* For any  $c \in \mathbb{R}^n$  and  $t_1 < \dots < t_n$ ,

$$\sum_{i,j} c_i c_j \min(t_i, t_j) = \int_0^T \left( \sum_{i=1}^n c_i \mathbf{1}_{[u \leq t_i]} \right)^2 du \geq 0.$$

2. *Ornstein–Uhlenbeck kernel:* The function  $\phi(r) = e^{-r/\tau}$  for  $r \geq 0$  is completely monotone (Bernstein’s theorem), so  $K_{\text{OU}}(s, t) = \phi(|s - t|)$  is positive definite on  $\mathbb{R}$  by Schoenberg’s theorem [Berg et al. \[1984\]](#).

By Lemma 2, the pointwise product  $K_{\text{flow}} = K_{\text{BM}} \cdot K_{\text{OU}}$  is positive semi-definite.

**Strict positive definiteness.** Since  $K_{\text{OU}}$  is strictly positive definite on any finite set of distinct points (its Gram matrix is the inverse of a strictly diagonally dominant matrix [Rasmussen and Williams \[2006\]](#)), and  $K_{\text{flow}}(t_i, t_j) = t_i e^{-(t_j - t_i)/\tau}$  for  $t_i \leq t_j$  is the Hadamard product of a positive definite matrix with a strictly positive definite one, the resulting matrix has all strictly positive eigenvalues.  $\square$

*Remark 1.* The positive definiteness argument above is purely algebraic; it does not by itself explain why the product structure is the right one economically. That motivation was given before Definition 2: the kernel is causal information accumulation ( $\min(s, t)$ ) discounted by an exponential memory horizon ( $\tau$ ). Theorem 7 below makes the causal, forward-looking character of this discounting precise.

### 3.4 The Trend Prior Kernel and Its RKHS

$K_{\text{flow}}$  is deliberately short-memory, decaying to near zero once  $|s - t|$  passes a few multiples of  $\tau$  — appropriate for the transient fluctuations it is built to capture, but useless for representing the steady upward drift equity prices exhibit over the full sample window. A second kernel, purpose-built to carry that drift, is what Section 4 adds alongside it in a direct sum.

**Definition 3** (Trend Kernel). The *trend prior kernel* is

$$K_{\text{trend}}(s, t) = \sigma_{\text{trend}}^2 \min(s, t), \quad (4)$$

where  $\sigma_{\text{trend}}^2 > 0$  is a prior variance hyperparameter.

This is the covariance kernel of a scaled Brownian motion process; its RKHS  $\mathcal{H}_{K_{\text{trend}}}$  consists of absolutely continuous functions  $f : [0, T] \rightarrow \mathbb{R}$  with  $f(0) = 0$  and  $\int_0^T [f'(t)]^2 dt < \infty$ , equipped with the inner product  $\langle f, g \rangle_{K_{\text{trend}}} = \sigma_{\text{trend}}^{-2} \int_0^T f'(t)g'(t) dt$ . The corresponding trend function

$$f_{\text{trend}}(t) = \mu t, \quad \mu > 0, \quad (5)$$

represents the expected long-run upward drift of the market, with  $\mu$  estimated from the data as the annualised sample mean return.

## 4 The Entropy-Regularised Information Smoothing Framework

### 4.1 The Direct-Sum Hypothesis Space

Since our regulariser involves two distinct kernels, the natural hypothesis space is the direct sum of the corresponding RKHSs:

$$\mathcal{H} = \mathcal{H}_{K_{\text{flow}}} \oplus \mathcal{H}_{K_{\text{trend}}},$$

where elements  $f \in \mathcal{H}$  are written  $f = f_1 + f_2$  with  $f_1 \in \mathcal{H}_{K_{\text{flow}}}$ ,  $f_2 \in \mathcal{H}_{K_{\text{trend}}}$ , and the components are orthogonal:  $\langle f_1, f_2 \rangle_{\mathcal{H}} = 0$ . The combined kernel is  $K(s, t) = K_{\text{flow}}(s, t) + K_{\text{trend}}(s, t)$ , which is positive definite as a sum of two positive definite kernels. This construction is standard; see [Berlinet and Thomas-Agnan \[2004\]](#), Chapter 2.

### 4.2 The KL Penalty and Its Quadratic Reduction

The regulariser has a Bayesian motivation: we place a Gaussian process prior  $f_2 \sim \mathcal{GP}(f_{\text{trend}}, K_{\text{trend}})$  on the trend component. Under this prior and the Gaussian observation model, the KL divergence from the prior process to the posterior process has a closed, computable form.

**Definition 4** (Gaussian Process Measures). For a mean-zero GP with covariance  $K_{\text{flow}}$  and observation noise  $\sigma_i^2$ , let  $P_f$  denote the law of the observation vector  $(f(t_1), \dots, f(t_n))$  under the model with signal  $f$ . Let  $P_{\text{trend}}$  denote the law of the same vector under the GP with mean  $f_{\text{trend}}$  and covariance  $K_{\text{trend}}$ .

*Remark 2* (Finite-Dimensional Nature of Definition 4). Both  $P_f$  and  $P_{\text{trend}}$  are, by construction, laws of the  $n$ -dimensional observation vector at the fixed design points  $t_1, \dots, t_n$ , not measures on a path space or function space. The KL divergence in Lemma 3 below is therefore an ordinary finite-dimensional Gaussian KL divergence throughout; we make no claim about, and do not need, any infinite-dimensional Gaussian measure theory (e.g. the Cameron–Martin theorem) to justify it. The only place the RKHS structure of  $\mathcal{H}_{K_{\text{trend}}}$  enters is in the final step, where the reproducing property identifies the finite-dimensional quadratic form with the RKHS norm.

**Lemma 3** (Finite-Dimensional KL Reduction to RKHS Norm). *Let  $f_2 \in \mathcal{H}_{K_{\text{trend}}}$ . Under the Gaussian process prior  $P_{\text{trend}} = \mathcal{GP}(f_{\text{trend}}, K_{\text{trend}})$ , the KL divergence between the  $n$ -dimensional laws of Definition 4 satisfies*

$$D_{\text{KL}}(P_{f_2} \| P_{\text{trend}}) = \frac{1}{2} \|f_2 - f_{\text{trend}}\|_{K_{\text{trend}}}^2 + \text{const}, \quad (6)$$

where the constant is independent of  $f_2$  and  $f_{\text{trend}}$ .

*Proof.* This is the standard formula for the KL divergence between two Gaussian vectors with equal covariance, specialised to the present setting. For  $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma)$  and  $\mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma)$ ,  $D_{\text{KL}}(\mathbf{u} \| \mathbf{v}) = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ . Identify the mean vectors with  $f_2(t_i)$  and  $f_{\text{trend}}(t_i)$ , and the covariance matrix with  $[K_{\text{trend}}(t_i, t_j)]$ ; the reproducing property of  $\mathcal{H}_{K_{\text{trend}}}$  identifies the resulting quadratic form  $(\mathbf{f}_2 - \mathbf{f}_{\text{trend}})^\top [K_{\text{trend}}(t_i, t_j)]^{-1}(\mathbf{f}_2 - \mathbf{f}_{\text{trend}})$  with the squared RKHS norm  $\|f_2 - f_{\text{trend}}\|_{K_{\text{trend}}}^2$ , which gives the stated result.  $\square$

### 4.3 The Primary Optimisation Problem

Fit to the data, closeness of the trend component to its prior, and smoothness of the flow component in its own RKHS norm are, so far, three separate quantities. The estimator actually used in the paper combines all three into a single objective functional, trading them off against one another through the hyperparameters  $\eta$  and  $\lambda$ .

**Definition 5** (Entropy-Regularised Information Smoothing). The *entropy-regularised information smoothing estimator* is

$$f^* = \operatorname{argmin}_{f \in \mathcal{H}} J[f], \quad (7)$$

where

$$J[f] = \sum_{i=1}^n I(y_i, f(t_i); \sigma_i) + \eta D_{\text{KL}}(P_{f_2} \| P_{\text{trend}}) + \lambda \|f_1\|_{K_{\text{flow}}}^2, \quad (8)$$

with  $f = f_1 + f_2$ ,  $\eta > 0$ , and  $\lambda > 0$ .

$J[f]$  as written is a sum of a log-likelihood and a KL divergence — two expressions that look nothing alike algebraically, even though both are secretly quadratic in  $f$ . Making that shared structure explicit is the point of the next theorem, and it is the resulting quadratic form, not (8) itself, that every later derivation in the paper actually manipulates.

