

Modelarea piețelor financiare

Capitolul 20 (capitol special): Semnătura traiectoriilor și volatilitatea realizată — un studiu de cercetare

Daniel Traian PELE

Academia de Studii Economice din București
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Academia Română, Institutul de Prognoză Economică
MSCA Digital Finance

Îmbunătățește semnătura traiectoriei prognoza volatilității?



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

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

— *Lyons & McLeod (2022)*

- Gu et al. (2024) rezumă o fereastră de date prin semnătura traiectoriei (engl. *path signature*)
  - pentru tarifele de transport, ei raportează o acurateți de peste cinci ori mai bună decît o prognoză din industrie
- Capitolul transferă metoda la varianța realizată (RV) zilnică a 50 de active din baza de date VOLARE (*VOLatility Archive for Realized Estimates*, arhiva de măsuri realizate ale volatilității)
  - reperul: log-HAR, modelul HAR (Heterogeneous AutoRegressive) Corsi (2009) aplicat logaritmului RV
  - patru ipoteze fixate înaintea rezultatelor
- La o zi, trăsăturile de semnătură nu aduc cîștig: Sig-L are o pierdere QLIKE medie cu 0,4% mai mare decît log-HAR
  - QLIKE: pierderea de cvasi-verosimilitate; o valoare mai mare înseamnă o prognoză mai slabă
  - Sig-L este semnificativ mai bun pe 3 din 50 de active (pe 1 după corecția Holm); din întîmplare ne-am aștepta la circa 2,5
- Cel mai puternic efect din studiu este cel mai vechi: modelarea logaritmului RV
  - la o zi, log-HAR este semnificativ mai bun decît HAR în nivel pe 21 din 50 de active (pe 5 după corecția Holm)

▸ Graficul central: QLIKE relativ la HAR, activ cu activ

Îmbunătățește o metodă construită pentru tarife de transport prognoza volatilității?

Forma traiectoriilor și prognoza volatilității

- Un studiu de cercetare: metoda lui Gu et al. (2024), aplicată variantei realizate zilnice
  - metoda a fost publicată pentru tarifele de transport de la Amazon
  - datele: 50 de active din VOLARE Cipollini et al. (2026)
  - VOLARE: VOLatility Archive for Realized Estimates, o arhivă de măsuri realizate
- Exemplele detaliate și subiectele avansate sînt în Aprofundări, după Concluzii
  - butonul din subsol duce la ele, iar butonul Înapoi revine la slide-ul de plecare

Rezultatele învățării

- Calculați semnătura unei traiectorii liniare pe porțiuni și aria ei Lévy
- Calculați trăsături (variabile) de semnătură și ponderi de nucleu pentru o prognoză a volatilității
- Interpretați comparații de prognoze pe multe active
  - cu teste Diebold–Mariano (DM)
  - cu mulțimi de încredere pentru modele
  - cu corecții Holm și Benjamini–Hochberg (BH)
- Interpretați rezultatul unui transfer preînregistrat într-un domeniu nou
  - inclusiv un rezultat nul, împreună cu mecanismul lui
- Aprofundări și Anexa: lectură pentru familiarizare

De ce prognozăm varianța realizată?

- Volatilitatea intră în aproape toți indicatorii de risc din acest curs
  - Măsuri ale riscului de coadă (*tail risk*) (Capitolul 7), prețuri de opțiuni (Capitolul 12), ponderi de portofoliu (Capitolul 4)
- Datele intraday fac varianța zilnică aproape observabilă
  - varianța realizată (Realised Variance, RV, Capitolul 9)
  - Prognoza volatilității devine o regresie cu o țintă zgomotoasă, calculată fără un model parametric
- Reperul este modelul HAR Corsi (2009) (Capitolul 9)
  - HAR: Heterogeneous AutoRegressive, autoregresiv heterogen
  - Trei medii ale RV trecute (zilnică, săptămînală, lunară), estimate prin OLS (Ordinary Least Squares, metoda celor mai mici pătrate)
- Orice metodă nouă trebuie să depășească HAR în afara eșantionului (*out-of-sample*), pe multe active, cu un test corect

Două componente de testat: trăsăturile de semnătură și ponderile de nucleu

- Gu et al. descriu o fereastră recentă a unei serii ca o **traiectorie**
  - o rezumă prin **semnătura traiectoriei** (engl. *path signature*)
  - Trăsăturile de semnătură intră într-o regresie liniară cu puțini coeficienți nenuli (LASSO, Least Absolute Shrinkage and Selection Operator, aplicat în doi pași)
  - Perioadele trecute a căror traiectorie seamănă cu cea curentă primesc o pondere mai mare (un nucleu de semnătură)
- Ei raportează o creștere de peste cinci ori a acurateței
  - față de o prognoză din industrie pentru tarifele de transport
  - Întrebare: îmbunătățește aceeași metodă modelele de tip HAR pentru varianța realizată?
- Fiecare componentă este testată separat
  - astfel încît un eventual cîștig poate fi atribuit trăsăturilor sau ponderilor



Wall Street și Bursa din New York, 25 martie 2020

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

Varianța realizată și variația pătratică (1/2): modelul prețului și RV

- **Întrebare**: ce anume prognozăm și cum măsurăm, din date intraday, varianța unei zile?
- Modelul prețului: prețul logaritmic se modifică printr-un drift, un termen de difuzie și salturi

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

  - p_t: prețul logaritmic (timpul în zile); μ_t: drift-ul; σ_t: volatilitatea instantanee; W_t: o mișcare browniană
  - J_t: un proces de salturi, cu ΔJ_s mărimea saltului la momentul s
- Varianța realizată: suma pătratelor randamentelor intraday din ziua t Barndorff-Nielsen & Shephard (2002)

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

  - M: numărul de randamente logaritmice intraday $r_{t,i}$ din ziua t (78 de randamente la cinci minute pentru o acțiune americană)
  - Limita este variația pătratică a zilei t: varianța integrată plus suma pătratelor salturilor

Varianța realizată și variația pătratică (2/2): eroarea de măsurare

- Fără salturi și fără zgomot, pe o grilă echidistantă, limita este varianța integrată
  - eroarea are o distribuție Normală mixtă

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

  - IV_t: varianța integrată, adică varianța adevărată a zilei t; IQ_t: cvarticitatea integrată
  - $\mathrm{MN}(0, V)$: distribuția Normală mixtă, cu varianță aleatoare V; condiționat de traiectoria volatilității, eroarea are varianța $2\,\mathrm{IQ}_t/M$
- RV este o măsurare zgomotoasă a lui IV_t
  - zgomotul este cel mai mare în zilele volatile (IQ_t mare)
  - Este eroarea de măsurare pe care o corectează HARQ (*HAR model with Realised Quarticity*), modelul HAR cu cvarticitatea realizată (secțiunea despre familia HAR)
- Zgomotul de microstructură deplasează estimarea la frecvențe foarte mari
  - randamentele la cinci minute sînt compromisul standard (Capitolul 9)

Semivariațele și cvarticitatea realizată din VOLARE

- Semivariațele Barndorff-Nielsen, Kinnebrock & Shephard (2010): RV descompusă după semnul randamentelor intraday

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

  - RV_t^+: varianța din creșteri, RV_t^-: varianța din scăderi
  - Suma lor este RV: $\mathrm{RV}_t = \mathrm{RV}_t^+ + \mathrm{RV}_t^-$
- Cvarticitatea realizată (RQ, *Realised Quarticity*): o estimare a lui IQ_t, scala erorii de măsurare a RV

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

  - Factorul $M/3$: un randament Normal cu varianța s^2 are momentul de ordinul patru $3s^4$, iar fiecare varianță la cinci minute este aproximativ IV_t/M
  - Limita cere un preț continuu, fără zgomot, și o grilă echidistantă
  - Un salt ΔJ adaugă aproximativ $M(\Delta J)^4/3$, deci RQ brută este sensibilă la salturi

Ținte și funcții de pierdere (1/2): ținta și modelele logaritmice

- Ținta la orizontul h este media RV pe următoarele h zile

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

  - t: originea prognozei (ultima zi observată); $h = 1, 5, 22$: o zi, o săptămînă și o lună de tranzacționare
- Modelele logaritmice prognozează $y = \ln \mathrm{RV}_{t+\mathbf{1}:t+h}$ și revin la nivel cu o corecție de varianță

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

  - F_t: prognoza lui $\mathrm{RV}_{t+\mathbf{1}:t+h}$ făcută în t; $\hat{y}_t$: prognoza logaritmică; $\hat{s}^{\mathbf{2}}$: varianța reziduurilor; $+\hat{s}^{\mathbf{2}}/2$ vine din $\mathbb{E}[e^Z] = e^{\mu+s^{\mathbf{2}}/\mathbf{2}}$ pentru $Z \sim N(\mu, s^{\mathbf{2}})$
  - Corecția este exactă dacă eroarea logaritmică are condiționat distribuția Normală, cu varianță constantă, și aproximativă în rest
- Pentru $h > 1$ țintele se suprapun
  - dependență serială pînă la lagul $h-1$ prin construcție, posibil și mai departe
- De aceea fiecare test folosește o varianță HAC
  - HAC: Heteroskedasticity and Autocorrelation Consistent, robustă la heteroscedasticitate și autocorelare

Cum asociem prognoza cu ținta

- În definiția țintei, indicele t este originea prognozei
- În formulele pierderilor și ale testelor, indicele marchează o pereche de evaluare
  - ținta realizată și prognoza făcută înainte ca ținta să fie observabilă
- Orizontul de o zi: informația pînă la închiderea de luni produce prognoza RV de marți
  - prognoza este stocată alături de RV realizată marți
- Orizontul de cinci zile: aceeași origine prognozează media de marți pînă lunea următoare
  - originea următoare produce o țintă suprapusă
- Testele de comparare a prognozelor folosesc pe fiecare rînd aceeași țintă realizată pentru ambele modele
- Faptul că ținta stă pe rîndul prognozei nu permite folosirea ei la estimare

