








Modelling Financial Markets

Chapter 20 (special chapter): Path Signatures and Realised Volatility — a Research Case Study



Daniel Traian PELE

Bucharest University of Economic Studies
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Romanian Academy, Institute for Economic Forecasting
MSCA Digital Finance

Do Path Signatures Improve Volatility Forecasts?



Terry Lyons. **Photo: Katrin Schmid (2024); CC BY-SA 2.0 DE; Wikimedia Commons**

"Signature-based techniques give mathematical insight into the interactions between complex streams of evolving data."

— *Lyons & McLeod (2022)*

- Gu et al. (2024) summarise a window of data by its path signature
  - for freight rates, they report a more than fivefold accuracy gain over an industry forecast
- The chapter transfers the method to the daily realised variance (RV) of 50 assets from the VOLARE database (VOLatility Archive for Realized Estimates)
  - benchmark: log-HAR, the HAR (Heterogeneous AutoRegressive) model Corsi (2009) applied to log RV
  - four hypotheses fixed before the results
- At one day, signature features bring no gain: Sig-L has a mean QLIKE loss 0.4% higher than log-HAR
  - QLIKE: the quasi-likelihood loss; a higher value means a worse forecast
  - Sig-L is significantly better on 3 of 50 assets (1 after the Holm correction); about 2.5 are expected by chance
- The strongest effect in the study is the oldest one: model the logarithm of RV
  - at one day, log-HAR is significantly better than HAR in levels on 21 of 50 assets (5 after the Holm correction)

▸ Central chart: QLIKE Relative to HAR, Asset by Asset

Does a Method Built for Freight Rates Improve Volatility Forecasts?

Path Shape and Volatility Forecasts

- A research case study: the method of Gu et al. (2024), transferred to daily realised variance
  - the method was published for freight rates at Amazon
  - data: 50 assets from VOLARE (VOLatility Archive for Realized Estimates) Cipollini et al. (2026)
- Detailed examples and advanced topics are in Further Topics, after the Conclusions
  - the footer button leads there; the Back button returns to the starting slide

What you should be able to do

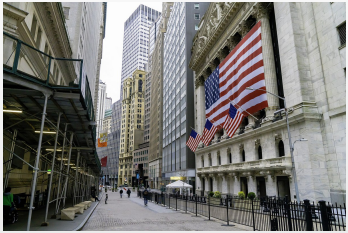
- Compute the signature of a piecewise-linear path and its Lévy area
- Compute signature features and kernel weights for a volatility forecast
- Interpret forecast comparisons on many assets
  - with Diebold–Mariano (DM) tests and model confidence sets
  - with Holm and Benjamini–Hochberg (BH) corrections
- Interpret the result of a pre-registered method transfer, including a null result and its mechanism
- Further Topics and Appendix: reading for familiarity

Why Forecast Realised Variance?

- Volatility is an input to almost every risk measure in this course
  - Tail-risk measures (Chapter 7), option prices (Chapter 12), portfolio weights (Chapter 4)
- Intraday data make daily variance almost observable
  - the measure is realised variance (RV, Chapter 9)
  - So volatility forecasting becomes a regression problem with a noisy but model-free target
- The benchmark is hard to beat
  - the HAR (heterogeneous autoregressive) model Corsi (2009) (Chapter 9)
  - Three averages of past RV; linear; estimated by OLS (ordinary least squares) in a second
- Any new method must beat HAR out of sample, on many assets, with a proper test

Two Ingredients to Test: Signature Features and Kernel Weights

- Gu et al. describe a recent window of a series as a **path** and summarise it by its **signature**
  - Signature features enter a sparse linear regression (LASSO, least absolute shrinkage and selection operator, in two steps)
  - Past periods whose path looks like today's get more weight (a signature kernel)
- They report a more than fivefold accuracy gain over an industry forecast of freight rates
  - Question: does the same recipe improve on HAR-type models for realised variance?
- Each ingredient is tested separately, so that a gain can be attributed to the features or to the weights



Wall Street and the New York Stock Exchange, 25 March 2020

Photo: Billie Grace Ward (2020); CC BY 2.0; Wikimedia Commons

Realised Variance and Quadratic Variation (1/2)

- **Question**: what exactly do we forecast, and how do intraday data measure the variance of a day?
- The price model: the log price moves by a drift, a diffusion term and jumps

$$dp_t = \mu_t\, dt + \sigma_t\, dW_t + dJ_t$$

  - p_t: log price at time t (time in days); μ_t: drift; σ_t: spot volatility; W_t: Brownian motion; J_t: pure-jump process, ΔJ_s = jump size at time s
- Realised variance: the sum of squared intraday returns of day t Barndorff-Nielsen & Shephard (2002)

$$\mathrm{RV}_t = \sum_{i=1}^{M} r_{t,i}^2 \xrightarrow{M\to\infty} \int_{t-1}^{t} \sigma_s^2\, ds + \sum_{t-1<s\le t} (\Delta J_s)^2$$

  - M: number of intraday returns in day t (for a US stock, 78 five-minute returns in the 6.5-hour session); $r_{t,i}$: the i-th intraday log return of day t
  - The limit is the quadratic variation of day t: the integrated variance plus the sum of squared jumps

Realised Variance and Quadratic Variation (2/2)

- Without jumps and noise, on an equally spaced grid, the limit is the integrated variance
  - the error is mixed Normal

$$\sqrt{M}\,(\mathrm{RV}_t - \mathrm{IV}_t) \Rightarrow \mathrm{MN}(0,\ 2\,\mathrm{IQ}_t)\,, \qquad \mathrm{IV}_t = \int_{t-1}^{t} \sigma_s^2\, ds, \quad \mathrm{IQ}_t = \int_{t-1}^{t} \sigma_s^4\, ds$$

  - IV_t: integrated variance, the true variance of day t; IQ_t: integrated quarticity
  - $\mathrm{MN}(0, V)$: mixed Normal, a Normal with a random variance V; given the volatility path, the error is Normal with variance $2\,\mathrm{IQ}_t/M$
- Reading: RV is a noisy measurement of IV_t
  - the noise is largest on volatile days (where IQ_t is large)
  - This is the measurement error that HARQ, the HAR model with realised quarticity, corrects for (HAR section)
- Microstructure noise makes very fine sampling biased
  - five-minute returns are the standard compromise (Chapter 9)

Semivariances and Quarticity in VOLARE

- Semivariances Barndorff-Nielsen, Kinnebrock & Shephard (2010): RV split by the sign of the intraday returns

$$\mathrm{RV}_t^+ = \sum_{i=1}^{M} r_{t,i}^2 \, 1\{r_{t,i} > 0\}, \qquad \mathrm{RV}_t^- = \sum_{i=1}^{M} r_{t,i}^2 \, 1\{r_{t,i} < 0\}$$

  - RV_t^+: variance from up-moves, RV_t^-: variance from down-moves
  - They add up: $\mathrm{RV}_t = \mathrm{RV}_t^+ + \mathrm{RV}_t^-$
- Realised quarticity (RQ): an estimate of IQ_t, the scale of the measurement error of RV

$$\mathrm{RQ}_t = \frac{M}{3} \sum_{i=1}^{M} r_{t,i}^4 \;\to\; \mathrm{IQ}_t$$

  - The factor $M/3$: a Normal return with variance s^2 has fourth moment $3s^4$, and each five-minute variance is about IV_t/M
  - The limit needs a continuous, noise-free price and an equally spaced grid; a jump ΔJ adds about $M(\Delta J)^4/3$, so raw RQ is jump-sensitive

Targets and Losses (1/2): the Target

- The target at horizon h is the average RV over the next h days

$$\mathrm{RV}_{t+\mathbf{1}:t+h} = \frac{1}{h}\sum_{j=\mathbf{1}}^{h} \mathrm{RV}_{t+j}, \qquad h \in \{1, 5, 22\}$$

  - t: the forecast origin (the last day observed); $h = 1, 5, 22$: one day, one week, one month of trading days
- Log models forecast $y = \ln \mathrm{RV}_{t+\mathbf{1}:t+h}$ and return to levels with a variance correction

$$F_t = \exp(\hat{y}_t + \hat{s}^{\mathbf{2}}/2)$$

  - F_t: forecast of $\mathrm{RV}_{t+\mathbf{1}:t+h}$ made at t; $\hat{y}_t$: fitted log forecast; $\hat{s}^{\mathbf{2}}$: variance of the regression residuals
  - $+\hat{s}^{\mathbf{2}}/2$: since $\mathbb{E}[e^Z] = e^{\mu + s^{\mathbf{2}}/\mathbf{2}}$ for $Z \sim N(\mu, s^{\mathbf{2}})$; exact if the log error is conditionally Normal with constant variance, otherwise an approximation
- For $h > 1$ targets overlap: this induces serial dependence
  - through lag $h - 1$ by construction, possibly longer
  - so every test uses a HAC (heteroskedasticity and autocorrelation consistent) variance

Matching a Forecast to Its Target

- In the target definition, the subscript t is the forecast origin
- In loss and test formulas, the subscript labels a matched evaluation row
  - the realised target
  - the forecast made before that target became observable
- One-day horizon: information through Monday close forecasts Tuesday RV
  - stored next to the Tuesday realisation
- Five-day horizon: the same origin forecasts the average from Tuesday to the following Monday
  - the next origin gives an overlapping target
- Forecast comparison tests use the same realised target for both models in each row
- Sharing a row with the forecast does not allow the target to be used in estimation

Targets and Losses (2/2): Robust Losses

- A loss is **robust** Patton (2011)
  - for a conditionally unbiased proxy, the *expected*-loss ranking with RV_t equals that with the true variance
  - Conditionally unbiased: $\mathbb{E}_{t-\mathbf{1}}[\mathrm{RV}_t] = \sigma_t^{\mathbf{2}}$, with $\mathbb{E}_{t-\mathbf{1}}$ the expectation given information up to $t-1$ and $\sigma_t^{\mathbf{2}}$ the true variance of day t
- The two robust losses used here (written for $h = 1$; for $h > 1$ replace RV_t by the target)

$$\mathrm{MSE}_t = (\mathrm{RV}_t - F_t)^{\mathbf{2}}, \qquad \mathrm{QLIKE}_t = \frac{\mathrm{RV}_t}{F_t} - \ln \frac{\mathrm{RV}_t}{F_t} - 1 \;\geq\; 0$$

  - MSE (mean squared error): the squared error; QLIKE (quasi-likelihood loss) is zero only if $F_t = \mathrm{RV}_t$ and grows faster when the forecast is too low than when it is too high
  - Example with $\mathrm{RV}_t = 1$: $F_t = 0.5$ gives $\mathrm{QLIKE}_t = 2 - \ln 2 - 1 = 0.307$, $F_t = 2$ gives $0.5 - \ln 0.5 - 1 = 0.193$
- QLIKE depends on the ratio RV_t / F_t
  - the same proportional error has the same loss on a calm day as on a stormy day
- MSE is dominated by a few crisis days
  - QLIKE is the primary loss, MSE the secondary one

The Panel: 50 Assets in Three Classes

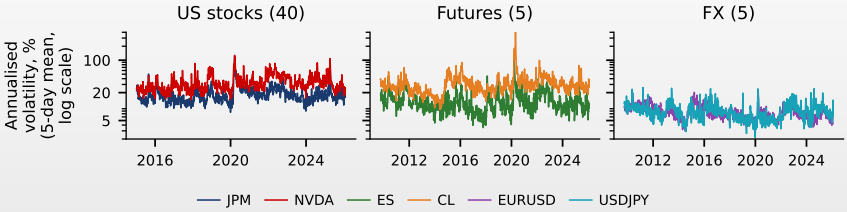
Class	Assets	Sample	Observations	Median volatility
American large-cap stocks	40	2 Jan 2015 – 30 Jan 2026	111 440	17.1%
Futures (C, CL, ES, GC, NG)	5	28 Sep 2009 – 30 Jan 2026	21 059	19.7%
FX (AUD, EUR, GBP, CAD, JPY vs USD)	5	25 Sep 2009 – 30 Jan 2026	21 203	7.3%

- **Question**: on which data do we test the transfer, and why on three asset classes?
- Assets
  - Stocks: 40 members of the Dow Jones 30 and S&P 100 (Standard & Poor's), labelled by ticker, from AAPL (Apple) to XOM (Exxon Mobil)
  - Futures: C = corn, CL = crude oil, ES = E-mini S&P 500, GC = gold, NG = natural gas
  - FX (foreign exchange): Australian dollar, euro, British pound, Canadian dollar and Japanese yen against the US dollar