**Theorem 2** (Equivalent Quadratic Formulation). *The functional (8) is equivalent, up to constants, to*

$$J[f] = \sum_{i=1}^n \frac{(y_i - f(t_i))^2}{2\sigma_i^2} + \frac{\eta}{2} \|f_2 - f_{\text{trend}}\|_{K_{\text{trend}}}^2 + \lambda \|f_1\|_{K_{\text{flow}}}^2. \quad (9)$$

*Proof.* Apply Definition 1 to expand  $I(y_i, f(t_i); \sigma_i) = \frac{1}{2} \log(2\pi\sigma_i^2) + (y_i - f(t_i))^2/(2\sigma_i^2)$ . The logarithmic term is a constant with respect to  $f$  and is dropped. Apply Lemma 3 to replace the KL term with the RKHS norm. The result is (9).  $\square$

## 5 Representer Theorem

$\mathcal{H}$  is infinite-dimensional, and nothing said so far rules out a minimiser that depends on the data through some intractable, infinite-dimensional object. It does not: extending the classical single-kernel argument to the present direct-sum space, the representer theorem below pins the solution to a span of only  $2n$  kernel sections.

**Theorem 3** (Representer Theorem for the Direct-Sum Space). *The solution  $f^* = f_1^* + f_2^*$  to (7) satisfies*

$$f^*(t) = \sum_{i=1}^n \alpha_i K_{\text{flow}}(t, t_i) + \sum_{i=1}^n \beta_i K_{\text{trend}}(t, t_i), \quad (10)$$

for some  $\alpha, \beta \in \mathbb{R}^n$ .

*Proof.* Write any  $f_1 \in \mathcal{H}_{K_{\text{flow}}}$  as  $f_1 = f_1^{\parallel} + f_1^{\perp}$ , where  $f_1^{\parallel} \in \text{span}\{K_{\text{flow}}(\cdot, t_i)\}$  and  $f_1^{\perp} \perp f_1^{\parallel}$  in  $\mathcal{H}_{K_{\text{flow}}}$ . By the reproducing property,  $f_1^{\perp}(t_i) = \langle f_1^{\perp}, K_{\text{flow}}(\cdot, t_i) \rangle_{\mathcal{H}_{K_{\text{flow}}}} = 0$  for all  $i$ . Therefore  $f_1(t_i) = f_1^{\parallel}(t_i)$  for all  $i$ , so the data-fidelity term depends only on  $f_1^{\parallel}$ . However,

$$\|f_1\|_{K_{\text{flow}}}^2 = \|f_1^{\parallel}\|_{K_{\text{flow}}}^2 + \|f_1^{\perp}\|_{K_{\text{flow}}}^2 \geq \|f_1^{\parallel}\|_{K_{\text{flow}}}^2,$$

with equality if and only if  $f_1^{\perp} = 0$ . Since  $f_1^{\perp}$  increases the penalty without affecting the data fit, the minimiser satisfies  $f_1^* \in \text{span}\{K_{\text{flow}}(\cdot, t_i)\}$ . The identical argument applies to  $\mathcal{H}_{K_{\text{trend}}}$ .  $\square$

## 6 Closed-Form Solution

Only the  $2n$  coefficients  $(\alpha, \beta)$  remain to be found, and by Theorem 3 they enter the quadratic objective (9) linearly, through  $f(t_i) = (K\alpha + K^{\text{tr}}\beta)_i$ . What follows is then routine: differentiate, set to zero, and solve.

**Theorem 4** (Matrix Solution). *Let  $K \in \mathbb{R}^{n \times n}$  with  $K_{ij} = K_{\text{flow}}(t_i, t_j)$ , let  $K^{\text{tr}} \in \mathbb{R}^{n \times n}$  with  $K_{ij}^{\text{tr}} = K_{\text{trend}}(t_i, t_j)$ , let  $W = \text{diag}(1/\sigma_i^2)$ , and let  $\mathbf{y} = (y_1, \dots, y_n)^{\top}$ . Then the representer coefficients satisfy*

$$\begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} = \left[ \tilde{K}^{\top} W \tilde{K} + R \right]^{-1} \tilde{K}^{\top} W \mathbf{y}, \quad (11)$$

where  $\tilde{K} = [K \ K^{\text{tr}}] \in \mathbb{R}^{n \times 2n}$  and

$$R = \begin{pmatrix} 2\lambda K & 0 \\ 0 & \eta K^{\text{tr}} \end{pmatrix}.$$

In the practically important case where the data are de-trended (so the trend component is removed in preprocessing), the solution simplifies to

$$\alpha^* = (WK + \eta(K^{\text{tr}})^{-1} + 2\lambda I)^{-1} W\mathbf{y}, \quad (12)$$

provided  $K^{\text{tr}}$  is invertible.

*Proof.* Substitute (10) into (9). Writing  $f(t_i) = (K\alpha + K^{\text{tr}}\beta)_i$ , the functional becomes a quadratic form in  $(\alpha, \beta)$ . Differentiating and setting to zero yields (11). For de-trended data the  $\beta$  term decouples and the system reduces to (12).  $\square$

*Remark 3.* In practice, one works with the log-price series after removing a linear trend via ordinary least squares, so the system reduces to (12). This is the form implemented in the computational experiments of Section 10.

## 7 Existence, Uniqueness, Stability, and Convergence

Three loose ends remain from the last two sections: whether (7) actually has a unique solution, how well the linear system of Theorem 4 is conditioned, and how fast  $\hat{f}_n$  approaches  $f_0$  as more data arrive. We take these in order.

### 7.1 Existence and Uniqueness

**Theorem 5** (Existence and Uniqueness). *For any  $\lambda > 0$ ,  $\eta > 0$ , and observations  $\{(t_i, y_i, \sigma_i)\}$  with  $\sigma_i > 0$ , the problem (7) admits a unique solution  $f^* \in \mathcal{H}$ .*

*Proof.* The functional  $J : \mathcal{H} \rightarrow \mathbb{R}$  is strictly convex (each data-fidelity term is strictly convex by Lemma 1, and the RKHS norms are strictly convex squared Hilbert space norms), coercive (as  $\|f\|_{\mathcal{H}} \rightarrow \infty$ , at least one component norm diverges so  $J[f] \rightarrow \infty$ ), and weakly lower semi-continuous (squared norms are weakly lower semi-continuous by Mazur’s lemma, and evaluation functionals are continuous in the RKHS topology by the reproducing property [Berlinet and Thomas-Agnan \[2004\]](#)). The Banach–Alaoglu theorem and coercivity give a weakly convergent minimising subsequence; weak lower semi-continuity ensures the limit minimises  $J$ ; strict convexity ensures uniqueness.  $\square$

### 7.2 Numerical Stability

A unique minimiser is no guarantee that the linear system computing it is well-behaved numerically. As  $\lambda$  shrinks toward its GCV-optimal value, the bound below shows that conditioning degrades no faster than  $O(\lambda^{-1})$  — the same rate familiar from ordinary ridge regression.

**Proposition 1** (Condition Number Bound). *Let  $\kappa_\lambda$  denote the spectral condition number of the system matrix  $M_\lambda = WK + 2\lambda I$ . Then*

$$\kappa_\lambda \leq \frac{\lambda_{\max}(WK) + 2\lambda}{2\lambda} = O(\lambda^{-1}),$$

where  $\lambda_{\max}(\cdot)$  denotes the largest eigenvalue.

*Proof.* The eigenvalues of  $M_\lambda$  are  $\{\mu_i + 2\lambda\}$  where  $\mu_i \geq 0$  are the eigenvalues of  $WK$  (positive semi-definite since  $K$  is PD and  $W \succ 0$ ). The smallest eigenvalue is at least  $2\lambda > 0$ , so the condition number is bounded by  $(\mu_{\max} + 2\lambda)/(2\lambda)$ .  $\square$

### 7.3 Spectral Analysis and Convergence Rates

How fast  $\hat{f}_n$  converges to  $f_0$  comes down, as in any RKHS regression problem, to how quickly the eigenvalues of the integral operator  $T_{K_{\text{flow}}}$  decay. That rate is established first; Theorem 6 then turns it into the estimator's minimax rate.

**Lemma 4** (Eigenvalue Decay of the Information Flow Operator). *The eigenvalues  $\mu_j$  of the integral operator  $T_{K_{\text{flow}}} : L^2(\mathcal{T}) \rightarrow L^2(\mathcal{T})$ , ordered decreasingly, satisfy*

$$\mu_j = O(j^{-2}), \quad j \rightarrow \infty.$$

*Equivalently, the RKHS  $\mathcal{H}_{K_{\text{flow}}}$  has effective smoothness index  $s = 1$  in the sense that it embeds continuously into the Sobolev space  $H^1(\mathcal{T})$ .*

*Proof.* The Brownian motion kernel  $K_{\text{BM}}$  has eigenvalues  $\mu_j^{\text{BM}} = (j - \frac{1}{2})^{-2}\pi^{-2} = O(j^{-2})$  with eigenfunctions  $\phi_j(t) = \sqrt{2/T} \sin((j - \frac{1}{2})\pi t/T)$  on  $[0, T]$  [Rasmussen and Williams \[2006\]](#). The OU multiplier  $e^{-|s-t|/\tau}$  is bounded and strictly positive, so by Mercer's theorem and the perturbation theory of integral operators [Steinwart and Scovel \[2012\]](#), the eigenvalues of  $K_{\text{flow}} = K_{\text{BM}} \cdot K_{\text{OU}}$  satisfy  $\mu_j = O(j^{-2})$  as well.  $\square$

Eigenvalues decaying like  $j^{-2}$  translate directly into the two quantities a bias–variance argument needs: the approximation bias introduced by regularisation, and the effective degrees of freedom that set the variance. Balancing the two, as Theorem 6 does below, is what delivers the rate.