Ținte și funcții de pierdere (2/2): pierderi robuste

- O funcție de pierdere este **robustă** dacă păstrează ordinea pierderilor *așteptate* Patton (2011)
  - pentru un proxy condiționat nedeplasat, ordinea pierderilor cu RV_t coincide cu cea cu varianța adevărată
  - Condiționat nedeplasat: $\mathbb{E}_{t-1}[\mathrm{RV}_t] = \sigma_t^2$, cu $\mathbb{E}_{t-1}$ media condiționată de informația pînă la $t-1$ și σ_t^2 varianța condiționată a zilei t, distinctă de varianța integrată realizată ex post
- Cele două pierderi robuste folosite (scrise pentru $h = 1$; pentru $h > 1$, RV_t se înlocuiește cu ținta)

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

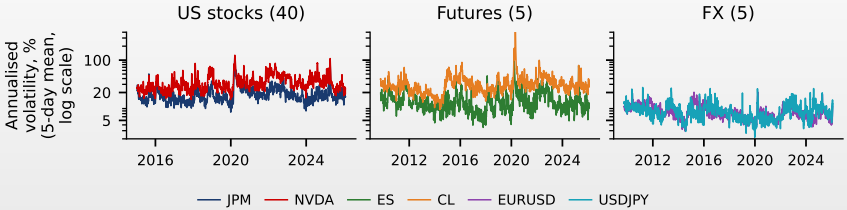
  - QLIKE (quasi-likelihood, pierderea de tip cvasi-verosimilitate) este zero doar dacă $F_t = \mathrm{RV}_t$ și crește mai repede cînd prognoza este prea mică decît cînd este prea mare
  - Exemplu cu $\mathrm{RV}_t = 1$: $F_t = 0{,}5$ dă $\mathrm{QLIKE}_t = 2 - \ln 2 - 1 = 0{,}307$, iar $F_t = 2$ dă $0{,}5 - \ln 0{,}5 - 1 = 0{,}193$
- QLIKE depinde de raportul RV_t / F_t
  - aceeași eroare proporțională produce aceeași pierdere într-o zi calmă ca într-o zi volatilă
- MSE (*Mean Squared Error*, eroarea pătratică medie) este dominată de cîteva zile de criză
  - de aceea QLIKE este pierderea principală, iar MSE cea secundară

Panelul: 50 de active în trei clase

Clasa	Active	Eșantion	Observații	Volatilitate mediană
Acțiuni americane cu capitalizare mare	40	2 ian. 2015 – 30 ian. 2026	111 440	17,1%
Futures (C, CL, ES, GC, NG)	5	28 sept. 2009 – 30 ian. 2026	21 059	19,7%
FX (AUD, EUR, GBP, CAD, JPY față de USD)	5	25 sept. 2009 – 30 ian. 2026	21 203	7,3%

- **Întrebare**: pe ce date testăm transferul și de ce pe trei clase de active?
- Trei clase de active: un cîștig care apare doar pe o clasă nu arată că metoda se transferă
- Acțiuni: 40 de componente ale Dow Jones 30 și S&P 100 (*Standard & Poor's*)
  - notate prin simbolul bursier, de la AAPL (Apple) la XOM (Exxon Mobil)
- Futures: C = porumb, CL = petrol, ES = E-mini S&P 500, GC = aur, NG = gaze naturale
- FX (foreign exchange, piața valutară): cinci valute față de dolarul american
  - dolarul australian, euro, lira sterlină, dolarul canadian și yenul japonez
- Volatilitatea mediană: $\sqrt{252\,\mathrm{mediana}(\mathrm{RV}_t)}$ pentru fiecare activ, apoi mediana clasei
  - 252 de zile de tranzacționare pe an
- Variabile: RV la cinci minute, semivariațele RV^+, RV^-, cvarticitatea realizată (RQ)
  - și prețurile zilnice de deschidere și de închidere

Volatilitatea realizată în clasele de active (1/2)



- Volatilitatea anualizată $\sqrt{252\,\mathrm{RV}_t}$ în %, medie pe cinci zile, scală logaritmică, două active pe clasă
  - acțiuni: JPM = JPMorgan Chase, NVDA = Nvidia; futures: ES = E-mini S&P 500, CL = petrol
  - FX: EURUSD = dolari SUA (Statele Unite ale Americii) pentru un euro, USDJPY = yeni pentru un dolar SUA
- Toate clasele au un vîrf în martie 2020
  - petrolul (CL) atinge maximul în aprilie 2020
  - un model de prognoză trebuie să se adapteze în cîteva zile

▶ Interactiv

Volatilitatea realizată în clasele de active (2/2)

- CL: seria continuă VOLARE leagă contracte succesive, iar prețurile seriei rămîn pozitive
  - prețul negativ din 20 aprilie 2020 a privit doar contractul pe mai, aproape de expirare
- Pe scală logaritmică volatilitatea este persistentă (memoria lungă aproximată de HAR)
- Vîrfurile se disipă în cîteva săptămîni, deci forma traiectoriei recente poate conține informație

MFM_ch20_case_study

HAR: trei medii ale trecutului (1/2)

- **Întrebare**: ce reper simplu trebuie depășit înainte de a crede într-o metodă nouă?
- Modelul HAR Corsi (2009): RV viitoare este o funcție liniară de trei medii ale RV trecute
  - mediile zilnică, săptămînală și lunară

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

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

  - $\beta_{\mathbf{0}}$: constanta; ε_{t+h}: eroarea de prognoză
  - $\beta_d, \beta_w, \beta_m$ (în practică ≥ 0): cît contează ziua de ieri, ultima săptămînă și ultima lună, estimați prin OLS separat pentru fiecare orizont h
  - $\beta_d + \beta_w + \beta_m$ apropiat de 1: persistență ridicată a volatilității
- Intuiția: participanții la piață tranzacționează la orizonturi zilnice, săptămînale și lunare
  - această cascadă aproximează memoria lungă cu doar trei parametri

HAR: trei medii ale trecutului (2/2): AR(22) restricționat și log-HAR

- HAR este un AR(22) restricționat: cei 22 de coeficienți ai lagurilor iau doar trei valori distincte

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

  - ϕ_j: coeficientul lui $\mathrm{RV}_{t+\mathbf{1}-j}$ într-un AR(22) (autoregresie cu 22 de laguri)

- log-HAR: aceeași regresie în logaritmi

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

  - Nu produce prognoze negative și este mai puțin dominat de vîrfuri
  - Logaritmul mediei nu este media logaritmilor, deci acest log-HAR nu este exact un AR(22) restricționat în $\ln \mathrm{RV}$ (Seminarul 20, A7)

Șapte modele într-un tabel (1/2)

Model	Țintă	Ponderi	Regresori
HAR	RV	egale	$\mathrm{RV}_t, \mathrm{RV}_t^{(w)}, \mathrm{RV}_t^{(m)}$
log-HAR	$\ln \mathrm{RV}$	egale	aceiași, în logaritmi
HARQ	RV	egale	HAR + $\mathrm{RV}_t \left(\mathrm{RQ}_t^{\mathbf{1/2}} - \overline{\mathrm{RQ}^{\mathbf{1/2}}}\right)$
SHAR	RV	egale	$\mathrm{RV}_t^+, \mathrm{RV}_t^-, \mathrm{RV}_t^{(w)}, \mathrm{RV}_t^{(m)}$
log-HAR-K	$\ln \mathrm{RV}$	nucleul de semnătură	regresorii log-HAR
Sig-L	$\ln \mathrm{RV}$	egale	log-HAR + 36 de termeni neconstanți ai semnăturii (39 înaintea eliminării constantelor), LASSO în doi pași
Sig-LK	$\ln \mathrm{RV}$	nucleul de semnătură	ca Sig-L: metoda lui Gu et al.

- Denumiri: L = LASSO, K = ponderi de nucleu
- Denumirile din grafice
  - Sig-L apare ca „Signature LASSO"
  - Sig-LK ca „Signature LASSO + kernel weights"
  - log-HAR-K ca „log-HAR + kernel weights"

Șapte modele într-un tabel (2/2)

- log-HAR, log-HAR-K, Sig-L și Sig-LK formează o analiză 2×2 pe componente
  - ablație: fiecare componentă inclusă sau exclusă pe rînd
  - trăsături de semnătură (da/nu) $\times$ ponderi de nucleu (da/nu)
- Log-HAR este reperul natural pentru modelele cu semnături
  - are aceeași țintă, iar regresorii lui se află printre candidați
- Totuși, LASSO poate elimina și termeni log-HAR
  - deci Sig-L față de log-HAR compară două proceduri, nu doar două mulțimi de variabile

O fereastră de date este o traiectorie

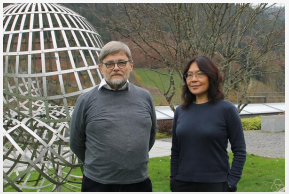
- **Întrebare**: cum rezumăm o fereastră de 22 de zile astfel încît să păstrăm ordinea mișcărilor, nu doar nivelurile lor?
- Considerăm ultimele L zile ale celor $d-1$ serii și unim punctele prin segmente
  - adăugăm timpul ca prim canal (augmentarea cu timpul)

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

  - L: lungimea ferestrei în zile (aici 22); $u_i \in [0,1]$: timpul rescalat al punctului i
  - d: numărul de canale (aici $d = 3$: timp, log RV, randament cumulat)
  - $r_j = \ln(C_j/O_j)$: randamentul logaritmic de la deschidere la închidere al zilei j, cu O_j, C_j prețurile de deschidere și de închidere
- O regresie pe fereastră folosește 22 de laguri
  - un rezumat al traiectoriei surprinde și ordinea, viteza și co-mișcarea lor
  - Exemplu: RV a crescut în timp ce prețurile scădeau, sau RV a crescut întîi și prețurile au scăzut mai tîrziu

