









Modelling Financial Markets

Chapter 9: Realised Volatility



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

How Do We Measure Latent Volatility?



Torben G. Andersen.
Photo: Soralum (2012);
CC BY-SA 4.0; Wikimedia Commons

"We provide a framework for integration of high-frequency intraday data into the measurement, modeling, and forecasting of daily and lower frequency return volatilities and return distributions."

— *Andersen, Bollerslev, Diebold & Labys (2003)*

- Realised variance (RV), the sum of squared 5-minute returns, measures each day's variance almost without a model
  - annualised volatility from the mean RV: 12.9% for SPY (the S&P 500 exchange-traded fund), 45.1% for Bitcoin
- Overnight returns carry 38% of SPY's daily variance, although the market is closed for 17.5 hours
- Jump tests on 1 489 SPY days: 236 rejections at 5% without correction
  - 17 after the periodicity correction and false discovery rate (FDR) control
- Out of sample, SPY 2022–2026: the log-HAR model (heterogeneous autoregressive) has a QLIKE (quasi-likelihood) loss of 0.310, against 0.343 for GARCH(1,1)-t
  - Diebold–Mariano $t = -2.30$, $p = 0.021$

▸ Central chart: Forecasts over Time (1/2)

From Intraday Prices to Variance Forecasts (1/2)

What we will explain

- Part I, **measurement and diagnosis**: realised variance and its precision, microstructure noise, jumps
- Part II, **modelling and forecasting**: stylised facts, the HAR model, forecast evaluation against GARCH
  - HAR: heterogeneous autoregressive
  - GARCH: generalised autoregressive conditional heteroskedasticity (Chapter 5)
- Extensions: rough volatility, a machine-learning case study, research assisted by AI
  - AI: artificial intelligence
- Appendix: asymptotic theory, noise- and jump-robust estimators, model extensions
- 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

From Intraday Prices to Variance Forecasts (2/2)

What you should be able to do

- ⊡ compute daily realised variance, its standard error and a confidence interval from intraday prices
- ⊡ interpret a signature plot as a diagnosis of microstructure noise
- ⊡ compute the ratio jump test with false discovery rate (FDR) control
- ⊡ compute the log-HAR estimates and interpret their persistence
- ⊡ derive the quasi-likelihood loss (QLIKE) from the Normal likelihood
- ⊡ interpret comparisons of variance forecasts: Diebold–Mariano, Mincer–Zarnowitz and the model confidence set
- ⊡ Further Topics and Appendix: reading for familiarity

Volatility Is Latent: Can We Measure It? (1/2)



Torben G. Andersen, 2012

Photo: Soralum (2012); CC BY-SA 4.0; Wikimedia Commons

- Volatility drives risk, option prices and portfolio weights, but it is never observed
  - risk: Chapters 7–8; option prices: Chapter 12; portfolio weights: Chapter 4
  - One daily return is a single draw from a distribution whose dispersion we want to estimate
  - GARCH (Chapter 5) **infers** today's variance from past daily returns through a model
- Idea: the sum of squared intraday price changes **measures** the day's variance almost without a model Andersen & Bollerslev (1998)
  - inside one day there are hundreds of price changes
  - Early idea: Merton (1980) showed that variance, unlike the mean, is estimated better with more frequent data

Volatility Is Latent: Can We Measure It? (2/2)

- Two different objects
  - realised variance: an **ex-post measurement** of day t, known after the close
  - GARCH variance: a **conditional forecast** of day t made with the information of day $t-1$
- Measure first, then model
  - Andersen, Bollerslev, Diebold & Labys (2001), Andersen, Bollerslev, Diebold & Ebens (2001) study the distribution of realised volatility
  - Andersen, Bollerslev, Diebold & Labys (2003) forecast it directly
  - textbook treatment: Franke, Härdle & Hafner (2019)

Reading Intraday Data

- SPY: an Exchange-Traded Fund (ETF) tracking the Standard & Poor's 500 stock index
  - ETF: a fund whose units trade on an exchange
- UTC means Coordinated Universal Time
  - Bitcoin daily intervals run midnight to midnight in that time zone
- VIX: the Volatility Index of the Chicago Board Options Exchange (CBOE)
  - an option-price measure of expected volatility
- All series end on 18 September 2026
- Returns $r_i = 100(\ln P_i - \ln P_{i-1})$ in %, P_i: the price at the end of bar i
- Charts and times quoted in this chapter mark the **end** of each 5-minute interval
  - 09:35 is the first SPY return, 09:30–09:35

▸ Further Topics: data series table

A Model for Intraday Prices (1/2)

- **Question**: how do we turn a day's 5-minute returns into a measurement of that day's variance?
- Notation: day t has M intervals of length $\Delta = 1/M$, returns $r_{t,i}$, daily return $r_t = \sum_i r_{t,i}$
  - Variances in $\%^2$
  - annualised volatility $\sqrt{252 \times \text{variance}}$ (SPY), $\sqrt{365 \times \text{variance}}$ (Bitcoin)
  - the SPY overnight return (close to next open) is outside the bars
- Log price in continuous time: $dp(s) = \mu(s)\,ds + \sigma(s)\,dW(s)$, with W a Brownian motion (Chapter 11)
  - $\sigma(s)$ is the **spot volatility**: it changes within the day and from day to day

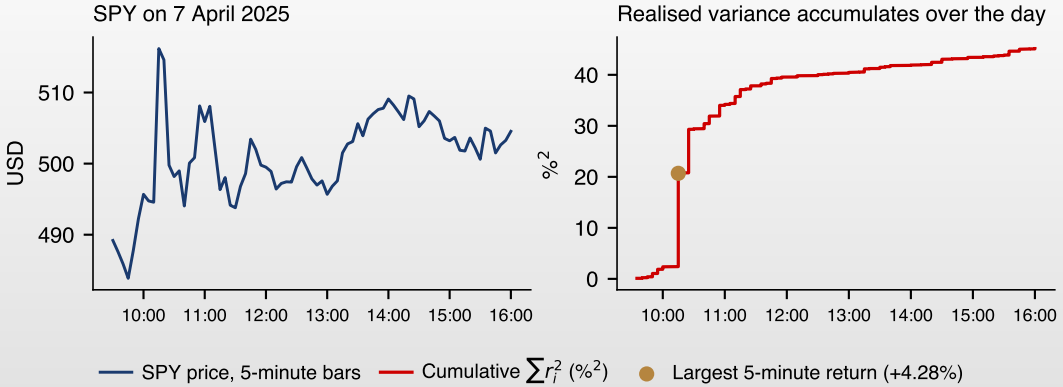
A Model for Intraday Prices (2/2)

- **Integrated variance** of day t: the quantity we want to measure
  - $\mathrm{IV}_t = \int_{t-1}^{t} \sigma^2(s)\, ds$: the spot variance accumulated over the day
  - in this chapter IV always means integrated variance, never implied volatility
- Given the path of σ, the daily return is Normal: $r_t \mid \sigma \sim N(\int \mu, \mathrm{IV}_t)$
  - assumption: σ is independent of W (no leverage effect: price falls do not raise volatility)
  - assumption: the drift is zero (or fixed by the path of σ)
  - With a leverage effect (Appendix: Itô semimartingales) this conditional normality fails, but the limit theory of realised variance (RV, next slide) still holds

Definition: Realised Variance

- **Realised variance**: the sum of squared intraday returns
  - $\mathrm{RV}_t = \sum_{i=1}^{M} r_{t,i}^2$, realised volatility $= \sqrt{\mathrm{RV}_t}$
- Why squares: the conditional means of the squared increments approximate $\int_{t-1}^{t} \sigma^2(s)\, ds$
  - over a short interval $\mathbb{E}[r_{t,i}^2 \mid \sigma] \approx \sigma^2(s_i)\,\Delta$
  - the observed sum of squares converges to the quadratic variation; it is not a Riemann sum
- **Quadratic variation** of day t, with $p = 100 \ln P$: $[p]_t - [p]_{t-1} = \operatorname{plim}_{M\to\infty} \sum_{i=1}^{M} r_{t,i}^2 = \mathrm{IV}_t$ for a price without jumps Barndorff-Nielsen & Shephard (2002)
  - The drift adds terms of order Δ per day, which vanish as $M \to \infty$; for SPY the mean 5-minute return is below 1% of its standard deviation
- No parameters to estimate: RV_t is a **model-free** measurement of one day's variance
- The squared daily return r_t^2 is the special case $M = 1$
  - unbiased (with zero drift) but extremely noisy

Five-Minute Returns on 7 April 2025 (1/2)



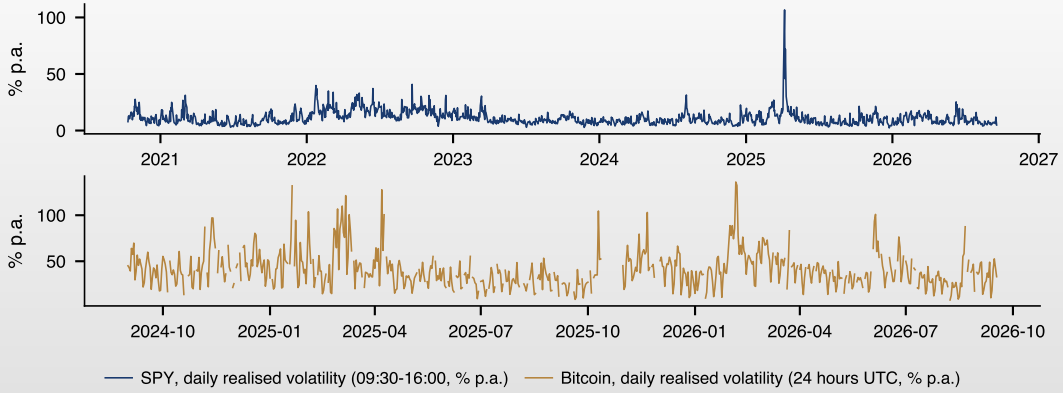
MFM_ch9_intraday_day ▶ Interactive

Five-Minute Returns on 7 April 2025 (2/2)

- Left: SPY on 5-minute bars on the most volatile day of the sample
- Right: running sum of squared returns
- SPY opened with an overnight log return of -3.27%, after tariff announcements
  - 78 five-minute returns followed
  - $\mathrm{RV} = 45.1\%^2$: annualised realised volatility $\sqrt{252 \times 45.1} = 107\%$
- One 5-minute return, of $+4.28\%$, contributed 41% of the day's realised variance
  - interval ending 10:15
- Open-to-close the day ended $+3.09\%$
  - the daily return hides the intraday swings that RV records

MFM_ch9_intraday_day

Daily Realised Volatility: SPY and Bitcoin (1/2)



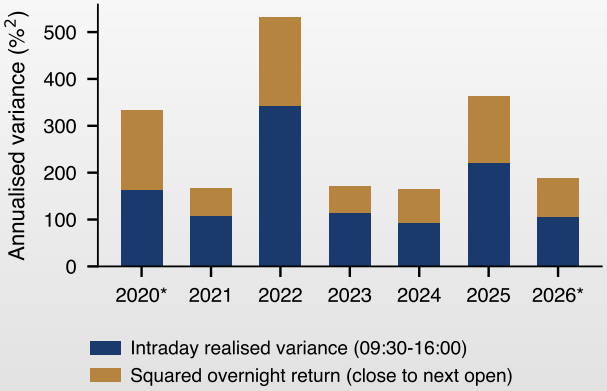
MFM_ch9_rv_series ▶ Interactive

Daily Realised Volatility: SPY and Bitcoin (2/2)

- Annualised, from 5-minute returns
  - SPY: regular US (United States) session
  - Bitcoin: 24 hours
  - the gap in October 2025 is missing data
- SPY: volatility from mean RV $\sqrt{252\,\overline{\mathrm{RV}}} = 12.9\%$ p.a.
  - median day 9.7%, range 2.3% to 106.7%
  - Volatility computed from mean variance exceeds the median: a few very volatile days raise the root-mean-square of daily volatilities
- Bitcoin: $\sqrt{365\,\overline{\mathrm{RV}}} = 45.1\%$ p.a.
  - median 37.8%
  - maximum 136.1% on 2026-02-05
- Both series show **volatility clustering**, as with GARCH (Chapter 5)
  - but every point is a measurement, not a model output

MFM_ch9_rv_series

Intraday and Overnight Variance (1/2)



MFM_ch9_overnight ▶ Interactive

Intraday and Overnight Variance (2/2)

- SPY, mean daily variance by year, annualised
  - intraday realised variance plus squared overnight return
  - 2020 and 2026 are partial years
- Overnight returns carry 38% of SPY's daily variance, although the market is closed for 17.5 hours
  - News (earnings, macro data, Asian and European sessions) accumulates and is priced at the open
- Type of evidence: description (it shows what happened in the data, without testing a cause)
- Annualised volatility
  - close-to-close 16.6%
  - open-to-close 13.6%
  - overnight 10.1%
- For daily risk we use **total** variance: $\mathrm{RV}_t + r^2_{\textbf{overnight},t}$
  - annualised mean: 16.4%, against 12.9% from intraday RV alone
  - Alternatives: scale intraday RV up, or weight the two parts optimally Hansen & Lunde (2005)

MFM_ch9_overnight

How Precise Is Realised Variance?

- Central limit theorem (CLT) for RV, in feasible form (computable from data) Barndorff-Nielsen & Shephard (2002)
  - $\sqrt{M}\,(\mathrm{RV}_t - \mathrm{IV}_t)/\sqrt{2\,\mathrm{RQ}_t} \to N(0,1)$
  - $\mathrm{IQ}_t = \int_{t-1}^{t} \sigma^4(s)\,ds$: the integrated quarticity, which sets the variance of RV
  - IQ_t is estimated by the realised quarticity $\mathrm{RQ}_t = \frac{M}{3}\sum_i r_{t,i}^4$
- Standard error: $\mathrm{se}(\mathrm{RV}_t) = \sqrt{\frac{2}{3}\sum_i r_{t,i}^4}$
- 95% confidence interval (CI): $\mathrm{RV}_t \exp\left(\pm 1.96\,\mathrm{se}(\mathrm{RV}_t)/\mathrm{RV}_t\right)$
  - on the log scale, so always positive
  - With constant volatility, $\mathrm{se}/\mathrm{RV} \approx \sqrt{2/M}$: 16% for $M = 78$
- The limiting variance $2\,\mathrm{IQ}_t$ is itself random (a mixed Normal limit)
  - replacing it by $2\,\mathrm{RQ}_t$ is legitimate because the convergence is **stable**
  - formal statement in the Appendix

Why Not Use Every Tick? (1/2)

- **Question**: why not use every trade?
  - more returns shrink the relative error of RV ($\sqrt{2/M}$)
- Observed prices are not the efficient price: $p_i^{\mathbf{obs}} = p_i^* + \varepsilon_i$
  - so $r_i = r_i^* + \varepsilon_i - \varepsilon_{i-\mathbf{1}}$
  - p_i^*: the efficient (true) log price, ε_i: the noise in the observed price, r_i^*: the efficient return
  - Sources of ε: bid–ask bounce, tick size (price discreteness), stale quotes (Chapter 10)

Why Not Use Every Tick? (2/2)

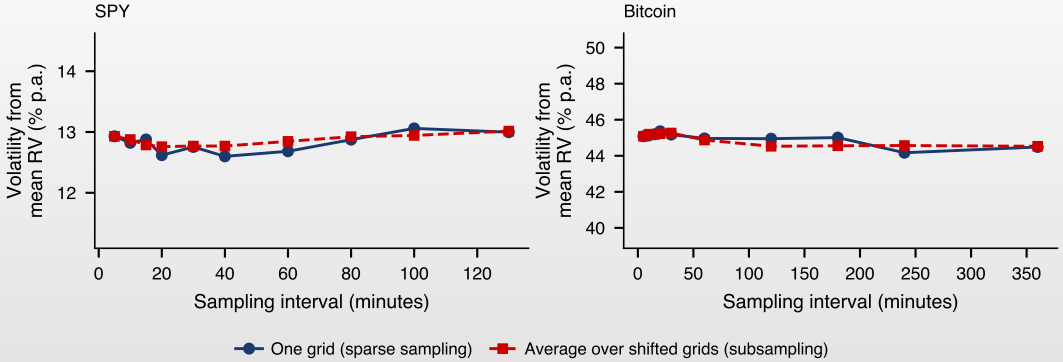
- With i.i.d. noise of variance ω^2 and n returns per day Zhang, Mykland & Aït-Sahalia (2005), Hansen & Lunde (2006)
  - $\mathbb{E}[\mathrm{RV}_t^{(n)} \mid \sigma] = \mathrm{IV}_t + 2n\omega^2$ (zero drift, no leverage): each squared return carries $2\omega^2$ of noise, so the bias **grows** with the sampling frequency
  - Adjacent returns share ε_{i-1} with opposite signs: $\mathrm{cov}(r_i, r_{i-1}) = -\omega^2$
- Trade-off: $\mathbb{E}[(\mathrm{RV}^{(n)} - \mathrm{IV})^2] \approx 2\,\mathrm{IQ}/n + 4n^2\omega^4$
  - discretisation error plus squared bias
  - MSE (mean squared error) optimal number of returns $n^* = (\mathrm{IQ}/(4\omega^4))^{1/3}$ Bandi & Russell (2008): more returns cut the first term, but raise the second; derivation in the Appendix
- Numbers, with $\mathrm{IV} = 1\%^2$ and $\omega = 0.01\%$, so $\omega^2 = 10^{-4}\%^2$
  - constant volatility within the day, so $\mathrm{IQ} = \mathrm{IV}^2 = 1\ (\%^4)$
  - the bias $2n\omega^2$ is 0.016 ($\%^2$) at 5 minutes ($n = 78$) and 4.680 at 1 second, larger than IV itself
- Optimal number of returns: $n^* = 292$ (one return every 80 s)

Further Topics: simulated signature plot

Why 5 Minutes?