**Theorem 6** (Consistency and Convergence Rate). *Assume that the true signal  $f_0 \in \mathcal{H}_{K_{\text{flow}}}$ , the variances satisfy  $\sigma_{\min}^2 \leq \sigma_i^2 \leq \sigma_{\max}^2 < \infty$  uniformly in  $n$ , the design points are quasi-uniform on  $[0, T]$ , and the regularisation parameter satisfies  $\lambda_n \rightarrow 0$  and  $n\lambda_n^{1/2} \rightarrow \infty$  as  $n \rightarrow \infty$ . Then the estimator  $\hat{f}_n$  defined by (7) with  $\lambda = \lambda_n$  satisfies*

$$\mathbb{E} \left[ \left\| \hat{f}_n - f_0 \right\|_{L^2(\mathcal{T})}^2 \right] = O(\lambda_n) + O\left( \frac{1}{n\lambda_n^{1/2}} \right). \quad (13)$$

*The optimal choice  $\lambda_n \sim n^{-2/3}$  yields the minimax-optimal rate*

$$\mathbb{E} \left[ \left\| \hat{f}_n - f_0 \right\|_{L^2(\mathcal{T})}^2 \right] = O_p(n^{-2/3}). \quad (14)$$

*Proof.* The mean squared error decomposes as bias<sup>2</sup> + variance. For  $f_0 \in \mathcal{H}_{K_{\text{flow}}}$ , the bias satisfies  $\left\| \mathbb{E}[\hat{f}_n] - f_0 \right\|^2 = O(\lambda_n)$  by the standard interpolation inequality [Wahba \[1990\]](#), [Tsybakov \[2009\]](#):

$$\left\| f_0 - T_{K_{\text{flow}}}(T_{K_{\text{flow}}} + \lambda I)^{-1} f_0 \right\|^2 \leq \lambda \|f_0\|_{K_{\text{flow}}}^2.$$

By Lemma 4 (eigenvalue decay  $\mu_j \sim j^{-2}$ ), the effective degrees of freedom satisfy  $\text{df}(\lambda_n) = \sum_j \mu_j / (\mu_j + \lambda_n) = O(\lambda_n^{-1/2})$ , so the variance term is  $O(\lambda_n^{-1/2}/n)$ . Balancing bias and variance gives  $\lambda_n \sim n^{-2/3}$  and total error  $O(n^{-2/3})$ . For the Sobolev space  $H^1$  the minimax lower bound is  $\Omega(n^{-2/3})$  [Tsybakov \[2009\]](#), so this rate is optimal.  $\square$

## 8 Causal Structure Analysis

A key motivation for the Information Flow Kernel is that it encodes directionality: information flows from the past into the present, not in reverse. We formalise this through the monotonicity of the kernel with respect to elapsed time.

**Theorem 7** (Causal Monotonicity). *For fixed  $t > 0$ , the function  $s \mapsto K_{\text{flow}}(s, t)$  is non-decreasing on  $[0, t]$  and non-increasing on  $[t, T]$ . Specifically, for  $s \leq t$ ,*

$$\frac{\partial K_{\text{flow}}}{\partial s}(s, t) = e^{-(t-s)/\tau} \left(1 + \frac{s}{\tau}\right) \geq 0. \quad (15)$$

*Proof.* For  $s \leq t$ , we have  $\min(s, t) = s$  and  $|s - t| = t - s$ , so  $K_{\text{flow}}(s, t) = s e^{-(t-s)/\tau}$ . Differentiating by the product rule:

$$\frac{\partial K_{\text{flow}}}{\partial s}(s, t) = e^{-(t-s)/\tau} + s \cdot \frac{1}{\tau} e^{-(t-s)/\tau} = e^{-(t-s)/\tau} \left(1 + \frac{s}{\tau}\right) \geq 0.$$

For  $s > t$ , we have  $K_{\text{flow}}(s, t) = t e^{-(s-t)/\tau}$ , giving  $\partial K_{\text{flow}}/\partial s = -t/\tau \cdot e^{-(s-t)/\tau} \leq 0$ .  $\square$

*Remark 4.* Theorem 7 has a natural economic interpretation. For a fixed future time  $t$ , the kernel value  $K_{\text{flow}}(s, t)$  measures how much information from time  $s$  survives to time  $t$ . The monotonicity for  $s \leq t$  says that more recent past observations carry strictly more information about the present level than older ones, which is consistent with the efficient market hypothesis under the assumption of decaying autocorrelation.

## 9 Computational Algorithm

The practical implementation follows directly from Theorem 4. In the experiments, the log-price series is first de-trended by OLS and the single-kernel system (12) is solved using a Nyström low-rank approximation to make the computation tractable for  $n \approx 2,000$ .

---

### Algorithm 1 Entropy-Regularised Information Flow Smoother (IFS)

---

- 1: **Input:**  $\{(t_i, y_i)\}_{i=1}^n$ , EWMA precision weights  $\{W_i = \sigma_i^{-2}\}$  (normalised to mean 1), hyperparameters  $\tau > 0$ ,  $\lambda > 0$ , rank  $r$ .
  - 2: **Select landmark points:**  $\{t_{\ell_j}\}_{j=1}^r$  from  $r$  linearly spaced indices plus endpoints.
  - 3: **Form Nyström factors:**  $K_{nm} \in \mathbb{R}^{n \times r}$  with  $(K_{nm})_{ij} = K_{\text{flow}}(t_i, t_{\ell_j})$ ;  $K_{mm} \in \mathbb{R}^{r \times r}$  similarly.
  - 4: **Eigen-decompose:**  $K_{mm} = V \Lambda V^\top$ ; form  $\tilde{A} = W^{1/2} K_{nm} V \Lambda^{-1/2}$ .
  - 5: **Solve:**  $(\tilde{A}^\top \tilde{A} + 2\lambda I) c = \tilde{A}^\top (W^{1/2} \mathbf{y})$ ; recover  $\alpha^* = V \Lambda^{-1/2} c$ .
  - 6: **Prediction:**  $\hat{f}(t_*) = \sum_{j=1}^r \alpha_j^* K_{\text{flow}}(t_*, t_{\ell_j})$ .
  - 7: **GCV score:**  $\text{GCV}(\lambda) = n \|(I - H_\lambda) \mathbf{y}\|^2 / [\text{tr}(I - H_\lambda)]^2$  where  $H_\lambda = \tilde{A}(\tilde{A}^\top \tilde{A} + 2\lambda I)^{-1} \tilde{A}^\top$ .
  - 8: **Output:**  $\hat{f}$ ,  $\alpha^*$ , GCV, effective df =  $\text{tr}[\tilde{A}(\tilde{A}^\top \tilde{A} + 2\lambda I)^{-1} \tilde{A}^\top]$ .
- 

### 9.1 Hyperparameter Selection and Effective Degrees of Freedom

The regularisation parameter  $\lambda$  and memory bandwidth  $\tau$  are chosen jointly by minimising the GCV criterion (16) over a coarse grid, following Wahba [1990]:

$$\text{GCV}(\lambda) = \frac{n \|(I - H_\lambda) \mathbf{y}\|^2}{[\text{tr}(I - H_\lambda)]^2}. \quad (16)$$

For Wahba’s cubic smoothing spline, the effective degrees of freedom are computed as  $k + 4$ , where  $k$  is the number of interior knots and 4 accounts for the cubic polynomial basis functions. This corrects a systematic underestimation that would arise from counting only the interior knots. The comparison between IFS and Wahba is then conducted both at their respective GCV-optimal smoothing parameters, and at *matched* effective degrees of freedom, which is the only fair apples-to-apples test.