Originile semnăturilor

- Kuo-Tsai Chen, geometria spațiilor de bucle
  - Chen (1957): integralele iterate ale unei traiectorii formează o algebră
  - Chen (1958): ele determină traiectoria pînă la reparametrizare și la porțiuni parcurse înapoi
- Lyons (1998): teoria traiectoriilor aspre (*rough paths*)
  - semnătura este mulțimea naturală de trăsături pentru ecuațiile diferențiale controlate
  - Soluția unei ecuații controlate de o traiectorie depinde de aceasta (în primă aproximare) doar prin semnătura ei
- Aplicații
  - în învățarea automată (Levin, Lyons & Ni (2013), Chevyrev & Kormilitzin (2016), Lyons & McLeod (2022))
  - în modelele de prețuri bazate pe semnături Cuchiero, Gazzani & Svaluto-Ferro (2023)



Terry Lyons (stînga), Oberwolfach, 2024

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

Semnătura: definiție (1/2)

- De ce integrale iterate
  - nivelul 1 reține doar incrementele totale
  - integralele pe momente ordonate rețin ordinea în care se mișcă canalele
- Un termen al semnăturii este o integrală iterată a traiectoriei de-a lungul unui cuvînt de canale

$$S(X)_{a,b}^{i_1 \cdots i_k} = \int_{a < u_1 < \cdots < u_k < b} dX_{u_1}^{i_1} \cdots dX_{u_k}^{i_k}$$

  - $X : [a, b] \to \mathbb{R}^d$: o traiectorie cu variație mărginită (lungime totală finită), de exemplu liniară pe porțiuni; X_u^i: canalul i la momentul u
  - Cuvîntul $(i_1, \ldots, i_k)$: o listă ordonată de k canale, cu litere din $\{1, \ldots, d\}$; k este **nivelul** termenului
  - dX_u^i: incrementul canalului i la momentul u (pe un segment liniar, panta înmulțită cu du), integrat pe momentele ordonate $u_1 < \cdots < u_k$

Semnătura: definiție (2/2): nivelurile

- Semnătura este întregul șir de niveluri

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

  - $S^{(k)}$: cei d^k termeni de nivelul k, unul pentru fiecare cuvînt de lungime k, aranjați ca un tablou k-dimensional (un tensor, iar $\otimes$ este produsul tensorial)
  - $\mathrm{Sig}^N(X)$: semnătura trunchiată, formată din nivelurile de la 0 la N, unde N este adîncimea
- Nivelul 1: incrementele totale ale canalelor

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

- Nivelul 2: incrementele canalului j ponderate cu cît s-a deplasat deja canalul i

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

  - Fiecare nivel adaugă o integrală și reține detalii tot mai fine ale ordinii mișcărilor

Segmente liniare și identitatea lui Chen (1/2)

- Un segment liniar cu incrementul $\Delta \in \mathbb{R}^d$ are ca semnătură exponențiala tensorială

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

  - $\Delta = X_b - X_a$: incrementul segmentului
  - $\Delta^{\otimes k}$: tabloul tuturor produselor $\Delta^{i_1} \cdots \Delta^{i_k}$, deci un termen de nivelul k este $\Delta^{i_1} \cdots \Delta^{i_k}/k!$

- **Identitatea lui Chen** Chen (1957)
  - semnătura a două traiectorii parcurse una după alta este produsul tensorial al semnăturilor lor

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

  - $X * Y$: concatenarea, în care Y începe unde se termină X
  - $A^{(k)}$, $B^{(k)}$: nivelul k al semnăturilor A, B, cu $A^{(\mathbf{0})} = B^{(\mathbf{0})} = 1$
  - Nivelul k al traiectoriei întregi adună toate modurile de a împărți un cuvînt de lungime k între prima și a doua bucată

Segmente liniare și identitatea lui Chen (2/2)

- O traiectorie liniară pe porțiuni cu m segmente, cu incrementele $\Delta_1, \ldots, \Delta_m$

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

  - Întregul algoritm: m actualizări, fiecare cu un cost de ordinul $O(d^N)$ operații la adîncimea N
- Ferestre mobile (*rolling windows*): identitatea lui Chen permite și eliminarea unei porțiuni inițiale

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

  - $X_{[s,t]}$: traiectoria restrînsă la $[s,t]$, cu $a \leq s \leq t$
  - $(\cdot)^{-1}$: inversul față de $\otimes$, adică semnătura porțiunii parcurse în sens invers

Exemplu rezolvat: o traiectorie cu două segmente

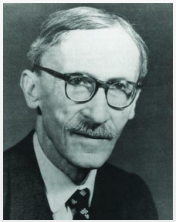
- Traiectoria $X : (0,0) \to (1,0) \to (1,2)$, cu incrementele $\Delta_1 = (1,0)$ și $\Delta_2 = (0,2)$
- Nivelul 1: $S^{(1)} = \Delta_1 + \Delta_2 = (1,2)$
- Nivelul 2 prin identitatea lui Chen: $S^{ij} = \frac{1}{2}\Delta_1^i \Delta_1^j + \Delta_1^i \Delta_2^j + \frac{1}{2}\Delta_2^i \Delta_2^j$
  - Δ_m^i: componenta i a incrementului segmentului m
- Rezultatul: $S^{11} = \frac{1}{2}$, $S^{12} = 0 + 1 \cdot 2 + 0 = 2$, $S^{21} = 0$, $S^{22} = \frac{1}{2} \cdot 2^2 = 2$
- Verificare: $S^{11} = (\Delta X^1)^2/2$ și $S^{22} = (\Delta X^2)^2/2$
  - identitatea shuffle: produsul a două coordonate de nivel 1 este suma celor două ordonări de nivel 2
- $S^{12} \neq S^{21}$: ordinea contează
  - canalul 1 se mișcă în primul segment, iar canalul 2 în al doilea
- Deplasarea totală (nivelul 1) nu distinge această traiectorie de cea parcursă întîi în sus și apoi la dreapta
  - ambele au $\Delta X = (1,2)$; nivelul 2 le distinge

Aria Lévy: ordinea exprimată numeric

- Aria Lévy A^{ij}: partea antisimetrică a nivelului 2

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

  - Este aria cu semn dintre traiectorie, proiectată pe canalele (i, j), și coarda ei (dreapta de la început la sfîrșit)
  - $A^{ij} > 0$: traiectoria se rotește în sens trigonometric în planul (i, j)
  - Citirea lead-lag (cine se mișcă primul) cere o asociere pozitivă: dacă i scade, apoi j crește, i se mișcă primul, dar $A^{ij} < 0$
- Partea simetrică, $\frac{1}{2}(S^{ij} + S^{ji}) = \frac{1}{2}\,\Delta X^i\,\Delta X^j$, nu aduce informație nouă
  - Exemplul rezolvat: $A^{\mathbf{12}} = (2 - 0)/2 = 1$, aria triunghiului dintre traiectorie și coardă
- Canalele (randament, log RV): două secvențe dau $A < 0$
  - o scădere a prețului urmată de o creștere a volatilității
  - o creștere a prețului urmată de o scădere a volatilității
- Aria Lévy este o statistică de tip lead-lag
  - secvențele de tip efect de levier (*leverage effect*) (Capitolul 5) sînt doar una dintre sursele unei arii negative
  - semnul singur nu le identifică



Paul Lévy (1886–1971)

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

Nucleul de semnătură

- Nucleu = produsul scalar al semnăturilor trunchiate Király & Oberhauser (2019)
  - un scor de asemănare între două traiectorii

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

  - X, Y: două traiectorii cu aceleași canale; w: un cuvînt de lungime $|w|$ (nivelul), cu termenul $S(X)^w$; termenul 1 corespunde nivelului 0
- Distanța indusă de nucleu (Gu et al., ec. 9)
  - pătratul distanței euclidiene dintre semnăturile trunchiate

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

  - $d_{\mathrm{Sig}} = 0$: aceeași semnătură pînă la adîncimea N; d_{Sig} mare: forme diferite
- Gu et al. și acest capitol folosesc acest nucleu **trunchiat**, cu același N ca pentru trăsături
  - Nucleul netrunchiat rezolvă o ecuație cu derivate parțiale (PDE, *Partial Differential Equation*) de tip Goursat Salvi et al. (2021), citată, dar nefolosită de Gu et al.
- Canalele se rescalează înaintea nucleului (▸ Anexă)
  - altfel domină canalul cu unitățile cele mai mari

Problema de la Amazon

- **Întrebare**: ce a făcut exact articolul și cît de solide sînt dovezile lui?
- Tarife săptămînale pe piața nord-americană de transport rutier Gu et al. (2024)
  - la nivel național și în cinci regiuni, 2018–2022
  - prognozate pe pînă la 12 săptămîni
  - Peste 100 de factori candidați, reduși la aproximativ 40
  - Metoda este folosită în operațiunile Amazon din noiembrie 2022 și a fost publicată la KDD 2024, conferința ACM (*Association for Computing Machinery*) de *Knowledge Discovery and Data Mining*
- Pentru fiecare orizont Δt (în săptămîni; notația articolului, nu un increment Δ): o regresie directă
  - a lui $y_{\tau+\Delta t}$ pe trăsăturile ξ_τ (slide-ul următor)



Un camion cu container pe o autostradă americană, 2025

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

Trăsăturile din articol

- Trăsăturile în săptămîna τ (Secțiunea 4.1, ec. 14)
  - factorii externi x_τ și semnătura traiectoriei țintei, augmentată cu timpul

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

  - y_{t_i}: tariful din săptămîna t_i, cu $\tau - l \leq t_i \leq \tau$
  - l: lungimea ferestrei în săptămîni
  - $\hat{y}_{\tau-l:\tau}$: traiectoria augmentată cu timpul a ultimelor l săptămîni
- Factorii externi aduc nivelul și informația din afara seriei
  - semnătura aduce forma traiectoriei recente

Metoda: LASSO în doi pași

- Pasul 1, selecție printr-un LASSO ponderat Tibshirani (1996) (ec. 14)

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

  - θ: vectorul coeficienților, $\|\theta\|_{1} = \sum_{j} |\theta_{j}|$; $w_{\tau} \geq 0$: ponderea săptămînii τ (egală sau dată de nucleu, slide-ul următor)
  - $\lambda \geq 0$: penalizarea; un λ mai mare favorizează, în general, soluții mai rare, iar $\lambda = 0$ înseamnă OLS ponderat
- Pasul 2, reestimare prin OLS ponderat doar pe suportul selectat (ec. 15)
  - suportul: coeficienții nenuli de la pasul 1