The Panel: Median Volatility and Variables

- Three asset classes
  - a gain that appears in one class only does not show that the method transfers
- Median volatility in the table: $\sqrt{252\,\mathrm{median}(\mathrm{RV}_t)}$ per asset, then the median over the class
  - 252 trading days a year turn a daily variance into an annual one
  - The data contain trading days only
- Variables
  - five-minute RV, semivariances RV^+, RV^-, realised quarticity (RQ)
  - daily open and close prices

Realised Volatility Across Asset Classes (1/2)



- Annualised volatility $\sqrt{252\,\mathrm{RV}_t}$ in %, five-day mean, log scale, two assets per class
  - stocks: JPM = JPMorgan Chase, NVDA = Nvidia
  - futures: ES = E-mini S&P 500, CL = crude oil
  - FX: EURUSD = US dollars per euro, USDJPY = yen per US dollar
- All classes share March 2020; crude oil (CL) peaks in April 2020
  - a forecaster must adapt within days
- CL: the continuous VOLARE series joins successive contracts
  - the negative price of 20 April 2020 concerned only the May contract, close to expiry
  - the series stays positive

▶ **Interactive**

Realised Volatility Across Asset Classes (2/2)

- On a log scale volatility is persistent (the long memory HAR approximates)
- Spikes decay within weeks, so the shape of the recent path may carry information

MFM_ch20_case_study

HAR: Three Averages of the Past (1/2)

- **Question**: which simple benchmark must be beaten before we believe a new method?
- The HAR model Corsi (2009): future RV is a linear function of the daily, weekly and monthly averages of past RV

$$\mathrm{RV}_{t+\mathbf{1}:t+h} = \beta_{\mathbf{0}} + \beta_d \,\mathrm{RV}_t + \beta_w \,\mathrm{RV}_t^{(w)} + \beta_m \,\mathrm{RV}_t^{(m)} + \varepsilon_{t+h}$$

$$\mathrm{RV}_t^{(w)} = \frac{1}{5}\sum_{j=\mathbf{0}}^{\mathbf{4}} \mathrm{RV}_{t-j}, \qquad \mathrm{RV}_t^{(m)} = \frac{1}{22}\sum_{j=\mathbf{0}}^{\mathbf{21}} \mathrm{RV}_{t-j}$$

- Notation
  - $\mathrm{RV}_t^{(w)}$, $\mathrm{RV}_t^{(m)}$: weekly (5-day) and monthly (22-day) averages of RV up to day t; ε_{t+h}: forecast error
  - $\beta_{\mathbf{0}}$: intercept; $\beta_d, \beta_w, \beta_m \geq 0$ in practice: how much yesterday, last week and last month weigh; estimated by OLS for each horizon h
  - $\beta_d + \beta_w + \beta_m$ close to 1: high persistence of volatility
- Traders at daily, weekly and monthly frequencies
  - a cascade that mimics long memory with three parameters

HAR: Three Averages of the Past (2/2)

- HAR is a restricted AR(22): the 22 lag coefficients take only three distinct values

$$\phi_{\mathbf{1}} = \beta_d + \frac{\beta_w}{5} + \frac{\beta_m}{22}, \qquad \phi_{\mathbf{2..5}} = \frac{\beta_w}{5} + \frac{\beta_m}{22}, \qquad \phi_{\mathbf{6..22}} = \frac{\beta_m}{22}$$

  - ϕ_j: the coefficient of $\mathrm{RV}_{t+\mathbf{1}-j}$ in an AR(22) (autoregression with 22 lags)

- log-HAR: the same regression in logs

$$\ln \mathrm{RV}_{t+\mathbf{1}:t+h} = \beta_{\mathbf{0}} + \beta_d \ln \mathrm{RV}_t + \beta_w \ln \mathrm{RV}_t^{(w)} + \beta_m \ln \mathrm{RV}_t^{(m)} + \varepsilon_{t+h}$$

  - No negative forecasts, less driven by spikes
  - Logs of averages are not averages of logs, so this log-HAR is not an exact restricted AR(22) in $\ln \mathrm{RV}$ (Seminar 20, A7)

Seven Models in One Table

Model	Target	Weights	Regressors
HAR	RV	equal	$\mathrm{RV}_t, \mathrm{RV}_t^{(w)}, \mathrm{RV}_t^{(m)}$
log-HAR	$\ln \mathrm{RV}$	equal	the same, in logs
HARQ	RV	equal	HAR + $\mathrm{RV}_t \left(\mathrm{RQ}_t^{\mathbf{1/2}} - \overline{\mathrm{RQ}^{\mathbf{1/2}}}\right)$
SHAR	RV	equal	$\mathrm{RV}_t^{+}, \mathrm{RV}_t^{-}, \mathrm{RV}_t^{(w)}, \mathrm{RV}_t^{(m)}$
log-HAR-K	$\ln \mathrm{RV}$	signature kernel	log-HAR regressors
Sig-L	$\ln \mathrm{RV}$	equal	log-HAR + 36 non-constant signature terms (39 before dropping constants), two-step LASSO
Sig-LK	$\ln \mathrm{RV}$	signature kernel	as Sig-L: the method of Gu et al.

- Names: L = LASSO, K = kernel weights
  - the charts call Sig-L "Signature LASSO", Sig-LK "Signature LASSO + kernel weights", log-HAR-K "log-HAR + kernel weights"
- log-HAR, log-HAR-K, Sig-L and Sig-LK form a 2×2 ablation
  - each ingredient switched on and off separately
  - signature features (yes/no) $\times$ kernel weights (yes/no)
- Log-HAR is the natural baseline for the signature models
  - same target, and its regressors are among the candidates
  - but the LASSO may also drop log-HAR terms
  - so Sig-L vs log-HAR compares two procedures, not only two feature sets

A Window of Data Is a Path

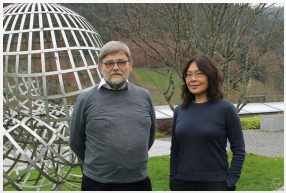
- **Question**: how do we summarise a 22-day window so that the order of the moves is kept, not only their levels?
- Take the last L days of $d-1$ series
  - add time as the first channel (time augmentation)
  - connect the points linearly

$$X : [0,1] \to \mathbb{R}^d, \qquad X_{u_i} = \left(u_i,\ \ln \mathrm{RV}_{t-L+1+i},\ \textstyle\sum_{j \le t-L+1+i} r_j\right), \qquad u_i = \tfrac{i}{L-1},\ i = 0, \dots, L-1$$

  - L: window length in days (here 22); d: number of channels (here $d = 3$: time, log RV, cumulative return); $u_i \in [0,1]$: rescaled time of the i-th point; $r_j = \ln(C_j/O_j)$: open-to-close log return of day j, with O_j, C_j the open and close prices
- A regression on the window sees 22 lags; a path summary sees their order, speed and co-movement
  - Example: RV rose while prices fell, or RV rose first and prices fell later

Where Signatures Come From

- Kuo-Tsai Chen, geometry of loop spaces
  - Chen (1957): the iterated integrals of a path form an algebra
  - Chen (1958): they determine the path up to reparametrisation and retraced pieces
- Lyons (1998): rough path theory
  - the signature is the natural feature set for controlled differential equations
  - A response to a path depends on it (to first order) only through its signature
- Machine learning: Levin, Lyons & Ni (2013), Chevyrev & Kormilitzin (2016), Lyons & McLeod (2022); finance: signature-based asset price models Cuchiero, Gazzani & Svaluto-Ferro (2023)



Terry Lyons (left), Oberwolfach, 2024

Photo: Katrin Schmid (2024); CC BY-SA 2.0 DE; Wikimedia Commons

The Signature: Definition (1/2)

- Why iterated integrals
  - level 1 keeps only the total increments
  - integrals over ordered times keep the order in which the channels move
- One signature term is an iterated integral of the path along a word of channels

$$S(X)_{a,b}^{i_{\mathbf{1}}\cdots i_k} = \int_{a<u_{\mathbf{1}}<\cdots<u_k<b} dX_{u_{\mathbf{1}}}^{i_{\mathbf{1}}} \cdots dX_{u_k}^{i_k}$$

- Notation
  - $X : [a, b] \to \mathbb{R}^d$: a path of bounded variation (finite total length), e.g. piecewise linear; X_u^i: its channel i at time u
  - Word $(i_{\mathbf{1}}, \ldots, i_k)$: an ordered list of k channels, letters in $\{1, \ldots, d\}$; k is the **level** of the term
  - $u_{\mathbf{1}} < \cdots < u_k$: the integral runs over ordered times; dX_u^i: increment of channel i at time u (for a linear segment, slope times du)

The Signature: Definition (2/2)

- The signature is the whole sequence of levels

$$\mathrm{Sig}(X) = (1,\ S^{(\mathbf{1})},\ S^{(\mathbf{2})},\ S^{(\mathbf{3})}, \dots), \qquad S^{(k)} \in (\mathbb{R}^d)^{\otimes k}$$

  - $S^{(k)}$: the d^k terms of level k, one for each word of length k, stored as a k-dimensional array (a tensor; $\otimes$: tensor product)
  - $\mathrm{Sig}^N(X)$: the truncated signature, levels 0 to N; N is the depth
- Level 1: the total increments of the channels

$$S^i = X_b^i - X_a^i$$

- Level 2: increments of channel j weighted by how far channel i has already moved

$$S^{ij} = \int_a^b (X_u^i - X_a^i)\, dX_u^j$$

  - Each level adds one integral and records finer details of the order of the moves

Linear Segments and Chen's Identity (1/2)

- A linear segment with increment $\Delta \in \mathbb{R}^d$ has the tensor exponential as signature

$$\mathrm{Sig}(\text{segment}) = \exp_{\otimes}(\Delta) = \left(1,\ \Delta,\ \tfrac{1}{2!}\Delta^{\otimes \mathbf{2}},\ \tfrac{1}{3!}\Delta^{\otimes \mathbf{3}}, \dots\right)$$

  - $\Delta = X_b - X_a$: the increment of the segment; $\Delta^{\otimes k}$: the array of all products $\Delta^{i_\mathbf{1}} \cdots \Delta^{i_k}$, so a level-$k$ term is $\Delta^{i_\mathbf{1}} \cdots \Delta^{i_k}/k!$

- **Chen's identity** Chen (1957)
  - the signature of two paths run one after the other is the tensor product of their signatures

$$\mathrm{Sig}(X * Y) = \mathrm{Sig}(X) \otimes \mathrm{Sig}(Y), \qquad (A \otimes B)^{(k)} = \sum_{p=\mathbf{0}}^{k} A^{(k-p)} \otimes B^{(p)}$$

  - $X * Y$: concatenation, Y starts where X ends; $A^{(k)}$, $B^{(k)}$: level k of the signatures A, B, with $A^{(\mathbf{0})} = B^{(\mathbf{0})} = 1$
  - Level k of the whole path collects every way of splitting a word of length k between the first and the second piece

Linear Segments and Chen's Identity (2/2)

- A piecewise-linear path with m segments, increments $\Delta_1, \ldots, \Delta_m$

$$\mathrm{Sig} = \exp_{\otimes}(\Delta_1) \otimes \cdots \otimes \exp_{\otimes}(\Delta_m)$$

  - This is the whole algorithm: m updates, each costing $O(d^N)$ at depth N
- Rolling windows: Chen also removes an initial piece

$$\mathrm{Sig}(X_{[s,t]}) = \mathrm{Sig}(X_{[a,s]})^{-1} \otimes \mathrm{Sig}(X_{[a,t]})$$

  - $X_{[s,t]}$: the path restricted to $[s,t]$, with $a \leq s \leq t$; $(\cdot)^{-1}$: inverse for $\otimes$, the signature of the piece run backwards

Worked Example: a Two-Segment Path

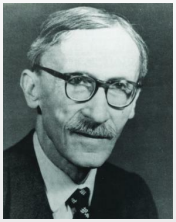
- The path $X : (0,0) \to (1,0) \to (1,2)$, with increments $\Delta_1 = (1,0)$ and $\Delta_2 = (0,2)$
- Level 1: $S^{(1)} = \Delta_1 + \Delta_2 = (1,2)$
- Level 2 by Chen's identity: $S^{ij} = \frac{1}{2}\Delta_1^i \Delta_1^j + \Delta_1^i \Delta_2^j + \frac{1}{2}\Delta_2^i \Delta_2^j$
  - Δ_m^i is the component i of the increment of segment m
- Result: $S^{11} = \frac{1}{2}$, $S^{12} = 0 + 1 \cdot 2 + 0 = 2$, $S^{21} = 0$, $S^{22} = \frac{1}{2} \cdot 2^2 = 2$
- Check: $S^{11} = (\Delta X^1)^2/2$ and $S^{22} = (\Delta X^2)^2/2$
  - shuffle identity: the product of two level-1 coordinates equals the sum of their two orderings at level 2
- $S^{12} \neq S^{21}$: the order matters, since channel 1 moves in the first segment and channel 2 in the second
- The total displacement (level 1) cannot tell this path from the one that goes up first and right second
  - both have $\Delta X = (1,2)$; level 2 tells them apart