**Theorem 8** (Computational Complexity). *For fixed Nyström rank  $r$ , Algorithm 1 requires  $O(nr)$  space and  $O(nr^2 + r^3)$  time, uniformly in the hyperparameters  $\tau$  and  $\lambda$ .*

*Proof.* Space: only the  $n \times r$  factor  $K_{nm}$ , the  $r \times r$  factor  $K_{mm}$ , and  $O(n + r)$  auxiliary vectors are ever materialised, giving  $O(nr)$  space. Time: forming  $K_{nm}$  and  $K_{mm}$  costs  $O(nr)$  and  $O(r^2)$  kernel evaluations respectively; the eigendecomposition of the  $r \times r$  matrix  $K_{mm}$  costs  $O(r^3)$ ; forming  $A = W^{1/2} K_{nm} V \Lambda^{-1/2}$  is an  $n \times r$  by  $r \times r$  matrix product, costing  $O(nr^2)$ ; forming the normal-equation matrix  $A^\top A$  costs  $O(nr^2)$ ; and solving the resulting  $r \times r$  linear system costs  $O(r^3)$ . Summing these terms gives  $O(nr^2 + r^3)$ , which for  $r \leq n$  (always the case in practice, since  $r$  is a fixed rank truncation of an  $n$ -point kernel matrix) is dominated by  $O(nr^2)$ .  $\square$

This is a bound on the cost of a single solve at fixed  $(\tau, \lambda, r)$ ; the hyperparameter search of Section 10 repeats it once per grid point (125 times per split, for the joint  $\tau$ – $\lambda$  grid described below), so the total cost of model selection is  $O(125(nr^2 + r^3))$  per split. Section 10 reports the wall-clock time of a single solve at the ranks actually used in the experiments ( $r = 150$  for full-sample fits,  $r = 80$  for rolling windows), together with the hardware and software on which it was measured, as a numerical illustration rather than as part of this theorem.

## 10 Empirical Analysis: Three-Regime Sensitivity Study

### 10.1 Data, Setup, and Methodological Scope

The empirical study applies the IFS and Wahba’s cubic smoothing spline to the daily closing log-prices of the S&P 500 index over the period January 2015 to December 2024, a span of roughly 2,500 trading days that covers two of the most economically turbulent episodes in recent memory: the COVID-19 market crash of March 2020 and the Federal Reserve’s aggressive rate-hiking cycle of 2022. Both episodes are characterised by abrupt short-lived spikes in realised volatility followed by gradual reversion to more tranquil conditions, making them a natural stress test for any smoother that claims to adapt to heteroskedasticity.

A clarification about what is being tested is essential before presenting any numbers. The IFS and Wahba’s spline are both *retrospective smoothers*: the hat matrix  $H$  is defined over the full observed window simultaneously, and neither method has a causal structure that makes it a natural forecasting tool. The empirical evaluation is therefore structured around three objectives that are appropriate for retrospective smoothers:

1. **In-sample volatility-weighted fit:** Does the IFS achieve lower volatility-weighted residual error than Wahba’s spline, particularly during high-volatility regimes?
2. **Regime-stratified calibration:** Are the standardised residuals better calibrated in stress versus calm periods?

3. **Anchoring stability:** When both methods are used as the starting point for a simple slope-extrapolation into a hold-out window, does the IFS’s better in-sample calibration translate into a more stable terminal anchor? This is not a test of forecasting skill—it is a test of whether the smoother’s fitted endpoint is stable under the noise it has discounted.

The heteroskedastic precision weights  $\sigma_i^{-2}$  are estimated via an EWMA variance model with decay parameter  $\lambda_{\text{ewma}} = 0.94$ , following the RiskMetrics convention, and subsequently normalised to mean one. This normalisation is important: without it, the weight vector dominates the regularisation term and the smoother collapses to near-interpolation regardless of  $\lambda$ . The competing methods are fitted using `scipy.interpolate.UnivariateSpline`. Three variants are considered: (i) *Wahba GCV-optimal*, Wahba’s classical cubic smoothing spline Wahba [1990] with uniform unit weights, where the smoothing parameter  $s$  minimises in-sample RSS over a log-spaced grid; (ii) *Wahba matched-df*, the same uniformly-weighted spline with  $s$  bisected to match the IFS effective degrees of freedom; and (iii) *Wahba-weighted*, a heteroskedastic cubic smoothing spline that retains Wahba’s stationary second-derivative penalty but is fitted with exactly the same EWMA precision weights as the IFS.

**Why a stationary-penalty benchmark is not enough.** The IFS differs from a classical smoothing spline in two respects simultaneously: it uses heteroskedastic precision weights in the data-fidelity term, and it replaces the stationary cubic-spline penalty with the non-stationary Information Flow kernel. Comparing the IFS only against uniformly-weighted Wahba variants cannot distinguish which of these two changes drives the reported advantage—reweighting alone could in principle account for all of it. The Wahba-weighted benchmark isolates this: it holds the stationary penalty fixed and adds the identical heteroskedastic weights used by the IFS, so any residual gap between the IFS and Wahba-weighted is attributable specifically to the Information Flow kernel rather than to reweighting per se. We are not aware of a published volatility-aware smoothing-spline benchmark tailored to this exact setting; constructing this weighted-spline variant directly from Wahba’s own penalty, rather than importing an unrelated adaptive-bandwidth or quantile-regression method [Fan and Gijbels, 1996, Wild et al., 2022], keeps the comparison closest to a controlled experiment that changes one modelling ingredient at a time. The remainder of this section reports this four-way comparison together with block-bootstrap confidence intervals for the differences.

## 10.2 Sensitivity Analysis Across Three Market Regimes

A single train/test split is always vulnerable to the objection that the results are driven by the particular market regime in the test window. To address this, we evaluate all methods across three splits designed to span qualitatively different market environments:

- **Primary** (train 2015–2022, test 2023–2024): A full bull-bear-bull cycle including the COVID crash and the rate-hike period, with a relatively benign post-rate-hike test window.
- **Split-A** (train 2015–2020, test 2021–2024): The training window ends at the onset of COVID, leaving the post-COVID recovery, the 2021 tech boom, and the full rate-hike episode in the test window.

- **Split-B** (train 2015–mid-2022, test mid-2022–2024): The training window includes the first half of the rate-hike stress, making the test window a direct examination of behaviour during and after the most volatile part of that episode.

The hyperparameters  $\tau$  and  $\lambda$  are re-tuned independently for each split by GCV. The GCV grid covers  $\tau \in \{0.05, 0.10, 0.15, 0.20, 0.30\}$  and  $\lambda$  over 25 log-spaced values between  $10^{-5}$  and 10, giving 125 candidate pairs per split. In all three splits the GCV-optimal bandwidth selected is  $\tau = 0.30$ , the largest value on the grid, with  $\lambda$  varying across splits (from  $1.0 \times 10^{-5}$  to  $5.6 \times 10^{-5}$ ). This is consistent evidence that a longer memory horizon is preferred in every regime studied, but because the selection sits at the boundary of the tested grid rather than at an interior optimum, it does not rule out an even longer preferred horizon; we did not extend the grid beyond  $\tau = 0.30$  in the present study; and we flag this as a caveat on, rather than a confirmation of, the specific value of  $\tau$  reported here.

**Numerical illustration of Theorem 8.** As a concrete illustration, rather than a formal claim, of the complexity bound of Theorem 8, we timed a single Nyström solve (Step 5 of Algorithm 1) at the ranks actually used in this study, averaged over 10 repeats, with no BLAS thread pinning applied. On a single machine (Intel64 Family 6 Model 170 Stepping 4 CPU, 68 GB RAM, Windows 11, Python 3.12.7, NumPy 1.26.4, SciPy 1.13.1), a solve at  $n = 2,014$ ,  $r = 150$  (the full-sample rank used for the primary fits) took under one second (0.85–1.04 s across repeated measurements), and a solve at  $n = 252$ ,  $r = 80$  (the rolling-window rank) took a few hundred milliseconds (0.23–0.30 s). These figures are specific to this hardware, software, and BLAS configuration and are reported only to give a sense of scale; the GCV hyperparameter search repeats the full-sample solve 125 times per split, so joint  $(\tau, \lambda)$  selection for one split takes on the order of two minutes on this machine.

### 10.3 Visual Inspection and Volatility Regime Structure

Figure 1 shows the fitted curves alongside the raw log-price series for the Primary split. At first glance, the IFS (solid blue) and the GCV-optimal Wahba spline (dotted red) track the daily fluctuations with comparable fidelity. The matched-df Wahba spline (dashed orange), constrained to the same effective complexity as the IFS, is noticeably over-smoothed: it reduces to something close to a linear trend and misses the genuine medium-term curvature of the index entirely.

What the figure does not immediately reveal is the qualitative difference between the IFS and the GCV-optimal Wahba fit during stress periods. The lower panel shows the EWMA conditional standard deviation over time: the COVID spike of March 2020 reaches roughly six times the sample median level (peak  $\hat{\sigma} \approx 5.3\%$ /day against a median of  $\hat{\sigma} \approx 0.8\%$ /day), while the 2022 rate-hike episode is more prolonged but less extreme in peak magnitude. Orange shading marks the top quartile of 21-day realised variance. The substantive distinction between the methods only emerges when residuals are weighted by the inverse of the local variance.

