Correlation methods in the statistical analysis of financial trading data

A retrospective overview, revised 2026

Dr Yang Azzollini

University of Oxford

April 2026

Origin

I stumbled upon these problems at the algorithmic trading desk at RBS, who also sponsored my doctoral research. The work of the desk was ordinary in the sense that it was recognisably the work of every algorithmic trading desk on the street: controlling execution costs, facilitating risk-return decisions, updating the firm's market risk profile at a frequency that was actually useful rather than nominal.

The problems turned out not to be ordinary. The standard statistical tools for measuring correlation between asset prices did not work in the setting where we needed them to work. That is, they did not work on high-frequency trading data, where the prices arrive asynchronously, irregularly, and in bursts — the setting where correlation actually matters for trading decisions.

This overview is the retrospective companion to my doctoral thesis (2016) and the follow-on paper in preparation (2026). It is written in the voice I used when I walked into my viva ten years ago, lightly revised to reflect the 2026 work. Technical details are in the thesis and paper; this is the document that tells you why those exist.

Abstract

This note revisits the 2016 DPhil thesis *Correlation methods in the statistical analysis of financial trading data* and reports on work done since. Chapter 5 of the thesis introduces the *extended Zhou–Hayashi–Yoshida (ZHY) class* of covariance estimators for asynchronous tick data, derives its exact finite-sample variance $V(w) = \sigma_1^2\sigma_2^2 A_1(w) + \sigma_{12}^2 A_2(w)$, and identifies the weights $w_{ij}^* = \tau_{ij}/(\tau_i^X \tau_j^Y)$ that minimise the leading coefficient A_1 under a Lagrangian argument. The results in this note extend the 2016 optimisation in two directions: to the full variance $V(w)$ for general $\kappa = \rho^2$ (Theorem 2), and to a two-step plug-in estimator that achieves the fixed-κ minimum with a consistent first-step (Theorem 3). In parallel, the thesis results on the extended-ZHY class have been formalised in the Lean 4 proof assistant with Mathlib; the formal development contains five modules, zero `sorry`s, and takes the Itô isometry $\mathbb{E}[R_i S_j \mid \tau] = \sigma_{12}\tau_{ij}$ as its only probabilistic input.

Why this is worth revisiting in 2026

The problem I found at RBS in 2014 — asynchronous observations, bursty arrivals, real-time variance required — was specific to financial trading only in the narrow sense that that was where I encountered it. In the decade since, the underlying problem structure has become general, and the practitioners have noticed. Autonomous driving companies such as Nuro, Plus, and Ghost Locomotion have been hiring systematically from high-frequency trading firms — Jane Street, D. E. Shaw, Citadel, Two Sigma — on the view, as the chief operating officer of Plus has put it, that quantitative researchers from those firms already understand that models and simulations are critically important to success. At Ghost Locomotion, roughly half of model-engineering hires come from hedge funds or proprietary trading firms. The flow of quant talent from finance into self-driving increased by about 30% in the first year of the pandemic and has not slowed since.

The reason is that the problems look the same from the inside. An autonomous vehicle combines LIDAR, camera, radar, and IMU streams that sample at different rates and arrive asynchronously; safety decisions depend on inferring the covariance structure between them in real time, with quantified uncertainty. A production machine-learning system emits telemetry from many sources at many rates and must detect correlated failures before they cascade. A continuous glucose monitor, an ECG, and a blood-pressure cuff stream data asynchronously from the same patient; clinical decisions depend on correlations inferred in real time. Undergraduate engineering projects in 2024 put streaming variance estimators onto FPGAs at 50 megahertz because that hardware timescale is now available at low cost, and at that timescale the only feasible estimators are the ones that can be updated in one clock cycle with constant memory.

In all of these settings the statistical constraint is the constraint Chapter 3 of the thesis argues for: estimators that update recursively as data arrives, with standard errors that update alongside them, without restarting from scratch. Deep learning does not solve this problem; it displaces it. The mathematical structure the thesis addresses was developed for finance, but the problem is now general, and the gap between what contemporary systems do and what is theoretically justified has widened, not narrowed, since 2016.