- **Question**: which frequency, when n^* depends on ω^2 and IQ_t?
  - both are unknown and different across assets
- The optimal frequency falls fast as the noise grows
  - $n^* = 292$ with $\omega = 0.01\%$
  - $n^* = 14$ (one return about every 29 minutes) with $\omega = 0.1\%$
- **Sparse sampling** at 5 minutes ($n = 78$ in a 6.5-hour session): a robust compromise Liu, Patton & Sheppard (2015)
  - the bias $2n\omega^2$ stays small, and the relative discretisation error is $\sqrt{2/78} = 16\%$
  - the cost: most trades are discarded
- The course data are 5-minute bars, the finest grid available
  - the signature plot (mean RV by sampling interval) checks on real data whether 5 minutes is sparse enough

The Signature Plot on Real Data (1/2)



MFM_ch9_signature ▶ Interactive

The Signature Plot on Real Data (2/2)

- Annualised volatility from mean RV by sampling interval
  - one grid vs the average over all shifted grids (subsampling)
- Approximately flat in the sampling interval: compatible with a small noise bias, not proof of no noise
  - SPY 12.9% at 5 minutes, 12.8% at 30, 13.0% at 130
  - Bitcoin 45.1% at 5 minutes, 44.5% at 6 hours
  - First-order autocorrelation of 5-minute returns: -0.008 (SPY), 0.005 (Bitcoin): i.i.d. noise would make it negative ($-\omega^2$), so no visible noise at 5 minutes
- Limit of this dataset: 5-minute bars say nothing about noise below 5 minutes
- Noise-robust estimators (Further Topics), as ratios to RV on SPY: differences between estimators, not a measured noise bias
  - two-scale RV on the standard grids: 0.84; with the day's first and last price in every grid: 0.97
  - realised kernel with jittered end points: 0.92; with the raw end prices: 0.95
  - with 78 bars, the theoretical versions drop part of the first and last intervals of the day, the most volatile ones
- Consistent with Liu, Patton & Sheppard (2015): almost 400 estimators on 31 assets rarely beat 5-minute RV
  - noise-robust estimators pay off when noise is large relative to the signal (tick or second-level data, illiquid assets)
- Type of evidence: description

MFM_ch9_signature

Prices Jump: Bipower Variation (1/2)



Ole E. Barndorff-Nielsen, 2007

Photo: Thomas Steiner (2007); CC BY-SA 2.5; Wikimedia Commons

- **Question**: how much of a day's variance comes from jumps, and how do we separate jumps from diffusion (the continuous part, $\sigma\, dW$)?
- Add jumps to the model: $dp = \mu\, ds + \sigma\, dW + dJ$
  - J: a jump process (news, Chapter 11)
  - Quadratic variation now has two parts: $\mathrm{RV}_t \to \mathrm{IV}_t + \sum_j \kappa_j^2$, with κ_j the jump sizes in day t
  - Why separate them? Jumps are less persistent than diffusive volatility, and they price differently (options, Chapter 12)

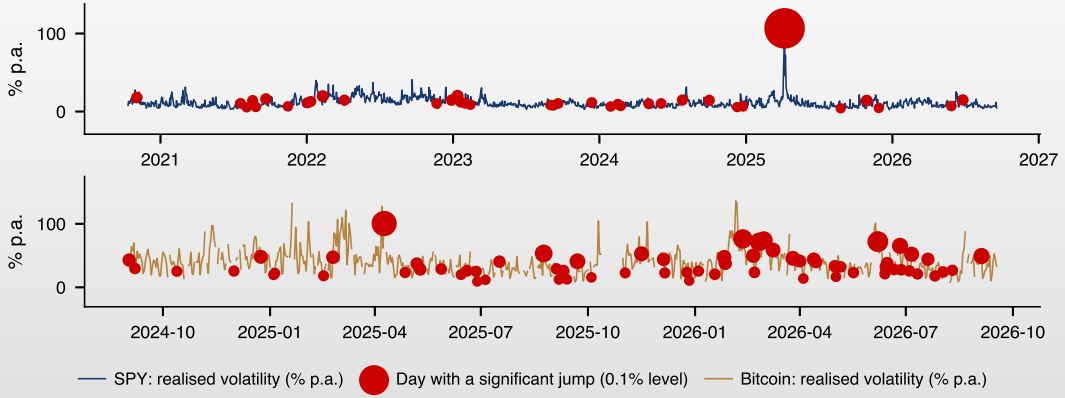
Prices Jump: Bipower Variation (2/2)

- **Bipower variation** (BV) Barndorff-Nielsen & Shephard (2004): products of adjacent absolute returns
  - $\mathrm{BV}_t = \mu_1^{-2} \frac{M}{M-1} \sum_{i=2}^{M} |r_{t,i}|\,|r_{t,i-1}| \to \mathrm{IV}_t$
  - $\mu_1 = \mathbb{E}|Z| = \sqrt{2/\pi} = 0.798$ for $Z \sim N(0,1)$, so $\mu_1^{-2} = \pi/2 = 1.571$; $\frac{M}{M-1}$: finite-sample normalisation
- Base case: jumps of finite activity (finitely many jumps per day) and no noise; other settings need extra conditions
- A jump then enters one return, which is multiplied by a small neighbour
  - its effect on BV vanishes as M grows
- The jump part of the day is $\max(\mathrm{RV}_t - \mathrm{BV}_t, 0)$
  - Example, $M = 4$ returns $0.1, 0.1, 0.1, 1.0$ (%): $\mathrm{RV} = 1.03$, $\mathrm{BV} = \frac{\pi}{2} \cdot \frac{4}{3}(0.01 + 0.01 + 0.10) = 0.25$, jump part 0.78 ($\%^2$)

A Test for Jumps

- Ratio statistic Huang & Tauchen (2005), Barndorff-Nielsen & Shephard (2006)
  - $z_t = \dfrac{(\mathrm{RV}_t - \mathrm{BV}_t)/\mathrm{RV}_t}{\sqrt{\theta\, M^{-1} \max\left(1, \mathrm{TQ}_t/\mathrm{BV}_t^2\right)}}$
  - $\theta = \pi^2/4 + \pi - 5 = 0.609$; tripower quarticity $\mathrm{TQ}_t = M\mu_{4/3}^{-3} \frac{M}{M-2} \sum_i |r_{t,i}|^{4/3} |r_{t,i-1}|^{4/3} |r_{t,i-2}|^{4/3}$, $\mu_{4/3} = \mathbb{E}|Z|^{4/3}$
  - Numerator: the relative jump share $(\mathrm{RV}_t - \mathrm{BV}_t)/\mathrm{RV}_t$; denominator: its standard error
  - θ: the asymptotic variance constant; TQ (Tripower Quarticity) estimates IQ_t robustly to jumps
- Under no jumps $z_t \to N(0,1)$
  - a jump makes $\mathrm{RV} > \mathrm{BV}$ and z_t large and positive
- We use $\alpha = 0.1\%$ (critical value 3.09) for the jump days plotted on the next slide
- Jump part (estimates $\sum_j \kappa_j^2$ of the process J): $J_t = \mathrm{RV}_t - \mathrm{BV}_t$ on a significant day, otherwise $J_t = 0$
  - continuous part $C_t = \mathrm{RV}_t - J_t$

Jump Days: SPY and Bitcoin (1/2)



MFM_ch9_jumps ▶ Interactive

Jump Days: SPY and Bitcoin (2/2)

- Red points: days with a significant jump at the 0.1% level
  - point size proportional to the jump part J_t
- SPY: 36 jump days of 1 489 (2.4%)
  - jumps are 2.8% of total realised variance
  - on a jump day the jump is on average 36% of that day's RV
  - largest statistic: 2021-08-27 ($z = 5.2$), 10:00–10:05: the Jackson Hole speech of the Fed chair
- Bitcoin: 67 jump days of 659 (10.2%), but only 1.8% of variance
  - many small jumps on quiet days (weekends), where one isolated move dominates the day
- The turbulent days of April 2025 are mostly **not** jump days
  - many large returns in a row raise BV as well

MFM_ch9_jumps

Multiple Testing and Spurious Jump Detections

- 1 489 daily tests at 5%: 236 rejections; which of them are false?
  - against 74 expected under no jumps
- Part of the excess is intraday seasonality, higher volatility at the open and close (Further Topics)
  - dividing returns by the periodicity factor f_i Boudt, Croux & Laurent (2011) already cuts the days rejected at 5% from 236 to 169
- Benjamini & Hochberg (1995): sort the p-values and reject the k smallest
  - k: the largest index with $p_{(k)} \leq kq/N$; $p_{(k)}$: k-th smallest p-value, N: number of tests, q: target FDR
  - here each day's p-value is the one-sided $1 - \Phi(z_t)$ of the ratio statistic
  - the FDR, $\mathbb{E}$[false/rejections], stays $\leq q$ (false/rejections $:= 0$ when nothing is rejected)
  - The guarantee needs valid p-values that are independent or positively regression dependent on the subset of true nulls (PRDS) Benjamini & Yekutieli (2001)
  - daily jump tests are only approximately so

Jump Counts under Multiplicity Control

- SPY, 1 489 days
  - without the periodicity correction, FDR at 5%: 43 jump days; at 1%: 14
  - Bonferroni (each day tested at 5%/1 489; family-wise error rate, FWER, 5%): 7
  - FDR 5% after the periodicity correction: 17
  - from 236 unadjusted rejections to 17: the drop does not identify which rejections were false
- Type of evidence: description
- Bajgrowicz, Scaillet & Treccani (2016), on their sample and tests: without multiplicity control most detected jumps are spurious
  - the true ones are rare and cluster on news
- Christensen, Oomen & Podolskij (2014), on their tick data: most 5-minute "jumps" are steep but continuous bursts of volatility
  - at tick frequency (fact or friction)
- A daily test answers "did day t contain a jump?"
  - locating the jump within the day needs an intraday test (Lee–Mykland, Appendix)

Check Yourself: Measurement, Noise and Jumps (1/2)

- Why does the standard error of RV use fourth powers of returns?
  - The variance of RV depends on integrated quarticity IQ_t, estimated by $\mathrm{RQ}_t = \frac{M}{3} \sum_i r_{t,i}^4$
- Why is SPY's daily risk understated if you use intraday RV alone?
  - Overnight returns bring 38% of SPY daily variance
- Why does sampling every tick overstate variance?
  - Every squared return adds $2\omega^2$ of noise, so the bias $2n\omega^2$ grows with the number of returns

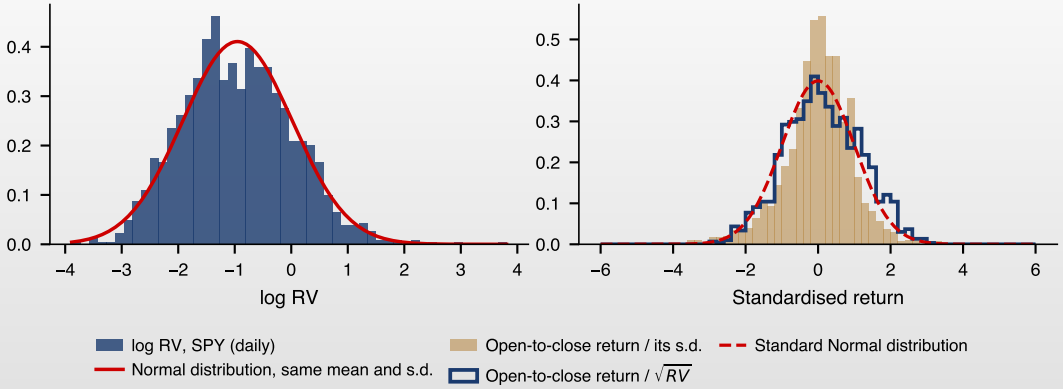
Check Yourself: Measurement, Noise and Jumps (2/2)

- What does a flat signature plot tell you?
  - It is compatible with a small noise bias at 5 minutes and coarser; it does not rule out noise
- What does it not tell you?
  - It says nothing about noise below 5 minutes
- Why does bipower variation stay close to IV on a jump day while RV does not?
  - With finite-activity jumps and no noise, the jump enters a single return: squared in RV, multiplied by a small neighbour in BV
- With 1 489 daily jump tests, why is "$p < 0.05$" not evidence of a jump?
  - At 5%, 1 489 jump-free days would give about 74 false rejections
- What does the FDR control?
  - The FDR controls the expected share of false rejections among rejections

What Regularities Does Realised Volatility Show?

- **Question**: which properties of RV must a forecasting model reproduce?
- Distribution: is RV roughly lognormal? If so, we model $\ln \mathrm{RV}$
- Returns divided by $\sqrt{\mathrm{RV}_t}$: how much of the fat tails is due to time-varying volatility?
- Persistence: how slowly does the autocorrelation of $\ln \mathrm{RV}$ decay?
- Implied volatility: does the VIX forecast realised volatility over the next month?

Distributions: log RV and Standardised Returns (1/2)



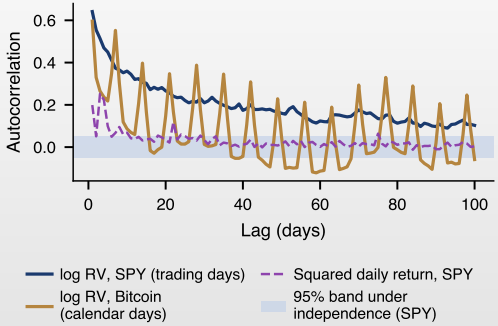
MFM_ch9_distributions ▶ Interactive

Distributions: log RV and Standardised Returns (2/2)

- Left: daily log RV of SPY and the Normal density
- Right: open-to-close returns scaled
  - by their standard deviation (filled)
  - by $\sqrt{\mathrm{RV}_t}$ (outline)
- SPY skewness and kurtosis (the Normal distribution: 0 and 3)
  - RV: skewness 20.0, kurtosis 550
  - log RV: skewness 0.29, kurtosis 3.21
  - Bitcoin log RV: skewness -0.27, kurtosis 3.33: model volatility in logs
- Open-to-close returns: kurtosis 15.73, and 2.66 after dividing by $\sqrt{\mathrm{RV}_t}$
  - the sample kurtosis can be computed even when the population fourth moment is infinite; it then estimates no finite population kurtosis (Chapter 1)
  - standard deviation after dividing: 1.05
  - Bitcoin: from 7.56 to 2.77
- Heavy tails of daily returns (Chapter 1) come mostly, ex post, from changing volatility
  - approximately a **mixture** of Normal distributions with variances IV_t
  - exact only without leverage and jumps (section Realised Variance)
- Type of evidence: description

MFM_ch9_distributions

Persistence of Realised Volatility (1/2)



- Autocorrelations up to 100 days
  - log RV of SPY (total variance) and Bitcoin
  - squared daily SPY returns
- SPY log RV: **highly persistent** dependence
  - 0.64 at lag 1, 0.24 at one month, still 0.10 at 100 days
  - an autoregressive AR(1) model would decay geometrically: $0.64^{22} \approx 0.00005$
  - a sample autocorrelation function (ACF) is consistent with long memory (autocorrelations decaying more slowly than geometrically) but cannot prove it