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

- Motivul reestimării: OLS după selecție elimină contracția (*shrinkage*) indusă de penalizare pe suportul selectat Belloni & Chernozhukov (2013)
  - LASSO micșorează spre zero coeficienții păstrați
  - reestimarea nu elimină deplasarea (*bias*) de selecție sau de omitere
  - Articolul îl numește „adaptiv" din cauza ponderilor w_{τ}, nu în sensul LASSO-ului adaptiv al lui Zou (2006)
- Alegerea lui λ în acest capitol: BIC minim de-a lungul traiectoriei LARS Efron et al. (2004) (▸ Anexă)
  - BIC: Bayesian Information Criterion; LARS: Least Angle Regression
  - suportul se alege în fiecare fereastră de antrenare

Metoda: ponderi date de nucleu

- Distanța fiecărei ferestre trecute față de ultima fereastră cu țintă observată (ec. 11)

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

  - $z_\tau = (x_\tau, y_{\tau+\Delta t})$: *perechile* de predictori și țintă ulterioară, nu trăsăturile ξ_τ (Algoritmul 1), cu traiectoria $z_{\tau-l:\tau}$ pe l săptămîni
  - t: săptămîna curentă; $t-\Delta t$: ultima săptămînă a cărei țintă y_t este deja observată
- Ponderi softmax cu temperatura $\gamma \geq 0$ (ec. 12)
  - normalizate pe toate săptămînile de antrenare s
  - ferestrele asemănătoare (δ_τ mic) primesc o pondere mai mare

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

  - Cazuri-limită: $\gamma = 0$ dă regresia obișnuită (ponderi egale), iar $\gamma \to \infty$ păstrează doar cea mai asemănătoare fereastră trecută
- O idee de tip cel mai apropiat vecin într-o regresie
  - învățăm din perioadele trecute care semănau cu prezentul
  - Costul ideii: mai puține observații efective, $n_{\mathrm{eff}} = 1/\sum_\tau w_\tau^2$ (Kish), cu $n_{\mathrm{eff}} = n$ pentru ponderi egale $1/n$ și $n_{\mathrm{eff}} = 1$ dacă o singură săptămînă ia toată ponderea

Rezultatele articolului: ce raportează

- Tarife naționale, cu trei luni înainte, aprilie–noiembrie 2021 (Tabelul 1)
  - eroarea relativă $|y - \hat{y}|/|y|$: aproximativ 1–2% față de 13–27% pentru prognoza din industrie, „de peste cinci ori"
  - y: tariful realizat, $\hat{y}$: prognoza lui
  - Orizonturi de 1–12 săptămîni (Tabelul 2): eroarea relativă este de 1,06–2,60% la nivel național și de 1,05–5,69% pe regiuni
  - Economia anuală estimată: aproximativ 50 de milioane de dolari
  - Ablația nucleului (Tabelul 3, șase valori săptămînale, 31 octombrie–5 decembrie 2021, o regiune): erori de pînă la 9,85% fără nucleu, față de pînă la 5,15% cu el
- Tipul de dovadă: comparație între proceduri (arată care metodă funcționează mai bine pe aceleași date, nu de ce)
  - față de o singură prognoză comercială, fără erori standard (slide-ul următor)

Rezultatele articolului: limitele dovezilor

- **Un singur reper**: o prognoză comercială al cărei furnizor nu este precizat
- Lipsesc comparațiile cu
  - un model ARIMA (AutoRegressive Integrated Moving Average)
  - un LASSO fără semnături
  - un mers aleator (*random walk*)
- **Evaluare scurtă**: cîteva luni, cîteva serii
  - fără erori standard, fără un test de acuratețe egală
- **Calibrare nepublicată**: l, N, γ, λ și modul lor de alegere
- Aceste limite nu invalidează metoda, dar slăbesc dovezile
- De aceea o testăm în alt context
  - replicarea înaintea extinderii (Capitolul 19)

Transferul la volatilitate: ce se preia și ce se adaptează

Se transferă

- Semnătura traiectoriei recente a țintei
  - augmentată cu timpul
- LASSO în doi pași cu ponderi
  - regresie directă pe orizont
- Ponderi softmax $\exp(-\gamma\,\delta_\tau)$
  - cu nucleul de semnătură trunchiat

Trebuie adaptat

- Țintă zilnică și zgomotoasă în loc de tarife săptămînale
  - de aici logaritmii, inferența HAC și pierderile robuste
- Factorii externi x_τ înlocuiți de termenii log-HAR
  - aceștia aduc informația despre nivel
- Nucleul compară traiectoriile (timp, $\ln \mathrm{RV}$, randament)
  - care se termină în τ și în t, actualizat la fiecare 5 zile
  - nu perechile $(x_\tau, y_{\tau+\Delta t})$ care se termină în $t - \Delta t$
- Deciziile nepublicate (I, N, γ, λ) fixate dinainte
  - unica prognoză comercială este înlocuită de repere puternice și teste formale

Patru ipoteze fixate înaintea rezultatelor

- **Întrebare**: ce fixăm înaintea rezultatelor, astfel încît un cîștig să nu poată fi găsit prin căutare?
- **H1 (trăsături)**: termenii semnăturii aduc informație dincolo de log-HAR
  - Test: Sig-L față de log-HAR, aceeași țintă, aceleași ponderi egale
- **H2 (ponderi)**: ponderile de nucleu îmbunătățesc o regresie dată
  - Teste: Sig-LK față de Sig-L și log-HAR-K față de log-HAR
- **H3 (metoda)**: metoda completă a lui Gu et al. depășește familia HAR
  - Teste: Sig-LK față de HAR și de log-HAR
  - mulțimea de încredere pentru modele (Model Confidence Set, MCS), pentru toate cele șapte modele
- **H4 (stres)**: orice cîștig este mai mare în perioadele de stres
  - crahul din pandemia COVID-19 și șocul tarifelor vamale americane din 2–30 aprilie 2025
  - acolo ferestrele trecute asemănătoare ar trebui să conteze cel mai mult

Temperatura nucleului: o constantă aleasă pe un bloc de validare

- Temperatură fără unități, fixată în fiecare fereastră de antrenare

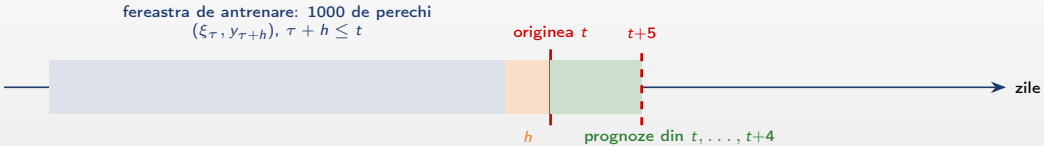
$$\gamma = \frac{c}{\mathrm{median}_\tau(\delta_\tau)}$$

  - δ_τ: distanțele nucleului dintre zilele de antrenare și fereastra curentă; $c > 0$: o constantă fără unități, singurul parametru de calibrare
  - Cu $c = 1$, o fereastră aflată la distanța mediană primește ponderea e^{-1} față de una identică, iar un c mai mare concentrează ponderile

- Valoarea lui c din grila $\{0{,}5; 1; 2; 4\}$ se alege separat pentru fiecare orizont și clasă de active
  - pe primele 250 de zile de prognoză ale fiecărui activ, cumulate la nivelul clasei
  - Criteriul: QLIKE(c) al fiecărui activ, împărțit la media QLIKE a activului pe toată grila, apoi mediat pe clasă
  - Evaluarea începe după ultima țintă de validare a clasei, deci nicio țintă folosită la calibrare nu este ulterioară unei origini evaluate

- Pentru cele 40 de acțiuni Sig-LK alege la fiecare orizont $c = 0{,}5$
  - ponderile cele mai apropiate de cele egale
  - pentru futures și FX alegerea se schimbă cu orizontul, semn că criteriul este plat și zgomotos
  - toate valorile sînt în Anexă

Estimarea pe fereastră mobilă, pas cu pas



- Reestimăm la fiecare 5 zile de tranzacționare pe ultimele 1 000 de perechi complete
- Între reestimări coeficienții sînt ficși, dar regresorii se actualizează zilnic
- Distanțele nucleului se recalculează la fiecare reestimare
  - față de fereastra care se termină la originea t

Protocolul de evaluare (1/3): testul DM

- **Întrebare**: este diferența de pierdere dintre două modele mai mare decît zgomotul din eșantion?
- Pe activ și orizont: QLIKE și MSE medii, comparate prin testul DM Diebold & Mariano (1995)
  - DM (Diebold–Mariano): testul de acuratețe predictivă egală
  - eroarea relativă din articol este raportată doar descriptiv (Anexă)

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

  - L: funcția de pierdere (QLIKE sau MSE); F_t^A, F_t^B: prognozele modelelor A și B, iar d_t este diferența dintre pierderile lor
  - $\bar{d}$: media lui d_t pe cele T zile de evaluare; $\hat{\Omega}$: varianța de termen lung a lui d_t (Aprofundări)
- Sub $H_0 : \mathbb{E}[d_t] = 0$, DM are aproximativ distribuția $N(0,1)$
  - $\mathrm{DM} < -1{,}645$ arată că A este semnificativ mai precis la 5% (test unilateral)
- Estimarea se face pe fereastră mobilă
  - deci testul se interpretează ca testul GW (Giacomini–White) necondiționat
  - Testul GW condiționat verifică în plus dacă d_{t-h} prezice d_t (cu instrumentele $(1, d_{t-h})$) Giacomini & White (2006)