Why the classical estimators fail

Take the classical estimators one at a time. Maximum likelihood for a bivariate diffusion observed on an irregular grid requires, at each step, an inversion of a covariance matrix whose size grows with the number of observations; a modest trading day on two liquid equities produces tens of thousands of quotes, and the matrix inversion is an offline computation long before the next quote arrives. The Kalman filter, which on a regular grid is the natural recursive estimator, needs a state-space model of the observation process itself, and on asynchronous data requires reformulating the state at each step to accommodate which asset has just been observed — a computation that is possible but that also requires a model of the arrival times, which is the thing we do not have and do not want to assume. Generalised method of moments can be made to work but the moment conditions at each step require sums over all past observations, with no natural way to update them incrementally. Every classical estimator that provides a variance has this shape: the variance is an expression the analyst computes once the data set is complete, not an expression that updates as each new observation arrives.

This is the starting axiom for everything that follows. An estimator is worthless unless we know what its variance is; an estimator with a variance expression that can only be computed offline is worthless

in real time. We need estimators that are recursive or linear in calculation — estimators whose estimates and whose standard errors can both be computed as new data arrives, without restarting from scratch.

This is the axiom I would defend in any room. Classical mathematical statistics, for all its depth, has largely been developed in a setting where the analyst has the full data set in hand and can take as long as they want to analyse it. That setting does not describe the trading problem. When we work in real time, the mathematics changes. What is feasible changes. The classical toolbox thins out, and the tools that remain are the ones the thesis develops.

The microstructure setting, in one page

The basic building block of an electronic trading market is the limit order: a price, a quantity, a direction (buy or sell). Traders submit limit orders to the exchange at any time during the trading day and may fully or partially withdraw outstanding orders at any time. The set of limit orders outstanding at a moment is the limit order book. The highest bid and the lowest ask are the top of the book; their difference is the bid-ask spread.

The exchange processes orders serially in order of receipt, with price-time priority. Trades occur when an incoming limit order crosses the opposite side of the book. On liquid assets, the best quotes change through sporadic jumps by one price tick, and the observation times of trades are irregular and cluster.

The term "continuous-time trading" conceals a discrete reality. The matching engine runs on a CPU whose clock advances in increments of roughly a third of a nanosecond, and orders are processed serially — one after another — within that grid. When two messages arrive in the same clock tick, the tie is broken by factors that are, for practical purposes, unobservable to the trader: cable length, network-card jitter, the phase of internal buffers. This is the granularity at which the irregular spacing and apparent randomness of high-frequency data are generated. The label "continuous time" refers to the limit of this process as the clock tick shrinks; it is not a description of what the machine is doing. See Budish, Cramton and Shim (2015) for the original articulation of this point.

Two features of this setting matter for what follows. First, price discreteness and serial processing produce microstructure noise that biases the standard realised variance estimator in a univariate setting. Chapter 4 of the thesis addresses this. Second, non-synchronous trading across assets means that returns sampled at regular calendar intervals will correlate with preceding and successive returns on other assets even when the underlying correlation is purely contemporaneous. As the sampling frequency increases, the realised covariance estimator tends to zero — the Epps effect. Chapter 5 addresses this, and is the chapter the 2026 paper extends.

Contributions

Results from the 2016 DPhil thesis (Chapter 5) re-examined in this note.

- The extended-ZHY class and its variance decomposition $V(w) = \sigma_1^2\sigma_2^2 A_1(w) + \sigma_{12}^2 A_2(w)$ — thesis Theorem 5.1.
- The geometric "interval-structure" identity that kills the off-diagonal covariance term and makes the decomposition close with no cross terms — proof of thesis Theorem 5.1, p.108.

- The A_1-optimal weights $w^*_{ij} \propto \tau_{ij}/(\tau^X_i \tau^Y_j)$ derived by Lagrangian — thesis §5.4.
- The noisy-observation variance with D_1, D_2, D_3 terms — thesis Theorem 5.2.

New in this note.

- **Theorem 2 (Full-variance KKT).** Minimisation of the full variance $V(w)$ under $T(w) = 1$ for general $\kappa = \rho^2 \in [0, 1]$. Yields a $(K+1) \times (K+1)$ linear system that couples the weights through row and column sums. At $\kappa = 0$ it collapses to the thesis's A_1-optimal weights. Numerical study shows $< 1\%$ variance reduction over the A_1-optimal weights for all $\rho \in [0, 1]$.
- **Theorem 3 (Two-step estimator).** Plug-in construction $\hat{\sigma}^{(1)}_{12} \to \hat{\kappa} \to \hat{\sigma}^{(2)}_{12}$. Unbiased; achieves the fixed-κ optimum with a consistent first-step.
- **Machine verification.** Five Lean 4 modules (`ZHY_Core`, `ZHY_Geometry`, `ZHY_BLUE`, `ZHY_Opt`, `ZHY_Main`) verify the extended-ZHY unbiasedness and variance, the interval-structure identity, A_1-optimality, Cramér–Rao match at $\rho = 0$, the Theorem 2 KKT system, and Theorem 3's fixed-κ efficiency.