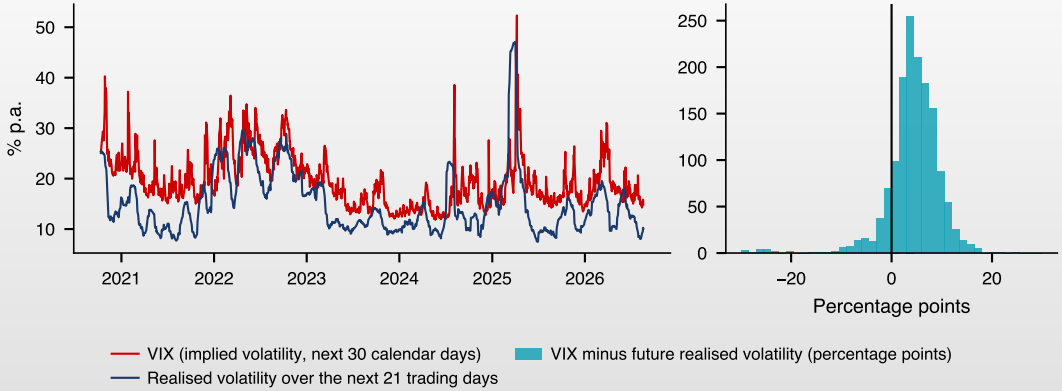
▶ Interactive

Persistence of Realised Volatility (2/2)

- Squared daily returns show the same persistence, hidden by noise
  - autocorrelation at lag 1: 0.20
  - the autocorrelation of r_t^2 requires a finite fourth moment, which Chapter 1 questions
- Bitcoin: peaks at 7, 14, 21, ... days: a **weekly cycle**
  - weekend RV is only 0.41 of weekday RV (Saturday 27% vs Wednesday 48% p.a.)
  - no equity or macro news, fewer institutional traders

MFM_ch9_acf

Implied vs Realised Volatility (1/2)



MFM_ch9_vix ▶ Interactive

Implied vs Realised Volatility (2/2)

- Left: VIX and SPY realised volatility over the next 21 trading days
- Right: the gap between them
- VIX mean 19.52% vs 15.15% realised afterwards
- VIX was higher on 88% of 1 468 days, by 4.36 percentage points on average: the mean ex-post volatility gap
- **Variance risk premium** (VRP) Carr & Wu (2009), Bollerslev, Tauchen & Zhou (2009)
  - $Q_t = \frac{252}{21} \sum_{j=1}^{21} V_{t+j}$: annualised total variance over the next 21 trading days
  - $\mathrm{VRP}_t = \mathbb{E}_t^{\mathbb{Q}}[Q_t] - \mathbb{E}_t^{\mathbb{P}}[Q_t]$: risk-neutral expectation ($\mathbb{Q}$, from option prices) minus the real-world one ($\mathbb{P}$); VIX^2 approximates the first term (Chapter 12)
  - $\mathrm{VIX}_t^2 - Q_t \approx \mathrm{VRP}_t + (\mathbb{E}_t^{\mathbb{P}}[Q_t] - Q_t)$: the ex-post gap adds a forecast error to the premium
  - the sign of the realised gap does not identify the sign of the ex-ante premium
- Type of evidence: description

MFM_ch9_vix

The HAR-RV Model (1/2)

- **Question**: how do we forecast tomorrow's RV when the autocorrelation of $\ln \mathrm{RV}$ decays very slowly?
  - the autocorrelation is still 0.10 at 100 days
- Corsi (2009): $\mathrm{RV}_{t+1} = \beta_0 + \beta_d \mathrm{RV}_t + \beta_w \mathrm{RV}_t^{(w)} + \beta_m \mathrm{RV}_t^{(m)} + \varepsilon_{t+1}$
  - $\mathrm{RV}_t^{(w)} = \frac{1}{5} \sum_{j=0}^{4} \mathrm{RV}_{t-j}$ (week), $\mathrm{RV}_t^{(m)} = \frac{1}{22} \sum_{j=0}^{21} \mathrm{RV}_{t-j}$ (month); Bitcoin trades every day, so we use 7 and 30 days
  - β_d, β_w, β_m: weights of yesterday's RV and of last week's and last month's average RV; β_0: constant; ε_{t+1}: forecast error
- Heterogeneous market hypothesis Müller et al. (1997): daily, weekly and monthly traders
  - the "H" (heterogeneous) of HAR-RV (Heterogeneous AutoRegressive model of Realised Volatility), the HAR model of realised variance
  - a few components with different persistence approximate the observed persistence over finite horizons
  - a stationary HAR is a finite AR: short memory in the asymptotic sense

The HAR-RV Model (2/2)

- An AR(22) with only 3 free slopes
  - weight on lag 1: $\beta_d + \beta_w/5 + \beta_m/22$
  - weight on lags 2–5: $\beta_w/5 + \beta_m/22$
  - weight on lags 6–22: $\beta_m/22$
- Estimation by ordinary least squares (OLS)
  - errors are heteroskedastic and autocorrelated, so we use Newey–West standard errors
  - Newey–West standard errors are HAC: heteroskedasticity- and autocorrelation-consistent

The log-HAR Model

- Why logs
  - $\ln \mathrm{RV}$ is close to Normal, whereas RV is very heavy-tailed (slide *Distributions*)
  - the forecast $\exp(\cdot)$ is always positive
  - extreme days no longer dominate the OLS fit
- $\ln \mathrm{RV}_{t+1} = \beta_0 + \beta_d \ln \mathrm{RV}_t + \beta_w \overline{\ln \mathrm{RV}}_t^{(w)} + \beta_m \overline{\ln \mathrm{RV}}_t^{(m)} + \varepsilon_{t+1}$
  - with $\overline{\ln \mathrm{RV}}_t^{(w)} = \frac{1}{5} \sum_{j=0}^{4} \ln \mathrm{RV}_{t-j}$ and $\overline{\ln \mathrm{RV}}_t^{(m)} = \frac{1}{22} \sum_{j=0}^{21} \ln \mathrm{RV}_{t-j}$
  - geometric means of variances
- Variance forecast: $\exp(\hat{y} + \hat{s}^2/2)$
  - $\hat{y}$: the fitted $\ln \mathrm{RV}$; $\hat{s}^2$: the residual variance
  - the conditional mean if the errors are $N(0, s^2)$ (lognormal approximation)
  - for non-Normal errors use the smearing factor $\frac{1}{T} \sum_t e^{\hat{\varepsilon}_t}$ Duan (1983)
  - with heteroskedastic errors both are approximations

HAR Estimates: SPY and Bitcoin

	SPY, HAR		Bitcoin, HAR	
	levels	log	levels	log
β_0	0.273	−0.079	1.496	0.223
	(0.064)	(0.027)	(0.412)	(0.094)
β_d	0.399	0.362	0.412	0.380
	(0.089)	(0.037)	(0.089)	(0.041)
β_w	0.222	0.336	0.409	0.465
	(0.096)	(0.048)	(0.122)	(0.080)
β_m	0.114	0.168	−0.065	−0.022
	(0.053)	(0.044)	(0.097)	(0.098)
R^2	0.31	0.47	0.34	0.36

- Full samples
  - SPY: total daily variance (intraday RV plus squared overnight return), 1 467 days
  - Bitcoin: 24-hour RV, 547 days on which all three HAR regressors are available
- Newey–West standard errors in brackets
  - HAC errors need the regressors orthogonal to the error and a finite long-run covariance of $x_t\varepsilon_t$; in levels this involves the fourth moment of RV, not of returns (Chapter 1)
  - HAC corrects for autocorrelation and heteroskedasticity, not for endogeneity

Reading the HAR Estimates

- SPY: all three components are significant
  - log-HAR persistence $\hat{\beta}_d + \hat{\beta}_w + \hat{\beta}_m = 0.87$: shocks fade slowly, as in GARCH (Chapter 5)
- $R^2 = 0.47$ for log variance vs 0.31 for variance in levels
  - not comparable: the dependent variables differ
  - the out-of-sample losses on a common target (section Forecasting: HAR vs GARCH) decide
- Bitcoin: daily and weekly components matter
  - the monthly coefficient is close to 0 (-0.022, SE 0.098)
  - Short sample (547 usable days out of about two years of data) and a strong weekly cycle: the week absorbs most of the memory
- Type of evidence: association (the variables move together; this does not prove causation)
  - an in-sample predictive fit; the out-of-sample losses decide

HARQ and SHAR

- **HARQ** (HAR model with Realised Quarticity) Bollerslev, Patton & Quaedvlieg (2016): yesterday's RV is measured with error, and the error attenuates the daily slope
  - $\mathrm{Var}(\mathrm{RV}_t - \mathrm{IV}_t) \approx 2\,\mathrm{IQ}_t / M$ (Appendix)
  - $\mathrm{RV}_{t+1} = \beta_0 + (\beta_d + \beta_Q \sqrt{\mathrm{RQ}_t})\,\mathrm{RV}_t + \beta_w \mathrm{RV}_t^{(w)} + \beta_m \mathrm{RV}_t^{(m)} + \varepsilon_{t+1}$
  - $\beta_Q < 0$: yesterday's RV gets less weight on days with high realised quarticity, that is, with large measurement error
- **SHAR** (Semivariance HAR model) Patton & Sheppard (2015): realised semivariances split RV by the sign of the return (Appendix)
  - $\mathrm{RS}_t^+ = \sum_i r_{t,i}^2 \mathbb{1}\{r_{t,i} > 0\}$, $\mathrm{RS}_t^- = \sum_i r_{t,i}^2 \mathbb{1}\{r_{t,i} < 0\}$, so $\mathrm{RS}_t^+ + \mathrm{RS}_t^- = \mathrm{RV}_t$
  - $\mathrm{RV}_{t+1} = \beta_0 + \beta^+ \mathrm{RS}_t^+ + \beta^- \mathrm{RS}_t^- + \beta_w \mathrm{RV}_t^{(w)} + \beta_m \mathrm{RV}_t^{(m)} + \varepsilon_{t+1}$
  - why: "bad" volatility, from negative returns, predicts more future volatility than "good" volatility, so $\beta^- > \beta^+$

Which Model Forecasts Realised Variance Best?

- Target: tomorrow's daily variance, forecast only with data available the evening before
  - SPY: total variance, 3 January 2022 to 18 September 2026 (1 180 days)
  - Bitcoin: 1 March 2025 onwards (422 days)
- Five models
  - log-HAR and HAR on realised variance from intraday data, re-estimated daily on an expanding window
  - GARCH(1,1) with Student-t errors on daily returns since 2000 (SPY) or 2014 (Bitcoin), re-estimated every 21 days (GARCH: Bollerslev (1986); Student-t errors: Bollerslev (1987))
  - EWMA (exponentially weighted moving average) on daily returns with decay factor $\lambda = 0.94$ (Chapter 5)
  - "yesterday's RV" (random walk: the forecast is the last observed RV)
- Models, inputs, history lengths and re-estimation frequencies all differ: the results compare these implemented procedures, not the HAR architecture as such

Evaluating Against a Noisy Target

- True variance is unobserved
  - we compare forecasts F_t with the proxy RV_t
  - A loss is **robust** if the ranking by *expected* loss with a proxy that is conditionally unbiased (given the forecasts' information) equals the ranking with true variance Patton (2011)
- MSE $= (\mathrm{RV}_t - F_t)^2$: robust
  - measures errors in variance units
  - is dominated by the most volatile days
- QLIKE $= \dfrac{\mathrm{RV}_t}{F_t} - \ln \dfrac{\mathrm{RV}_t}{F_t} - 1 \geq 0$: robust
  - depends only on the ratio
  - penalises under-prediction more
  - Where QLIKE comes from: for $r \sim N(0, F)$, $-2 \ln$ density is $\ln F + r^2/F$ plus a constant; replacing r^2 by RV_t and subtracting the minimum, reached at $F_t = \mathrm{RV}_t$, gives the formula above
- Common mistake: MAE, or MSE on volatility instead of variance, is **not** robust
  - MAE: mean absolute error
  - Andersen & Bollerslev (1998): against r_t^2, GARCH forecasts explained a few % of the variation; against RV the same forecasts explain far more: the proxy was noisy, not the model

Comparing Two Forecasts: Diebold–Mariano

- Diebold & Mariano (1995) (DM) test: $t = \bar{d}/\mathrm{se}_{\mathrm{NW}}(\bar{d}) \to N(0,1)$
  - loss difference: $d_t = L(\mathrm{RV}_t, F_t^A) - L(\mathrm{RV}_t, F_t^B)$
  - $L(\mathrm{RV}_t, F_t)$: the loss (QLIKE or MSE) of forecast F_t; $\bar{d}$: mean of d_t; $\mathrm{se}_{\mathrm{NW}}$: Newey–West standard error
  - $t < 0$: model A has the lower loss
  - HAC standard error because d_t is autocorrelated (Chapter 8)
- What drives t: the mean gap **and** how stable the gap is from day to day
  - Check, ignoring autocorrelation, $T = 100$ days: a gap with mean 0.01 and standard deviation 0.02 gives $t = 0.01/(0.02/\sqrt{100}) = 5$
  - a gap with mean 0.03 and standard deviation 0.30 gives $t = 1$
  - two forecasts that move together have a stable gap: a small mean difference can then be highly significant
- Conditions: a central limit theorem for d_t, with a positive long-run variance
  - QLIKE: $d_t = \mathrm{RV}_t(1/F_t^A - 1/F_t^B) + \ln(F_t^A/F_t^B)$, so the moments of d_t depend also on the forecasts in the denominators
  - nested models (HAR inside SHAR or HARQ) with estimated parameters: under equal accuracy d_t can degenerate, and $N(0,1)$ is only a conventional approximation (Clark & McCracken)
  - valid alternatives: rolling windows of fixed length with the Giacomini–White (GW) test, or a bootstrap that re-estimates the models

Calibration: Mincer–Zarnowitz

- Mincer & Zarnowitz (1969) (MZ): $\mathrm{RV}_t = a + b\,F_t + u_t$
  - regress the realised target on the forecast
  - a: intercept, b: slope, u_t: error
  - a calibrated forecast satisfies $a = 0$, $b = 1$ in this linear regression; $\mathbb{E}[\mathrm{RV}_t - F_t] = 0$ alone does not imply it
  - $b < 1$: high forecasts come with realised means below the 45° line
  - $b > 1$: high forecasts come with realised means above the 45° line
  - Check: a forecast of 2 on calm days and 4 on busy days, with realised means 1.5 and 4.5, has $b = (4.5 - 1.5)/(4 - 2) = 1.5 > 1$ and $a = 1.5 - 1.5 \times 2 = -1.5$
  - Joint Wald test with Newey–West covariance: needs u_t orthogonal to $(1, F_t)$ and a finite long-run covariance of $(1, F_t)\,u_t$
  - not rejecting $a = 0$, $b = 1$ does not prove calibration
  - R^2 measures how much of the variation the forecast explains
  - R^2 is capped by the noise in RV: even the true variance would not reach 1 Andersen & Bollerslev (1998)

Results: SPY, 2022–2026 (1/2)

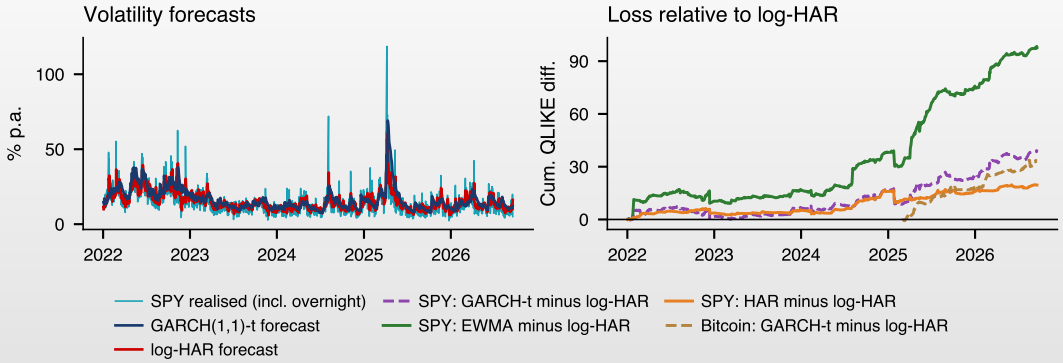
Model	QLIKE	$\mathrm{DM}_{\mathrm{QLIKE}}$ t (p)	MSE (%^{4})	$\mathrm{DM}_{\mathrm{MSE}}$ p	MZ b	MZ R^2
log-HAR	0.310	–	4.15	–	1.16	0.30
HAR	0.326	-1.95 (0.051)	4.48	0.117	0.75	0.26
GARCH(1,1)-t	0.343	-2.30 (0.021)	4.87	0.188	0.70	0.20
EWMA	0.393	-4.02 (< 0.001)	5.14	0.050	0.74	0.14
Yesterday's RV	0.502	-5.95 (< 0.001)	5.41	0.060	0.54	0.29