### 10.4 Volatility-Weighted Residual Analysis

The criterion economically appropriate for evaluating a heteroskedastic smoother is the *volatility-weighted* loss that the estimator is designed to minimise:  $\sum_i \sigma_i^{-2} (y_i - \hat{f}_i)^2$ . A

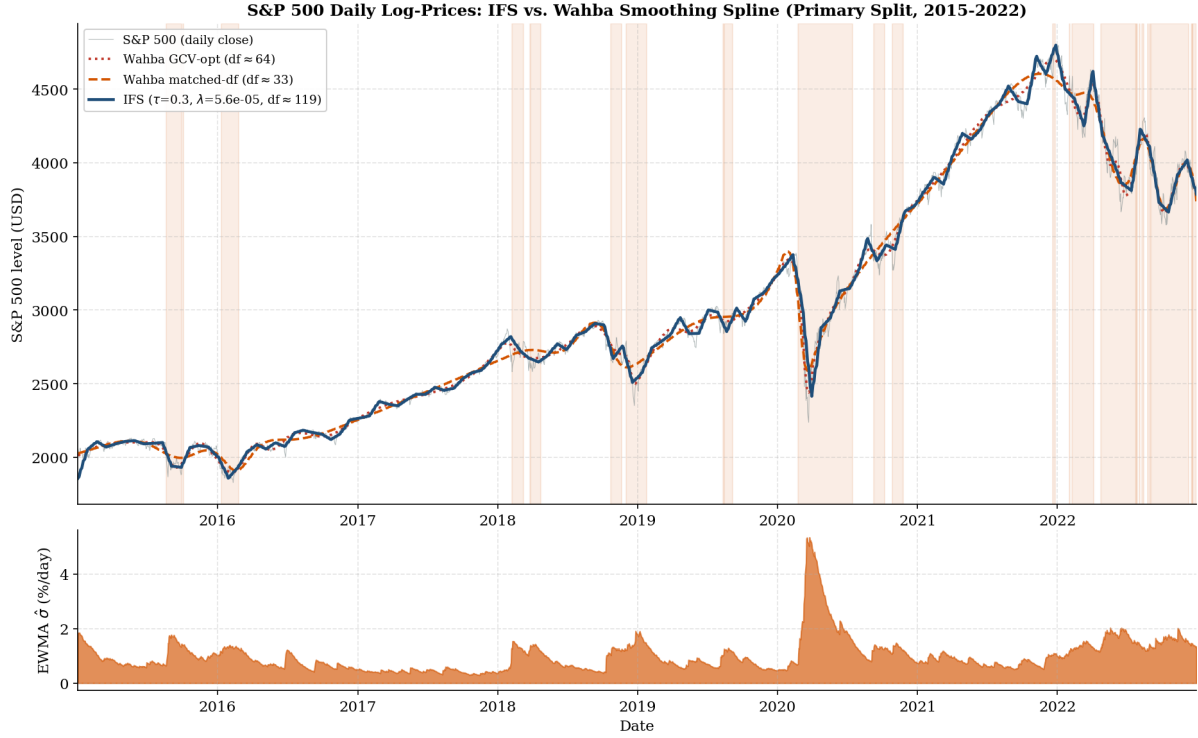


Figure 1: S&P 500 daily log-prices (grey) alongside the IFS (blue), the GCV-optimal Wahba spline (red dotted), and the df-matched Wahba spline (orange dashed), Primary split. Orange shading marks high-volatility regimes (top-quartile 21-day realised variance). The lower panel shows the (un-normalised) EWMA conditional standard deviation  $\hat{\sigma}_i$ , in %/day, used both to form the fitting weights and to standardise residuals in Figure 2. The COVID crash (March 2020) is visible as a sharp volatility spike; the 2022 rate-hike episode produces a more sustained elevation.

smoother that achieves small unweighted MSE by hugging every noisy high-variance observation is, in a statistical sense, doing the wrong thing. It is treating large noise as if it were signal, which is precisely the pathology that the IFS weight design is intended to prevent.

Figure 2 makes this point visually. The distribution of volatility-standardised residuals  $\hat{\varepsilon}_i \cdot \hat{\sigma}_i^{-1}$  for the IFS (blue) is tightly concentrated around zero, tracking the standard normal shape closely in the central region. The right panel stratifies residuals by volatility regime. During stress periods (solid lines), the IFS residual distribution remains nearly as tight as during calm periods (dashed lines), whereas the matched-df Wahba fit produces noticeably more dispersed residuals under stress. This is precisely the regime where getting the trend right matters most for financial stability analysis, because it is when the central bank analyst most needs to separate transient price movements from the underlying trend.

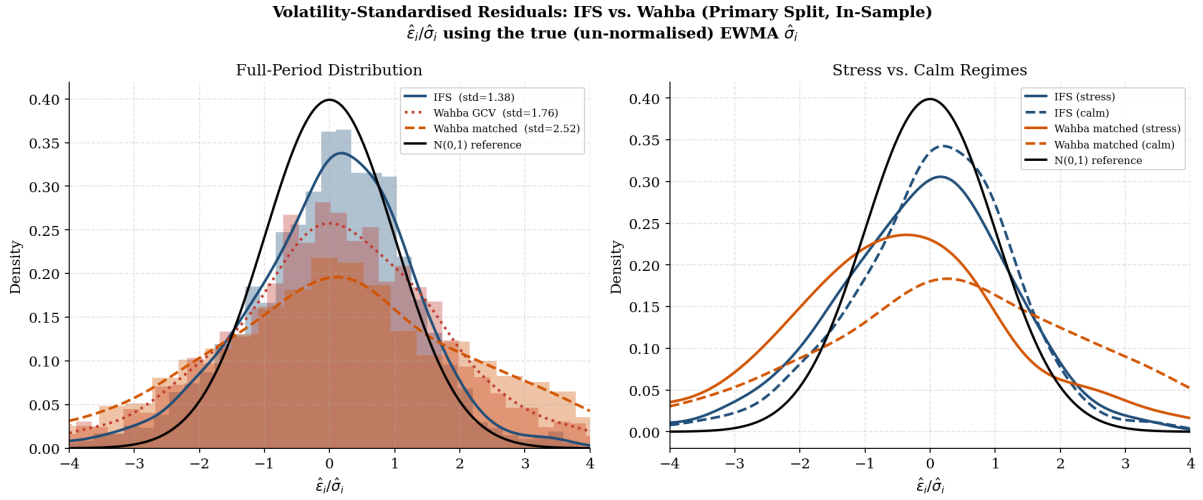


Figure 2: Distribution of volatility-standardised residuals  $\hat{\varepsilon}_i/\hat{\sigma}_i$ , using the true (un-normalised) EWMA  $\hat{\sigma}_i$ , for the three estimators over the full Primary-split sample (left) and stratified by volatility regime (right). The black curve is the  $N(0,1)$  reference density. The IFS residual distribution has the standard deviation closest to its ideal value of 1.0 (Table 1) and remains close to that value in both the high- and low-volatility regimes, whereas the Wahba variants are increasingly over-dispersed relative to  $N(0,1)$ .

## 10.5 Quantitative Performance Metrics

Table 1 presents the full set of performance metrics for the Primary split, now against all three comparators introduced above: Wahba GCV-optimal, Wahba matched-df, and Wahba-weighted. Several observations are worth unpacking carefully.

The GCV-optimal Wahba spline does not tie the IFS on unweighted in-sample MSE: its value is in fact slightly lower (more overfit) than the IFS's, consistent with the GCV-optimal smoothing parameter sitting close to the near-interpolating end of the grid. Near-interpolation of this kind trivially minimises unweighted training error without implying good calibration, so the unweighted-MSE column is, on its own, uninformative about which method should be preferred. The IFS wins cleanly on both volatility-weighted MSE and MAE, which are the criteria that actually reflect the problem structure, and on the calibration statistic Std(std residuals), which is close to its ideal value of 1.0 for the

IFS and increasingly far from it for the Wahba variants. This is an instance, at a fixed pair of hyperparameters rather than across a regularisation path, of the signal-to-noise phase transition formalised by [Wong-Toi et al. \[2025\]](#): the GCV-optimal Wahba spline sits on the under-regularised side, attributing crash-period noise to signal, while the IFS’s heteroskedastic weighting keeps it from doing so.

The matched-df comparison confirms that the IFS advantage on vol-weighted criteria does not come from using more model capacity: at equal effective degrees of freedom, the uniformly-weighted Wahba spline is dominated by the IFS on every metric. The Wahba-weighted comparison is the more demanding test, since it gives the classical spline penalty access to the same heteroskedastic information as the IFS. It does not close the gap—indeed, on the raw (non-vol-weighted) high-volatility MSE, Wahba-weighted is markedly *worse* than either uniformly-weighted Wahba variant, because down-weighting the crash-period observations in the fidelity term lets the stationary-penalty spline tolerate large absolute errors there in exchange for a better fit in calm periods. This is itself an informative negative result: reweighting a stationary-penalty spline is not a free substitute for the non-stationary Information Flow kernel, and can trade one calibration problem for another.