The Lévy Area: Order as a Number

- The Lévy area A^{ij}: the antisymmetric part of level 2

$$A^{ij} = \tfrac{1}{2}(S^{ij} - S^{ji})$$

  - Equals the signed area between the path, projected on channels (i, j), and its chord (the straight line from start to end)
  - $A^{ij} > 0$: the path turns anticlockwise in the (i, j) plane
  - The lead-lag reading (who moves first) needs a positive association: if i falls and then j rises, i moves first, yet $A^{ij} < 0$
- Symmetric part: $\tfrac{1}{2}(S^{ij} + S^{ji}) = \tfrac{1}{2}\,\Delta X^i\,\Delta X^j$, no new information
  - Worked example: $A^{\mathbf{12}} = (2 - 0)/2 = 1$, the area of the triangle between the path and the chord
- Channels (return, log RV): a price fall followed by a volatility rise gives $A < 0$
  - so does a price rise followed by a volatility fall
- A lead–lag statistic
  - leverage-type sequences (Chapter 5) are one contributor to a negative area
  - the sign alone does not identify them



Paul Lévy (1886–1971)

Photo: Konrad Jacobs; CC BY-SA 2.0 DE; Wikimedia Commons

The Signature Kernel

- Kernel = inner product of the truncated signatures Király & Oberhauser (2019): a similarity score between two paths

$$k_N(X, Y) = \langle \mathrm{Sig}^N(X), \mathrm{Sig}^N(Y) \rangle = 1 + \sum_{k=1}^{N} \sum_{|w|=k} S(X)^w S(Y)^w$$

  - X, Y: two paths with the same channels; w: a word, $|w|$: its length (level); $S(X)^w$: the term of word w; the leading 1 is level 0
- Distance induced by the kernel (Gu et al., Eq. 9)
  - squared Euclidean distance between the truncated signatures

$$d_{\mathrm{Sig}}(X, Y) = k_N(X, X) - 2\, k_N(X, Y) + k_N(Y, Y) = \|\mathrm{Sig}^N(X) - \mathrm{Sig}^N(Y)\|^2$$

  - $d_{\mathrm{Sig}} = 0$: the two windows have the same signature up to depth N; large d_{Sig}: different shapes
- Gu et al. and this chapter use this **truncated** kernel, with N as for the features
  - The untruncated kernel solves a Goursat PDE (partial differential equation) Salvi et al. (2021); the paper cites it but does not use it
- Channels are rescaled before the kernel (▸ Appendix): otherwise the channel with the largest units dominates

The Problem at Amazon

- **Question**: what exactly did the paper do, and how strong is its evidence?
- Weekly North American truckload rates, national and five regions, 2018–2022 Gu et al. (2024)
  - forecast up to 12 weeks ahead
  - More than 100 candidate factors screened to about 40
  - In Amazon operations since November 2022; KDD 2024 (Knowledge Discovery and Data Mining, the conference of the ACM, Association for Computing Machinery)
- One direct regression per horizon Δt (in weeks; the paper's notation, not an increment Δ), $y_{\tau+\Delta t}$ on the features ξ_τ (next slide)



A container truck on an American interstate, 2025

Photo: DanTD (2025); CC BY 4.0; Wikimedia Commons

The Features of the Paper

- Features at week τ (Section 4.1, Eq. 14)
  - external factors x_τ
  - the signature of the time-augmented target path

$$\xi_\tau = \left[x_\tau,\ \mathrm{Sig}^N(\hat{y}_{\tau-l:\tau})\right], \qquad \hat{y}_{\tau-l:\tau} = \{(t_i, y_{t_i})\}$$

  - y_{t_i}: the rate in week t_i, $\tau - l \leq t_i \leq \tau$
  - l: lookback in weeks
  - $\hat{y}_{\tau-l:\tau}$: the time-augmented path of the last l weeks
- The external factors bring the level and information from outside the series
  - the signature brings the shape of the recent path

The Method: Two-Step LASSO

- Step 1, selection by a weighted LASSO Tibshirani (1996) (Eq. 14)

$$\hat{\theta}_{\mathrm{L}} = \arg\min_{\theta} \sum_{\tau} w_{\tau} \left(y_{\tau+\Delta t} - \xi_{\tau}^{\top}\theta\right)^{2} + \lambda \|\theta\|_{1}$$

  - θ: coefficient vector, $\|\theta\|_{1} = \sum_{j} |\theta_{j}|$; $w_{\tau} \geq 0$: weight of week τ (equal, or from the kernel, next slide)
  - $\lambda \geq 0$: penalty; a larger λ generally favours sparser solutions; $\lambda = 0$ is weighted OLS
- Step 2, refit by weighted OLS on the selected support only, the nonzero coefficients of step 1 (Eq. 15)

$$\hat{\theta} = \arg \min_{\mathrm{supp}(\theta) = \mathrm{supp}(\hat{\theta}_{\mathrm{L}})} \sum_{\tau} w_{\tau} \left(y_{\tau+\Delta t} - \xi_{\tau}^{\top}\theta\right)^{2}$$

- Why refit: LASSO shrinks the kept coefficients towards zero
  - OLS after selection removes the penalty-induced shrinkage on the selected support Belloni & Chernozhukov (2013)
  - the refit does not remove selection or omitted-variable bias
  - The paper calls it "adaptive" because of the weights w_{τ}, not in the sense of the adaptive LASSO of Zou (2006)
- Choice of λ in this chapter: the support with the smallest BIC Efron et al. (2004)
  - BIC: Bayesian Information Criterion
  - along the LARS (Least Angle Regression) path
  - in each training window (Appendix)

The Method: Kernel Weights

- Distance of each past window to the last window with an observed target (Eq. 11)

$$\delta_\tau = d_{\mathrm{Sig},N}\left(z_{\tau-l:\tau},\ z_{t-\Delta t-l:t-\Delta t}\right)$$

  - $z_\tau = (x_\tau, y_{\tau+\Delta t})$: the *pairs* of predictors and later target, not the features ξ_τ (Algorithm 1); $z_{\tau-l:\tau}$: their path over l weeks
  - t: the current week; $t-\Delta t$: the last week whose target y_t is already observed; $d_{\mathrm{Sig},N}$: the signature distance at depth N
- Softmax weights with temperature $\gamma \geq 0$ (Eq. 12): similar windows (small δ_τ) get more weight

$$w_\tau = \frac{\exp(-\gamma\,\delta_\tau)}{\sum_s \exp(-\gamma\,\delta_s)}$$

  - s: runs over all training weeks, so the weights sum to 1; $\gamma = 0$: ordinary (equal-weight) regression; $\gamma \to \infty$: only the most similar past window
- A nearest-neighbour idea inside a regression: learn from the past periods that looked like today
  - Cost of the idea: fewer effective observations, $n_{\mathrm{eff}} = 1/\sum_\tau w_\tau^2$ (Kish); $n_{\mathrm{eff}} = n$ for equal weights $1/n$, $n_{\mathrm{eff}} = 1$ if one week takes all the weight

What the Paper Reports

- National rates, three months ahead, April–November 2021 (Table 1)
  - relative error $|y - \hat{y}|/|y|$ (y: realised rate, $\hat{y}$: its forecast)
  - about 1–2% vs 13–27% for the industry forecast, "more than fivefold"
  - Horizons 1–12 weeks (Table 2): 1.06–2.60% nationally, 1.05–5.69% by region; estimated annual saving of about 50 million dollars
  - Kernel ablation (Table 3, six weekly values, 31 October–5 December 2021, one region): errors up to 9.85% without the kernel vs up to 5.15% with it
- Type of evidence: comparison of procedures (it shows which method works better on the same data, not why)
  - against a single commercial forecast, without standard errors (next slide)

What the Paper Reports: Limits of the Evidence

- **One benchmark**: an unnamed commercial forecast
  - no ARIMA (autoregressive integrated moving average) model
  - no LASSO without signatures, no random walk
- **Short evaluation**: a few months, a few series, no standard errors, no test of equal accuracy
- **Undisclosed tuning**: l, N, γ, λ and how they were chosen
- None of this makes the method wrong; it makes the evidence thin
  - which is why it is tested elsewhere
  - replication before extension (Chapter 19)

What Transfers to Volatility: Kept and Adapted

Transfers

- Signature of the time-augmented recent path of the target
- Two-step LASSO with weights, direct regression per horizon
- Softmax weights $\exp(-\gamma\,\delta_\tau)$ with the truncated signature kernel

Must be adapted

- Daily, noisy target instead of weekly rates: logs, HAC inference, robust losses
- External factors x_τ replaced by the log-HAR terms (they carry the level)
- The kernel compares the paths (time, $\ln \mathrm{RV}$, return) ending at τ and at t
  - updated every 5 days
  - not the pairs $(x_\tau, y_{\tau+\Delta t})$ ending at $t - \Delta t$
- Undisclosed choices (l, N, γ, λ) fixed ex ante
  - strong benchmarks and formal tests replace the single commercial forecast

Four Hypotheses Fixed Before the Results

- **Question**: what do we fix before the results, so that a gain cannot be found by searching?
- **H1 (features)**: signature terms add information beyond log-HAR
  - Test: Sig-L vs log-HAR, same target, same equal weights
- **H2 (weights)**: kernel weights improve a given regression
  - Tests: Sig-LK vs Sig-L and log-HAR-K vs log-HAR
- **H3 (method)**: the full method of Gu et al. beats the HAR family
  - Sig-LK vs HAR and vs log-HAR
  - model confidence set, MCS, of all seven models
- **H4 (stress)**: any gain is larger in stress periods
  - COVID-19 crash, 2–30 April 2025, the US tariff shock
  - where similar past windows should matter most

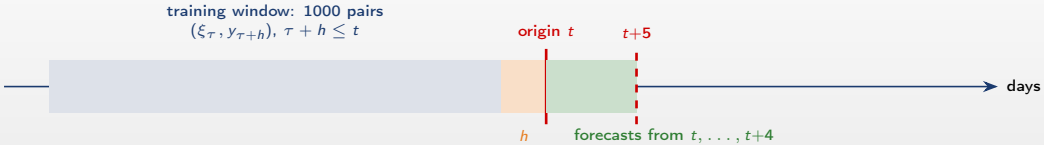
Kernel Temperature: One Choice per Class, on a Validation Block

- Scale-free temperature, set in every training window

$$\gamma = \frac{c}{\mathrm{median}_\tau(\delta_\tau)}$$

  - δ_τ: kernel distances of the training days to the current window; $c > 0$: a dimensionless constant, the only tuning parameter
  - $c = 1$: a window at the median distance gets weight e^{-1} relative to an identical one; larger c: more concentrated weights
- Grid $c \in \{0.5, 1, 2, 4\}$, chosen per horizon and asset class
  - on the first 250 forecast days of each asset, pooled within the class
  - Criterion: mean over the class of QLIKE(c) divided by the asset's average over the grid
  - Evaluation starts after the last validation target of the class, so no tuning target postdates an evaluated origin
- For the 40 stocks the smallest temperature, $c = 0.5$, wins at every horizon for Sig-LK
  - the most nearly equal weights
  - for futures and FX the choice changes with the horizon
  - a flat, noisy criterion (all values in the Appendix)

Rolling Estimation, Step by Step



- Re-estimate every 5 trading days on the last 1000 complete pairs
  - between re-estimations the coefficients are fixed
  - the regressors are updated daily
- Kernel distances are recomputed at each re-estimation, against the window ending at the origin t

Evaluation Protocol (1/3): the DM Test

- **Question**: is the loss difference between two models larger than sampling noise?
- Per asset and horizon: mean QLIKE and MSE Diebold & Mariano (1995)
  - compared by the DM (Diebold–Mariano) test of equal predictive accuracy
  - the paper's relative error is reported descriptively only (Appendix)

$$d_t = L(\mathrm{RV}_t, F_t^A) - L(\mathrm{RV}_t, F_t^B), \qquad \mathrm{DM} = \frac{\bar{d}}{\sqrt{\hat{\Omega}/T}}$$