- 1 180 one-day-ahead forecasts of total daily variance
  - DM against log-HAR: $d_t = L(\text{log-HAR}) - L(\text{model})$ with L = QLIKE or MSE, Newey–West
  - $t < 0$ favours log-HAR
  - conditions for DM and MZ: on the loss differences and on the MZ errors (Diebold–Mariano slide), not on the moments of returns
- Of these procedures, log-HAR has the lowest QLIKE and MSE
  - QLIKE is 10% below GARCH(1,1)-t
  - Against EWMA and yesterday's RV the gain is highly significant; against HAR in levels only marginally
- MSE differences are not significant at 5% (EWMA borderline, $p = 0.050$)
  - a few crisis days (April 2025) dominate the squared errors

Results: SPY, 2022–2026 (2/2)

- Mincer–Zarnowitz: $a = 0, b = 1$ is not rejected for log-HAR ($p = 0.555$)
  - its R^2 is higher than GARCH's
  - log-HAR $b > 1$: high forecasts slightly below the realised means (0.66 standard errors from 1, SE 0.25)
  - GARCH $b < 1$: high forecasts slightly above the realised means, also not significant ($b = 0.70$, SE 0.24; joint MZ test: $p = 0.355$)
- Yesterday's RV has a high R^2 but $b = 0.54$
  - right direction; high forecasts far above the realised means
- Type of evidence: comparison of procedures (it shows which method works better on the same data, not why)
  - for Bitcoin (Further Topics), the gain of log-HAR over HAR in levels is not significant ($p = 0.223$)

Forecasts over Time (1/2)



MFM_ch9_forecasts ▶ Interactive

Forecasts over Time (2/2)

- Left: SPY realised volatility (with overnight) and the log-HAR and GARCH forecasts, annualised
- Right: cumulative QLIKE loss of each model minus log-HAR (rising = log-HAR better)
- After a shock both forecasts jump, but GARCH stays high longer
  - GARCH persistence is close to 1, while log-HAR returns faster
  - April 2025: GARCH above 40% p.a. until late April
  - realised volatility below 20% from mid-April
- The cumulative loss difference rises steadily, with jumps at turbulent episodes
  - the gain is not driven by one day
- Of the three models plotted against log-HAR, EWMA loses most
  - a fixed $\lambda = 0.94$ reacts too much to single days

MFM_ch9_forecasts

The Model Confidence Set

Model	QLIKE	MCS p-value	Eliminated at step	In MCS (90%)
log-HAR	0.310	1.000	–	✓
SHAR	0.320	0.223	–	✓
HAR	0.326	0.013	3	
HARQ	0.344	0.095	4–5	
GARCH(1,1)-t	0.343	0.095	4–5	
EWMA	0.393	0.012	2	
Yesterday's RV	0.502	< 0.001	1	

- Model confidence set (MCS) Hansen, Lunde & Nason (2011)
  - on the five models plus SHAR and HARQ
  - test equal predictive ability, drop the worst model, repeat
  - range statistic $T_R = \max_{i,j} |\bar{d}_{ij}| / \widehat{\mathrm{se}}(\bar{d}_{ij})$, with $\bar{d}_{ij}$ the mean loss gap between surviving models i and j
  - distribution of T_R: block bootstrap with 20-day blocks and 10 000 replications, so the smallest reportable p-value is of order 10^{-4}
  - as for DM, the bootstrap needs a central limit theorem for the QLIKE loss differences
  - SHAR and HARQ nest HAR: for these pairs the reference distribution is a conventional approximation
  - the surviving set contains the best model with asymptotic probability $\geq 90\%$
- SPY, 1 180 days from 2022: the 90% set is {log-HAR, SHAR}
- A multiple-comparison procedure
  - many pairwise DM tests without size control would overstate the evidence (Chapter 8)

▸ Further Topics: reading MCS p-values

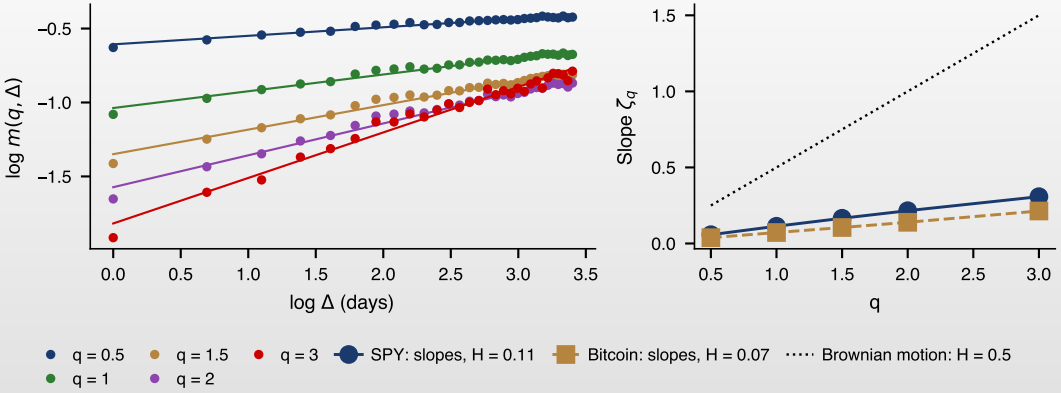
Is Volatility Rough?

- **Question**: are log-volatility paths rougher than those of a Brownian motion?
- Gatheral, Jaisson & Rosenbaum (2018): log volatility behaves like a **fractional Brownian motion**
  - Hurst exponent $H \approx 0.1$
  - $H = 0.5$: ordinary Brownian motion
  - $H < 0.5$: increments are negatively correlated, paths are rougher
  - $H > 0.5$: smoother, trending
- Rough models reproduce the shape of the implied volatility smile
  - for short maturities (Chapter 12)

Fractional Brownian Motion and the Scaling Law

- Fractional Brownian motion B_H, $0 < H < 1$: a Gaussian process with $B_H(0) = 0$ and mean 0
  - $\mathrm{Cov}(B_H(t), B_H(s)) = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right)$
  - $H = 1/2$: the covariance is $\min(t, s)$, ordinary Brownian motion
  - increments: $B_H(t+\Delta) - B_H(t) \sim N(0, \Delta^{2H})$ for every t
- Model $\ln \sigma_t = \text{const} + \nu B_H(t)$
  - the change over Δ days is $\nu \Delta^H Z$ in distribution, $Z \sim N(0,1)$
  - Scaling law: $m(q, \Delta) = \frac{1}{T}\sum_t |\ln \sigma_{t+\Delta} - \ln \sigma_t|^q$, with $\sigma_t = \sqrt{\mathrm{RV}_t}$, estimates $\nu^q \, \mathbb{E}|Z|^q \, \Delta^{qH}$, so the slope of $\ln m(q, \Delta)$ on $\ln \Delta$ is $\zeta_q = qH$
- Estimate: slopes ζ_q for $q \in \{0.5, 1, 1.5, 2, 3\}$
  - then regress ζ_q on q through the origin: $\hat{H} = \sum_q q\,\zeta_q / \sum_q q^2$
- Check with $q = 2$
  - for Brownian motion $m(2, \Delta) = \nu^2 \Delta$, so doubling Δ doubles m and $\zeta_2 = 1 = 2 \times 0.5$
  - with $H = 0.1$, doubling Δ multiplies $m(2, \Delta)$ only by $2^{0.2} \approx 1.15$

Scaling of Log Volatility Increments (1/2)



MFM_ch9_rough ▶ Interactive

Scaling of Log Volatility Increments (2/2)

- Left: $\ln m(q, \Delta)$ against $\ln \Delta$ for SPY, $\Delta = 1, \ldots, 30$ days
- Right: slopes ζ_q against q
- The log–log points lie close to straight lines and the slopes are close to linear in q
  - consistent with the scaling law, which the plot alone cannot establish
  - $\hat{H} = 0.11$ for SPY (block bootstrap 95% CI $[0.08; 0.13]$), 0.07 for Bitcoin: far below 0.5
  - Bitcoin uses only observed pairs at each calendar lag; interpolating the missing days gives 0.08
- The catch: RV measures volatility with error
  - the next slides show how measurement error lowers $\hat{H}$
- Type of evidence: description

MFM_ch9_rough

Roughness and Measurement Noise: the Identification Problem

- The estimate $H \approx 0.1$ comes from RV, a noisy proxy of the latent σ_t
  - Cont & Das (2024): the scaling estimator applied to realised volatility finds roughness even when spot volatility is smooth: an estimation artefact
  - Bolko, Christensen, Pakkanen & Veliyev (2023) (generalised method of moments, GMM, that models the RV error) and Fukasawa, Takabatake & Westphal (2022) (consistent as the sampling frequency grows) still find $H < 0.5$
- Formal question: $\ln \mathrm{RV}_t^{1/2} = \ln \sigma_t + e_t$
  - $\ln \sigma_t$ is fractional of index H, e_t is the measurement error of RV
  - test H_0: $H = 1/2$ (classical diffusive volatility) against H_1: $H < 1/2$
  - Identification problem: a small H and a large $\mathrm{Var}(e_t)$ both flatten $m(2, \Delta)$ at short lags (next slide)
- Short-run roughness can coexist with long memory Bennedsen, Lunde & Pakkanen (2022)
- Stakes: the short-maturity skew and volatility forecasts depend on H (Chapter 12)

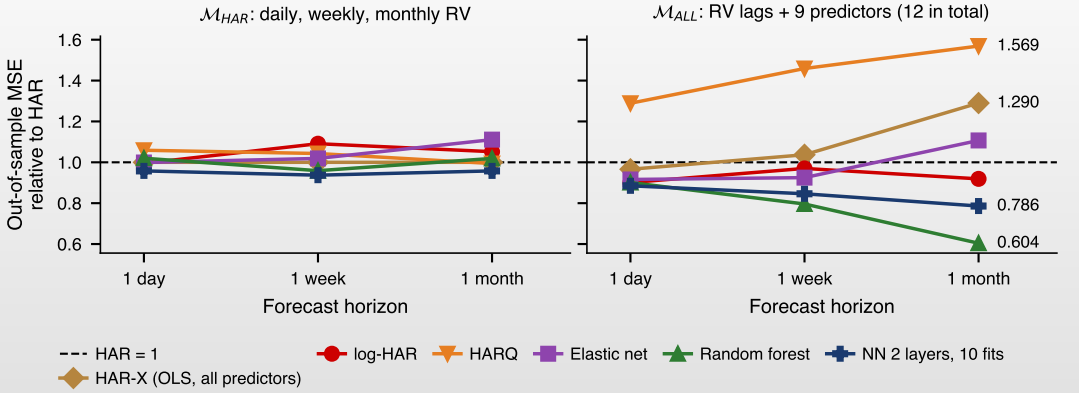
How Measurement Error Flattens the Scaling Law

- Observed increment = latent increment + error increment
  - $(\ln \sigma_{t+\Delta} - \ln \sigma_t) + (e_{t+\Delta} - e_t)$
  - if e_t is i.i.d. and independent of σ, the variances add: $\mathrm{Var}(e_{t+\Delta} - e_t) = 2\,\mathrm{Var}(e_t)$ for every $\Delta \geq 1$
  - hence $m(2, \Delta) = c\,\Delta^{2H} + k$, $c = \nu^2$, $k = 2\,\mathrm{Var}(e_t)$: a constant that does not grow with Δ
- Toy: $H = 0.5$, $c = 1$, $k = 1$
  - $m(2,1) = 1 + 1 = 2$, $m(2,2) = 2 + 1 = 3$: log–log slope $\ln(3/2)/\ln 2 = 0.58$, so $\hat{H} = 0.58/2 = 0.29$ instead of 0.5
  - at long lags the constant matters less: from $\Delta = 20$ to 40, slope $\ln(41/21)/\ln 2 = 0.97$, $\hat{H} = 0.48$
- Testable implication: under i.i.d. error, $\hat{H}$ rises when short lags are dropped
  - the mini-case in section AI for Scientific Discovery runs exactly this check on SPY

Case Study: Can Machine Learning Beat HAR?

- Christensen, Siggaard & Veliyev (2023): does off-the-shelf machine learning (ML) forecast daily realised variance better than the HAR family?
  - if so, why?
  - why ask: HAR Corsi (2009) is linear in three RV averages, while RV is right-skewed and persistent, so a nonlinear map, or predictors beyond past RV, may help
- Data: 29 of the 30 stocks of the Dow Jones Industrial Average (DJIA)
  - January 2001 – December 2017, $T = 4\,257$ days
  - 5-minute RV ($n = 78$)
- Information sets: $\mathcal{M}_{\text{HAR}}$ and $\mathcal{M}_{\text{ALL}}$
  - $\mathcal{M}_{\text{HAR}}$: daily, weekly, monthly RV
  - $\mathcal{M}_{\text{ALL}}$ adds 9 predictors: implied volatility, VIX, volume, momentum, macro and uncertainty indices

Case Study: Gains by Horizon (1/2)



MFM_ch9_ml_vs_har ▶ Interactive

Case Study: Gains by Horizon (2/2)

- Source: Christensen, Siggaard & Veliyev (2023)
  - out-of-sample MSE relative to HAR, averaged over 29 DJIA stocks (test set: the last 20% of 2001–2017)
  - left: RV lags only; right: all 12 predictors
  - HAR-X: OLS on all predictors
- RV lags only: the six models shown stay between 0.937 and 1.111 times the HAR MSE
  - at all three horizons
  - three RV averages leave little for a flexible model to learn
- All 12 predictors: the gain of trees and networks grows with the horizon
  - the authors attribute this to the higher persistence of the ML fits
  - random forest $0.901 \to 0.796 \to 0.604$
  - network with 2 hidden layers, averaged over 10 fits (NN_2^{10}): $0.885 \to 0.846 \to 0.786$
- Linear models with the same predictors deteriorate
  - HAR-X $0.966 \to 1.290$
  - HARQ $1.289 \to 1.569$
  - Among the linear models, only log-HAR stays below 1 at all horizons (0.901, 0.970, 0.919)
- Type of evidence: comparison of procedures
  - means of per-stock MSE ratios, not a significance test of the ranking (Further Topics)
  - the ranking depends on the split into training and test samples (Appendix)

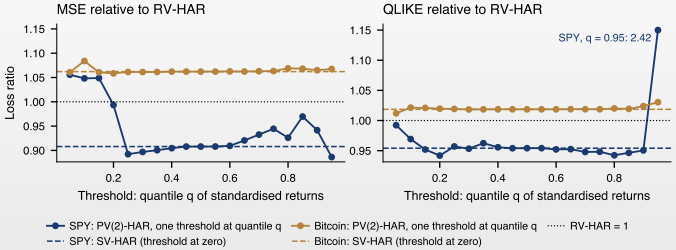
MFM_ch9_ml_vs_har

▸ Further Topics: ranking of 21 models

Case Study: Is Zero the Best Threshold?