Aprofundări: varianța HAC

Protocolul de evaluare (2/3): mulțimea de încredere a modelelor

- De ce o mulțime de modele: testul DM compară doar două modele
  - aici sînt șapte, fără un reper ales dinainte
- MCS la 90% Hansen, Lunde & Nason (2011)
  - mulțimea de modele care conține cel mai bun model cu probabilitate de cel puțin 90%
  - Construită prin eliminare: cît timp statistica $T_{\max}$ (cea mai mare diferență standardizată dintre pierderea medie a unui model și media modelelor aflate încă în mulțime) respinge acuratețea egală, se elimină cel mai slab model
  - Valorile p (*p-values*) se obțin prin bootstrap staționar Politis & Romano (1994), cu 1 000 de replicări și lungimea medie a blocului de $\max(h, 10)$ zile
  - Seriile de pierdere se reeșantionează împreună, pentru a păstra dependența dintre modele

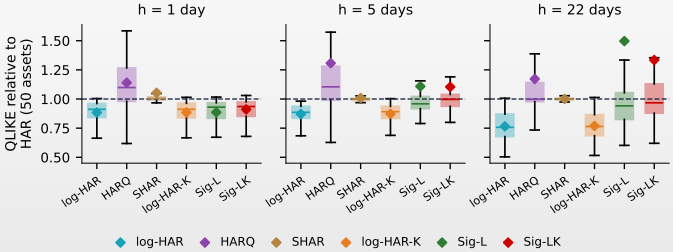
Protocolul de evaluare (3/3): testul pe panel și corecțiile

- Testul pe panel: testul DM aplicat diferenței medii de pierdere pe secțiune transversală
  - o singură serie, cu o valoare pe dată
- Numărăm activele (din 50) cu rezultate semnificativ mai bune
  - Corecția Holm Holm (1979) controlează FWER (Family-Wise Error Rate, probabilitatea a cel puțin unei respingeri false)
  - Corecția BH Benjamini & Hochberg (1995) controlează FDR (False Discovery Rate, proporția așteptată de respingeri false printre respingeri)
- Trei subperioade
  - 20 feb.–30 apr. 2020 (crahul COVID-19)
  - 2–30 apr. 2025 (șocul tarifelor americane)
  - toate celelalte zile (perioadele de evaluare pe clase: ▸ Anexă)

O abatere declarată

- Preînregistrat: filtrul BPQ („insanity filter") pe toate modelele: o prognoză în afara intervalului țintei din eșantionul de estimare devine media acestuia
  - Un test al codului pe JPM (un singur activ, înaintea oricărui rezultat pe panel) a arătat că înlocuiește prognoze corecte din perioada crahului cu media din eșantionul de estimare, dominată de perioada calmă
  - Pe 12 martie 2020, RV observată în ziua următoare pentru JPM a fost de aproximativ 40 de ori mai mare decît prognoza HAR filtrată
- Modificare făcută înainte de estimarea pe panelul de 50 de active
  - rezultatele principale folosesc doar regula de pozitivitate (o prognoză ≤ 0 devine media din eșantionul de estimare)
  - Filtrul BPQ rămîne ca verificare de robustețe și este raportat complet

QLIKE relativ la HAR, activ cu activ (1/2)



- **Întrebare**: depășește metoda lui Gu et al. modelele HAR și, mai ales, log-HAR?
- Raportul pentru fiecare activ: QLIKE medie a modelului împărțită la QLIKE medie a HAR
  - un raport sub 1 înseamnă un model mai precis decît HAR
  - Fiecare cutie acoperă intervalul intercuartilic pe cele 50 de active, linia din cutie este mediana, rombul este media, iar mustățile se întind pe 1,5 intervale intercuartilice

▶ Interactiv

QLIKE relativ la HAR, activ cu activ (2/2)

- Rapoartele se mediază pe active cu ponderi egale
  - aici, în ablație și în rezultatele filtrate
  - media rapoartelor diferă de raportul mediilor
  - deci la schimbarea reperului se recalculează pe fiecare activ
- Medianele descriu un activ tipic, iar mediile rămîn sensibile la cîteva pierderi mari

MFM_ch20_case_study

Interpretarea comparației QLIKE

- La orizontul de o zi ($h = 1$), log-HAR, log-HAR-K și Sig-L au rapoarte apropiate
  - 0,883, 0,885 și, respectiv, 0,886
  - Doar trăsăturile (Sig-L față de log-HAR): +0,4% la QLIKE, deci cele 36 de trăsături în plus nu aduc niciun cîștig
  - Trăsături și ponderi (Sig-LK față de log-HAR): +4,5%
- La o săptămînă și la o lună, Sig-LK rămîne în urma log-HAR
  - diferența relativă de QLIKE: +28,0% la $h = 5$
  - +92,9% la $h = 22$
  - într-o variantă fără canalul randamentelor, Sig-LK față de HAR (0,878 și 0,770) se apropie de log-HAR (0,871 și 0,765)
  - Medianele rapoartelor față de HAR (0,997 la $h = 5$ și 0,968 la $h = 22$) sînt mult mai bune decît mediile: pierderea provine de la cîteva active și zile
- HARQ și SHAR (*Semivariance HAR model*, model HAR cu semivarianțe) sînt aici inferioare modelului HAR în nivel
  - deoarece specificația principală nu aplică filtrul BPQ
  - Anexa arată că filtrul ameliorează HARQ
- Primul verdict: metoda nu depășește cel mai simplu model logaritmic
  - o egalitate aproximativă nu dovedește acuratețe egală
- Tipul de dovadă: comparație între proceduri

Comparații activ cu activ, cu corecție pentru testare multiplă

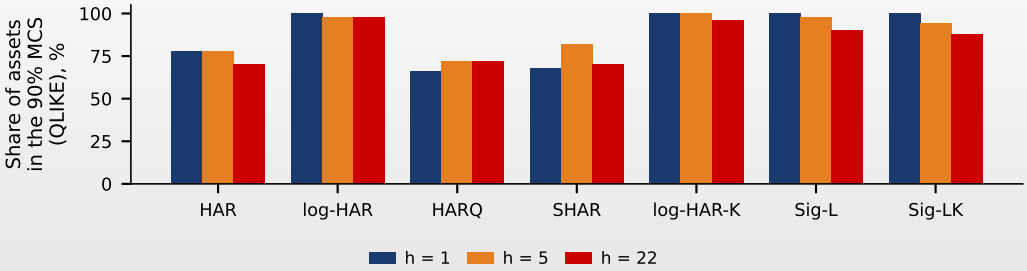
Comparația (QLIKE)	$h = 1$	$h = 5$	$h = 22$
Sig-LK față de log-HAR	2/0/0 | 5/0/0	0/0/0 | 15/0/2	0/0/0 | 27/1/16
Sig-L față de log-HAR	3/1/1 | 5/0/0	0/0/0 | 9/0/0	1/0/0 | 21/0/0
log-HAR-K față de log-HAR	3/0/0 | 3/0/0	3/0/0 | 1/0/0	5/0/0 | 6/0/0
Sig-LK față de Sig-L	2/0/0 | 5/0/0	0/0/0 | 5/0/0	2/0/0 |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

- Fiecare celulă are două părți, separate prin |
  - stînga: numărul de active semnificativ mai bune (fără corecție/Holm/BH)
  - dreapta: numărul de active semnificativ mai slabe (fără corecție/Holm/BH)
- Test DM unilateral la 5%, pe o familie de 50 de active
  - testul DM cere momente finite de ordin mai mare decît 2 și dependență slabă pentru diferența de pierdere QLIKE d_t, nu pentru randamente; prognozele apropiate de zero pot încălca aceste condiții
- Holm controlează FWER sub orice dependență
  - BH controlează FDR sub independență sau PRDS Benjamini & Yekutieli (2001)
  - PRDS: Positive Regression Dependence on a Subset, dependență pozitivă de regresie pe o submulțime

Comparații activ cu activ: lectura tabelului

- H1 (semnături): la $h = 1$, Sig-L este superior lui log-HAR pe puține active
  - pe 3 din 50 de active fără corecție
  - pe 1 după corecția Holm
  - Doar din întîmplare ne-am aștepta la aproximativ 2,5 respingeri necorectate
- Tot la $h = 1$, Sig-L este inferior lui log-HAR pe 5 active fără corecție și pe niciunul după Holm
- H2 (ponderi): nicio diferență semnificativă după Holm, la niciun orizont
  - log-HAR-K față de log-HAR
  - Sig-LK față de Sig-L
- Alegeți reperul care izolează ideea
  - față de HAR la $h = 1$, Sig-LK este superior pe 12 active
  - dar log-HAR singur este superior pe 21

Mulțimi de încredere pentru modele (1/2)



- Proporția activelor (din 50) pentru care fiecare model aparține MCS de 90% sub QLIKE
  - log-HAR, la $h = 1$, 5 și 22: 100%, 98% și 98%
  - Sig-LK: 100%, 94% și 88%
  - MCS pe QLIKE cere aceleași condiții ca testul DM pentru diferențele de pierdere dintre modele, nu pentru randamente

▶ Interactiv

Mulțimi de încredere pentru modele (2/2)

- MCS separă rar cele patru modele logaritmice
  - în schimb HAR, HARQ și SHAR sînt excluse pentru o cincime pînă la o treime dintre active
  - ținta în nivel este clar cea mai slabă alegere
- Sub MSE aproape fiecare model rămîne în mulțime
  - diferențele de pierdere medie au erori standard mari

MFM_ch20_case_study

Teste pe panel: cumularea dovezilor

Comparația: DM t (p unilateral)	$h = 1$	$h = 5$	$h = 22$
Sig-LK față de log-HAR	+1,45 (0,927)	+1,68 (0,953)	+1,80 (0,964)
Sig-L față de log-HAR	-1,91 (0,028)	+1,48 (0,931)	+1,51 (0,935)
log-HAR-K față de log-HAR	+1,89 (0,971)	+0,40 (0,654)	+1,92 (0,973)
Sig-LK față de Sig-L	+1,74 (0,959)	+0,98 (0,837)	-0,77 (0,221)
log-HAR față de HAR	-3,09 (<0,001)	-3,95 (<0,001)	-4,50 (<0,001)