  - L: the loss (QLIKE or MSE); F_t^A, F_t^B: forecasts of models A and B; d_t: their loss differential
  - $\bar{d}$: the mean of d_t over the T evaluation days; $\hat{\Omega}$: the long-run variance of d_t (Further Topics)
- Under $H_0 : \mathbb{E}[d_t] = 0$, DM is approximately $N(0, 1)$
  - $\mathrm{DM} < -1.645$ means model A is significantly more accurate at 5% (one-sided)
- Rolling estimation, so the test is read as the unconditional GW (Giacomini–White) test
  - The conditional GW test also asks whether d_{t-h} predicts d_t (instruments $(1, d_{t-h})$) Giacomini & White (2006)

Further Topics: HAC variance

Evaluation Protocol (2/3): the Model Confidence Set

- Why a set of models: the DM test compares two models
  - here there are seven and no benchmark chosen in advance
- MCS at 90% Hansen, Lunde & Nason (2011): the set of models that contains the best one with probability at least 90%
  - Built by elimination: while the $T_{\max}$ statistic (the largest standardised gap between a model's mean loss and the average over the models still in the set) rejects equal accuracy, drop the worst model
  - p-values by the stationary bootstrap Politis & Romano (1994), 1000 replications, mean block length $\max(h, 10)$ days
  - The loss series are resampled jointly, to keep the dependence across models

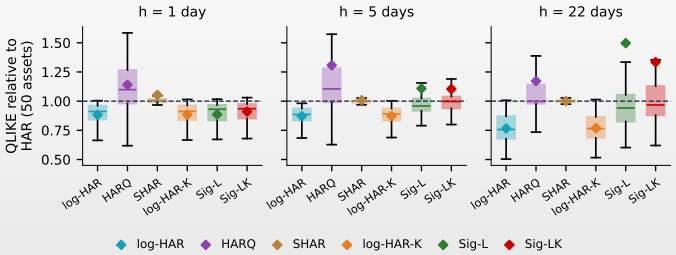
Evaluation Protocol (3/3): Panel Test and Corrections

- Panel: DM test on the cross-sectional mean loss differential (one series, one value per date)
- Count the assets (of 50) with significantly better results
  - The Holm correction Holm (1979) controls the FWER (family-wise error rate, the probability of at least one false rejection)
  - BH Benjamini & Hochberg (1995) controls the FDR (false discovery rate, the expected share of false rejections among the rejections)
- Subperiods (evaluation periods by class: ▸ Appendix)
  - 20 Feb–30 Apr 2020 (COVID-19 crash)
  - 2–30 Apr 2025 (US tariff shock)
  - all other days

A Disclosed Deviation

- Pre-registered: the BPQ (Bollerslev–Patton–Quaedvlieg) insanity filter on all models: a forecast outside the in-sample range of the target becomes the in-sample mean
  - A test run of the code on JPM (one asset, before any panel result) showed it replacing correct crash forecasts by the calm in-sample mean
  - On 12 March 2020 the realised next-day RV of JPM was about 40 times the filtered HAR forecast
- Changed before estimating the models on the 50-asset panel
  - primary results with positivity only
  - a forecast ≤ 0 becomes the in-sample mean
  - The BPQ filter is kept as a robustness check and reported in full

QLIKE Relative to HAR, Asset by Asset



- **Question**: does the method of Gu et al. beat the HAR models and, above all, log-HAR?
- Ratio for each asset: mean QLIKE of the model over mean QLIKE of HAR; below 1: more accurate than HAR
  - Boxes: interquartile range over the 50 assets; line: median; diamond: mean; whiskers: 1.5 interquartile ranges
- Ratios are averaged over assets with equal weights
  - here, in the ablation and in the filtered results
  - the mean of the ratios is not the ratio of the means
  - so a new benchmark means new per-asset ratios
- Medians describe a typical asset; means remain sensitive to a few large losses

MFM_ch20_case_study ▶ Interactive

Reading the QLIKE Comparison

- $h = 1$: log-HAR, log-HAR-K and Sig-L are close (0.883, 0.885, 0.886)
  - Features alone (Sig-L vs log-HAR): +0.4% in QLIKE, so 36 extra features bring no gain
  - Features and weights (Sig-LK vs log-HAR): +4.5%
- $h = 5, 22$: Sig-LK falls behind log-HAR (+28.0%, +92.9%)
  - Its medians relative to HAR (0.997, 0.968) are much better than its means: a few assets and days drive the loss
  - in a variant without the return channel, Sig-LK relative to HAR (0.878, 0.770) comes close to log-HAR (0.871, 0.765)
- HARQ and SHAR (Semivariance HAR model) lose to HAR in levels here
  - because the main design does not apply the BPQ filter
  - the filter rescues HARQ (Appendix)
- First verdict: the method does not beat the simplest log model
  - a near-tie is not a proof of equal accuracy
- Type of evidence: comparison of procedures

Asset by Asset: Wins and Losses, Corrected

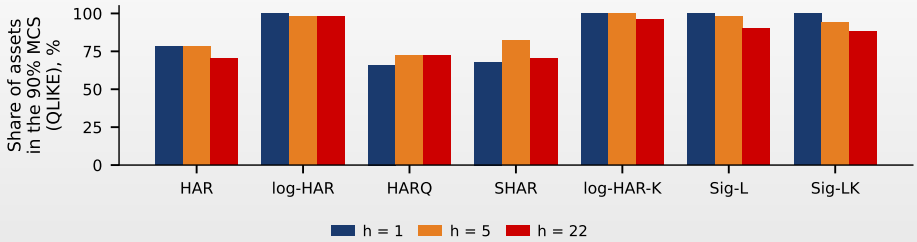
Comparison (QLIKE)	$h = 1$	$h = 5$	$h = 22$
Sig-LK vs log-HAR	2/0/0 | 5/0/0	0/0/0 | 15/0/2	0/0/0 | 27/1/16
Sig-L vs log-HAR	3/1/1 | 5/0/0	0/0/0 | 9/0/0	1/0/0 | 21/0/0
log-HAR-K vs log-HAR	3/0/0 | 3/0/0	3/0/0 | 1/0/0	5/0/0 | 6/0/0
Sig-LK vs Sig-L	2/0/0 | 5/0/0	0/0/0 | 5/0/0	2/0/0 | 10/0/0
Sig-LK vs HAR	12/5/6 | 0/0/0	7/1/1 | 0/0/0	8/2/4 | 1/0/0
log-HAR vs HAR	21/5/17 | 0/0/0	23/13/20 | 0/0/0	42/25/42 | 0/0/0

- Cell: significantly better on raw/Holm/BH assets | significantly worse on raw/Holm/BH assets
  - one-sided DM at 5%, family of 50 assets
  - the DM test needs finite moments of order greater than 2 and weak dependence of the QLIKE loss differential d_t, not of the returns; forecasts close to zero can violate them
  - Holm controls the FWER under any dependence
  - BH controls the FDR under independence or PRDS (positive regression dependence on a subset) Benjamini & Yekutieli (2001)

Asset by Asset: Reading the Table

- H1 (signatures): Sig-L beats log-HAR on 3 of 50 assets at $h = 1$ raw, 1 after Holm
  - about 2.5 raw wins are expected by chance
- Also at $h = 1$, Sig-L loses to log-HAR on 5 assets raw and on none after Holm
- H2 (weights): log-HAR-K vs log-HAR and Sig-LK vs Sig-L: no Holm-significant difference at any horizon
- Against HAR at $h = 1$, Sig-LK wins on 12 assets, but log-HAR alone wins on 21
  - choose the benchmark that isolates the idea

Model Confidence Sets (1/2)



- Share of the 50 assets in which each model belongs to the 90% MCS under QLIKE ($h = 1, 5, 22$)
  - log-HAR 100%, 98%, 98%
  - Sig-LK 100%, 94%, 88%
  - the MCS on QLIKE needs the same conditions as the DM test on the loss differentials between models, not on the returns
- The MCS rarely separates the four log models
  - HAR, HARQ and SHAR are excluded for a fifth to a third of the assets
  - the level target is the clearest loser

▶ Interactive

Model Confidence Sets (2/2)

- Under MSE almost every model stays in the set
  - mean-loss differences have large standard errors

MFM_ch20_case_study

Panel Tests: Pooling the Evidence

Comparison: DM t (one-sided p)	$h = 1$	$h = 5$	$h = 22$
Sig-LK vs log-HAR	+1.45 (0.927)	+1.68 (0.953)	+1.80 (0.964)
Sig-L vs log-HAR	-1.91 (0.028)	+1.48 (0.931)	+1.51 (0.935)
log-HAR-K vs log-HAR	+1.89 (0.971)	+0.40 (0.654)	+1.92 (0.973)
Sig-LK vs Sig-L	+1.74 (0.959)	+0.98 (0.837)	-0.77 (0.221)
log-HAR vs HAR	-3.09 (<0.001)	-3.95 (<0.001)	-4.50 (<0.001)

- Series: the mean QLIKE differential over the assets available that day (unbalanced panel)
  - negative t = first model better
  - the DM test needs finite moments of order greater than 2 and weak dependence of the QLIKE loss differential d_t, not of the returns; forecasts close to zero can violate them
- Pooled, only the log transform is clearly significant
- Sig-L vs log-HAR at $h = 1$: significant in a single one-sided test ($p = 0.028$) in the unbalanced panel
  - in the balanced panel the result disappears (next slide)
  - For the 40 stocks alone the sign flips: $t = +1.72$ against Sig-L
- After a Holm correction over the three horizons (threshold $0.05/3 \approx 0.017$) it is no longer significant
- Conditional GW test, Sig-LK vs log-HAR: 0, 0 and 0 Holm rejections out of 50

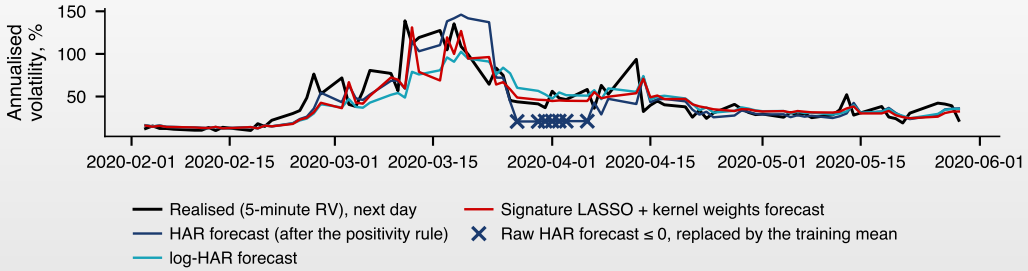
The Balanced Panel: the Same Assets Every Day

Comparison: DM t (one-sided p)	$h = 1$	$h = 5$	$h = 22$
Sig-LK vs log-HAR	$+1.19$ (0.884)	$+1.61$ (0.946)	$+1.41$ (0.921)
Sig-L vs log-HAR	$+0.75$ (0.773)	$+1.49$ (0.932)	$+1.24$ (0.892)
log-HAR vs HAR	-2.21 (0.014)	-2.51 (0.006)	-4.76 (<0.001)

- The unbalanced panel averages the assets available each day: futures and FX from 2014, the stocks only from 27 Jan 2020
  - differences between classes are thus mixed with differences between periods
- The balanced panel: all 50 assets, only the days on which all of them have forecasts (1511 days at $h = 1$, 27 Jan 2020 – 29 Jan 2026)
- Sig-L vs log-HAR at $h = 1$: $t = +0.75$ ($p = 0.773$), against $t = -1.91$ ($p = 0.028$) in the unbalanced panel
  - on the common calendar only the panel of the 5 futures favours Sig-L ($t = -4.03$)
- The advantage of log-HAR over HAR remains: t between -4.76 and -2.21

MFM_ch20_case_study

Inside the COVID-19 Crash: JPM, Next Day (1/2)



- Annualised volatility in % of the forecasts, $\sqrt{252 F_t}$, and of the target, $\sqrt{252\,\mathrm{RV}_{t+\mathbf{1}}}$
  - February–May 2020
  - all models lag the first jump
- Sig-LK reacts slightly faster than log-HAR in mid-March
  - when the path has already turned
  - after the peak both decay at similar speeds