- Bollerslev, Medeiros, Patton & Quaedvlieg (2022), *Journal of Econometrics*: is the sign of the return, i.e. the threshold 0 of semivariances, the best way to split yesterday's RV for forecasting?
  - SHAR Patton & Sheppard (2015) splits yesterday's RV into the negative and the positive semivariance
  - the threshold 0 has an economic motivation, loss aversion, but statistically it is arbitrary
- **Partial variance** (PV): the part of RV contributed by the returns between two thresholds
  - $\mathrm{PV}_t^{(g)} = \sum_i r_{t,i}^2 \, 1\{c_{g-1} < r_{t,i} \leq c_g\}$, with $c_0 = -\infty$ and $c_G = +\infty$, so $\sum_{g=1}^{G} \mathrm{PV}_t^{(g)} = \mathrm{RV}_t$
  - $G = 2$ and $c_1 = 0$: the semivariances; $G = 1$: RV
- **PV(G)-HAR**: in HAR, yesterday's RV is replaced by the G partial variances
  - the weekly and monthly components stay RV; in the paper's notation, SV-HAR (with realised semivariances) is SHAR and RV-HAR (with realised variance) is HAR
- The paper's claims, on SPY and 28 DJIA stocks, 1997–2013
  - with a single threshold, the threshold 0 is hard to beat
  - two thresholds, one in the left tail and one in the right, improve the forecasts significantly
  - for large covariance matrices, the threshold 0 remains the robust choice

Case Study: Loss as a Function of the Threshold, SPY and Bitcoin



- Mean out-of-sample loss of PV(2)-HAR divided by that of RV-HAR, for the threshold q on the horizontal axis
  - q: the quantile of the 5-minute returns divided by $\sqrt{\mathrm{RV}_t}$; below 1, better than RV-HAR
  - dashed line: SV-HAR, with the threshold exactly 0
- SPY, 1,180 days (Jan 2022 – Sep 2026): the MSE ratio stays between 0.892 and 0.909 for q from 0.25 to 0.60, around the SV-HAR value (0.908)
- Bitcoin, 422 days (Mar 2025 – Sep 2026): every threshold gives an MSE between 1.058 and 1.084 of that of RV-HAR

MFM_ch9_partial_variances ▶ Interactive

Case Study: Method and Replication on the Course Data

- The threshold as a quantile: $c_{t,g} = \sqrt{\mathrm{RV}_t}\ \hat{Q}(z;\, q_g)$, where $z = r_{t,i}/\sqrt{\mathrm{RV}_t}$
  - the same q follows the level of volatility; $q = 0.5$ gives practically the semivariances
- The paper: SPY, 4,202 days, forecasts on rolling windows of 1,000 days
  - PV(3)-HAR: the two thresholds, chosen by 5-fold cross-validation (CV) on the grid $q = 0.05, 0.10, \ldots, 0.95$
  - MSE and QLIKE losses; one-sided Diebold–Mariano (DM) tests
- The replication: the same construction on the 5-minute bars of the course
  - SPY: the target is intraday RV (09:30–16:00), as in the paper; the HAR slides of the chapter add the overnight return
  - Bitcoin: 24-hour UTC RV, components over 7 and 30 calendar days
  - expanding window, re-estimated daily, quantiles only from the days before the forecast
  - PV(3)-HAR: the pair of quantiles chosen by CV (MSE) every 21 days
  - a forecast below the smallest RV in the window is replaced by it (insanity filter)
- Differences: a sample almost 3 times shorter, no individual DJIA stocks, two-sided DM tests

Case Study: the Zero Threshold and Gains Concentrated in a Few Days

Loss relative to RV-HAR (MSE / QLIKE)	Paper: SPY 1997–2013	SPY 2022–2026	Bitcoin 2025–2026
SV-HAR (threshold 0)	0.967 / 0.981	0.908 / 0.954	1.062 / 1.019
PV(3)-HAR (two thresholds, CV)	0.950 / 0.971	0.815 / 1.013	1.145 / 1.061

- SPY: SV-HAR has a lower QLIKE than RV-HAR, DM $t = -1.79$ ($p = 0.073$); the paper (one-sided test): $p = 0.005$
  - over the whole sample, intraday-RV target: $\beta^- = 1.44$ (SE 0.63) and $\beta^+ = -0.34$ (SE 0.39); the Appendix (total-variance target) differs
- PV(3)-HAR does not beat SV-HAR on QLIKE: DM $t = +1.66$ ($p = 0.10$)
  - in the paper, PV(3)-HAR has smaller losses than SV-HAR: $p = 0.011$ (MSE) and 0.072 (QLIKE)
  - the chosen pair changes: the most frequent, (0.05, 0.95), in 15 of 57 selections; mean chosen quantiles 0.23 and 0.74 (the paper: 0.14 and 0.70)
- The 10 days with the largest RV bring 89% of the squared error of RV-HAR on SPY
  - without them, SV-HAR has an MSE of 1.240 and a QLIKE of 0.937 of those of RV-HAR: the MSE gain disappears

Case Study: What the Result Supports

- Type of evidence: a comparison of forecasting procedures on SPY
  - SV-HAR: the QLIKE gain is reproduced, not significant at 5%
  - PV(3)-HAR: the QLIKE gain is not reproduced
  - PV(3)-HAR: MSE falls by 18.5%, but the gain is concentrated on extreme days and is not significant in the conventional DM approximation ($t = -0.97$, $p = 0.33$)
- Inference for nested models: RV-HAR is SV-HAR or PV(3)-HAR with equal daily coefficients
  - the DM p-values above (expanding window) are the conventional approximation
  - rolling window of fixed length, 287 days (Giacomini–White, GW): against RV-HAR, MSE has $p = 0.34$ for SV-HAR and $p = 0.35$ for PV(3)-HAR
  - with the same window, QLIKE is higher than for RV-HAR: the ratio is 1.70 for SV-HAR and 3.23 for PV(3)-HAR
  - the tests need a central limit theorem for the loss differences; with QLIKE their moments depend also on the forecasts in the denominators
- Limits: forecasts over 4.7 years for SPY and 1.5 years for Bitcoin; MSE dominated by a few extreme days

The Discovery Loop on the Roughness Question (1/2)

- Question: is $\hat{H} \approx 0.1$ a property of σ_t or of the RV error e_t?
- One pass of the loop with an AI assistant (reusable prompts in the Appendix)
  - Literature: tabulate the published estimates of H (estimator, frequency, treatment of the RV error)
  - check every DOI (digital object identifier) on Crossref
  - Competing hypotheses, each with its shape of $m(2, \Delta)$: fractional with $H = 0.1$; $H = 0.5$ plus i.i.d. error; two factors, one fast mean-reverting
  - Replicate the chapter's $\hat{H} = 0.11$ before extending
  - fix lag windows, measures and the fit with a constant noise term k in advance

The Discovery Loop on the Roughness Question (2/2)

- Checks the researcher owns
  - Proxy error: simulate $H = 0.5$ plus error of the same size; the procedure must not return $H \approx 0.1$
  - Overlapping increments invalidate OLS standard errors of ζ_q: use a block bootstrap
  - report every window tried
  - Assistants quote "$H \approx 0.1$ for all assets": check each number against the cited table

Mini-Case: Is $\hat{H}$ an Artefact of i.i.d. Error? (1/2)

- A typical assistant proposal: $\hat{H} \approx 0.1$ comes from i.i.d. measurement error in RV
  - testable implication: $\hat{H}$ **rises** when short lags are dropped

	$\Delta \in [1, 30]$	$[5, 30]$	$[10, 50]$	Fit with k, $[1, 50]$
Simulated $H = 0.5$ plus error	0.18	0.24	0.26	0.39
SPY, total RV	0.106	0.089	0.077	0.096
SPY, intraday RV	0.130	0.094	0.078	0.114

- First column replicates the chapter
- The simulated row shows the mechanism: $\hat{H}$ rises
  - the fit recovers most of the gap to 0.5
  - one path, not a power study

Mini-Case: Is $\hat{H}$ an Artefact of i.i.d. Error? (2/2)

- SPY: $\hat{H}$ **falls** with the window and the fitted k is 0
  - fit with k: $\hat{H} = 0.096$, block bootstrap 95% CI $[0.058; 0.268]$
- Interpretation: an illustrative diagnostic, inconsistent with the i.i.d.-error explanation
  - a formal rejection needs a simulation-calibrated test
  - the non-i.i.d. mechanism of Cont & Das (2024) is not tested by this design
- Type of evidence: description

MFM_ch9_ai_discovery_case

Your Turn: A Project Seed

- Question: does the roughness of SPY and Bitcoin volatility survive an estimator that models the RV measurement error?
- Replicate first (Seminar 9, B9)
  - $\hat{H} = 0.11$ (SPY) and 0.07 (Bitcoin) from the scaling law
  - $\hat{H} = 0.18$ for simulated $H = 0.5$ plus error
- Deliverables
  - A pre-registration (hypotheses, estimators, lag windows, tests, simulation design) dated before the first run
  - The GMM estimator of Bolko, Christensen, Pakkanen & Veliyev (2023) on the course RV series, with Monte Carlo size and power at $H \in \{0.1, 0.5\}$
  - Code, a two-page report with every specification tried, and a referee-style critique of your own result
- Log: github.com/danpele/MFM/tree/main/project
  - every prompt that shaped a design choice in `AI_USE.md`
  - every error the assistant made in `AI_ERRORS.md`

Which Tool When?

Situation	Suggested tool
Liquid asset, 1–5 minute bars	RV at 5 minutes, subsampled
Tick data, or an illiquid stock (Bucharest Stock Exchange, BVB)	Pre-averaging or flat-top realised kernel ($n^{-1/4}$; Parzen $n^{-1/5}$; Appendix: rates of convergence); check the signature plot
Daily risk of an exchange-traded asset	Total variance: RV plus squared overnight return
Separate news jumps from diffusion	BV or MedRV (median realised variance; Appendix: jump-robust estimators), ratio test with FDR control; Lee–Mykland for the jump time
Daily or weekly variance forecast	log-HAR (plus jumps, weekends, RQ); benchmark GARCH
Comparing forecasts	QLIKE and MSE on variance, Diebold–Mariano, Mincer–Zarnowitz, model confidence set

- Strongest result: out of sample, log-HAR has a lower QLIKE than GARCH(1,1)-t in both markets
  - MSE comparisons are inconclusive
- The 5-minute rule is justified only if the signature plot is flat at 5 minutes and coarser
- The jump counts depend on FDR control and on the periodicity correction

Check Yourself, and Next: Chapter 10

Retrieval questions

- Why is R^2 of log RV not comparable with R^2 of RV in levels?
  - slide *Reading the HAR Estimates*
- Why is QLIKE robust to a noisy RV proxy while MSE on volatility is not?
  - slide *Evaluating Against a Noisy Target*
- Why does the MCS state a set of models rather than a winner?
  - slide *The Model Confidence Set*
- How can measurement error in RV make volatility look rough?
  - slide *How Measurement Error Flattens the Scaling Law*

Chapter 10: market microstructure

- Where does the noise ε of this chapter come from
  - limit order books, bid–ask spreads, tick sizes
- Liquidity, market impact and high-frequency trading, on equity and crypto exchanges

The Practical Decision

- **What I choose**
  - for daily risk, total variance: 5-minute RV plus the squared overnight return (the night brings 38% of SPY's daily variance)
  - for tomorrow's variance, log-HAR: QLIKE 0.310 against 0.343 for GARCH(1,1)-t on SPY; it stays in the 90% set, with SHAR
  - for jump days, the ratio test with the periodicity correction and FDR control: 17 days of 1 489 for SPY, not 236
- **Under which conditions**
  - a liquid asset whose signature plot is flat from 5 minutes up; tick data or an illiquid stock need pre-averaging or a realised kernel
  - the ranking holds for these implemented procedures and for the QLIKE loss; MSE comparisons are inconclusive
  - the DM and MCS p-values need a central limit theorem for the QLIKE loss differences; for nested pairs (SHAR, HARQ against HAR) they are a conventional approximation
- **What result would make me drop it**
  - the 5-minute RV: a signature plot that rises as the interval shrinks to 5 minutes, the fingerprint of noise
  - log-HAR: on a new sample, a DM test on QLIKE that no longer favours it over GARCH (here $t = -2.30$, $p = 0.021$), or its exit from the 90% set

Data: SPY and Bitcoin Intraday Series

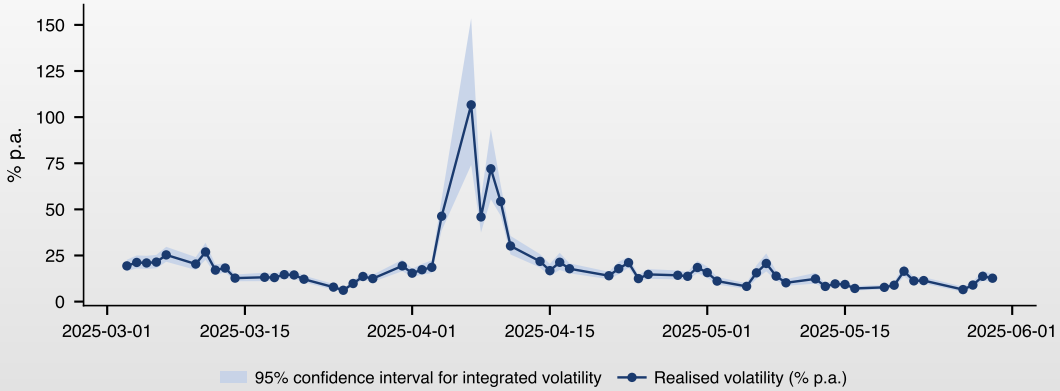
Series	Frequency	Period	Days
SPY (ETF)	5-minute bars, 09:30–16:00 New York time	2020–2026	1 489
Bitcoin	5-minute bars, 24-hour days (UTC)	2024–2026	659
SPY, Bitcoin	daily closes	2000/2014–2026	for GARCH
VIX	daily close	2020–2026	implied volatility

- SPY: 78 returns per full day, 42 (09:30–13:00) on the 12 early-close days
  - the bar labelled 13:00 starts after the close: it is not an interval; its first print is the 13:00 price only when the 12:55 bar is missing
  - bars off the 5-minute grid removed
  - missing bars inside a day carry the last price: 30 bars on 16 days
- Bitcoin: 288 returns per day
  - a missing bar carries the last price, so its interval merges with the next: 1 347 bars on 245 of 659 days
  - the day starts at the 00:00 price (the open of the first bar or the previous 23:55 close) and ends at the last observed price
  - without the incomplete days, annualised volatility: SPY 12.97% instead of 12.93%, Bitcoin 46.17% instead of 45.08%
  - days with fewer than 95% of the bars dropped: 88 of 747 calendar days, mostly a data gap in October 2025

Adding Up Day and Night Variance

- SPY, annualised variances (%^{2} per year)
  - intraday RV: 12.9%, variance 167
  - total variance $\mathrm{RV}_t + r^2_{\text{overnight},t}$: 16.4%, variance 270
  - overnight part: $270 - 167 = 103$, i.e. $\sqrt{103} \approx 10.1\%$; share $103/270 = 38\%$
- Why not $13.6^2 + 10.1^2$?
  - the open-to-close 13.6% squares the whole session return, not each 5-minute return
  - $(\sum_i r_{t,i})^2 = \mathrm{RV}_t + 2\sum_{i<j} r_{t,i}\, r_{t,j}$: the cross products are zero only in expectation; in this sample they add about 17 ($13.6^2 \approx 184$)
  - close-to-close (16.6%) also contains the cross term $2\, r_{\text{overnight},t}\, r_{\text{open-close},t}$

Confidence Intervals for Daily Volatility (1/2)



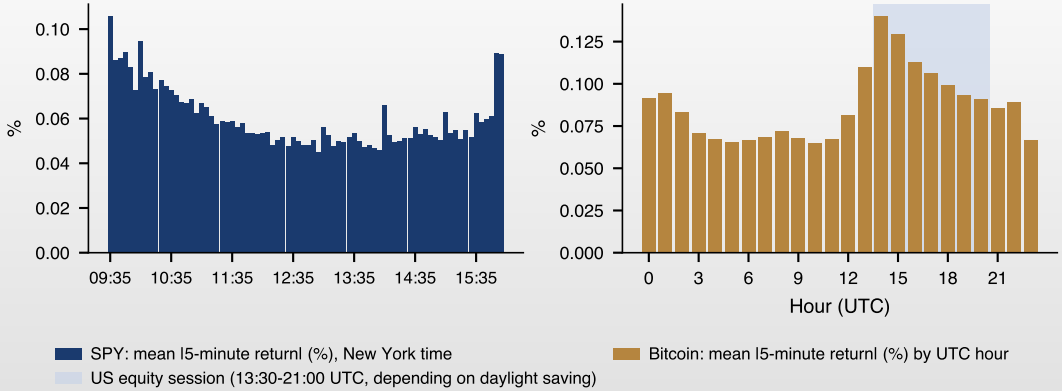
MFM_ch9_rv_ci

Confidence Intervals for Daily Volatility (2/2)

- SPY, March–May 2025
  - band: 95% confidence interval for $\sqrt{252\,\mathrm{IV}_t}$ from realised quarticity
- The volatility band takes the square root of the variance endpoints
  - $\sqrt{\mathrm{RV}_t}\exp\left(\pm 0.98\,\mathrm{se}(\mathrm{RV}_t)/\mathrm{RV}_t\right)$: the log half-width is halved
- SPY: a single day's volatility is known to about $\pm 20\%$
  - median $\mathrm{se}/\mathrm{RV} = 0.19$, so the volatility factors are $[0.83, 1.20]$
  - median relative width: 37%
- 9 April 2025: volatility 72%, CI $[55\%, 93\%]$
  - the day the tariff pause was announced
  - The interval widens when a few returns dominate the sum: $\sum r^4$ grows faster than $(\sum r^2)^2$
- The band assumes a continuous price and negligible noise
  - on a jump day, $\sum r^4$ keeps the term κ^4
  - so the band has no nominal coverage for either the total variation or IV_t (section Jumps)
- RV is still far more precise than r_t^2
  - the relative error of r_t^2 is of order $\sqrt{2} \approx 141\%$ under the Normal distribution

MFM_ch9_rv_ci

Intraday Seasonality (1/2)



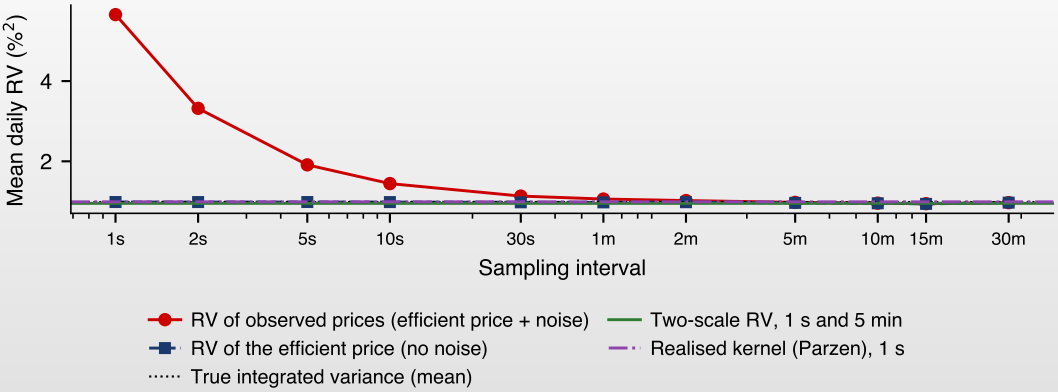
MFM_ch9_seasonality

Intraday Seasonality (2/2)