- Seria testată: diferența medie de QLIKE pe activele disponibile în acea zi (panel neechilibrat)
  - un t negativ înseamnă că primul model este mai bun
  - testul DM cere momente finite de ordin mai mare decît 2 și dependență slabă pentru diferența de pierdere QLIKE d_t, nu pentru randamente; prognozele apropiate de zero pot încălca aceste condiții
- La nivel agregat, doar transformarea logaritmică este clar semnificativă
- Sig-L față de log-HAR la $h = 1$ este semnificativ într-un singur test unilateral
  - $p = 0{,}028$, în panelul neechilibrat
  - pe panelul echilibrat rezultatul dispare (slide-ul următor)
  - Doar pentru cele 40 de acțiuni semnul se inversează: $t = +1{,}72$ împotriva Sig-L
- După corecția Holm pe cele trei orizonturi (pragul $0{,}05/3 \approx 0{,}017$), rezultatul nu mai este semnificativ
- Testul GW condiționat pentru Sig-LK față de log-HAR, după corecția Holm
  - 0 respingeri din 50 de active la fiecare orizont

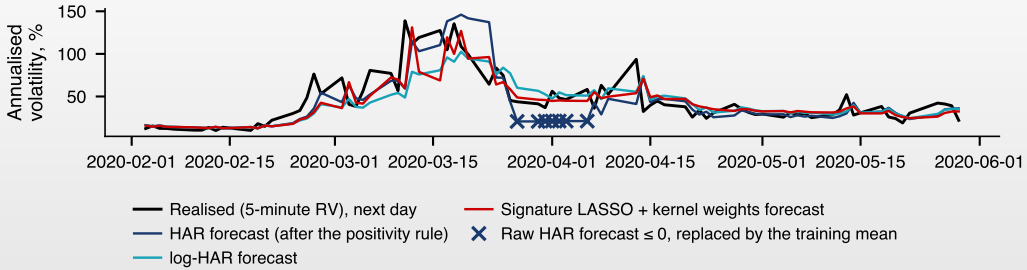
Panelul echilibrat: aceleași active în fiecare zi

Comparația: DM t (p unilateral)	$h = 1$	$h = 5$	$h = 22$
Sig-LK față de log-HAR	+1,19 (0,884)	+1,61 (0,946)	+1,41 (0,921)
Sig-L față de log-HAR	+0,75 (0,773)	+1,49 (0,932)	+1,24 (0,892)
log-HAR față de HAR	−2,21 (0,014)	−2,51 (0,006)	−4,76 (<0,001)

- Panelul neechilibrat mediază activele disponibile în fiecare zi: futures și FX din 2014, acțiunile abia din 27 ian. 2020
  - diferențele dintre clase se amestecă astfel cu diferențele dintre perioade
- Panelul echilibrat: toate cele 50 de active, doar zilele în care toate au prognoze (1 511 de zile la $h = 1$, 27 ian. 2020 – 29 ian. 2026)
- Sig-L față de log-HAR la $h = 1$: $t = +0{,}75$ ($p = 0{,}773$), față de $t = -1{,}91$ ($p = 0{,}028$) în panelul neechilibrat
  - pe calendarul comun, doar panelul celor 5 contracte futures favorizează Sig-L ($t = -4{,}03$)
- Avantajul log-HAR față de HAR se păstrează: t între $-4{,}76$ și $-2{,}21$

MFM_ch20_case_study

În crahul din pandemia COVID-19: JPM, ziua următoare (1/2)



- Toate modelele reacționează cu întîrziere la primul salt
  - volatilitatea anualizată în % a prognozelor, $\sqrt{252 F_t}$, și a țintei, $\sqrt{252\,\mathrm{RV}_{t+1}}$
  - perioada: februarie–mai 2020

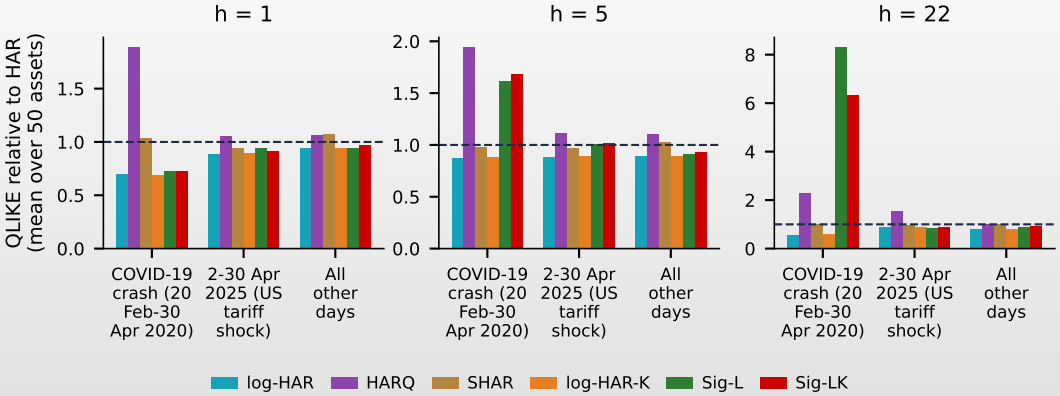
▶ Interactiv

În crahul din pandemia COVID-19: JPM, ziua următoare (2/2)

- Sig-LK reacționează puțin mai repede decît log-HAR la mijlocul lui martie
  - atunci traiectoria și-a schimbat deja direcția
  - după vîrf ambele scad cu viteze asemănătoare
- HAR în nivel depășește ținta, apoi prognoza HAR brută devine negativă la sfîrșitul lui martie
  - graficul arată valoarea după regula de pozitivitate: media țintei în fereastra de antrenare (segmentul plat)
  - cele 7 origini înlocuite, 27 martie – 6 aprilie 2020, sînt marcate cu $\times$

MFM_ch20_case_study

Perioade de stres față de perioade normale (H4) (1/2)



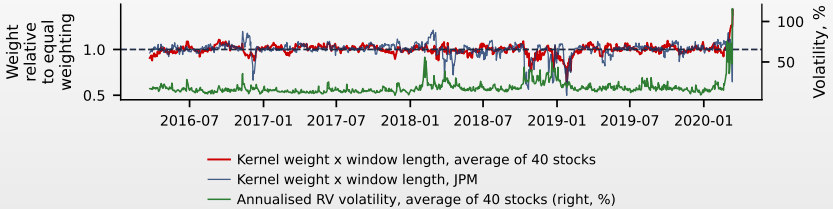
MFM_ch20_case_study ▶ Interactiv

Perioade de stres față de perioade normale (H4) (2/2)

- Media pe active a QLIKE relativ la HAR în fiecare subperioadă fixată înaintea evaluării
- La $h = 1$, în crahul COVID-19: log-HAR 0,70, Sig-LK 0,73, HARQ 1,89
- În 2–30 aprilie 2025 (șocul tarifelor americane): log-HAR 0,88, Sig-LK 0,92
- În celelalte zile: log-HAR 0,94, Sig-LK 0,97
- La $h = 22$, în crahul COVID-19, modelele cu semnături au cele mai slabe rezultate
  - pierderea Sig-LK este de 6,35 ori pierderea HAR
  - pierderea Sig-L este de 8,30 ori pierderea HAR
  - tocmai acolo unde ferestrele trecute asemănătoare ar trebui să ajute
- H4 nu este susținută
- Tipul de dovadă: descriere (arată ce s-a întîmplat în date, fără un test al cauzei)
  - fără un test formal de interacțiune

MFM_ch20_case_study

Ponderile nucleului la originea din 16 martie 2020 (1/2)



- Ponderile Sig-LK ($h = 1$) la originea din 16 martie 2020 pe zilele de antrenare
  - zilele de antrenare: 24 mar. 2016 – 13 mar. 2020
  - ponderile sînt înmulțite cu lungimea ferestrei, astfel încît valoarea 1 corespunde ponderării egale
- Ponderile sînt aproape egale, între 0,61 și 1,43 ori $1/n$
  - mărimea efectivă a eșantionului este 975 din 1 000 (media celor 40 de acțiuni)

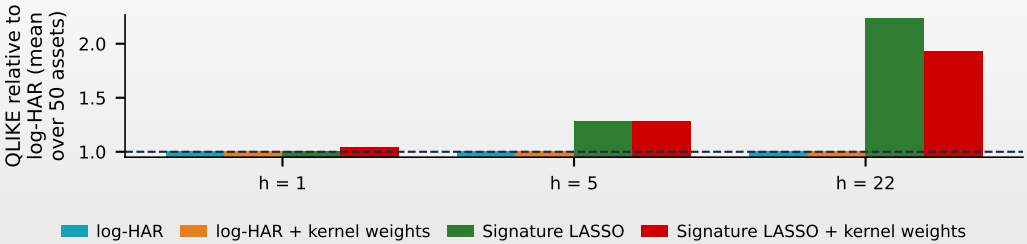
▶ Interactiv

Ponderile nucleului la originea din 16 martie 2020 (2/2)

- Cele mai mari ponderi revin ultimelor zile dinaintea originii
- Ferestrele recente de 22 de puncte se suprapun cu cea curentă
  - asemănarea poate reflecta parțial suprapunerea, nu un regim economic distinct
- Semnătura depinde doar de incremente (invarianța la translație): reflectă forma, nu nivelul
  - scăderile de la sfîrșitul lui 2018 primesc o pondere *mai mică*
  - ponderile sînt aproape necorelate cu log-volatilitatea (corelația este +0,04)

MFM_ch20_case_study

Ablația 2×2



- **Întrebare**: ce componentă produce rezultatul, trăsăturile de semnătură sau ponderile nucleului?
- Media pe cele 50 de active a QLIKE relativ la log-HAR
  - trăsături (semnătură sau nu) încrucișate cu ponderi (nucleu sau egale)
- Linia punctată marchează valoarea 1 (log-HAR)
  - o bară peste ea înseamnă o prognoză mai slabă decît log-HAR

MFM_ch20_case_study ▶ Interactiv

Ablația 2×2: valorile

- Doar ponderile (log-HAR-K): 1,002 la $h = 1$, 1,002 la $h = 5$ și 1,004 la $h = 22$, deci niciun efect detectabil asupra pierderii medii
- Doar trăsăturile (Sig-L): 1,004 la $h = 1$, 1,285 la $h = 5$ și 2,238 la $h = 22$: termenii de nivelul 3 explodează în crahuri (Aprofundări)
- Ambele componente (Sig-LK): 1,045 la $h = 1$, 1,280 la $h = 5$ și 1,929 la $h = 22$
  - Mărimea efectivă medie a eșantionului cu ponderile Sig-LK: 816 din 1 000
  - Adăugarea ponderilor la Sig-L reduce pierderea la $h = 22$, fără a o elimina
  - În schimb, înrăutățesc rezultatul la $h = 1$ din cauza petrolului (raportul pentru CL este 2,71), spre deosebire de cîștigul adus de nucleu în Tabelul 3 al lui Gu et al.