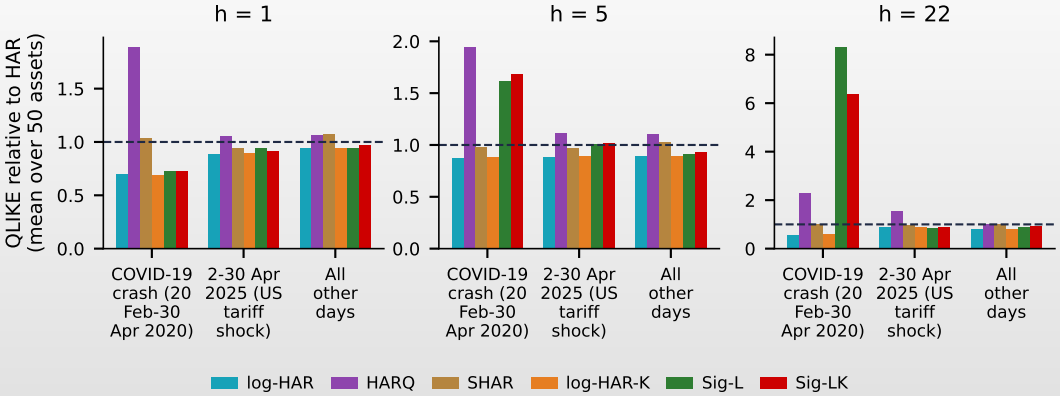
▶ **Interactive**

Inside the COVID-19 Crash: JPM, Next Day (2/2)

- HAR in levels overshoots, then the raw HAR forecast turns negative in late March
  - the chart shows the value after the positivity rule: the mean target of the training window (the flat segment)
  - the 7 replaced origins, 27 March – 6 April 2020, are marked with $\times$

MFM_ch20_case_study

Stress Periods vs Normal Times (H4) (1/2)



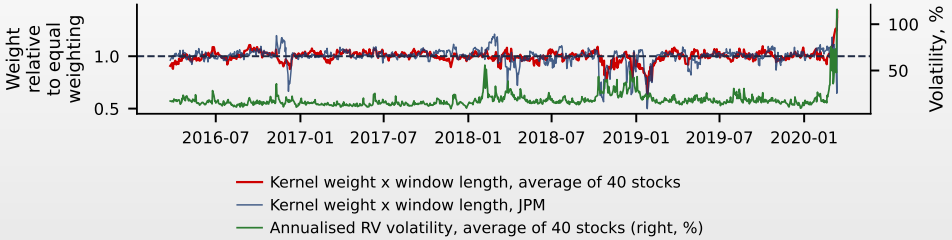
MFM_ch20_case_study ▶ Interactive

Stress Periods vs Normal Times (H4) (2/2)

- Mean over assets of QLIKE relative to HAR within each subperiod fixed before the evaluation
- $h = 1$, COVID-19 crash: log-HAR 0.70, Sig-LK 0.73, HARQ 1.89
  - 2–30 April 2025 (US tariff shock): 0.88 and 0.92
  - other days: 0.94 and 0.97
- $h = 22$, COVID-19 crash: Sig-LK 6.35 and Sig-L 8.30 times the loss of HAR
  - exactly where similar past windows should help, the signature models fail most
  - H4 not supported
- Type of evidence: description (it shows what happened in the data, without testing a cause)
  - no formal interaction test

MFM_ch20_case_study

Kernel Weights at the Origin 16 March 2020 (1/2)



- Weights of Sig-LK ($h = 1$) at the origin 16 March 2020 over the training days
  - training days: 24 Mar 2016 – 13 Mar 2020
  - weights times the window length: 1 = equal weighting
- Nearly equal: between 0.61 and 1.43 times $1/n$
  - effective sample size 975 of 1000 (average over the 40 stocks)

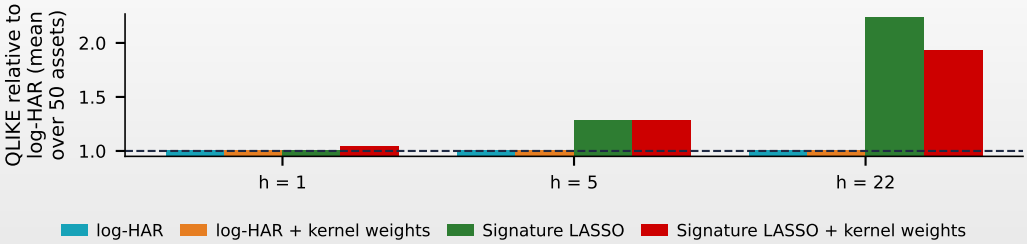
▶ Interactive

Kernel Weights at the Origin 16 March 2020 (2/2)

- The highest weights go to the last days before the origin
- Recent 22-point windows overlap the current window
  - similarity may partly reflect overlap, not a distinct economic regime
- The late-2018 sell-off gets *less* weight
  - weights are almost uncorrelated with log volatility (+0.04)
  - the translation-invariant signature sees shape, not level

MFM_ch20_case_study

The 2×2 Ablation



- **Question**: which ingredient drives the result, the signature features or the kernel weights?
- Mean over the 50 assets of QLIKE relative to log-HAR
  - features (signature or not) crossed with weights (kernel or equal)
- The dashed line marks 1 (log-HAR)
  - a bar above it means a worse forecast than log-HAR

MFM_ch20_case_study ▶ Interactive

The 2×2 Ablation: the Numbers

- Weights alone (log-HAR-K): 1.002, 1.002, 1.004 at $h = 1, 5, 22$: no detectable effect on the mean loss
- Features alone (Sig-L): 1.004, 1.285, 2.238: level-3 terms explode in crashes (Further Topics)
- Both (Sig-LK): 1.045, 1.280, 1.929
  - Average effective sample size with the Sig-LK weights: 816 of 1000
  - Adding the weights to Sig-L reduces the loss at $h = 22$ without removing it
  - They worsen $h = 1$ through crude oil (CL: 2.71), unlike the kernel gain in Table 3 of Gu et al.

By Asset Class

Sig-L relative to log-HAR (QLIKE)	$h = 1$	$h = 5$	$h = 22$
Stocks (40)	1.009	1.332	2.505
Futures (5)	0.965	1.084	1.206
FX (5)	0.996	1.111	1.134

- The only gain: futures at $h = 1$ (-3.5%) and, marginally, FX
  - Five futures and one horizon: a candidate for a new, pre-registered study, not a finding of this one
- Stocks, 80% of the panel, show no gain
  - the mean of the ratios over assets is dominated by the stocks
  - the DM test on the unbalanced panel, however, mixes classes with periods: before 2020 it holds only futures and FX

Limitations of This Study

- One design out of many: a null result for this design is not one for signatures in general
  - the design: $L = 22$, $N = 3$, three channels, BIC, a four-point grid for c
- Truncated kernel only
- No external factors (implied volatility, macro news), central in Gu et al.
- Realized GARCH Hansen, Huang & Shek (2012) and machine-learning benchmarks are outside the comparison
  - GARCH: generalised autoregressive conditional heteroskedasticity (Chapter 5)
  - machine-learning benchmarks: Chapter 13
- Stocks are evaluated over six years only

Open Directions

- A level-aware kernel: basepoint augmentation or the level of $\ln \mathrm{RV}$ in the distance (Seminar 20, C2)
- Signatures of the intraday path (5-minute returns)
- The PDE kernel Salvi et al. (2021)
- Log-signatures and bounded (tanh) channels against explosive level-3 terms Morrill et al. (2020)
- Cross-asset paths, where Lévy areas measure lead–lag across markets
- Futures at $h = 1$ on a longer panel, with this design fixed in advance

The Question: The Signature as a Lead-Lag Metric

- **Question**: does one asset move before another, and can the signature of their joint path measure it?
  - lead-lag relations show in which market information appears first (price discovery)
- Bennett, Cucuringu & Reinert (2022), *Machine Learning*
  - the pairwise lead-lag relations form a directed network, which is then split into clusters of leaders and followers
  - data: the 434 NYSE (New York Stock Exchange) stocks with the highest volume, daily returns 2000–2019
- One of the metrics compared: twice the Lévy area of the path of normalised log prices over $[t-m, t]$

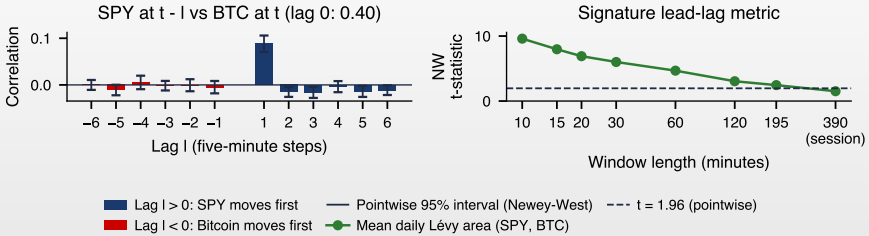
$$S^{ij} = 2A^{ij} = \iint_{t-m<u<v<t} \left(dX_u^i \, dX_v^j - dX_u^j \, dX_v^i\right)$$

  - $S^{ij} > 0$: moves of i are followed by moves of j in the same direction
  - unlike cross-correlations, it needs no choice of a maximum lag
  - it assumes a positive association: it cannot tell "i leads, with negative association" from "j leads, with positive association"

What the Paper Found, and with Which Procedure

- The empirical results of the paper come from a metric other than the signature (published version, Sections 5–6)
  - the metric: ccf-auc (area under the cross-correlation function) with distance correlation, with lags of up to 5 days
  - the network clustering: Hermitian RW (random walk)
- The lead-lag clusters are significant in a permutation test that shuffles the order of the days
  - the permutation destroys both the lagged dependence between assets and the autocorrelation of each asset: on its own, the test does not separate them
- In the authors' comparisons the clusters do not match SIC (Standard Industrial Classification) sectors
  - their correlations with volume and with market capitalisation are small
- The paper's numbers (Section 6.3): the signal built on these clusters has an annualised Sharpe ratio of 0.62
  - the S&P 500 has 0.40 over the same period
  - both are before transaction costs
- The next slides: an illustration of the Lévy area on intraday data, not the paper's procedure

SPY and Bitcoin at Five Minutes: Who Moves First?



- SPY (the S&P 500 ETF; ETF: Exchange-Traded Fund) and Bitcoin (BTC, in USD): five-minute bars, New York time, 09:30–16:00
  - 489 days with all 78 bars for both assets (out of 513 common days), 3 Sep 2024 – 18 Sep 2026
- Left: mean over days of the correlation between the SPY return at $t - l$ and the BTC return at t; bars: pointwise 95% intervals with the Newey-West (NW) standard error of the tests
- Right: NW t-statistic of the mean daily Lévy area (SPY, BTC), for windows from 10 minutes to the whole session; the line at 1.96 is a pointwise threshold

MFM_ch20_leadlag ▶ Interactive

The Method in Brief and How the Replication Differs

- The path of the day: cumulative sums of the SPY and BTC returns, each return divided by the standard deviation of its asset on that day
  - 78 returns: from the open of the 09:30 bar to the close of each bar 09:30–15:55
  - the normalisation the authors recommend: otherwise the area grows with volatility
- The Lévy area of a window of m steps is a sum of lagged products of the returns x_i (SPY) and y_i (BTC)

$$A_m = \tfrac{1}{2} \sum_{l=1}^{m-1} \sum_i \left(x_i \, y_{i+l} - y_i \, x_{i+l} \right)$$

  - lag l appears $m - l$ times: a window of 2 steps (10 minutes) keeps lag 1 only
  - a long window adds many lags without effect, hence noise
- Daily statistic: the mean area over all windows of m steps in the day
  - the test: NW t with 8 lags on the 489 daily values
- Differences from the paper
  - two assets instead of 434, and five-minute bars instead of daily returns
  - no network clustering and no trading signal

The Result: SPY Leads Bitcoin by Five Minutes; Sensitivity to the Window Length

- SPY moves before Bitcoin by one five-minute step
  - the correlation at lag $+1$ is 0.088 (NW $t = 9.8$), and at lag -1 it is -0.005 ($t = -0.71$)
  - the direct test of direction: the difference between the correlations at lags $+1$ and -1 is 0.093 (NW $t = 9.5$)
  - the contemporaneous correlation is 0.40; at lags $+2$ to $+6$ the correlations are small and negative (between -0.016 and -0.004)
- The signature metric finds the effect only in short windows
  - 10 minutes: $t = 9.6$, and the area is positive on 70% of the days
  - 2 hours: $t = 3.0$; the whole session (390 minutes): $t = 1.51$, area positive on 51% of the days
- The analysis is exploratory: 20 tests (12 non-zero lags and 8 windows)
  - Holm-adjusted p: 0.032 for the 2-hour window, 0.157 for 195 minutes
- The effect weakens in 2026: the lag $+1$ correlation is 0.11 in 2024, 0.12 in 2025 and 0.04 in 2026 (January–September)

What the Lead-Lag Result Supports, and Its Limits

- Type of evidence: a five-minute lead-lag association between two assets, not a causal relation and not a profitable signal
  - no trading costs and no out-of-sample evaluation
- The sign of the area measures the orientation of the path; reading it as "SPY leads" assumes a positive association between the two assets
- Limits
  - a one-step lag also arises from unsynchronised timestamps of the two sources; separating the two needs tick data
  - the window length m plays the role of the maximum lag: the signature does not remove this choice
  - the NW t-test needs finite moments and weak dependence of the daily statistic tested (bounded correlations, normalised areas), not of the raw returns

When Does a Method Transfer Across Domains?