- Left: SPY, mean $|r_{t,i}|$ by 5-minute interval of the day
  - full days only
  - ticks mark the end of the interval, New York time
- Right: Bitcoin by UTC hour
- SPY: a **U-shape** Andersen & Bollerslev (1997)
  - 0.106% in the first 5 minutes, 0.050% around midday, 0.089% in the last 5 minutes
  - The opening auction incorporates overnight news; the close concentrates index and fund rebalancing
- Bitcoin trades 24/7 but is most volatile at 14:00 UTC
  - 14:00 UTC: near the US equity opening (09:30 New York: 13:30 UTC in summer, 14:30 UTC in winter)
  - quietest at 10:00 UTC (ratio 2.2)
- RV sums over the day, so seasonality does not bias it
  - it does bias bipower variation (BV, defined in section Jumps) and makes the jump test over-reject
  - the remedy, dividing returns by a periodicity factor f_i Boudt, Croux & Laurent (2011), is applied to the jump test in section Jumps (Appendix)

MFM_ch9_seasonality

The Volatility Signature Plot: Simulation (1/2)



MFM_ch9_noise_simulation

The Volatility Signature Plot: Simulation (2/2)

- A signature plot shows mean RV against the sampling interval
  - here: 100 simulated days, one price per second
  - efficient price with random daily volatility plus i.i.d. noise, $\omega = 0.01\%$
- True mean IV $= 0.98$; RV of observed prices:
  - 5.66 at 1 second, 1.06 at 1 minute, 0.98 at 5 minutes
  - The bias matches $2n\omega^2$ (slide *Why Not Use Every Tick?*)
- Without noise, RV is flat in the sampling interval
  - the rise at high frequency is the fingerprint of noise
- Noise-robust estimators on 1-second data: two-scale RV 0.95, realised kernel 0.99
  - error standard deviation across days: 5-minute RV 0.187, two-scale 0.150, kernel 0.045
  - using all data pays off only with a noise correction

MFM_ch9_noise_simulation

Subsampling and Two-Scale RV

- **Subsampling**: compute RV on K shifted grids and average
  - offsets $0, 1, \ldots, K-1$; $\mathrm{RV}_t^{(k)}$: RV on the grid with offset k
  - $\mathrm{RV}_t^{\mathsf{avg}} = \frac{1}{K}\sum_{k=0}^{K-1} \mathrm{RV}_t^{(k)}$: same bias as one sparse grid, smaller variance
- **Two-scale RV** (TSRV) Zhang, Mykland & Aït-Sahalia (2005): remove the remaining bias with the all-data RV
  - $\mathrm{TSRV}_t = \mathrm{RV}_t^{\mathsf{avg}} - \frac{\bar{n}}{n}\mathrm{RV}_t^{\mathsf{all}}$, $\bar{n} = K^{-1}\sum_k n_k$, n_k: returns on grid k; $\mathrm{RV}_t^{\mathsf{all}}$: RV on all n returns, which estimates the noise term $2n\omega^2$
  - standard grids (grid k uses the prices $k, k+K, \ldots$): $\bar{n} = (n-K+1)/K$, e.g. $73/6 = 12.17$ for $n = 78$, $K = 6$
  - adding the day's first and last price to every grid gives $83/6 = 13.83$, but the common end-point noise then does not vanish: a finite-sample approximation, not the consistent TSRV
  - In small samples, divide TSRV by $1 - \bar{n}/n$

Realised Kernels



Neil Shephard, 2004

Photo: Renate Schmid (2004); CC BY-SA 2.0 de; Wikimedia Commons

- **Realised kernel** Barndorff-Nielsen, Hansen, Lunde & Shephard (2008): add weighted autocovariances of returns
  - $\mathrm{RK}_t = \gamma_0 + 2\sum_{h=1}^{H} k\left(\frac{h}{H+1}\right)\gamma_h$, $\gamma_h = \sum_i r_{t,i}\, r_{t,i-h}$
  - Parzen weight function k: 1 at 0, falling smoothly to 0 at 1; keeps RK (Realised Kernel) positive
- Noise makes γ_1 negative
  - the kernel adds it back and cancels the bias
- Bandwidth H (here not the Hurst exponent) grows with the noise-to-signal ratio
  - $H \approx 3.5\,\xi^{4/5} n^{3/5}$, with $\xi^2 = \omega^2/\sqrt{\mathrm{IQ}}$ the noise-to-signal ratio
  - course code (a heuristic pilot): $\hat{\omega}^2 = \mathrm{RV}/(2n)$, and the 20-minute RV in place of $\sqrt{\mathrm{IQ}}$
- Jittering: the first and last prices are replaced by the averages of the first and last two prices, so the end-point noise does not stay in RK

The Benjamini–Hochberg Rule on Five Tests

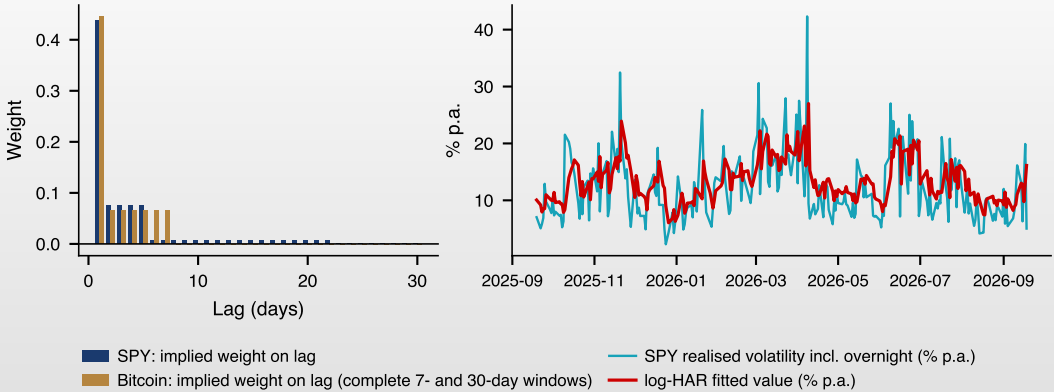
- Five days, sorted p-values $0.003, 0.012, 0.024, 0.045, 0.60$
  - with $q = 0.05$ and $N = 5$ the thresholds kq/N are $0.01, 0.02, 0.03, 0.04, 0.05$
- The first three pass, the fourth fails: $k = 3$, three days rejected
  - $0.024 \leq 0.03$, but $0.045 > 0.04$
- Same data, number of rejections
  - no correction ($p \leq 0.05$): 4
  - BH: 3
  - Bonferroni ($p \leq 0.05/5 = 0.01$): 1
- The procedure is step-up
  - with p-values $0.003, 0.012, 0.032, 0.038, 0.60$, the third fails its own threshold ($0.032 > 0.03$)
  - but the fourth passes ($0.038 \leq 0.04$), so $k = 4$ and four days are rejected

Implied Volatility as a Variance Forecast

- As a forecast: regress $\frac{252}{21}\mathrm{RV}_{t,t+21}$ on VIX_t^2
  - $\frac{252}{21}\mathrm{RV}_{t,t+21}$: total variance, annualised $\%^2$
  - VIX_t^2: annualised $\%^2$; approximations: 30 calendar days $\approx$ 21 trading days, S&P 500 index options (SPX, Cboe) against SPY
  - slope $0.56 < 1$, standard error (SE) 0.06: Newey–West with 20 lags, since the 21-day targets overlap; 0.05 with 7 lags, 0.06 with 42
  - the slope below 1 does not identify an over-reaction of the market or the premium mechanism
  - $R^2 = 0.23$: VIX is informative
  - the level is a separate question: the mean ex-post gap is 4.36 percentage points in volatility and 138.2 in variance (annualised $\%^2$)
  - neither gap identifies the conditional premium: both add the forecast error $\mathbb{E}_t^{\mathbb{P}}[Q_t] - Q_t$
  - the same slope-versus-level distinction returns in the Mincer–Zarnowitz test

MFM_ch9_vix

Implied Lag Weights and In-Sample Fit (1/2)



MFM_ch9_har

Implied Lag Weights and In-Sample Fit (2/2)

- Left: weights of log-HAR on each past day
  - Bitcoin: complete 7- and 30-day windows
- Right: SPY, last 250 days, realised volatility and log-HAR fitted values
- SPY, weights by lag
  - lag 1: $\hat{\beta}_d + \hat{\beta}_w/5 + \hat{\beta}_m/22 = 0.362 + 0.067 + 0.008 = 0.437$
  - each of lags 2–5: $0.067 + 0.008 = 0.075$
  - each of lags 6–22: 0.008
  - Three coefficients mimic the slow decay of the autocorrelation function
- Bitcoin: yesterday 0.446, days 2–7 about 0.066 each, almost nothing beyond a week
- In-sample fitted values follow the level of volatility, not the day-to-day noise
- HAR is an AR(22) with 19 restrictions: $\phi_2 = \ldots = \phi_5$, $\phi_6 = \ldots = \phi_{22}$
  - Newey–West Wald test on SPY: $W = 18.2$, $p = 0.51$
  - not rejected (derivation in the Appendix)

MFM_ch9_har

Extensions of HAR

- **HAR-CJ** (HAR model with Continuous and Jump components) Andersen, Bollerslev & Diebold (2007): separate continuous (C) and jump (J) parts as regressors
  - $\ln \mathrm{RV}_{t+1} = \beta_0 + \beta_{Cd} \ln C_t + \beta_{Cw} \ln C_t^{(w)} + \beta_{Cm} \ln C_t^{(m)}$
  - $\quad + \beta_{Jd} \ln(1 + J_t) + \beta_{Jw} \ln(1 + J_t^{(w)}) + \beta_{Jm} \ln(1 + J_t^{(m)}) + \varepsilon_{t+1}$
  - $C_t^{(w)}, J_t^{(w)}$ ($C_t^{(m)}, J_t^{(m)}$): arithmetic means over the last 5 (22) days, as $\mathrm{RV}_t^{(w)}$; logs are taken after averaging
- Weekend or holiday dummies
- Leverage effect: negative returns raise future RV
- Implied volatility as an extra regressor
- Multi-step: regress $\sum_{j=1}^{h} \mathrm{RV}_{t+j}$ directly on the same regressors

Results: Bitcoin, 2025–2026 (1/2)

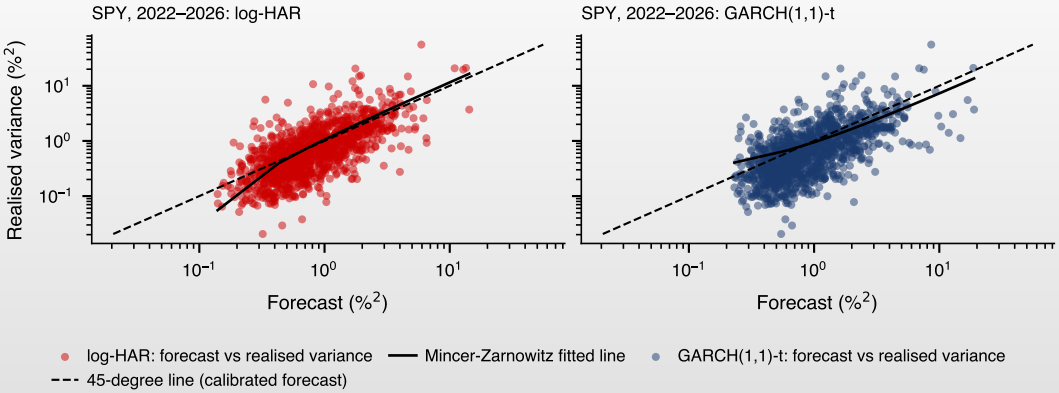
Model	QLIKE	DM_{QLIKE} t (p)	MSE ($\%^4$)	DM_{MSE} p	MZ b	MZ R^2
log-HAR	0.282	–	26.6	–	0.82	0.38
HAR	0.408	-1.22 (0.223)	27.6	0.431	0.89	0.34
GARCH(1,1)-t	0.360	-4.32 (< 0.001)	42.2	0.055	0.55	0.23
EWMA	0.346	-2.96 (0.003)	36.3	0.033	0.64	0.18
Yesterday's RV	0.466	-5.38 (< 0.001)	35.0	0.061	0.58	0.33

- 422 one-day-ahead forecasts of 24-hour realised variance
  - same models and tests as for SPY
- log-HAR again has the lowest QLIKE: 22% below GARCH(1,1)-t
  - against HAR in levels the gap is not significant ($p = 0.223$)
  - GARCH on daily data is miscalibrated: MZ $b = 0.55$ ($p < 0.001$): high forecasts come with lower realised means

Results: Bitcoin, 2025–2026 (2/2)

- HAR in levels is worse than GARCH under QLIKE (0.408 vs 0.360)
  - after calm weekends it predicts too low a variance
  - QLIKE punishes under-prediction
- Under MSE only EWMA is significantly worse than log-HAR at 5% ($p = 0.033$)
  - GARCH: $p = 0.055$
- Bitcoin and SPY MSE are not comparable
  - MSE is in $(\%^2)^2 = \%^4$ and grows with the square of the variance level
  - QLIKE depends only on RV_t / F_t and is comparable across markets

Mincer–Zarnowitz Regressions (1/2)



MFM_ch9_forecasts

Mincer–Zarnowitz Regressions (2/2)

- SPY, 2022–2026: realised variance against the forecast (log scales)
  - dashed: 45-degree line
  - black: fitted regression in levels
- log-HAR forecasts span a wider range
  - they separate calm from turbulent days better
- On calm days GARCH overstates the variance
  - GARCH forecasts rarely fall below $0.26\%^2$ per day (1% of days)
  - log-HAR goes down to $0.14\%^2$
- The cloud is wide for both
  - one day's variance is hard to predict
  - RV itself is measured with error

MFM_ch9_forecasts

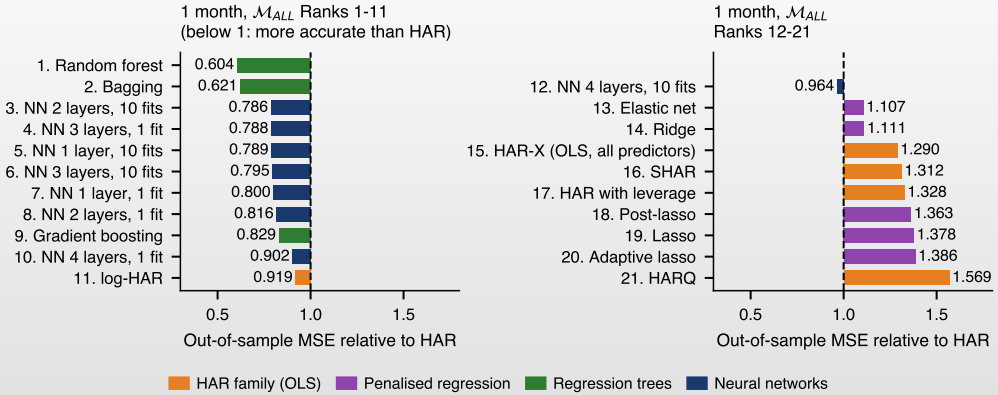
Reading the Model Confidence Set

- An MCS p-value is a running maximum
  - if a model is eliminated at step k, its MCS p-value is $\max_{j \le k} p_j$
  - p_j: the p-value of the equal-ability test at step j
  - so the MCS p-values rise along the elimination order, which the table reads off
  - equal values (GARCH, HARQ) mean the later model's own step p-value was not larger, so their relative order is not identified from the table
  - a model stays in the 90% set if its MCS p-value exceeds 0.10
- Why is HAR (QLIKE 0.326) dropped before GARCH (0.343)?
  - each step eliminates the model with the largest **standardised** loss gap to the other survivors
  - not the largest mean loss
  - SHAR differs from HAR only in its daily term, so their losses move together and the gap is very stable: DM SHAR vs HAR $t = -3.39$, $p < 0.001$
  - GARCH loses more on average, but its gaps are noisier: against log-HAR $t = -2.30$, smaller in size than the gap that separates HAR from SHAR (as in the check on the Diebold–Mariano slide)
  - the DM table compares each model only with log-HAR, so it does not show the HAR–SHAR pair, the likely reason HAR goes early

Case Study: Can Machine Learning Beat HAR?