Pe clase de active

Sig-L relativ la log-HAR (QLIKE)	$h = 1$	$h = 5$	$h = 22$
Acțiuni (40)	1,009	1,332	2,505
Futures (5)	0,965	1,084	1,206
FX (5)	0,996	1,111	1,134

- Singurul cîștig: futures la $h = 1$ (-3,5%) și, marginal, FX
  - Rezultatul privește doar cinci contracte futures și un orizont: este un candidat pentru un nou studiu preînregistrat, nu o concluzie a acestuia
- Acțiunile, 80% din panel, nu arată niciun cîștig
  - media rapoartelor pe active este dominată de acțiuni
  - testul DM pe panelul neechilibrat amestecă însă clasele cu perioadele: înainte de 2020 conține doar futures și FX

Limitele acestui studiu

- Am testat o singură specificație din multe
  - $L = 22$, $N = 3$, trei canale, BIC, o grilă de patru puncte pentru c
  - un rezultat nul pentru această specificație nu este unul pentru semnături în general
- Este folosit doar nucleul trunchiat
- Lipsesc factorii externi (volatilitatea implicită, știrile macro), esențiali în Gu et al.
- Realized GARCH Hansen, Huang & Shek (2012) și modelele de referință din învățarea automată sînt în afara comparației
  - GARCH: Generalised AutoRegressive Conditional Heteroskedasticity, Capitolul 5
  - învățarea automată: Capitolul 13
- Acțiunile sînt evaluate doar pe șase ani

Direcții deschise

- Un nucleu sensibil la nivel (Seminarul 20, C2)
  - augmentarea cu punct de bază sau nivelul $\ln \mathrm{RV}$ în distanță
- Semnături ale traiectoriei intraday (randamente la 5 minute)
- Nucleul PDE Salvi et al. (2021)
- Semnături logaritmice (*log-signatures*) și canale mărginite (tanh) împotriva termenilor explozivi de nivelul 3 Morrill et al. (2020)
- Traiectorii între active, unde ariile Lévy măsoară relațiile lead-lag între piețe
- Futures la $h = 1$ pe un panel mai lung, cu această specificație fixată dinainte

Întrebarea: semnătura ca măsură lead-lag

- **Întrebare**: se mișcă un activ înaintea altuia, iar semnătura traiectoriei lor comune poate măsura acest lucru?
  - relațiile lead-lag arată pe ce piață apare mai întîi informația (descoperirea prețului, *price discovery*)
- Bennett, Cucuringu & Reinert (2022), *Machine Learning*
  - relațiile lead-lag dintre perechi de active formează o rețea orientată, împărțită apoi în grupuri de lideri și de urmăritori
  - datele: 434 de acțiuni NYSE (New York Stock Exchange) cu cel mai mare volum, randamente zilnice 2000–2019
- Una dintre măsurile comparate: dublul ariei Lévy a traiectoriei log-prețurilor normalizate pe $[t-m, t]$

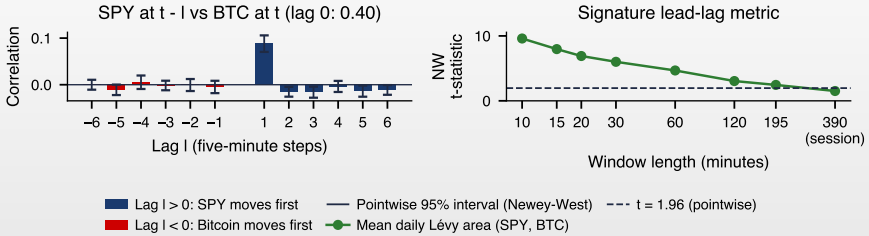
$$S^{ij} = 2A^{ij} = \iint_{t-m<u<v<t} \left(dX_u^i \, dX_v^j - dX_u^j \, dX_v^i\right)$$

  - $S^{ij} > 0$: mișcările lui i sînt urmate de mișcări în același sens ale lui j
  - spre deosebire de corelațiile încrucișate, nu cere alegerea unui lag maxim
  - presupune o asociere pozitivă: nu distinge „i conduce, cu asociere negativă" de „j conduce, cu asociere pozitivă"

Ce a găsit articolul și cu ce procedură

- Rezultatele empirice ale articolului vin din altă metrică decît semnătura (versiunea publicată, secțiunile 5–6)
  - metrica: ccf-auc (aria de sub funcția de corelație încrucișată) cu corelația de distanță (*distance correlation*), cu laguri de pînă la 5 zile
  - gruparea rețelei: Hermitian RW (RW: mers aleator)
- Grupurile lead-lag sînt semnificative la un test de permutare a ordinii zilelor
  - permutarea distruge atît dependența decalată dintre active, cît și autocorelarea fiecărui activ: singur, testul nu le separă
- În comparațiile autorilor, grupurile nu coincid cu sectoarele SIC (Standard Industrial Classification)
  - corelațiile cu volumul și cu capitalizarea sînt mici
- Cifrele articolului (secțiunea 6.3): semnalul construit pe aceste grupuri are un raport Sharpe anualizat de 0,62
  - S&P 500 are 0,40 în aceeași perioadă
  - ambele cifre sînt fără costuri de tranzacționare
- Slide-urile următoare: o ilustrare a ariei Lévy pe date intraday, nu procedura din articol

SPY și Bitcoin la 5 minute: cine se mișcă primul?



- SPY (ETF-ul pe S&P 500; ETF: Exchange-Traded Fund) și Bitcoin (BTC, în USD): bare de 5 minute, ora New York, 09:30–16:00
  - 489 de zile cu toate cele 78 de bare pentru ambele active (din 513 de zile comune), 3 sep. 2024 – 18 sep. 2026
- Stînga: corelația medie pe zile dintre randamentul SPY la $t - l$ și randamentul BTC la t; barele: intervale de 95% punctuale, cu eroarea standard Newey-West (NW) a testelor
- Dreapta: statistica t NW a mediei zilnice a ariei Lévy (SPY, BTC), pentru ferestre de la 10 minute pînă la întreaga sesiune; pragul 1,96 este punctual

MFM_ch20_leadlag ▶ Interactiv

Metoda pe scurt și diferențele față de articol

- Traiectoria zilei: sumele cumulate ale randamentelor SPY și BTC, fiecare randament împărțit la abaterea standard a activului în acea zi
  - 78 de randamente: de la deschiderea barei de 09:30 la închiderea fiecărei bare 09:30–15:55
  - normalizarea recomandată de autori: altfel aria crește odată cu volatilitatea
- Aria Lévy a unei ferestre de m pași este o sumă de produse decalate ale randamentelor x_i (SPY) și y_i (BTC)

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

  - lagul l apare de $m - l$ ori: fereastra de 2 pași (10 minute) păstrează doar lagul 1
  - o fereastră lungă adaugă multe laguri fără efect, deci zgomot
- Statistica zilnică: media ariei pe toate ferestrele de m pași din zi
  - testul: t NW cu 8 laguri pe cele 489 de valori zilnice
- Diferențe față de articol
  - două active în loc de 434 și bare de 5 minute în loc de randamente zilnice
  - fără gruparea rețelei și fără semnal de tranzacționare

Rezultatul: SPY precedă Bitcoin cu 5 minute; sensibilitatea la lungimea ferestrei

- SPY se mișcă înaintea Bitcoin cu un pas de 5 minute
  - corelația la lagul $+1$ este 0,088 (t NW $= 9{,}8$), iar la lagul -1 este $-0{,}005$ ($t = -0{,}71$)
  - testul direct al direcției: diferența dintre corelațiile la lagurile $+1$ și -1 este 0,093 (t NW $= 9{,}5$)
  - corelația contemporană este 0,40; la lagurile $+2$ pînă la $+6$ corelațiile sînt mici și negative (între $-0{,}016$ și $-0{,}004$)
- Măsura de semnătură găsește efectul doar pe ferestre scurte
  - 10 minute: $t = 9{,}6$, iar aria este pozitivă în 70% din zile
  - 2 ore: $t = 3{,}0$; întreaga sesiune (390 de minute): $t = 1{,}51$, aria pozitivă în 51% din zile
- Analiza este exploratorie: 20 de teste (12 laguri nenule și 8 ferestre)
  - p ajustat Holm: 0,032 pentru fereastra de 2 ore, 0,157 pentru 195 de minute
- Efectul scade în 2026: corelația la lagul $+1$ este 0,11 în 2024, 0,12 în 2025 și 0,04 în 2026 (ianuarie–septembrie)

Ce susține rezultatul lead-lag și limitele lui

- Tipul de dovadă: o asociere lead-lag de 5 minute între două active, nu o relație cauzală și nici un semnal profitabil
  - fără costuri de tranzacționare și fără evaluare în afara eșantionului
- Semnul ariei măsoară orientarea traiectoriei; citirea „SPY conduce” presupune o asociere pozitivă între cele două active
- Limite
  - un decalaj de un singur pas apare și din marcaje de timp nesincronizate între cele două surse; separarea cere date tick
  - lungimea ferestrei m joacă rolul lagului maxim: semnătura nu elimină această alegere
  - testul t NW cere momente finite și dependență slabă pentru statistica zilnică testată (corelații mărginite, arii normalizate), nu pentru randamentele brute

Cînd se transferă o metodă dintr-un domeniu în altul?