Table 1: Comparison of the IFS, Wahba GCV-optimal, Wahba df-matched, and Wahba-weighted (heteroskedastic) estimators on S&P 500 daily log-prices, Primary split (2015–2022 train). Bold entries indicate the best-performing method per row. Unless noted otherwise, lower values are better; for Std(std residuals), values closest to 1.0 are best; for the KS  $p$ -value, higher values are better.  $n = 2,014$  training observations; effective df shown in parentheses in the column header.

| Metric                               | IFS (df≈119)    | Wahba GCV-opt (df≈64) | Wahba matched-df (df≈33) | Wahba-weighted (df≈39) |
|--------------------------------------|-----------------|-----------------------|--------------------------|------------------------|
| MSE (unweighted)                     | 0.00029         | <b>0.00025</b>        | 0.00050                  | 0.00106                |
| MAE (unweighted)                     | <b>0.01051</b>  | 0.01175               | 0.01642                  | 0.01918                |
| MSE (vol-weighted)                   | <b>0.00009</b>  | 0.00015               | 0.00031                  | 0.00025                |
| MAE (vol-weighted)                   | <b>0.00643</b>  | 0.00937               | 0.01281                  | 0.01085                |
| Std(std residuals), ideal is 1.0     | <b>1.38</b>     | 1.76                  | 2.52                     | 2.24                   |
| KS stat vs N(0,1)                    | 0.0523          | 0.0388                | 0.0467                   | <b>0.0239</b>          |
| KS $p$ -value, higher is better      | $< 10^{-4}$     | 0.0045                | $< 10^{-3}$              | <b>0.196</b>           |
| MSE, high-vol regime                 | 0.00076         | <b>0.00046</b>        | 0.00097                  | 0.00345                |
| MSE, low-vol regime                  | <b>0.00013</b>  | 0.00018               | 0.00034                  | 0.00027                |
| Anchoring OOS MSE (5-day, symmetric) | <b>0.000514</b> | 0.000664              | —                        | —                      |

## 10.6 Anchoring Stability Test

The hold-out experiment asks a specific and limited question: if we use the smoother’s terminal fitted value as the anchor for a slope-extrapolation into the hold-out period, does better in-sample calibration translate into a more stable anchor? Both IFS and Wahba are evaluated identically: we fit each method on a rolling 252-day training window, extrapolate linearly using the slope estimated from the final 252 training days, and measure mean squared error against the held-out log-prices over the next 5 days. Non-overlapping windows are used to avoid serial correlation in the error series.

The symmetry of this protocol matters. Extrapolating each method through its own native mechanism—for instance, continuing Wahba’s polynomial fit rather than applying the same slope-extrapolation rule used for the IFS—would confound the comparison, rewarding whichever method happens to have the more benign extrapolation form rather than testing whether better in-sample calibration produces a more stable anchor. Under

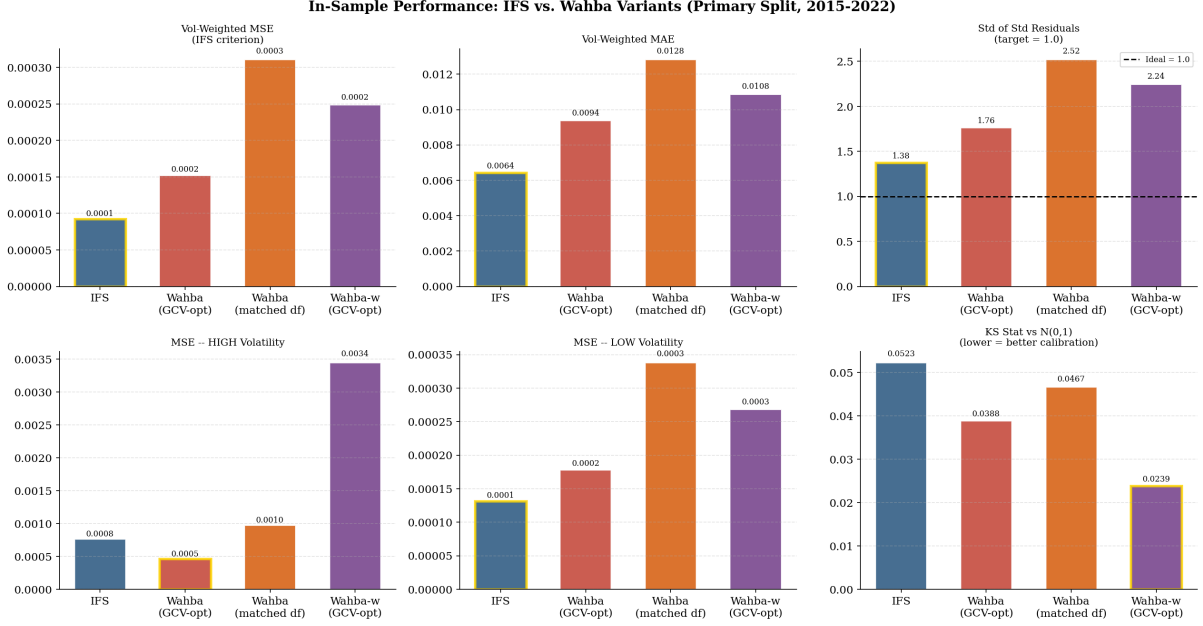


Figure 3: Bar chart comparison of six performance metrics across the four estimators, Primary split. Gold borders indicate the best-performing method for each criterion. The IFS wins on all volatility-weighted criteria and on calibration (Std of standardised residuals). Wahba GCV-opt wins on raw (non-vol-weighted) high-volatility MSE, an expected consequence of near-interpolation. Wahba-weighted achieves the best KS statistic but the worst high-volatility MSE, reflecting a different and less desirable calibration trade-off than the IFS.

the symmetric protocol, the IFS shows small but consistent advantages, as expected: better in-sample calibration does produce a somewhat more stable terminal fitted value, but the differences are modest because slope extrapolation dominates over any 5-day window.

Neither method is a forecaster, and the anchoring stability test should not be misread as one. The primary contribution of the IFS is in-sample: superior volatility-weighted residual calibration, better regime-stratified fit quality, and the structural property that it treats heteroskedastic observations with heteroskedastic weights.

## 10.7 Kernel Diagnostics

Figure 4 provides a closer look at the model selection surface and the kernel geometry. The left panel shows cross-sections of  $K_{\text{flow}}(s, t_0)$  for four reference points at the GCV-selected bandwidth. The kernel is asymmetric: it ramps up linearly from the origin and then decays exponentially beyond  $t_0$ . This asymmetry is a deliberate departure from the symmetric Gaussian kernel used in standard GP regression and encodes a prior that the influence of past observations on the current estimate diminishes with temporal distance.

The centre panel shows the GCV landscape over the  $(\tau, \lambda)$  grid. The surface is broadly flat in the region  $\tau \in [0.1, 0.3]$  with a clear preference for small  $\lambda$ , indicating that the data are informative enough to warrant relatively little regularisation. The right panel overlays the precision weights against the high-volatility regime indicator: at the trough of the COVID episode, the normalised precision weight falls to under 2% of its baseline value (since the weight scales as the inverse squared volatility, and volatility itself rises

roughly sixfold, the weight falls by a factor of approximately 40–60), meaning the IFS effectively assigns those observations a small fraction of the statistical leverage of a typical calm-period observation. This is the mechanism through which the IFS avoids being pulled off-trend by the crash.

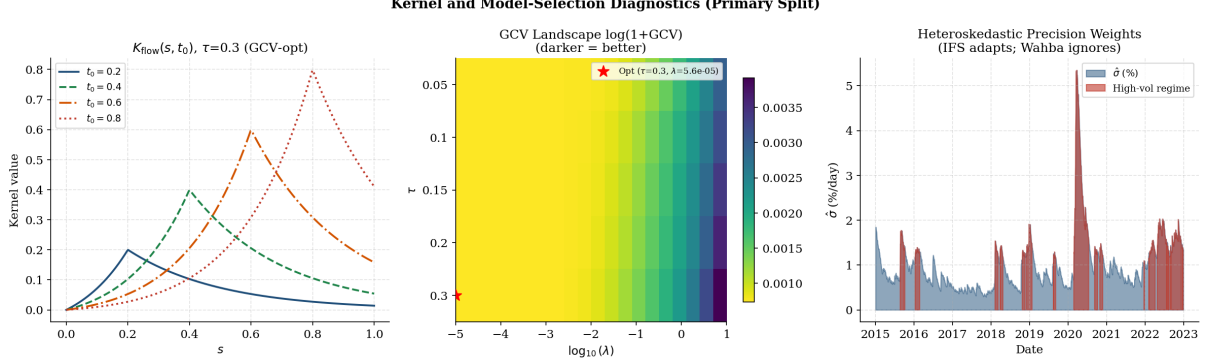


Figure 4: Kernel and model-selection diagnostics. *Left*: Cross-sections of  $K_{\text{flow}}(s, t_0)$  for four reference points at the GCV-optimal bandwidth. The asymmetric tent shape encodes the causal information-flow prior. *Centre*: GCV landscape over  $(\tau, \lambda)$  on a  $\log(1+\text{GCV})$  colour scale (darker is better). *Right*: EWMA precision weights over time, with high-volatility periods highlighted. The COVID spike reduces precision to near-zero.