- Models, with minimal tuning
  - the HAR family by OLS and penalised regressions
  - tree ensembles (bagging, random forest, boosting) and neural networks
  - full catalogue in the Appendix
- Evaluation: out-of-sample MSE relative to HAR
  - per stock and then averaged
  - DM tests per stock and MCS at 90%

Case Study: Ranking of 21 Models, One Month Ahead (1/2)



MFM_ch9_ml_vs_har

Case Study: Ranking of 21 Models, One Month Ahead (2/2)

- Source: Christensen, Siggaard & Veliyev (2023)
  - MSE relative to HAR, one month ahead, all 12 predictors, mean over 29 DJIA stocks
  - NN_k^e: k hidden layers, e networks averaged
- Tree ensembles lead: random forest 0.604, bagging 0.621
  - all eight networks beat HAR (0.786–0.964), but depth does not help
- Penalised regressions and linear HAR extensions lose to HAR (1.107–1.569)
- Each bar averages per-stock ratios: not a significance test of the ranking

MFM_ch9_ml_vs_har

Itô Semimartingales and Stable Convergence

- Itô semimartingale: $dp = \mu\, ds + \sigma\, dW + dJ$
  - μ, σ adapted (known from the past at each time)
  - σ càdlàg (right-continuous with left limits, so it may jump)
  - σ may depend on W (leverage)
  - Without jumps: $\sqrt{M}(\mathrm{RV}_t - \mathrm{IV}_t) \xrightarrow{\mathcal{L}\text{-}s} \mathrm{MN}(0, 2\,\mathrm{IQ}_t)$, a **mixed** Normal distribution whose variance $2\,\mathrm{IQ}_t$, with integrated quarticity IQ, $\mathrm{IQ}_t = \int \sigma^4(s)\, ds$, is itself random Barndorff-Nielsen & Shephard (2002)
- **Stable** convergence ($\mathcal{L}$-s) Mykland & Zhang (2009)
  - the limit holds jointly with every variable of the path's σ-field
  - Hence $(\mathrm{RV}_t - \mathrm{IV}_t)/\sqrt{2\,\mathrm{RQ}_t/M} \to N(0,1)$: studentising by $\sqrt{2\,\mathrm{RQ}_t/M}$, with $\mathrm{RQ}_t \to \mathrm{IQ}_t$ random, is legitimate only under stable convergence
- Finite samples: the t-statistic for $\ln \mathrm{RV}_t$ has better coverage than the one in levels Barndorff-Nielsen & Shephard (2005)
  - the confidence intervals of the lecture use it

Back

Deriving the Noise Bias and the Optimal Frequency

- Observed return $r_i = r_i^* + \varepsilon_i - \varepsilon_{i-1}$, with ε i.i.d. $(0, \omega^2)$ and independent of p^*
  - Given the path of σ (zero drift, no leverage: σ independent of W):
    $\mathbb{E}[\mathrm{RV}^{(n)} \mid \sigma] = \sum_i \mathbb{E}[r_i^{*2} \mid \sigma] + \sum_i \mathbb{E}[(\varepsilon_i - \varepsilon_{i-1})^2] = \mathrm{IV} + 2n\omega^2$
  - $\mathrm{cov}(r_i, r_{i-1}) = -\mathbb{E}[\varepsilon_{i-1}^2] = -\omega^2$
  - By a law of large numbers for the noise terms, $\mathrm{RV}^{(n)}/(2n) \xrightarrow{p} \omega^2$ as $n \to \infty$: the all-tick RV estimates the noise, not IV
    Bandi & Russell (2008)
- MSE: $\mathbb{E}[(\mathrm{RV}^{(n)} - \mathrm{IV})^2] \approx \underbrace{2\,\mathrm{IQ}/n}_{\text{discretisation}} + \underbrace{4n^2\omega^4}_{\text{squared bias}} + O(n)$
  - First-order condition $-2\,\mathrm{IQ}/n^2 + 8n\omega^4 = 0$ gives $n^* = (\mathrm{IQ}/(4\omega^4))^{1/3}$ Bandi & Russell (2008)

Back

Intraday Seasonality and the Jump Test

- Seasonality biases BV (not RV)
  - $\mathbb{E}|r_i|\,|r_{i-1}| \propto \sigma_i \sigma_{i-1} \leq (\sigma_i^2 + \sigma_{i-1}^2)/2$
  - Fast changes of scale (open, close) push $\mathrm{RV} - \mathrm{BV}$ up, so the ratio test (section Jumps) over-rejects
- Periodicity factor f_i: standard-deviation estimator of Boudt, Croux & Laurent (2011)
  - scale each return by $\sqrt{\mathrm{BV}_t / M_t}$
  - take the root mean square by interval of the day
  - normalise the mean of f_i^2 to 1
  - SPY: $f = 1.78$ at the open, 0.73 at the minimum, 1.52 at the close
- Dividing $r_{t,i}$ by f_i before the test, at the 0.1% level 21 SPY rejections remain
  - for the 5% level: main text

Back

Rates of Convergence under Noise

Estimator	Error of order	Positive	Source
RV on all ticks	inconsistent (bias $2n\omega^2$)	yes	Bandi & Russell (2008)
Sparse RV, fixed 5-minute grid	$M^{-1/2}$, M fixed; bias stays	yes	Liu, Patton & Sheppard (2015)
Two-scale RV	$n^{-1/6}$	no	Zhang, Mykland & Aït-Sahalia (2005)
Multi-scale RV	$n^{-1/4}$	no	Zhang (2006)
Realised kernel, flat-top / Parzen	$n^{-1/4}$ / $n^{-1/5}$	no / yes	Barndorff-Nielsen, Hansen, Lunde & Shephard (2008)
Pre-averaging	$n^{-1/4}$	no	Jacod, Li, Mykland, Podolskij & Vetter (2009)

- $n^{-1/4}$ is the optimal rate with i.i.d. noise
  - the price of noise is half the usual exponent
- With $n = 23\,400$ seconds: $n^{1/4} \approx 12.4$, against $\sqrt{78} \approx 8.8$ for 5-minute RV
  - a heuristic only: rates ignore constants, noise size and the bias of sparse RV
  - the rates hold for the theoretical estimators: standard TSRV grids, kernels with jittered end points
  - consistent with the modest gains in Liu, Patton & Sheppard (2015)

Back

Pre-Averaging (1/2)

- Local averages of $k_n \approx \theta\sqrt{n}$ returns with a weight g on $[0,1]$ Jacod, Li, Mykland, Podolskij & Vetter (2009)
  - θ: a tuning constant, not the jump-test θ
  - e.g. $g(x) = \min(x, 1-x)$
  - $\bar{r}_i = \sum_{j=1}^{k_n - 1} g(j/k_n)\, r_{i+j}$: noise variance shrinks like $1/k_n$, the signal keeps order k_n/n
- Estimator, with $\psi_1 = \int_0^1 g'(x)^2 dx$ and $\psi_2 = \int_0^1 g(x)^2 dx$
  - $\psi_1 = 1$ and $\psi_2 = 1/12$ for the triangular g; g': the derivative of g
  - $\widehat{\mathrm{IV}} = \frac{1}{\theta\, \psi_2 \sqrt{n}} \sum_i \bar{r}_i^2 - \frac{\psi_1}{2\,\theta^2 \psi_2\, n} \sum_i r_i^2$ (second term: noise bias correction)

Back

Pre-Averaging (2/2)

- Rate $n^{-1/4}$
  - the same device gives jump-robust (pre-averaged bipower) and covariance estimators
- Built for tick data
  - on 5-minute bars little noise is left to average out (empirical signature plot, section Microstructure Noise)

Back

Realised Covariance and Asynchronous Trading

- Realised covariance $\sum_i r^{(1)}_{t,i} r^{(2)}_{t,i} \to \int \sigma_{12}(s)\, ds$ needs prices of both assets on the **same** grid
  - $r^{(1)}, r^{(2)}$: the returns of assets 1 and 2, $\sigma_{12}(s)$: their spot covariance
  - Epps effect Epps (1979): measured correlation falls towards 0 as sampling gets finer, because trades are asynchronous and the last price is stale
- Hayashi & Yoshida (2005): sum the products of **all** pairs of returns whose time intervals overlap
  - $\sum_{i,j} r^{(1)}_i r^{(2)}_j 1\{(t_{i-1}, t_i] \cap (s_{j-1}, s_j] \neq \emptyset\}$: $(t_{i-1}, t_i]$, $(s_{j-1}, s_j]$: the intervals between consecutive trades of each asset
  - unbiased without synchronisation (in the absence of noise)
- With noise as well: pre-averaged or multivariate kernel versions
  - application: intraday portfolio risk (Chapter 4)

Back

Locating Jumps Within the Day, Lee–Mykland (1/2)

- Lee & Mykland (2008): $L_{t,i} = r_{t,i}/\hat{\sigma}_{t,i}$
  - $\hat{\sigma}^2_{t,i}$: the bipower average of the previous $K = 270$ returns (their value for 5-minute data)
  - Under no jumps, $\max_i |L_i|$ over n returns has a Gumbel limit: reject if $(|L| - C_n)/S_n > -\ln(-\ln(1-\alpha))$
  - $C_n = \sqrt{2\ln n} - \frac{\ln\pi + \ln\ln n}{2\sqrt{2\ln n}}$ and $S_n = 1/\sqrt{2\ln n}$: the centring and scaling constants of the Gumbel limit
  - our $\hat{\sigma}^2$ includes the factor $\pi/2$, so the constant $c = \sqrt{2/\pi}$ of Lee & Mykland (2008) becomes 1
  - SPY, returns divided by f_i, $\alpha = 1\%$, $n = 115\,441$ (the whole sample): threshold $|L| > 5.41$

Back

Locating Jumps Within the Day, Lee–Mykland (2/2)

- 169 jump returns (5-minute returns, not days) on 105 days, 62 of them on Wednesdays between 14:00 and 15:00
  - when Federal Open Market Committee (FOMC) statements (14:00) and press conferences (14:30) are released
  - Largest: $L = -22.9$ on 2025-10-29
  - only 10% of these days are also flagged by the daily ratio test
- The two tests ask different questions
  - Lee–Mykland: one extreme return relative to local volatility
  - ratio test: a daily gap between RV and BV

Back

Jump-Robust Estimators of Integrated Variance (1/2)

- Truncated RV Mancini (2009): $\text{TRV}_t = \sum_i r_{t,i}^2 \, 1\{|r_{t,i}| \leq c \, \hat{\sigma}_t f_i \, \Delta^{\varpi}\}$, $\varpi \in (0, 1/2)$
  - Keep only the squared returns below a threshold: c: a multiple, $\hat{\sigma}_t$: a jump-robust daily volatility, f_i: the periodicity factor, ϖ: the exponent of the interval length Δ
  - Diffusive increments are $O_p(\Delta^{1/2})$ (of order $\sqrt{\Delta}$) and end up below the threshold; jumps are $O(1)$ and are cut off

Back

Jump-Robust Estimators of Integrated Variance (2/2)

- MedRV (Median Realised Variance) Andersen, Dobrev & Schaumburg (2012): $\frac{\pi}{6-4\sqrt{3}+\pi}\frac{M}{M-2}\sum_i \mathrm{med}(|r_{i-1}|,|r_i|,|r_{i+1}|)^2$
  - med: the median of three adjacent absolute returns; the first fraction makes MedRV consistent for IV
  - An isolated return larger than both neighbours never enters the median (which shifts to a neighbour), so a large isolated jump is excluded, though its effect is not removed exactly in finite samples
  - MedRV is less sensitive than BV to isolated zero returns
  - two adjacent jumps can still enter the median
- Jump share of SPY RV, summed over all days without a test
  - jump part: RV minus the robust estimator
  - BV 6.3%, MedRV 7.2%, TRV (Truncated Realised Variance) 12.5%
  - significant jump days only: 2.8%
- TRV with $c = 3$, $\varpi = 0.49$: truncation bites on 44% of days
  - c is a tuning choice
- Aït-Sahalia & Jacod (2009): compare power variations $\sum |r|^p$, $p > 2$, at Δ and $k\Delta$
  - the ratio tends to 1 with jumps and to $k^{p/2-1}$ without
  - known limits under both hypotheses

Back

HAR as a Restricted AR(22), the Wald Test

- Unrestricted: $\ln \mathrm{RV}_{t+1} = \phi_0 + \sum_{k=1}^{22} \phi_k \ln \mathrm{RV}_{t+1-k} + \varepsilon_{t+1}$
  - HAR imposes $\phi_2 = \ldots = \phi_5$ (3 restrictions) and $\phi_6 = \ldots = \phi_{22}$ (16): $R\phi = 0$ with 19 rows
- Wald: $W = (R\hat{\phi})^\top (R\hat{V}_{\mathrm{NW}} R^\top)^{-1} R\hat{\phi} \to \chi^2_{19}$ under the HAR restrictions
  - R: the 19×23 restriction matrix, $\hat{V}_{\mathrm{NW}}$: Newey–West covariance of $\hat{\phi}$
  - large W = restrictions rejected
- SPY log total variance: $W = 18.2$, $p = 0.51$
  - R^2 0.477 (AR(22)) vs 0.470 (HAR)
  - The step restrictions are not rejected
  - HAR saves 19 parameters at almost no in-sample cost
  - less estimation noise helps out of sample

Back

Measurement Error in the Regressors and HARQ (1/2)

- $\mathrm{RV}_t = \mathrm{IV}_t + u_t$, with $\mathbb{E}[u_t \mid \sigma] = 0$ and $\mathrm{Var}(u_t \mid \sigma) = 2\,\mathrm{IQ}_t/M$
  - the CLT of the section Realised Variance
  - If $\mathrm{IV}_{t+\mathbf{1}} = \beta_{\mathbf{0}} + \beta\,\mathrm{IV}_t + \eta_{t+\mathbf{1}}$, OLS of $\mathrm{RV}_{t+\mathbf{1}}$ on RV_t gives plim $\hat{\beta} = \beta\,\dfrac{\mathrm{Var}(\mathrm{IV})}{\mathrm{Var}(\mathrm{IV}) + \mathbb{E}[2\,\mathrm{IQ}/M]} < \beta$ for $\beta > 0$: attenuation ($\eta_{t+\mathbf{1}}$: the shock to IV; plim: probability limit)
- If the signal variance does not rise with IQ_t, the signal-to-noise ratio of RV_t falls when IQ_t is high
  - a working approximation
  - HARQ approximates the falling weight by $\beta_d + \beta_Q\sqrt{\mathrm{RQ}_t}$, $\beta_Q < 0$ Bollerslev, Patton & Quaedvlieg (2016)

Back

Measurement Error in the Regressors and HARQ (2/2)

- SPY, levels, Newey–West SE
  - intraday RV: $\hat{\beta}_Q = -0.0088$ (0.0042)
  - total variance: $\hat{\beta}_Q = -0.0022$ (0.0024)
  - Daily slope (intraday RV) 1.13 at the median $\sqrt{\text{RQ}}$, 1.09 at its 99th percentile: the sign predicted by the theory, a small effect
- Read $\hat{\beta}_d$ together with the other slopes
  - the daily slope above 1 (1.135) is partly offset by a negative weekly slope (-0.120)
  - the three RV averages overlap by construction (each contains yesterday's RV), so the split is imprecise (SE of $\hat{\beta}_d$: 0.494)
  - the slopes sum to 1.04 (total variance: 0.80), so the intraday levels fit is an in-sample diagnostic for the sign of β_Q, not a forecasting model
- Like for like, total variance: HAR $\hat{\beta}_d = 0.399$, HARQ $\hat{\beta}_d = 0.542$
  - letting the weight fall with RQ (Realised Quarticity) raises the base daily slope, in the direction the attenuation argument predicts
  - the difference is within one SE

Back

Beyond HAR, Semivariances and Joint Return–RV Models (1/2)

- Realised semivariances enter SHAR Patton & Sheppard (2015): $\beta_d \mathrm{RV}_t$ becomes $\beta^+ \mathrm{RS}_t^+ + \beta^- \mathrm{RS}_t^-$
  - $\mathrm{RS}_t^{\pm} = \sum_i r_{t,i}^2 \, 1\{r_{t,i} \gtrless 0\}$, $\mathrm{RS}^+ + \mathrm{RS}^- = \mathrm{RV}$
  - SPY, total-variance target, levels (overnight return counted as one more return of the day): $\hat{\beta}^- = 0.56$ (0.23), $\hat{\beta}^+ = 0.21$ (0.18); the case study (intraday-RV target) gives 1.44 and -0.34
  - Wald $\beta^+ = \beta^-$: $p = 0.35$
  - Bad volatility is more persistent, as in Patton & Sheppard (2015), but six years of data do not separate the two slopes
  - In the MCS comparison, HARQ and SHAR are estimated in levels with the filter of Bollerslev, Patton & Quaedvlieg (2016): forecasts outside the [min, max] range of the estimation-sample values are replaced by the sample mean