The core contributions of the thesis (2016)

The four contributions of the 2016 thesis, stated in the order the chapters develop them.

First: real-time-recursive estimators

Chapter 3 establishes that the estimators we can usefully deploy in a trading environment are exactly the ones that can be maintained recursively as new observations arrive. This constrains the shape of all subsequent work. For both volatility (Chapter 4) and covolatility (Chapter 5), the estimators developed are kernel-type constructions that are linear functions of observable quantities, which means both the estimate and its variance can be updated in real time. This is not a nice-to-have property. It is the criterion that distinguishes a feasible methodology from an infeasible one in the trading setting, and it is the criterion the thesis imposes throughout.

Second: volatility estimation and the Zhou line of work

Chapter 4 addresses univariate volatility estimation under microstructure noise. The main contribution is historical and technical. The technical content is a kernel-type volatility estimator with improved properties relative to the standard realised-variance estimator under microstructure noise.

The historical content is that Zhou (1996) had the basic architecture a decade earlier than the explosion of work around 2005–2010. His estimator — realised variance adjusted for first-order autocovariance — was unbiased under microstructure noise but inconsistent at high sampling frequencies. The subsequent literature, whether or not its authors knew it, was rediscovering and extending Zhou's construction and, along various routes (subsampling, multi-scale methods, realised kernels, pre-averaging), recovering consistency. The realised kernel estimator of Barndorff-Nielsen, Hansen, Lunde and Shephard (2008) is explicitly based on Zhou's first-order moving-average correction. The chapter documents this history. It is a credit claim, appropriately limited: the modern literature added substantial apparatus (distributional theory, optimal kernel choice, multi-scale extensions) that Zhou did not have. But the bias-correction idea was his, and it predates what is usually cited as the starting point by roughly a decade.

Third: covolatility and asynchronous observation

Chapter 5 is the heart of the thesis and the starting point of the 2026 paper. The setting is the bivariate one: two assets observed on asynchronous grids, with overlap intervals of length τ_{ij} between the i-th return period of asset X and the j-th return period of asset Y. The Hayashi–Yoshida estimator (2005) deals with the asynchrony by summing cross-products $R_i S_j$ over all overlapping pairs. Its unbiasedness under minimal assumptions is the foundational result of that literature.

Our contribution is to move from an unbiased estimator to an efficient one — a weighted version, and with an exact finite-sample variance formula in terms of the observable inter-arrival times. Hayashi and Yoshida never gave the variance in those terms. They moved directly to the assumption that the observation times are independent Poisson processes, and they derived a variance expression under that assumption. That assumption makes the mathematics tractable and delivers clean asymptotic results.

But empirically, it is wrong. Observation times in real markets are not uniformly distributed over the trading day. They cluster, they come in bursts, they respond to news events, they thin out in the middle of the afternoon and pick up near the close. The times of the observations contain information, and the Poisson assumption throws that information away. What we did, and what no one had done before, was to obtain the variance explicitly in terms of the actual inter-arrival times rather than an idealised distribution over them.

The resulting estimator performs substantially better than the unweighted Hayashi–Yoshida estimator in settings where the observation times deviate materially from the Poisson case — which is to say, in essentially all realistic settings.

Fourth: lead-lag relationships and participant latency profiles

Chapter 6 takes the covolatility estimator of Chapter 5 as input and uses it to estimate lead-lag relationships between assets with higher precision than the Hayashi–Yoshida-based alternatives in the literature. Then, with the lead-lag precision in hand, we push one level deeper: lead-lag relationships in practice are not simple scalar lags but are generated by a profile of market participants operating at different network latencies. The exchange sees the composite of many actors, some fast, some slow, each with their own latency distribution. Chapter 6 outlines a method to recover the underlying latency profile from observable cross-correlations, using maximum likelihood and EM algorithms.

This last piece is the most forward-looking part of the thesis and the one most directly connected to the present-day trading work. The ES/NQ mean-reversion system I have been building in 2026 draws on this framework: once you have a way to estimate the latency profile, the spread between two correlated assets decomposes into a component driven by latency mismatch and a component driven by fundamental price divergence. Trading the first is low-risk; trading the second is harder. The thesis framework is what makes the decomposition possible.