- Methods travel faster than evidence
  - a KDD paper on freight rates becomes a candidate for volatility, credit, energy
  - Which properties of the source problem made the method work?
  - Does the target share them?
- Finance adds a hard filter Menkveld et al. (2024)
  - 164 teams on the same data disagree more than their standard errors suggest
- Criterion: a transfer succeeds if the gain over the strongest simple benchmark survives
  - a pre-registered set $\mathcal{S}$ of designs, assets and horizons
  - with the family corrected
  - In this chapter: $\mathcal{S} = 50$ assets $\times$ 3 horizons $\times$ 4 variants; Holm/BH were applied per comparison and horizon (families of 50 assets), variants reported descriptively; no joint correction over all 600 cells

The Discovery Loop Applied to This Chapter

- **Read**: an assistant lists every equation, hyperparameter and benchmark of the paper
  - it marks what is not disclosed
  - each item is checked against the PDF (prompts)
- **Map**: for each component, does it transfer, what must change, which benchmark isolates it
- **Pre-register**: hypotheses, grid, tests and correction
  - fixed in a dated document before the first result
  - the assistant drafts, the team decides every number
- **Build and test the pieces**: signatures against identities, not against another library
  - level 1 = increment, Chen on a split path
- **Run the grid, then attack it**
  - Holm over 50 assets
  - a search for the single asset that would make a misleading headline (prompts)
- **Explain the null**, checked with data (checklist: Appendix)
  - here flat weights (translation invariance)
  - explosive level-3 terms (exponential link)

Mini-Case: Three Readings of the Same Results

- "Signatures beat HAR on 12 of 50 assets" (Sig-LK, one-day horizon, raw DM)
  - True, but log-HAR alone beats HAR on 21: the gain belongs to the log transform
- "Signatures improve E-mini S&P 500 volatility forecasts" (ES, $h = 1$, DM $t = -2.91$ for Sig-LK against log-HAR)
  - One cell out of 150 asset-horizon cells: after Holm, Sig-LK wins on 0 of 50 at $h = 1$ and none at $h = 5, 22$
- The headline of the pre-registered evaluation
  - "no robust advantage over log-HAR in the specification evaluated; worse at longer horizons; no detectable effect of the kernel weights"
  - An AI (Artificial Intelligence) assistant asked to "find where the method works" will produce the first two; the plan protects the third
- Type of evidence: comparison of procedures
  - the same runs, read without correction, on a single cell or by the pre-registered plan

Your Turn: Design a Transfer Study

- Pick one open direction of this chapter
  - level-aware kernel, intraday signatures, log-signatures, futures only
  - Write H_0/H_1, the benchmark that isolates it, and the result that would make you drop it
- Use an assistant to list what could go wrong in your design, then write one unit test per item
- Fix the pre-registration in a dated document before the first out-of-sample number
  - run the grid
  - report the distribution over $\mathcal{S}$
- Log every AI error you catch in `AI_ERRORS.md` (Seminar 20, C1 gives typical ones)

Key Takeaways

- Signatures turn a window into a feature vector with a theory behind it
  - Chen's identity; uniqueness and universality (details in Further Topics)
- Transferred to realised variance, the method of Gu et al. does not beat log-HAR
  - H1 (signatures add information): not supported, no gain at one day, a loss at a week and a month
  - H2 (kernel weights help): not supported, weights are almost flat and down-weight past stress; H3: only HAR in levels is beaten, through the log target; H4 (gains in stress): not supported
- The translation-invariant kernel sees shape, not level
  - level-3 terms explode in crashes
- The strongest effect in the study is the oldest one: model the log of RV
- The null result is informative
  - the ablations and the mechanism explain why the method does not transfer

Readings on Signatures and Realised Measures

Reading

- Signatures: Chevyrev & Kormilitzin (2016); Lyons & McLeod (2022)
  - realised measures: Franke, Härdle & Hafner (2019) and Chapter 9

Advanced courses on quantinar.com

- Statistics of Financial Markets
- Applied Time Series Analysis with Python
- Machine Learning in Financial Risk

Code and data

- Quantlets: github.com/danpele/MFM/tree/main/Quantlets/Ch_20
  - data: VOLARE, free for research at volare.unime.it

Check your understanding

- What information about order can a path signature retain that a simple average loses? (the two-segment example)
- Why compare with both a level target and a log target?
- Which ablation isolates the contribution of signatures? (the 2×2 ablation)
- Can one significant win among many assets establish superiority?
- What is the role of the correction for multiple comparisons? (evaluation protocol)
- How can a null result still be informative? (limitations and open directions)

The Practical Decision

- **What I choose**: log-HAR to forecast daily realised variance; the method of Gu et al. (Sig-LK) does not replace it
  - log-HAR is in the 90% MCS for 100%, 98% and 98% of the assets at $h = 1$, 5 and 22
  - on the panel, log-HAR beats HAR at every horizon (DM t from -3.09 to -4.50)
- **Under which conditions**
  - QLIKE as the main loss, with the log target and the variance correction $\hat{s}^2/2$
  - rolling estimation, re-estimated every 5 days on the last 1000 pairs
  - a panel dominated by stocks (40 of the 50 assets)
- **What result would make me drop it**
  - a gain of Sig-L or Sig-LK over log-HAR that survives the Holm correction in a pre-registered study
  - for example, the Sig-L gain on futures at $h = 1$ (-3.5% in QLIKE), confirmed on a longer panel

Why Five-Minute RV, and Where Do the Data Come From?

- Alternatives to RV
  - bipower variation (robust to jumps)
  - the realised kernel (robust to noise) Barndorff-Nielsen et al. (2008)
  - range-based estimators
- Target of this chapter: five-minute RV
  - across 400 assets, little beats it as a proxy of daily variance Liu, Patton & Sheppard (2015)
- Its semivariances and quarticity are on the same grid
  - so SHAR (the semivariance HAR model) and HARQ use consistent inputs
- Every model is evaluated on the same target
  - so loss differences come from the models, not from the volatility measure
- VOLARE is built at the Universities of Messina and Florence from tick data Cipollini et al. (2026)
  - with the same estimators for every asset class
- Free for research at volare.unime.it, not redistributed with this course
  - download the realised-variance files
  - point the notebooks to your local folder
- The authors request the citation
  - Cipollini, Cruciani, Gallo, Insana, Otranto & Spagnolo (2026), arXiv:2602.19732

Two Refinements (1/2): HARQ

- RV measures the true variance with error

$$\mathrm{RV}_t = \mathrm{IV}_t + \eta_t, \qquad \mathrm{Var}(\eta_t) \approx 2\,\mathrm{IQ}_t / M$$

  - η_t: measurement error; errors-in-variables attenuate (pull towards zero) β_d, most on noisy days
- HARQ Bollerslev, Patton & Quaedvlieg (2016): the weight on yesterday's RV falls when RV is measured less precisely

$$\mathrm{RV}_{t+\mathbf{1}:t+h} = \beta_{\mathbf{0}} + \left(\beta_d + \beta_Q(\mathrm{RQ}_t^{\mathbf{1/2}} - \overline{\mathrm{RQ}^{\mathbf{1/2}}})\right)\mathrm{RV}_t + \beta_w \mathrm{RV}_t^{(w)} + \beta_m \mathrm{RV}_t^{(m)} + \varepsilon_{t+h}$$

  - $\mathrm{RQ}_t^{\mathbf{1/2}}$: square root of realised quarticity, a measure of the error of RV_t; $\overline{\mathrm{RQ}^{\mathbf{1/2}}}$: its in-sample mean, so β_d is the weight on an average day
  - Expected $\beta_Q < 0$: on a noisy day (RQ_t above average) yesterday's RV gets less weight
- BPQ (Bollerslev–Patton–Quaedvlieg) filter ("insanity filter")
  - a forecast outside the in-sample range of the target is replaced by the in-sample mean

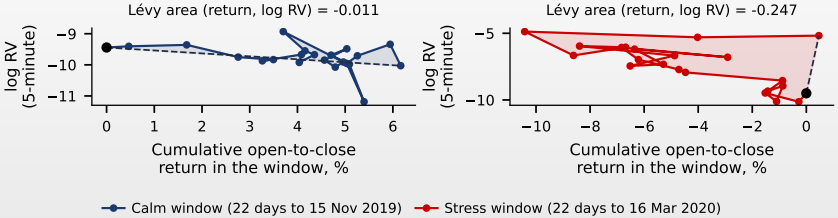
Two Refinements (2/2): SHAR

- SHAR Patton & Sheppard (2015): split the daily term by the sign of the intraday returns

$$\mathrm{RV}_{t+1:t+h} = \beta_0 + \beta^+ \mathrm{RV}_t^+ + \beta^- \mathrm{RV}_t^- + \beta_w \mathrm{RV}_t^{(w)} + \beta_m \mathrm{RV}_t^{(m)} + \varepsilon_{t+h}$$

  - RV_t^+, RV_t^-: the positive and negative semivariances of day t
  - $\beta^+ = \beta^-$ gives back HAR, since $\mathrm{RV}_t^+ + \mathrm{RV}_t^- = \mathrm{RV}_t$
- Patton and Sheppard find $\beta^- > \beta^+$
  - Variance from down-moves predicts more future variance than variance from up-moves: an intraday version of the leverage effect (Chapter 5)
- The signature can capture a related effect from daily data
  - the Lévy area between returns and log RV (Path Signatures section)

Two Windows of JPM as Paths (1/2)



- Path (cumulative open-to-close return, log RV) over 22 days
  - dashed chord from start to end
  - the shaded region gives the Lévy area
- Calm window (to 15 Nov 2019) vs stress window (to 16 Mar 2020)
  - calm: log RV moves by -0.58, the path wanders, Lévy area -0.011
  - stress: log RV rises by +4.32 (RV $\times$ 75) while the intraday return path falls and then recovers
  - stress: Lévy area -0.247 (returns in decimal units for the area, percent on the axis)

Two Windows of JPM as Paths (2/2)

- The net intraday return is only $+0.5\%$
  - the crash was largely overnight, outside this channel
  - no signature level recovers what the input omits

MFM_ch20_signatures

Why Signatures Work (1/2): Uniqueness and Shuffles

- **Uniqueness** Hambly & Lyons (2010): the full signature identifies the path
  - $\mathrm{Sig}(X)$ determines X up to tree-like equivalence (adding pieces that are retraced backwards) and translation
  - With a strictly increasing time channel nothing can be retraced, so only translation remains
  - A finite truncation Sig^N is not injective: two different paths can share levels $1..N$
- **Shuffle identity**: a product of signature terms is a sum of higher terms

$$S^i S^j = S^{ij} + S^{ji}, \qquad S^{ij} S^k = S^{ijk} + S^{ikj} + S^{kij}$$

  - The sum runs over all interleavings (shuffles) of the two words that keep the order inside each word
  - So a linear model in $\mathrm{Sig}^N(X)$ already contains polynomials of the increments

Why Signatures Work (2/2): Universality and Decay

- **Universality**: any continuous function of the path is approximately linear in the signature
  - Stone–Weierstrass, via the shuffle identity

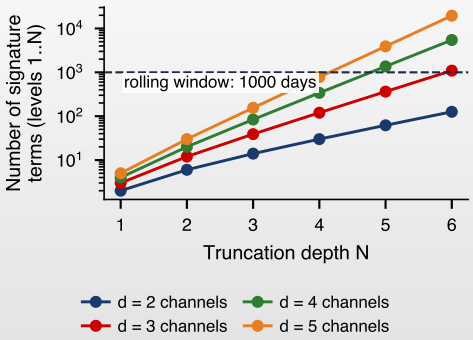
$$\forall \varepsilon > 0 \;\; \exists\, \ell \text{ linear}: \quad \sup_{X \in K} \left| f(X) - \langle \ell, \mathrm{Sig}(X) \rangle \right| < \varepsilon$$