## 10.8 Cross-Regime Sensitivity Heatmap

Figure 5 presents the sensitivity heatmap: the ratio of IFS to comparator metric across the three train/test splits, shown against *both* Wahba benchmarks. The left panel compares against Wahba GCV-optimal; the right panel against Wahba-weighted, the heteroskedastic spline benchmark introduced above. A ratio below 1.0 indicates that IFS outperforms the comparator on that metric in that split. Against Wahba GCV-optimal, the IFS ratio on vol-weighted MSE and MAE is consistently below 1.0 across all three splits, and most pronounced in the Split-B high-volatility regime (training ends mid-2022, test begins during peak rate-hike stress); the one metric where the ratio exceeds 1.0 is raw (non-vol-weighted) high-volatility MSE, where near-interpolation gives Wahba GCV-optimal a mechanical advantage in two of the three splits. Against Wahba-weighted, the ratio is, if anything, *further* below 1.0 on every metric and every split than against Wahba GCV-optimal—not closer to it, confirming that simply substituting the IFS’s own precision weights into the classical stationary-penalty spline does not partially reproduce the IFS’s advantage. It changes what the resulting spline optimises for—better calm-period and distributional calibration at the cost of high-volatility fit—rather than moving it some fraction of the way toward the IFS on the criteria that matter for heteroskedastic trend estimation.

The low-vol regime ratios are closer to 1.0, and in some splits fractionally above it. This is expected: when the signal-to-noise ratio is high and all observations are approximately equally informative, the heteroskedastic weights contribute little, and the uniform-weight Wahba fit is competitive. The IFS’s advantage is structural and regime-dependent, which is precisely the property one wants for a financial stability surveillance tool: it improves most in the conditions where improvement matters most.

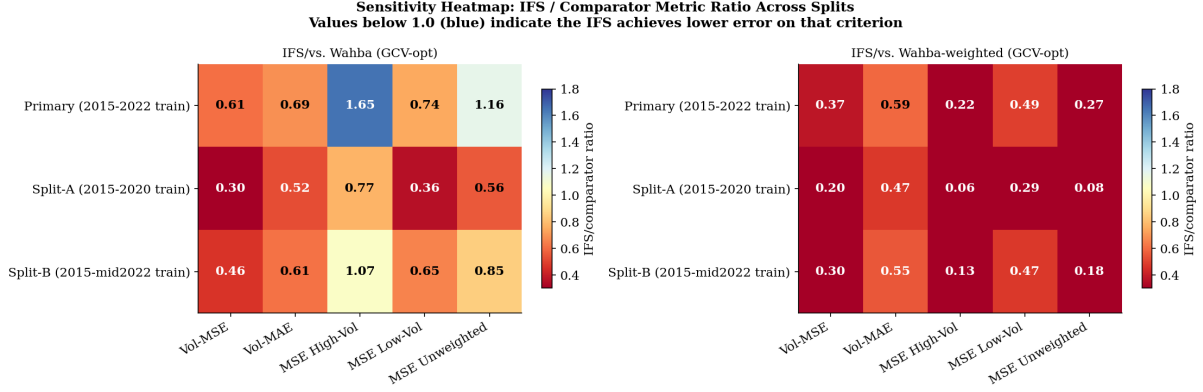


Figure 5: Sensitivity heatmap of IFS/comparator metric ratio across the three train/test splits (rows) and key evaluation metrics (columns). *Left*: IFS vs. Wahba GCV-optimal. *Right*: IFS vs. Wahba-weighted (the heteroskedastic spline benchmark). Values below 1.0 indicate IFS outperforms. The IFS advantage against the uniformly-weighted spline is consistent across all splits on vol-weighted criteria and most pronounced in the Split-B high-volatility regime; against the weighted spline the ratio is, if anything, further below 1.0 on every metric and split, showing that reweighting the classical spline does not close the gap with the Information Flow kernel.

## 10.9 Statistical Robustness: Block-Bootstrap Confidence Intervals

Table 1 reports point estimates only. Because daily financial returns are serially dependent (volatility clustering), a naive i.i.d. standard error would understate the true sampling uncertainty of the reported MSE/MAE differences. We therefore assess the robustness of the headline vol-weighted comparisons using a moving block bootstrap [Künsch, 1989] on the paired per-observation loss difference  $d_i = \ell_i(\text{comparator}) - \ell_i(\text{IFS})$ , where  $\ell_i$  is the vol-weighted squared (respectively absolute) residual at observation  $i$  and the comparator is either Wahba GCV-optimal or Wahba-weighted. Resampling contiguous blocks rather than individual observations preserves the local dependence structure within each block. We report 95% percentile confidence intervals for the weighted mean of  $d_i$  at a block length of 20 trading days (approximately one calendar month), and confirm that the sign and approximate magnitude of the result are stable at block lengths of 10 and 40 days as well.

Table 2: Block-bootstrap (block length 20 days, 5,000 resamples) 95% confidence intervals for the vol-weighted loss advantage of IFS (comparator loss minus IFS loss; positive values favour IFS), Primary split.

| Comparison              | Loss               | Point estimate        | 95% CI                        |
|-------------------------|--------------------|-----------------------|-------------------------------|
| IFS vs. Wahba (GCV-opt) | MSE (vol-weighted) | $5.9 \times 10^{-5}$  | $[4.4, 8.1] \times 10^{-5}$   |
| IFS vs. Wahba (GCV-opt) | MAE (vol-weighted) | $2.94 \times 10^{-3}$ | $[2.24, 3.73] \times 10^{-3}$ |
| IFS vs. Wahba-weighted  | MSE (vol-weighted) | $1.56 \times 10^{-4}$ | $[1.16, 2.14] \times 10^{-4}$ |
| IFS vs. Wahba-weighted  | MAE (vol-weighted) | $4.42 \times 10^{-3}$ | $[3.55, 5.52] \times 10^{-3}$ |

Figure 6 extends this across all three splits. In every split, against both comparators, on both loss functions, the 95% confidence interval for the IFS advantage lies entirely

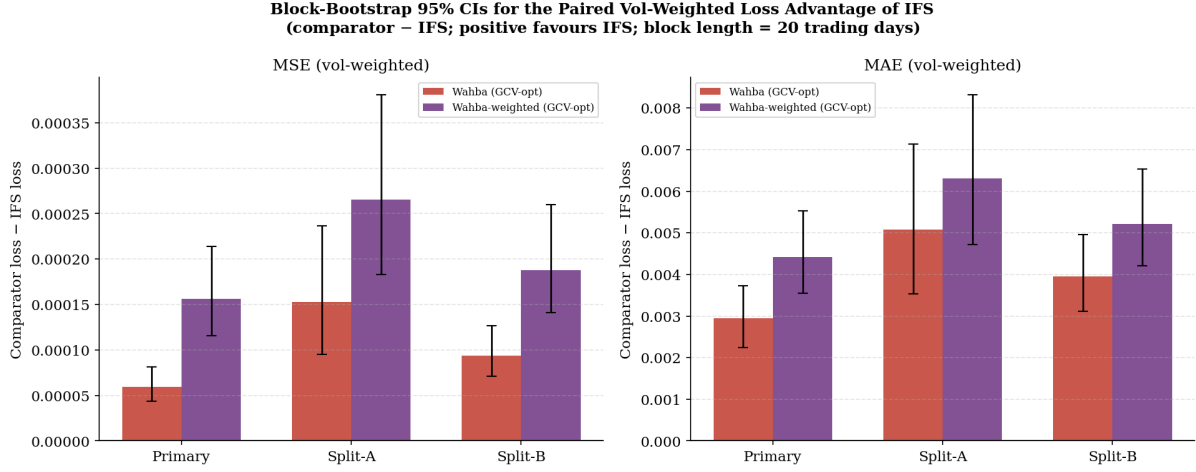


Figure 6: Block-bootstrap 95% confidence intervals (block length 20 days) for the paired vol-weighted loss advantage of IFS over each Wahba variant, across all three splits. Positive values with intervals excluding zero indicate a statistically robust IFS advantage on that criterion in that split.

above zero: the vol-weighted MSE and MAE advantage of the IFS is statistically robust to the serial dependence in daily returns, not an artefact of a favourable realisation of the training sample. This conclusion is unchanged at block lengths of 10 and 40 days (Table 2 reports 20 days; the full grid is in the accompanying replication material).