Back

Beyond HAR, Semivariances and Joint Return–RV Models (2/2)

- Realized GARCH Hansen, Huang & Shek (2012): a GARCH equation plus a **measurement equation**
  - the GARCH equation is driven by the realised measure x_t
  - $\ln h_t = \omega + \beta \ln h_{t-\mathbf{1}} + \gamma \ln x_{t-\mathbf{1}}$, $\ln x_t = \xi + \varphi \ln h_t + \tau(z_t) + u_t$, $r_t = \sqrt{h_t}\, z_t$: joint quasi-maximum likelihood
  - h_t: conditional variance; x_t: realised measure (e.g. RV); z_t: standardised shock; $\tau(z_t)$: leverage function
  - ω: here the GARCH constant (not the noise); ξ, φ: measurement-equation coefficients, u_t: its error
- HEAVY (high-frequency-based volatility) models Shephard & Sheppard (2010): consistent multi-step forecasts
  - one GARCH-type equation for the daily variance, driven by lagged RV
  - one equation for RV itself

Back

Case Study Data, 29 DJIA Stocks and 12 Predictors

- Data: 29 of the 30 DJIA stocks (Visa excluded), 29 January 2001 – 31 December 2017
  - $T = 4\,257$ days
  - 5-minute RV from cleaned New York Stock Exchange (NYSE) trade prices
  - $n = 78$ returns a day
- Split: training 70% (2 964 days), validation 10% (424), test 20% (847)
  - $2\,964 + 424 + 847 = 4\,235$: the first 22 days only feed the monthly RV average
  - robustness: training 1 000 or 2 000 days, validation 200
- $\mathcal{M}_{\mathbf{HAR}}$: daily, weekly, monthly RV
- $\mathcal{M}_{\mathbf{ALL}}$ adds 9 predictors
  - implied volatility, earnings-day dummy, 1-week momentum, change in log dollar volume
  - VIX, squared Hang Seng return, business-conditions index, change in the 3-month T-bill rate, policy-uncertainty index

Back

Case Study Models, from HAR to Neural Networks

- Models: HAR, log-HAR, HAR with leverage, SHAR, HARQ
  - by OLS on a rolling window
- Penalised: ridge, lasso, elastic net, adaptive and post-lasso
  - ridge: L_2 penalty; lasso: L_1; elastic net: a mix
  - penalties tuned on the validation set
- Trees: bagging, random forest, gradient boosting
  - random forest with Breiman's defaults: 500 trees, $P/3$ candidate predictors per split (P: number of predictors), minimum node size 5
- Neural networks NN_k^e: $k = 1, \ldots, 4$ hidden layers (2; 4-2; 8-4-2; 16-8-4-2 neurons)
  - leaky ReLU (Rectified Linear Unit), Adam, dropout, early stopping
  - $e = 1$: best of 100 seeds on validation, $e = 10$: average of the best 10
- Evaluation: out-of-sample MSE relative to HAR
  - averaged over stocks
  - one-sided DM test per stock Diebold & Mariano (1995)
  - MCS at 90% Hansen, Lunde & Nason (2011)

Back

Case Study, Further Results

- One day ahead, RV lags only
  - MSE relative to HAR: log-HAR 0.995, HARQ 1.059, random forest 1.020, neural networks 0.954–0.990
  - Past RV leaves little to learn
  - the networks give the clearest gains (1–5%), log-HAR barely beats HAR
- One day ahead, all 12 predictors: elastic net 0.916
  - the other models are on the horizon chart of the case study
  - Extra predictors pay most with shrinkage, nonlinearity or the log transform
  - HARQ overfits
- MCS at 90%, with all predictors
  - the networks stay in the set for about 80% of stocks
  - HAR stays for about 20–40% of stocks
  - Variable importance from accumulated local effects: daily RV, weekly RV and implied volatility dominate in all four models shown
  - Random forest and networks rank implied volatility higher than OLS with all predictors and elastic net

Back

Case Study, Replication, Limits and Extensions (1/2)

- Replicate on course data: SPY 5-minute bars
  - open-to-close RV with 78 returns a day
  - the authors report that SPY confirms the stock results
  - The authors' robustness split: 1 000 training days, 200 validation days, the rest to 18 September 2026 as test set
  - Predictors: VIX (computed from S&P 500 index options; a proxy for SPY's own implied volatility, which would need SPY options), squared Hang Seng return, SPY momentum and dollar volume, T-bill and policy uncertainty (Federal Reserve Economic Data, FRED)
  - no earnings dummy for a fund

Back

Case Study, Replication, Limits and Extensions (2/2)

- Limits stated by the authors
  - Minimal tuning: more tuning could improve ML further
  - the networks are trained once and not re-estimated in the test window
  - Rankings depend on the split: with 1 000 training days the financial crisis falls in the test set and the HAR extensions beat HAR
  - HAR models with a separate jump component are not included
  - accumulated local effects describe the average prediction and miss rare events such as earnings days
- Open extension: the paper ranks variance forecasts by MSE only
  - does the ML advantage survive QLIKE, which is robust to the noisy RV proxy Patton (2011)?
  - does it hold for Bitcoin's 24-hour RV?

Back

The Discovery Loop, Reusable Prompts (1/2)

- Tools: an AI assistant
  - a large language model (LLM), such as Claude, ChatGPT, Gemini or GitHub Copilot
  - literature search: Semantic Scholar, Elicit, Google Scholar
- **1. Literature**: tabulate the evidence
  - then check every DOI on Crossref (the title must match)
  - `Tabulate the estimates of H published since 2018: estimator, sampling frequency, treatment of the RV measurement error, point estimate and standard error, DOI.`
- **2. Competing hypotheses**, each with its implied shape of $m(2, \Delta)$
  - `Derive m(2, Delta) for three processes that give H near 0.1 with the scaling estimator: fractional log-volatility with H = 0.1; H = 0.5 plus i.i.d. error; a two-factor model with one fast mean-reverting factor.`
- **3. Data and code**: replicate the chapter's number first
  - `Write Python that builds daily total variance from the SPY 5-minute bars of the course data (5-minute RV plus the squared overnight return, previous close to today's open), computes m(q, Delta) for q in {0.5, 1, 1.5, 2, 3}, Delta = 1..30, and the slope H.` Target: $\hat{H} = 0.11$

Back

The Discovery Loop, Reusable Prompts (2/2)

- **4. Robustness**, pre-registered, with a power check
  - `Write a pre-registered plan: lag windows [1,30], [5,30], [10,50]; RV, bipower variation and 30-minute subsampled RV; a fit m(2, Delta) = c Delta^(2H) + k; block-bootstrap intervals; size and power by simulation at H = 0.1 and 0.5.`
- **5. Critique**: ask for the strongest counter-argument
  - `Act as a referee: which non-i.i.d. error structures in log RV (serially correlated, level-dependent, driven by overnight returns or jumps) would survive the constant-k correction, and how would you detect each?`
- **6. Report**: all windows and measures, bootstrap intervals, the simulation design, code and the AI log

Back

Further Reading

- Textbook and surveys
  - Franke, Härdle & Hafner (2019): statistics of financial markets, volatility models
  - Andersen, Bollerslev, Diebold & Labys (2003): modelling and forecasting realised volatility
  - Corsi (2009): the HAR model
- Crypto-assets
  - Intraday information can enter Bitcoin risk forecasts; the entropy of high-frequency returns has been proposed as one such signal Pele & Mazurencu-Marinescu-Pele (2019)
  - Crypto-assets form a distinct asset class by the features of their return distributions Pele et al. (2023)

Back

Further Study on Quantinar

Advanced courses on quantinar.com

- Nonparametric and Semiparametric Models
- Kalman Filter
- Applied Time Series Analysis with Python

Code and data

- Quantlets: github.com/danpele/MFM/tree/main/Quantlets/Ch_09

Back

Acronyms used in this material (1/2)

- **ACF**: AutoCorrelation Function
- **AI**: Artificial Intelligence
- **AR**: AutoRegressive (model)
- **BH**: Benjamini–Hochberg (procedure)
- **BV**: Bipower Variation
- **BVB**: Bursa de Valori București — *the Bucharest Stock Exchange*
- **CI**: Confidence Interval
- **CLT**: Central Limit Theorem
- **CV**: Cross-Validation
- **DJIA**: Dow Jones Industrial Average
- **DM**: Diebold–Mariano (test)
- **DOI**: Digital Object Identifier
- **ETF**: Exchange-Traded Fund
- **EWMA**: Exponentially Weighted Moving Average
- **FDR**: False Discovery Rate
- **FOMC**: Federal Open Market Committee
- **FRED**: Federal Reserve Economic Data
- **FWER**: Family-Wise Error Rate
- **GARCH**: Generalised AutoRegressive Conditional Heteroskedasticity
- **GMM**: Generalized Method of Moments
- **GW**: Giacomini–White (test)
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors)
- **HAR**: Heterogeneous AutoRegressive (model)
- **HAR-CJ**: HAR model with Continuous and Jump components
- **HAR-RV**: Heterogeneous AutoRegressive model of Realised Volatility
- **HAR-X**: Heterogeneous AutoRegressive model with additional regressors
- **HARQ**: HAR model with Realised Quarticity

Acronyms used in this material (2/2)

- **HEAVY**: High-frEquency-bAsed VolatilitY (models)
- **IQ**: Integrated Quarticity
- **LLM**: Large Language Model
- **MAE**: Mean Absolute Error
- **MCS**: Model Confidence Set
- **MedRV**: Median Realised Variance
- **ML**: Machine Learning
- **MSE**: Mean Squared Error
- **MZ**: Mincer–Zarnowitz (regression)
- **NYSE**: New York Stock Exchange
- **OLS**: Ordinary Least Squares
- **PRDS**: Positive Regression Dependence on a Subset (of p-values)
- **PV**: Partial Variance
- **QLIKE**: Quasi-LIKElihood (loss function)
- **ReLU**: Rectified Linear Unit
- **RK**: Realised Kernel
- **RQ**: Realised Quarticity
- **RS**: Realised Semivariance
- **RV**: Realised Variance
- **SE**: Standard Error
- **SHAR**: Semivariance HAR model
- **SV**: SemiVariance (SV-HAR = SHAR)
- **TQ**: Tripower Quarticity
- **TRV**: Truncated Realised Variance
- **TSRV**: Two-Scale Realised Variance
- **US**: United States
- **UTC**: Coordinated Universal Time (Temps Universel Coordonné)
- **VIX**: CBOE Volatility IndeX
- **VRP**: Variance Risk Premium

References (1/6)

- Aït-Sahalia, Y., Jacod, J. (2009). Testing for jumps in a discretely observed process. *Annals of Statistics*, 37(1), 184–222.
- Andersen, T.G., Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4(2–3), 115–158.
- Andersen, T.G., Bollerslev, T. (1998). Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4), 885–905.
- Andersen, T.G., Bollerslev, T., Diebold, F.X. (2007). Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility. *Review of Economics and Statistics*, 89(4), 701–720.
- Andersen, T.G., Bollerslev, T., Diebold, F.X., Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1), 43–76.
- Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453), 42–55.
- Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579–625.
- Andersen, T.G., Dobrev, D., Schaumburg, E. (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics*, 169(1), 75–93.
- Bajgrowicz, P., Scaillet, O., Treccani, A. (2016). Jumps in high-frequency data: spurious detections, dynamics, and news. *Management Science*, 62(8), 2198–2217.
- Bandi, F.M., Russell, J.R. (2008). Microstructure noise, realized variance, and optimal sampling. *Review of Economic Studies*, 75(2), 339–369.

References (2/6)

- 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., Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B*, 64(2), 253–280.
- Barndorff-Nielsen, O.E., Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1), 1–37.
- Barndorff-Nielsen, O.E., Shephard, N. (2005). How accurate is the asymptotic approximation to the distribution of realised variance? In D.W.K. Andrews, J.H. Stock (Eds.), *Identification and Inference for Econometric Models*, Cambridge University Press, 306–331.
- Barndorff-Nielsen, O.E., Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4(1), 1–30.
- Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series 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.
- Bennedsen, M., Lunde, A., Pakkanen, M.S. (2022). Decoupling the short- and long-term behavior of stochastic volatility. *Journal of Financial Econometrics*, 20(5), 961–1006.
- Bolko, A.E., Christensen, K., Pakkanen, M.S., Veliyev, B. (2023). A GMM approach to estimate the roughness of stochastic volatility. *Journal of Econometrics*, 235(2), 745–778.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.

References (3/6)

- Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *Review of Economics and Statistics*, 69(3), 542–547.
- Bollerslev, T., Medeiros, M.C., Patton, A.J., Quaedvlieg, R. (2022). From zero to hero: realized partial (co)variances. *Journal of Econometrics*, 231(2), 348–360.
- 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.
- Bollerslev, T., Tauchen, G., Zhou, H. (2009). Expected stock returns and variance risk premia. *Review of Financial Studies*, 22(11), 4463–4492.
- Boudt, K., Croux, C., Laurent, S. (2011). Robust estimation of intraweek periodicity in volatility and jump detection. *Journal of Empirical Finance*, 18(2), 353–367.
- Carr, P., Wu, L. (2009). Variance risk premiums. *Review of Financial Studies*, 22(3), 1311–1341.
- Christensen, K., Oomen, R.C.A., Podolskij, M. (2014). Fact or friction: jumps at ultra high frequency. *Journal of Financial Economics*, 114(3), 576–599.
- Christensen, K., Siggaard, M., Veliyev, B. (2023). A machine learning approach to volatility forecasting. *Journal of Financial Econometrics*, 21(5), 1680–1727.
- Cont, R., Das, P. (2024). Rough volatility: fact or artefact? *Sankhya B*, 86(1), 191–223.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196.
- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

References (4/6)

- Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383), 605–610.
- Epps, T.W. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74(366), 291–298.
- Franke, J., Härdle, W.K., Hafner, C.M. (2019). Statistics of Financial Markets: An Introduction, 5th ed. Springer.
- Fukasawa, M., Takabatake, T., Westphal, R. (2022). Consistent estimation for fractional stochastic volatility model under high-frequency asymptotics. *Mathematical Finance*, 32(4), 1086–1132.
- Gatheral, J., Jaisson, T., Rosenbaum, M. (2018). Volatility is rough. *Quantitative Finance*, 18(6), 933–949.
- 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. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7), 873–889.
- Hansen, P.R., Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, 24(2), 127–161.
- Hansen, P.R., Lunde, A., Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Hayashi, T., Yoshida, N. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 11(2), 359–379.

References (5/6)

- Huang, X., Tauchen, G. (2005). The relative contribution of jumps to total price variance. *Journal of Financial Econometrics*, 3(4), 456–499.
- Jacod, J., Li, Y., Mykland, P.A., Podolskij, M., Vetter, M. (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic Processes and their Applications*, 119(7), 2249–2276.
- Lee, S.S., Mykland, P.A. (2008). Jumps in financial markets: a new nonparametric test and jump dynamics. *Review of Financial Studies*, 21(6), 2535–2563.
- 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.
- Mancini, C. (2009). Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics*, 36(2), 270–296.
- Merton, R.C. (1980). On estimating the expected return on the market: an exploratory investigation. *Journal of Financial Economics*, 8(4), 323–361.
- Mincer, J., Zarnowitz, V. (1969). The evaluation of economic forecasts. In J. Mincer (Ed.), *Economic Forecasts and Expectations*, NBER, 3–46.
- Müller, U.A., Dacorogna, M.M., Davé, R.D., Olsen, R.B., Pictet, O.V., von Weizsäcker, J.E. (1997). Volatilities of different time resolutions: analyzing the dynamics of market components. *Journal of Empirical Finance*, 4(2–3), 213–239.
- Mykland, P.A., Zhang, L. (2009). Inference for continuous semimartingales observed at high frequency. *Econometrica*, 77(5), 1403–1445.
- Patton, A.J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.

References (6/6)

- 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.
- Pele, D.T., Mazurencu-Marinescu-Pele, M. (2019). Using high-frequency entropy to forecast Bitcoin's daily value at risk. *Entropy*, 21(2), 102.
- Pele, D.T., Wesselhöfft, N., Härdle, W.K., Kolossiatis, M., Yatracos, Y.G. (2023). Are cryptos becoming alternative assets? *European Journal of Finance*, 29(10), 1064–1105.
- Shephard, N., Sheppard, K. (2010). Realising the future: forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics*, 25(2), 197–231.
- Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach. *Bernoulli*, 12(6), 1019–1043.
- Zhang, L., Mykland, P.A., Aït-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472), 1394–1411.