  - K: a compact set of time-augmented paths with a common start (in the 1-variation topology); $f : K \to \mathbb{R}$: a continuous function of the path
  - $\langle \ell, \mathrm{Sig}(X) \rangle$: a weighted sum of finitely many signature terms, i.e. a linear regression on them
  - Sufficient depth permits nonlinear approximation; this does not guarantee accuracy at depth 3 or out of sample. The cost is d^N features, so sparsity is needed
- **Factorial decay**: high levels are small

$$\|S^{(k)}\| \leq \frac{\|X\|_{\text{1-var}}^k}{k!}$$

  - $\|X\|_{\text{1-var}}$: total length of the path (sum of the lengths of its segments); $\|S^{(k)}\|$: size (Euclidean norm) of level k
  - Truncating at a small depth N loses little for paths of moderate length, which is why channels are rescaled (Appendix)

How Many Features? The Truncation Dimension



- Number of terms in levels $1..N$ of a d-dimensional path (a geometric sum, since level k has d^k words)
  - $\dim(d, N) = \sum_{k=1}^{N} d^k = \frac{d^{N+1}-d}{d-1}$
- This chapter: $d = 3$ channels (time, log RV, cumulative return), depth $N = 3$: $3 + 9 + 27 = 39$ terms
  - Three are constant (words with only the time letter), so 36 vary
- $d = 5$, $N = 4$ would give 780 terms for 1000 observations: the LASSO selection would become unstable among many correlated terms
- Here $N = 3$ and $L = 22$ are fixed ex ante, with $N = 2$ as a robustness check

MFM_ch20_signatures

Two Invariances and the Basepoint

- **Translation invariance**: $\mathrm{Sig}(X + v) = \mathrm{Sig}(X)$ for any constant vector $v \in \mathbb{R}^d$
  - only increments enter
  - The signature does not see the **level** of volatility, so the level enters separately: the log-HAR terms play the role of the external factors x_τ of Gu et al.
- **Basepoint augmentation** Morrill et al. (2020): the path starts at the origin, then runs through the points of the window

$$\hat{X} = (0,\ X_{u_{\mathbf{0}}},\ X_{u_{\mathbf{1}}}, \ldots, X_{u_{L-\mathbf{1}}}), \qquad \mathrm{Sig}(\hat{X}) = \exp_{\otimes}(X_{u_{\mathbf{0}}}) \otimes \mathrm{Sig}(X)$$

  - Level 1 becomes $X_{u_{\mathbf{0}}} + (X_{u_{L-\mathbf{1}}} - X_{u_{\mathbf{0}}}) = X_{u_{L-\mathbf{1}}}$: the starting point does not appear on its own, but the higher-order terms depend on the absolute level (for example $\ln \mathrm{RV}$), so the signature becomes level-aware
  - Translating the path to start at 0 does not do this: it is itself a translation (Seminar 20, C2 uses the prepended segment)
- **Reparametrisation invariance**: $X_{\psi(u)}$, with ψ increasing, has the same signature
  - without a time channel, two moves with the same spatial shape, run at different speeds, have the same signature
  - the time channel tells them apart by referring them to a fixed clock; the augmented path stays invariant to a reparametrisation of the whole augmented path
  - $u_i = i/(L-1)$ maps every window to $[0, 1]$: the absolute duration (22 vs 2 days) needs physical time or a separate regressor
- In this chapter: time augmentation, no basepoint
  - so the kernel compares shapes, not levels
  - the consequence shows in the kernel weights (Results section)

Three Windows: Weights and Information

- Distances $\delta = (0.1, 0.4, 0.9)$ and $\gamma = 2$
- Unnormalised weights: $(0.819, 0.449, 0.165)$, with sum 1.433
- Normalised weights: $w = (0.571, 0.313, 0.115)$
- $n_{\text{eff}} = 1/\sum w_i^2 \approx 2.28$, compared with 3 under equal weights
- γ acts as an inverse temperature
  - a larger γ concentrates the weights, gaining similarity and losing information

Evaluation Protocol: the HAC Variance

- Why not the plain sample variance: d_t is autocorrelated
  - overlapping targets up to lag $h-1$
  - volatility clustering in the losses
  - so $\hat{\Gamma}_0/T$ understates the uncertainty of $\bar{d}$ (and over-rejects) when the sum of the autocovariances is positive, the typical case with overlapping targets
- The Newey–West estimator Newey & West (1987), heteroskedasticity and autocorrelation consistent (HAC)

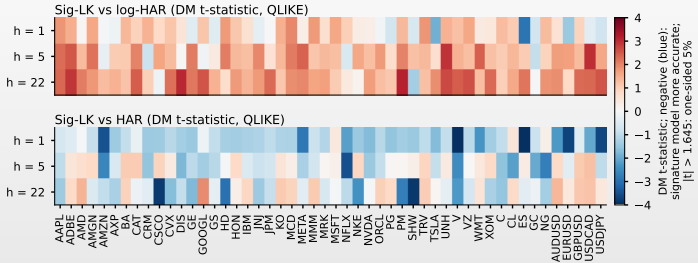
$$\hat{\Omega} = \hat{\Gamma}_0 + 2\sum_{j=1}^{q}\Big(1-\frac{j}{q+1}\Big)\hat{\Gamma}_j, \qquad \hat{\Gamma}_j = \frac{1}{T}\sum_{t=j+1}^{T}(d_t-\bar{d})(d_{t-j}-\bar{d})$$

  - $\hat{\Gamma}_j$: the sample autocovariance of d_t at lag j (not the kernel temperature γ)
  - The Bartlett weights $1-j/(q+1)$ decline linearly with the lag and guarantee $\hat{\Omega} \geq 0$
  - $q = \max(h, \lceil T^{1/3} \rceil)$ lags: at least h, for the dependence created by the overlap ($\lceil\cdot\rceil$: rounding up)

Two Numerical Examples: Holm, BH and the MCS

- Holm compares three sorted p-values, $(0.004, 0.023, 0.08)$, in turn with $(0.05/3, 0.05/2, 0.05)$
  - the first two pass, the third fails, so we stop
- BH compares them with $(0.05/3, 2 \cdot 0.05/3, 0.05)$ and keeps the largest passing rank: again two
- In a family of 50 tests the same p-values may no longer pass
  - which is why the family is fixed before the results
- MCS with three models A, B and C, standardised loss gaps 2.4, -0.5 and -0.7
  - the bootstrap p-value is 0.04, so at 10% we remove A
- Recompute the test for B and C only: the new p-value is 0.30, so we stop and B and C remain
- Failure to separate them is not proof that their population losses are equal
- The 90% coverage is a repeated-sample guarantee, not a posterior probability for the set obtained

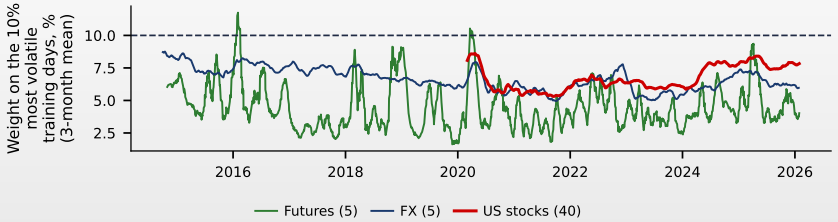
DM Statistics for Every Asset and Horizon



- Diebold–Mariano t-statistics of the QLIKE differential, one column per asset ticker
  - 40 stocks, 5 futures, 5 FX pairs
  - blue: Sig-LK more accurate, red: less accurate
- Against log-HAR (top): pale at $h = 1$, red at $h = 5, 22$
  - blue exceptions: ES, TSLA = Tesla, EURUSD
  - five assets significantly red before correction
- Against HAR (bottom): blue at $h = 1$ because Sig-LK inherits the log target

MFM_ch20_case_study

Does the Kernel Ever Favour Stress? (1/2)



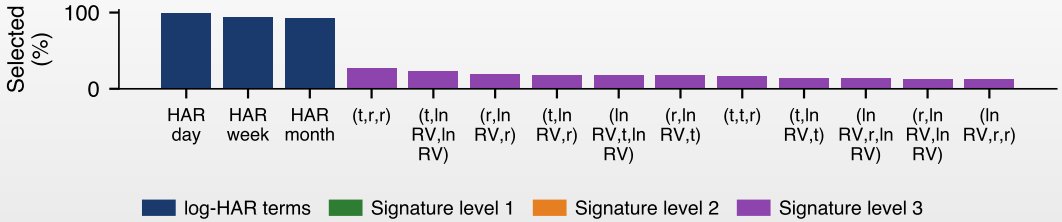
- Weight that Sig-LK ($h = 1$) puts on the 10% most volatile days of its training window
  - averaged by class, three-month mean
- Stocks, per re-estimation: mean 6.7%, below the 10% of equal weighting
  - between 4.4% and 9.9% (maximum on 17 Mar 2020)
- FX stays below 10% throughout
- Futures exceed 10% (three-month mean) only in two short episodes
  - early 2016 (up to 11.7%)
  - March–April 2020

Does the Kernel Ever Favour Stress? (2/2)

- The mechanism behind the null result for H2
  - the weights do not implement "learn from past crises"

MFM_ch20_case_study

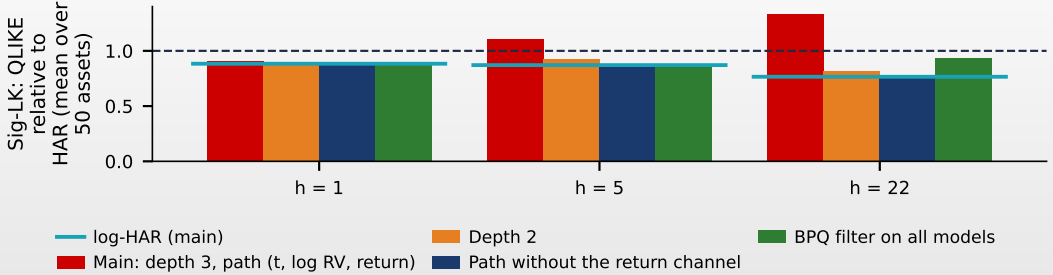
What the LASSO Selects



- Across all assets and horizons: share of re-estimations in which a regressor is in the support of Sig-LK
  - the support is selected by BIC
  - labels: HAR day/week/month = the three log-HAR terms, t = time, ln RV = $\ln \mathrm{RV}$, r = cumulative return
- The log-HAR terms are kept in 99%, 94% and 93% of the re-estimations
  - the LASSO rediscovers HAR
  - average support 8.7 regressors (Sig-L), 6.7 (Sig-LK)
- The most frequent signature term, (t,r,r) (27%), and all frequent ones are of level 3
  - cubic in the increments
  - far outside the training range in a crash, amplified by $\exp(\cdot)$

MFM_ch20_case_study

Robustness of Sig-LK



- Sig-LK relative to HAR (mean over 50 assets)
  - in the main design and in two pre-registered variants
  - with the BPQ filter (the disclosed deviation)
- The reference line marks log-HAR
  - a bar below 1 means a more accurate forecast than HAR

MFM_ch20_case_study

Robustness of Sig-LK: the Variants

- Depth 2: 0.922 and 0.820 at $h = 5, 22$ (main design: 1.104, 1.334)
- Path without returns: 0.878 and 0.770
  - the ablations suggest that the long-horizon failures are linked to terms that contain the return channel
- Depth 2 also reduces them: the ablations suggest a role of the level-3 terms
  - removing terms, however, also changes the selection, the kernel distances and the weights
- Without returns the model approaches log-HAR (0.871 and 0.765)
  - depth 2 trails it
- Both variants reduce instability
  - neither establishes H1

Features and Scaling

◂ Route ◂ Decay ◂ Kernel

- Regressors at day τ for the signature models: 3 log-HAR terms + 39 signature terms
  - $\xi_\tau = \left[\ln \mathrm{RV}_\tau,\ \ln \mathrm{RV}_\tau^{(w)},\ \ln \mathrm{RV}_\tau^{(m)},\ \mathrm{Sig}^3(X_{\tau-21:\tau})\right]$, path X = (time, $\ln \mathrm{RV}$, cumulative r), 22 points
- Each non-time channel q is multiplied by $\kappa_q = 1/(s_q\sqrt{L-1})$
  - so its typical total move over the window is about 1
  - s_q: standard deviation of the daily increments of channel q from the start of the training window up to the origin t ($W + h$ days)
  - $L - 1 = 21$: the number of increments in one path, used only in the factor $\sqrt{L-1}$
  - Time is not rescaled (constant increments, $s = 0$): it runs from 0 to 1; a word $(i_1..i_k)$ is rescaled by $\prod_j \kappa_{i_j}$
- Before the LASSO every column is standardised with weighted moments of the training window
  - constant columns (words of time only) are dropped
- Raw features at τ use data up to τ
  - scales and standardisation use data up to the origin t
  - the no-look-ahead cutoff is t