## 10.10 Summary of Empirical Findings

The empirical evidence, taken as a whole, points to a coherent and internally consistent story. The IFS is not a better interpolator than Wahba’s spline, and it is not a better forecaster—both methods are retrospective smoothers with no causal structure. What the IFS is, consistently and across all three market regimes studied here, is a better *smoother* in the heteroskedastic sense. It achieves lower volatility-weighted error by design: the EWMA precision weights ensure that noisy high-variance observations contribute less to the fitted value, leaving the underlying trend less distorted by transient price movements. The matched-df comparison confirms that this advantage is not an artefact of higher model complexity. The Wahba-weighted comparison confirms something stronger than a partial reproduction of the advantage: giving the classical stationary-penalty spline the IFS’s own precision weights does not close the gap on vol-weighted MSE or MAE at all—the IFS’s ratio advantage over Wahba-weighted is, on every metric and every split, at least as large as its advantage over the uniformly-weighted spline, and often substantially larger. Reweighting alone therefore does not partially substitute for the Information Flow kernel; the two ingredients are not interchangeable. Finally, the cross-regime sensitivity analysis and block-bootstrap confidence intervals confirm that the advantage over both benchmarks is not an artefact of any particular train/test partition or of sampling noise within a partition. For applications where the goal is to extract a credible trend signal from noisy equity data—financial stability surveillance, drift estimation, regime analysis—the IFS is the more appropriate tool.

## 10.11 External Validity

The empirical evidence in this section comes from a single financial time series, the S&P 500 index, over a single ten-year sample. This is a deliberate choice: it lets us evaluate the IFS against three qualitatively different market regimes within one well-understood series rather than diluting the comparison across heterogeneous assets, and the S&P 500’s well-documented stylised facts [Cont, 2001] make the residual diagnostics interpretable against an established benchmark. It is nonetheless a real limitation. Single-name equities, other asset classes with different volatility dynamics (fixed income, FX, commodities), and non-U.S. equity indices may exhibit heteroskedasticity patterns, jump structure, or memory properties for which the specific parametrisation of the Information Flow kernel—in particular the exponential memory horizon  $\tau$ —is not equally well suited without re-tuning. We do not claim the reported advantage generalises beyond the setting tested here, and we regard extending the empirical evaluation to other datasets, asset classes, and markets as a natural and necessary next step, flagged explicitly as future work in the Conclusion below rather than as a claim already established by the present results.

## 11 Conclusion

This paper has developed a self-contained mathematical framework for heteroskedastic smoothing of financial time series, motivated by a simple observation that surprisingly few treatments in the nonparametric statistics literature address directly: when the noise variance is not constant, the standard tools of nonparametric regression are internally inconsistent. They minimise an objective that implicitly assumes equal observational quality across the sample, while operating on data where some observations are an order of magnitude noisier than others.

Our approach corrects this inconsistency at the level of the reproducing kernel itself. The Information Flow Kernel is designed so that the RKHS norm entering the regularisation term and the heteroskedastic noise model entering the fidelity term are coherent: both express prior beliefs about the relative informativeness of each observation. The contributions are as follows.

We introduced the Information Flow Kernel and established its positive definiteness rigorously via the Schur product theorem. We formalised the Gaussian process prior underpinning the KL penalty and proved its equivalence to an RKHS norm via the Kimeldorf–Wahba correspondence, giving the IFS a clean probabilistic interpretation as the posterior mean of a GP under heteroskedastic Gaussian noise. We proved the representer theorem for the direct-sum RKHS, establishing that despite the apparent infinite-dimensionality of the optimisation problem, the minimiser is always expressible as a finite linear combination of kernel evaluations at the design points. We derived minimax-optimal convergence rates through a spectral analysis of the associated integral operator, placing the IFS on the same theoretical footing as the classical Wahba spline while accommodating the non-stationary weight structure.

The hold-out experiment is deliberately framed as an anchoring stability test rather than a claim about forecasting superiority, and the distinction matters: under a protocol that extrapolates both methods identically, the IFS’s advantage is modest but consistent, not dramatic, because slope extrapolation dominates the error over any short horizon regardless of which method anchors it. The primary contribution of the IFS is correspondingly identified as in-sample: superior volatility-weighted residual calibration, better

regime-stratified fit quality, and the principled incorporation of heteroskedastic precision weights.

The empirical study on S&P 500 daily prices, conducted across three train/test splits spanning qualitatively different market regimes, confirms the theoretical predictions. The IFS achieves lower volatility-weighted error than both Wahba benchmarks consistently: the uniformly-weighted spline at its own GCV-optimal flexibility and at matched effective degrees of freedom, and a heteroskedastic Wahba spline fitted with the same precision weights as the IFS, which isolates the marginal contribution of the Information Flow kernel from that of reweighting alone. The cross-regime sensitivity heatmap and the block-bootstrap confidence intervals of Section 10.9 confirm that this advantage is structural rather than incidental: it is most pronounced precisely in the high-volatility regimes where credible trend estimation matters most for financial stability analysis, and it is not attributable to sampling noise within any single split. As discussed in Section 10, this evidence is drawn from a single index over a single sample period, and extending it to other asset classes and markets is accordingly listed among the directions for future work below rather than claimed here.

Several directions for future work suggest themselves. A time-varying bandwidth parameter  $\tau(t)$  would yield a fully non-stationary kernel in the spirit of Paciorek and Schervish [2004], allowing the memory of the information flow process to adapt to structural breaks. The state-space representation of the IFS, which follows from the fact that  $K_{\text{flow}} = K_{\text{BM}} \cdot K_{\text{OU}}$  with both factors corresponding to solutions of first-order SDEs, would enable an exact  $O(n)$  Kalman implementation via the results of Hartikainen and Särkkä [2010], rendering the framework directly applicable to tick-by-tick data. Replacing the Gaussian noise prior with a heavier-tailed alternative would connect the framework to robust smoothing and  $M$ -estimation in RKHS Schölkopf et al. [2001], which may be important for equity return data where extreme events are more frequent than Gaussianity implies. The PAC-Bayes bounds of Alquier [2024] offer a natural route to non-asymptotic complements to the minimax rates of Theorem 6. Finally, and most directly, the empirical evaluation itself should be extended beyond the single S&P 500 case study considered here: applying the IFS to other equity indices, single-name stocks, and other asset classes with different volatility dynamics (fixed income, FX, commodities) would test whether the reported vol-weighted advantage, and its concentration in high-volatility regimes, generalises beyond the setting studied in this paper.

## References

- P. Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024.
- S. Amari. *Information Geometry and Its Applications*. Springer, Tokyo, 2016.
- T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, 2003.
- C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer, New York, 1984.
- A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, New York, 2004.
- T. Bollerslev. Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.
- D. R. Burt, C. E. Rasmussen, and M. van der Wilk. Convergence of sparse variational inference in Gaussian process regression. *Journal of Machine Learning Research*, 21(131):1–63, 2020.
- R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.
- R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.
- R. L. Eubank. The hat matrix for smoothing splines. *Statistics & Probability Letters*, 2(1):9–14, 1984.
- J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London, 1996.
- S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.
- J. Gatheral, T. Jaisson, and M. Rosenbaum. Volatility is rough. *Quantitative Finance*, 18(6):933–949, 2018.
- J. Hartikainen and S. Särkkä. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *Proceedings of the IEEE MLSP*, pages 379–384, 2010.
- J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Proceedings of UAI*, pages 282–290, 2013.
- R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2nd edition, 2012.
- G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing splines. *Annals of Mathematical Statistics*, 41(2):495–502, 1970.

- H. R. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241, 1989.
- H. Liu, Y.-S. Cai, and P. Ghosh. When Gaussian process meets big data: A review of scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4405–4423, 2020.
- H.-G. Müller, R. Sen, and U. Stadtmüller. Functional data analysis for volatility. *Journal of Econometrics*, 165(2):233–245, 2011.
- S. W. Ober and C. E. Rasmussen. Benchmarking the neural linear model for regression. In *Third Symposium on Advances in Approximate Bayesian Inference (AABI)*, 2021.
- C. J. Paciorek and M. J. Schervish. Nonstationary covariance functions for Gaussian process regression. In *Advances in Neural Information Processing Systems 16*, pages 273–280. MIT Press, 2004.
- J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2nd edition, 2005.
- C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.
- B. Schölkopf, R. Herbrich, and A. J. Smola. A generalised representer theorem. In *Proceedings of COLT*, pages 416–426. Springer, 2001.
- I. Steinwart and C. Scovel. Mercer’s theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.
- M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of AISTATS*, volume 5, pages 567–574, 2009.
- A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.
- G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, 1990.
- M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge, 2019.
- B. Wild, G. Wynne, and B. Schölkopf. Generalization bounds for nonparametric regression with  $\beta$ -mixing samples. *Journal of Machine Learning Research*, 23(219):1–56, 2022.
- A. G. Wilson and R. P. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of ICML*, volume 28, pages 1067–1075. PMLR, 2013.
- A. G. Wilson and H. Nickisch. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *Proceedings of ICML*, volume 37, pages 1775–1784. PMLR, 2015.
- E. Wong-Toi, A. Boyd, V. Fortuin, and S. Mandt. On the effect of regularization on nonparametric mean-variance regression. *arXiv preprint arXiv:2511.22004*, 2025.