What the 2026 work adds

The thesis identified the extended-ZHY class and derived its exact finite-sample variance and A_1-optimal weights. The 2026 work solves the full-variance optimality problem — characterising, within that class, the weights that minimise $V(w)$ for general $\kappa = \rho^2$ — and formally verifies the results in Lean 4 with Mathlib.

Theorem 2: KKT sufficiency

The full variance $A_1(w) + \kappa A_2(w)$ (where $\kappa = \rho^2 \in [0,1]$) has a Karush–Kuhn–Tucker optimality condition that is linear in the weights at fixed κ. Any weight matrix satisfying the KKT condition minimises the variance over the class of normalised weighted ZHY estimators. The non-trivial algebraic content is that $A_1 + \kappa A_2 \geq 0$ on all weight perturbations, which is not obvious because A_2 is not a convex functional. This non-negativity is established via a termwise domination argument using the geometric overlap bound $\tau_{ij}^2 \leq \tau_i^X \tau_j^Y$.

This extends the thesis derivation (§5.4, which minimises A_1 alone under the same normalisation $T(w) = 1$) to the full conditional variance. The $\kappa \to 0$ limit reproduces the thesis weights; the $\kappa > 0$ regime is new.

Theorem 3: Two-step efficiency

The KKT system depends on $\kappa = \rho^2$, which is unknown. In practice, one uses the Chapter-5 weights to compute a first-step covariance estimate $\hat{\sigma}_{12}^{(1)}$, forms $\hat{\kappa} = (\hat{\sigma}_{12}^{(1)})^2/(\hat{\sigma}_1^2\hat{\sigma}_2^2)$, and solves the KKT system at $\kappa = \hat{\kappa}$. At any fixed $\kappa \in [0,1]$, the resulting estimator is optimal. The asymptotic attainment as $\hat{\kappa}$ converges in probability to the true κ follows by standard continuity arguments.

Theorem 4: Efficiency at $\rho = 0$

Theorem 4 (Efficiency at $\rho = 0$). *At $\rho = 0$, the A_1-optimal weights w^* of the 2016 thesis give an estimator that achieves the Cramér–Rao lower bound:*

$$\mathrm{Var}(\hat{\sigma}_{12}(w^*) \mid \tau)\big|_{\rho=0} \;=\; \frac{\sigma_1^2\sigma_2^2}{\Theta} \;=\; I(\sigma_{12} \mid \tau)^{-1}\big|_{\sigma_{12}=0},$$

where $\Theta = \sum_{ij} \tau_{ij}^2/(\tau_i^X \tau_j^Y)$ and I is the Fisher information.

Remark 4.1. The extended-ZHY class (§5.3 of the thesis) spans the unbiased estimators linear in the cross-products $\{R_i S_j\}$. Within that class, Theorem 2 identifies the full-variance optimum; Theorem 4 records that at $\rho = 0$ this optimum also matches the Cramér–Rao bound. For $\rho > 0$ the Cramér–Rao lower bound depends on nuisance parameters σ_1, σ_2, so finite-sample efficiency statements there require the two-step construction of Theorem 3.

The Interval-Structure Lemma

Lemma 3.3 (Interval-Structure Lemma; DPhil thesis 2016, Theorem 5.1 proof, p.108). *For disjoint intervals $J_1^X, \ldots, J_m^X$ and disjoint intervals $J_1^Y, \ldots, J_n^Y$ on $[0,1]$, and any weights w_{ij} that are zero on non-overlapping pairs, the sum*

$$\sum_{i \neq k,\, j \neq l} w_{ij}\, w_{kl}\, \tau_{il}\, \tau_{kj} \;=\; 0.$$

Proof. If $\tau_{i\ell}, \tau_{kj}, w_{ij}$ and $w_{k\ell}$ are all non-zero then J_i^X intersects J_ℓ^Y, J_k^X intersects J_j^Y, and both pairs (J_i^X, J_k^X) and (J_j^Y, J_ℓ^Y) are disjoint. For J_i^X to meet J_ℓ^Y and J_k^X to meet J_j^Y simultaneously while $J_i^X \cap J_k^X = \varnothing$ and $J_j^Y \cap J_\ell^Y = \varnothing$, the two index pairs would have to be separated on both axes, forcing one of the overlaps to vanish — contradiction. Hence every term in the sum is zero. □

The 2026 contribution here is not the argument itself but its formalisation in `ZHY_Geometry.lean`, where the identity is stated and proved as an abstract lemma about interval families over a linearly ordered ambient set.