Estimation: Choosing λ Without Look-Ahead

- Inside each training window: the whole weighted LASSO path by LARS
  - For every support on the path: weighted OLS refit (step 2 of Gu et al.)
- Pick the support that minimises a BIC-type criterion with the Kish effective sample size
  - a heuristic of this design
  - $\mathrm{BIC}(S) = n_{\mathrm{eff}} \ln \hat{s}_S^2 + (|S| + 1) \ln n_{\mathrm{eff}}, \quad n_{\mathrm{eff}} = 1/\sum_\tau w_\tau^2$
  - S: a support (set of kept regressors), $|S|$: its size, $+1$ for the intercept; $\hat{s}_S^2$: weighted residual variance of the refit on S
- Why not cross-validation: 50 assets $\times$ 3 horizons $\times$ hundreds of re-estimations
  - BIC uses only the window
  - Kish size corrects for unequal weights, not for serial dependence (overlapping targets, $h > 1$): not a likelihood-based BIC for these data

Back

Evaluation Windows

Class	Evaluation starts	Ends	Days per asset (median)
Stocks	27 Jan 2020	29 Jan 2026	1 499
Futures	25 Sep 2014	29 Jan 2026	2 919
FX	27 Aug 2014	29 Jan 2026	2 952

- 265 929 forecasts per model over the three horizons: 89 893 ($h = 1$), 89 293 ($h = 5$), 86 743 ($h = 22$)
- Stocks start in 2015, so their evaluation begins in January 2020
  - the COVID-19 crash is in every stock sample
  - Futures and FX start in 2009 and are evaluated from 2014

Back

Chosen Kernel Temperatures

	Sig-LK			log-HAR-K		
Class	$h = 1$	$h = 5$	$h = 22$	$h = 1$	$h = 5$	$h = 22$
Stocks	0.5	0.5	0.5	0.5	0.5	1.0
Futures	2.0	1.0	0.5	2.0	1.0	4.0
FX	0.5	2.0	0.5	4.0	2.0	4.0

- The value of c from the grid $\{0.5, 1, 2, 4\}$, chosen on the validation block of each class and horizon

Back

The Paper's Metric on VOLARE

Median of asset mean errors $\lvert \mathrm{RV} - F \rvert / \mathrm{RV}$ (%)	$h = 1$	$h = 5$	$h = 22$
HAR	57.2	48.9	51.7
log-HAR	50.4	41.9	38.7
HARQ	55.0	49.0	50.7
SHAR	57.9	49.2	51.8
log-HAR-K	49.7	41.0	38.7
Sig-L	50.1	42.6	42.6
Sig-LK	50.6	43.6	43.7

- Median over assets of the mean relative error: about 40–60%
  - against 1–2% for freight rates in Gu et al.
- Daily RV is a noisy, heavy-tailed target
  - the evaluated models have much larger relative errors than those reported for freight rates
  - this does not establish a lower bound on attainable error
  - Sig-LK forecasts above ten times the in-sample maximum: 8 at $h = 1$, 29 at $h = 5$ (concentrated on 10 assets) and 36 at $h = 22$; log-HAR: none

Back

The Insanity Filter, Revisited

QLIKE relative to HAR, BPQ filter on all models	$h = 1$	$h = 5$	$h = 22$
HAR	1.000	1.000	1.000
log-HAR	0.763	0.654	0.778
HARQ	0.995	0.912	1.014
SHAR	1.014	1.003	1.000
log-HAR-K	0.772	0.659	0.779
Sig-L	0.871	0.819	0.904
Sig-LK	0.890	0.867	0.930

- The filter rescues HARQ relative to HAR (0.995 at $h = 1$ instead of 1.140), as BPQ intended
- Sig-LK relative to log-HAR, both filtered: 1.238, 1.424, 1.209
  - the gap widens at $h = 1, 5$ and narrows at $h = 22$
- Filtered forecasts, $h = 1$: HAR 180, HARQ 140, Sig-LK 143, log-HAR 29 (of 89 893)
  - the filter hits the signature models far more often than log-HAR

Back

What the Researcher Must Check

- **The method is the paper's**: kernel, weights, refit, as in the cited equations
  - Test: each component of the method points to the equation, table or section of the paper it comes from; undisclosed choices are labelled as choices of the study
- **No look-ahead**: λ, scales, c and features use only data available at the origin
  - Test: change every value after the origin (for example, multiply future RV by 10); the forecast made at the origin must not change
- **The benchmark isolates the idea**: the 2×2 ablation is reported, not only the full model against HAR
- **Losses and tests are the right way round**: $\mathrm{QLIKE}(\mathrm{RV}, \mathrm{RV}) = 0$ and $\mathrm{QLIKE}(\mathrm{RV}, F) > 0$ otherwise
  - The DM variance is HAC with bandwidth at least h

Back

The Discovery Loop, Prompts (1/2)

- **Read**: extract exactly what the paper did
  - `"From Gu et al. (2024, KDD), list every equation of the estimator, every hyperparameter and its reported value, and the benchmark of each table. Mark what is not disclosed."`
  - Check each item against the PDF: equation numbers, table numbers, the kernel actually used
- **Map**: what transfers, what must be adapted
  - `"Our target is daily realised variance. For each component of the method, say whether it transfers, what must change, and which benchmark isolates it."`
- **Pre-register**: hypotheses, grid, tests, correction
  - fixed in a dated document signed by the team before the first result

Back

The Discovery Loop, Prompts (2/2)

- **Build and test the pieces**: signatures against identities, not against another library
  - "Write unit tests for a depth-3 signature: level 1 equals the increment, the symmetric part of level 2 equals half the outer product, Chen's identity on a split path."
- **Run the grid, then attack it**
  - "Here are all 600 runs (50 assets × 3 horizons × 4 specifications: main, depth 2, no return channel, BPQ filter). Which conclusions survive Holm over 50 assets? Which single asset or subperiod would make a misleading headline?"
- **Explain the null**: find the mechanism, not an excuse

Back

Acronyms used in this material (1/2)

- **AAPL**: Apple Inc. (Nasdaq ticker)
- **ACM**: Association for Computing Machinery
- **AI**: Artificial Intelligence
- **AR**: AutoRegressive model
- **ARIMA**: AutoRegressive Integrated Moving Average (model)
- **AUD**: Australian Dollar (ISO 4217 code)
- **BH**: Benjamini–Hochberg (procedure)
- **BIC**: Bayesian Information Criterion
- **BPQ**: Bollerslev–Patton–Quaedvlieg (insanity filter of the HARQ paper)
- **CAD**: Canadian Dollar (ISO 4217 code)
- **CL**: Crude oil (WTI light sweet) futures, NYMEX ticker
- **COVID-19**: COronaVIrus Disease 2019
- **DM**: Diebold–Mariano (test)
- **ES**: E-mini S&P 500 futures (CME ticker)
- **ETF**: Exchange-Traded Fund
- **EURUSD**: Euro / US dollar exchange rate (FX pair code)
- **FDR**: False Discovery Rate
- **FWER**: Family-Wise Error Rate
- **FX**: Foreign eXchange
- **GARCH**: Generalised AutoRegressive Conditional Heteroskedasticity
- **GC**: Gold futures, COMEX ticker
- **GW**: Giacomini–White (test)
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors)

Acronyms used in this material (2/2)

- **HAR**: Heterogeneous AutoRegressive (model)
- **HARQ**: HAR model with Realised Quarticity
- **JPM**: JPMorgan Chase & Co. (NYSE ticker)
- **KDD**: Knowledge Discovery and Data Mining (ACM SIGKDD conference)
- **LARS**: Least-Angle RegreSsion (algorithm)
- **LASSO**: Least Absolute Shrinkage and Selection Operator
- **MCS**: Model Confidence Set
- **MSE**: Mean Squared Error
- **NG**: Natural Gas futures, NYMEX ticker
- **NVDA**: NVIDIA Corporation (Nasdaq ticker)
- **NW**: Newey–West (standard errors)
- **NYSE**: New York Stock Exchange
- **OLS**: Ordinary Least Squares
- **PDE**: Partial Differential Equation
- **PRDS**: Positive Regression Dependence on a Subset (of p-values)
- **QLIKE**: Quasi-LIKElihood (loss function)
- **RQ**: Realised Quarticity
- **RV**: Realised Variance
- **SHAR**: Semivariance HAR model
- **TSLA**: Tesla, Inc. (Nasdaq ticker)
- **US**: United States
- **USDJPY**: US dollar / Japanese yen exchange rate (FX pair code)
- **VOLARE**: VOLatility Archive for Realized Estimates (database)
- **WTI**: West Texas Intermediate (crude oil benchmark)
- **XOM**: Exxon Mobil Corporation (NYSE ticker)

References (1/3)

- Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6), 1481–1536.
- Barndorff-Nielsen, O.E., Kinnebrock, S., Shephard, N. (2010). Measuring downside risk: realised semivariance. In *Volatility and Time Series Econometrics*, Oxford University Press, 117–136.
- Barndorff-Nielsen, O.E., Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society B*, 64(2), 253–280.
- Belloni, A., Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521–547.
- Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1), 289–300.
- Benjamini, Y., Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4), 1165–1188.
- Bennett, S., Cucuringu, M., Reinert, G. (2022). Lead–lag detection and network clustering for multivariate time series with an application to the US equity market. *Machine Learning*, 111(12), 4497–4538; arXiv:2201.08283.
- Bollerslev, T., Patton, A.J., Quaedvlieg, R. (2016). Exploiting the errors: a simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1), 1–18.
- Chen, K.-T. (1957). Integration of paths, geometric invariants and a generalized Baker–Hausdorff formula. *Annals of Mathematics*, 65(1), 163–178.
- Chen, K.-T. (1958). Integration of paths—a faithful representation of paths by noncommutative formal power series. *Transactions of the American Mathematical Society*, 89(2), 395–407.
- Chevyrev, I., Kormilitzin, A. (2016). A primer on the signature method in machine learning. arXiv:1603.03788.
- Cipollini, F., Cruciani, G., Gallo, G.M., Insana, A., Otranto, E., Spagnolo, F. (2026). VOLatility Archive for Realized Estimates (VOLARE). arXiv:2602.19732.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196.

References (2/3)

- Cuchiero, C., Gazzani, G., Svaluto-Ferro, S. (2023). Signature-based models: theory and calibration. *SIAM Journal on Financial Mathematics*, 14(3), 910–957.
- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407–499.
- Franke, J., Härdle, W.K., Hafner, C.M. (2019). Statistics of Financial Markets: An Introduction, 5th ed. Springer.
- Giacomini, R., White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6), 1545–1578.
- Gu, H., Guo, X., Jacobs, T.L., Kaminsky, P., Li, X. (2024). Transportation marketplace rate forecast using signature transform. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4997–5005; arXiv:2401.04857.
- Hambly, B., Lyons, T. (2010). Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, 171(1), 109–167.
- Hansen, P.R., Huang, Z., Shek, H.H. (2012). Realized GARCH: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, 27(6), 877–906.
- Hansen, P.R., Lunde, A., Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Király, F.J., Oberhauser, H. (2019). Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31), 1–45.
- Levin, D., Lyons, T., Ni, H. (2013). Learning from the past, predicting the statistics for the future, learning an evolving system. arXiv:1309.0260.

References (3/3)

- Liu, L.Y., Patton, A.J., Sheppard, K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187(1), 293–311.
- Lyons, T. (1998). Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2), 215–310.
- Lyons, T., McLeod, A.D. (2022). Signature methods in machine learning. arXiv:2206.14674.
- Menkveld, A.J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., et al. (2024). Nonstandard errors. *Journal of Finance*, 79(3), 2339–2390.
- Morrill, J., Fermanian, A., Kidger, P., Lyons, T. (2020). A generalised signature method for multivariate time series feature extraction. arXiv:2006.00873.
- Newey, W.K., West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.
- Patton, A.J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.
- Patton, A.J., Sheppard, K. (2015). Good volatility, bad volatility: signed jumps and the persistence of volatility. *Review of Economics and Statistics*, 97(3), 683–697.
- Politis, D.N., Romano, J.P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313.
- Salvi, C., Cass, T., Foster, J., Lyons, T., Yang, W. (2021). The signature kernel is the solution of a Goursat PDE. *SIAM Journal on Mathematics of Data Science*, 3(3), 873–899.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1), 267–288.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.