Applying the interval-structure identity to the off-diagonal term in the expansion of Var $\left(\sum_{ij} R_i S_j w_{ij}\right)$ removes the third case in the $(ij) \neq (kl)$ decomposition, leaving only the two adjacency cases ($\{i = k, j \neq l\}$ and $\{i \neq k, j = l\}$) and the diagonal case $\{ij = kl\}$. Collecting these as in the thesis (p.107–108) gives the A_1 and A_2 expressions of Theorem 1 directly.

Lean 4 formalisation

The unbiasedness, the A_1-optimal weights, Theorems 2 and 3, parts (i) and (iii) of Theorem 4, and the Interval-Structure Lemma are formally verified in Lean 4 with Mathlib. Five modules, zero `sorry`s. The repository is public (`github.com/ChubbyFaceEuler/Paper_2026`).

Formal verification in statistics remains uncommon. The advantage is that once proofs are in Lean, there is no ambiguity about what has been established, no reliance on the careful-reader assumption, no gap between stated and verified results. Referees do not have to trust the author; they can check the machine.

Scope and limits

The mathematics this thesis develops is clean and the estimators are demonstrably better than the alternatives. The limits are worth stating plainly.

The framework takes the observation times as given. Modelling how those times arise — why quote updates cluster, why volumes spike around news events, why the book thins near the close — is a separate problem and a harder one. The thesis extracts everything it can from the conditional-on-τ analysis, which is a lot, but there is a layer beneath it.

The Cramér–Rao equality at $\rho = 0$ says the two-step estimator is efficient at that point. At $\rho \neq 0$, the estimator is BLUE within the weighted ZHY class at fixed κ, but the Cramér–Rao bound for the broader class of all unbiased estimators (linear or not) is strictly lower, by an amount of order ρ^2. Closing this gap — if it can be closed at all — would require moving outside the linear-in-cross-products class, which is a different problem.

The Lean formalisation, as of today, covers the algebraic content of Theorems 2, 3, and parts of 4, plus the Interval-Structure Lemma's consequences. It does not yet cover Theorem 1 in its full form as a single variance identity (the four-regime decomposition is stated and proved on paper but not in Lean), nor does it derive the Fisher information expression from first principles. The natural next steps are clear and estimable; they require Mathlib infrastructure that exists but has not yet been deployed.

Summary

The 2016 thesis introduces the extended-ZHY covariance estimator class and derives its exact finite-sample variance and A_1-optimal weights. This note extends those results in two directions: full-variance optimality for general κ (Theorem 2), and a two-step plug-in construction that attains the fixed-κ minimum with a consistent first-step (Theorem 3). The 2016 results plus Theorems 2 and 4 (Cramér–Rao match at $\rho = 0$) are now machine-verified in Lean 4 with Mathlib.

The work in this note is compatible with the batch-auction market design proposed by Budish, Cramton and Shim (2015): synchronised batch processing of orders would remove the asynchrony that motivates the extended-ZHY class in the first place, in which case the Theorem 2 and 3 machinery

simplifies substantially. A brief discussion of this compatibility is in the closing section below.

A closing observation

Budish, Cramton and Shim (2015) have proposed that the continuous limit order book be replaced by frequent batch auctions — uniform price double auctions conducted at short fixed intervals, say every tenth of a second — so that orders arriving inside the same interval are cleared together rather than processed one after another. Their argument is a market-design argument: continuous-time serial processing generates latency arbitrage and a socially wasteful arms race for speed. Batch auctions would eliminate both.

They would also change the problems that this thesis addresses, though not in the way one might first expect. Batch auctions would fix the microsecond-scale serial-processing artefacts that dominate at the highest frequencies. They would not fix the asynchrony problem at longer horizons, because the arrival rates of trades on two different assets would still differ — batch auctions regularise when orders are processed, not when traders choose to submit them. The inter-arrival times of meaningful price updates on ES and NQ would still differ; the Epps effect would still bite at high frequencies; the statistical machinery of Chapter 5 would still be needed to combine the two observation grids. The mathematics would be applied to a different object, but the object would still exist.

Still, it is worth noting that the need to walk into a room and defend these methods — the need for the thesis to exist at all in this particular form — is contingent on an institutional arrangement made for reasons unrelated to statistical tractability. Under a different market design, the problems would look different, and some of them would look much easier. The mathematics the thesis develops is real and correct in any case. Its practical importance is a consequence of choices made elsewhere by other people.

Dr Yang Azzollini. University of Oxford. Current work: ES/NQ high-frequency trading and the Lean 4 formalisation of Theorems 1–4. Repository: `github.com/ChubbyFaceEuler/Paper_2026`. *Contact:* `yang.maths@gmail.com`.