- Metodele circulă mai repede decît dovezile
  - un articol KDD despre tarife de transport devine candidat pentru aplicații în volatilitate, credit sau energie
  - Ce proprietăți ale problemei-sursă au făcut metoda să funcționeze?
  - Le are și problema-țintă?
- Finanțele adaugă un filtru sever Menkveld et al. (2024)
  - 164 de echipe care analizează aceleași date obțin estimări care diferă mai mult decît sugerează erorile lor standard
- Criteriu: un transfer reușește dacă cîștigul față de cel mai puternic reper simplu rămîne valabil
  - pe o mulțime preînregistrată $\mathcal{S}$ de specificații, active și orizonturi
  - cu o corecție pentru testarea multiplă aplicată familiei de teste
  - În acest capitol: $\mathcal{S} = 50$ de active $\times$ 3 orizonturi $\times$ 4 variante
  - Holm și BH s-au aplicat pe fiecare comparație și orizont (familii de 50 de active), iar variantele sînt raportate descriptiv, fără o corecție comună pe toate cele 600 de celule

Bucla descoperirii aplicată acestui capitol

- **Citim**: un asistent listează fiecare ecuație, hiperparametru și reper din articol (cereri)
  - marchează ce nu este publicat
  - fiecare element se verifică în PDF
- **Stabilim corespondența** pentru fiecare componentă
  - dacă se transferă, ce trebuie schimbat și ce reper o izolează
- **Preînregistrăm**: ipotezele, grila, testele și corecția, înaintea primului rezultat
  - sînt fixate într-un document datat
  - asistentul face ciorna, iar echipa decide fiecare cifră
- **Construim și testăm piesele**: semnăturile față de identități, nu față de altă bibliotecă
  - nivelul 1 = incrementul, Chen pe o traiectorie împărțită
- **Rulăm grila, apoi o supunem unor teste de rezistență** (cereri)
  - corecția Holm pe 50 de active
  - căutarea activului care ar da un titlu înșelător
- **Explicăm rezultatul nul** (lista de verificări: Anexă)
  - aici ponderi plate (invarianța la translație)
  - și termeni explozivi de nivelul 3 (legătura exponențială)
  - ambele verificate cu date

Mini studiu de caz: trei interpretări ale acelorași rezultate

- „Semnăturile depășesc HAR pe 12 din 50 de active" (Sig-LK, orizont de o zi, DM necorectat)
  - Adevărat, dar log-HAR singur este superior lui HAR pe 21: cîștigul aparține transformării logaritmice
- „Semnăturile îmbunătățesc prognozele de volatilitate pentru E-mini S&P 500"
  - ES, $h = 1$, DM $t = -2{,}91$ pentru Sig-LK față de log-HAR
  - O celulă din 150 de celule activ-orizont: după Holm, Sig-LK nu are niciun cîștig din 50 la $h = 1$ și niciunul la $h = 5, 22$
- Titlul rezultat din evaluarea preînregistrată:
  - „niciun avantaj robust față de log-HAR în specificația evaluată; mai slab la orizonturi mai lungi; niciun efect detectabil al ponderilor nucleului"
  - Un asistent AI (*Artificial Intelligence*, inteligență artificială) rugat să „găsească unde funcționează metoda" le va produce pe primele două; planul îl protejează pe al treilea
- Tipul de dovadă: comparație între proceduri
  - aceleași rulări, citite fără corecție, pe o singură celulă sau după planul preînregistrat

Aplicație pentru proiect: planificați un transfer

- Alegeți o direcție deschisă din acest capitol
  - nucleu sensibil la nivel, semnături intraday, semnături logaritmice, doar futures
  - Scrieți H_0/H_1, reperul care o izolează și rezultatul care v-ar face să renunțați la ea
- Folosiți un asistent pentru a lista ce poate merge prost în studiul dumneavoastră
  - apoi scrieți un test unitar pentru fiecare element
- Fixați preînregistrarea într-un document datat înaintea primului rezultat în afara eșantionului
- Rulați grila și raportați distribuția rezultatelor pe $\mathcal{S}$
- Documentați fiecare eroare AI detectată în `AI_ERRORS.md` (Seminarul 20, C1 dă erori tipice)

Idei de reținut

- Semnăturile transformă o fereastră într-un vector de trăsături cu fundament teoretic
  - identitatea lui Chen; unicitate și universalitate (detalii în Aprofundări)
- Aplicată variantei realizate, metoda lui Gu et al. nu depășește log-HAR
  - H1 (semnăturile aduc informație) nu este susținută: niciun cîștig la o zi, iar la o săptămînă și la o lună termenii de nivelul 3 iau valori explozive în crahuri
  - H2 (ponderile nucleului ajută) nu este susținută: nucleul invariant la translație reflectă forma, nu nivelul, iar ponderile sînt aproape plate
  - H3 (metoda depășește familia HAR): doar HAR în nivel este depășit, iar cel mai puternic efect din studiu este cel mai vechi: modelați logaritmul RV
  - H4 (cîștiguri mai mari în perioadele de stres) nu este susținută
- Rezultatul nul este informativ: ablațiile și mecanismul explică de ce metoda nu se transferă

Lecturi despre semnături și măsuri realizate

Lecturi

- Semnături: Chevyrev & Kormilitzin (2016); Lyons & McLeod (2022)
  - măsuri realizate: Franke, Härdle & Hafner (2019) și Capitolul 9

Cursuri avansate pe quantinar.com

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

Cod și date

- Quantlets: github.com/danpele/MFM/tree/main/Quantlets/Ch_20
  - date: VOLARE, gratuit pentru cercetare la volare.unime.it

Verificarea înțelegerii

- Ce informație despre ordine păstrează semnătura unei traiectorii și pierde o medie simplă? (exemplul cu două segmente)
- De ce comparăm atît cu o țintă în nivel, cît și cu una în logaritmi?
- Ce ablație izolează contribuția semnăturilor? (ablația 2×2)
- Dovedește un singur avantaj semnificativ între multe active superioritatea metodei?
- Ce rol are corecția pentru comparații multiple? (protocolul de evaluare)
- Cum poate fi informativ un rezultat nul? (limite și direcții deschise)

Decizia practică

- **Ce aleg**: log-HAR pentru prognoza variației realizate zilnice; metoda lui Gu et al. (Sig-LK) nu îl înlocuiește
  - log-HAR este în MCS de 90% pentru 100%, 98% și 98% dintre active, la $h = 1$, 5 și 22
  - pe panel, log-HAR bate HAR la fiecare orizont (DM t între $-3{,}09$ și $-4{,}50$)
- **În ce condiții**
  - QLIKE ca pierdere principală, cu ținta în logaritmi și corecția de variață $\hat{s}^2/2$
  - estimare pe fereastră mobilă, reestimată la fiecare 5 zile pe ultimele 1 000 de perechi
  - un panel dominat de acțiuni (40 din 50 de active)
- **Ce rezultat m-ar face să renunț**
  - un cîștig al Sig-L sau Sig-LK față de log-HAR care rămîne după corecția Holm, într-un studiu preînregistrat
  - de exemplu, cîștigul Sig-L pe futures la $h = 1$ ($-3{,}5\%$ la QLIKE), confirmat pe un panel mai lung

De ce RV la cinci minute și de unde provin datele?

- Alternative la RV
  - variația bipower (robustă la salturi)
  - nucleul realizat (robust la zgomot) Barndorff-Nielsen et al. (2008)
  - estimatorii bazați pe amplitudinea zilnică a prețului (*range*)
- Ținta acestui capitol este RV la cinci minute Liu, Patton & Sheppard (2015)
  - pe 400 de active, puțini estimatori o depășesc ca proxy al variației zilnice
- Semivariațele și cvarticitatea sînt calculate pe aceeași grilă temporală
  - deci SHAR (modelul HAR cu semivariațe) și HARQ folosesc date de intrare coerente
- Toate modelele sînt evaluate pe aceeași țintă
  - deci diferențele de pierdere provin din modele, nu din măsura volatilității
- VOLARE este construită la universitățile din Messina și Florența Cipollini et al. (2026)
  - din date tick cu tick
  - cu aceiași estimatori pentru toate clasele de active
- Arhiva este gratuită pentru cercetare la volare.unime.it
  - nu este redistribuită împreună cu acest curs
  - descărcați fișierele de variață realizată și setați în notebook-uri calea către folderul local
- Autorii cer citarea
  - Cipollini, Cruciani, Gallo, Insana, Otranto și Spagnolo (2026), arXiv:2602.19732

Două rafinări (1/2): HARQ

- RV este o măsurare cu eroare a variației adevărate

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

  - η_t: eroarea de măsurare; erorile în variabile atenuează β_d (îl trag spre zero), mai ales în zilele zgomotoase
- HARQ Bollerslev, Patton & Quaedvlieg (2016): ponderea RV de ieri scade cînd RV este măsurată mai puțin precis

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

  - $\mathrm{RQ}_t^{\mathbf{1/2}}$: rădăcina pătrată a cvarticității realizate, o măsură a erorii lui RV_t; $\overline{\mathrm{RQ}^{\mathbf{1/2}}}$: media ei în eșantion (*in-sample*), deci β_d este ponderea într-o zi obișnuită
  - Se așteaptă $\beta_Q < 0$: într-o zi zgomotoasă (RQ_t peste medie) RV de ieri primește o pondere mai mică
- Filtrul BPQ (Bollerslev, Patton și Quaedvlieg), „insanity filter"
  - o prognoză în afara intervalului țintei în eșantion este înlocuită cu media în eșantion

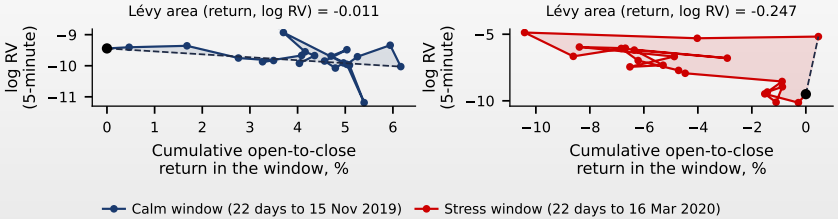
Două rafinări (2/2): SHAR

- SHAR Patton & Sheppard (2015): descompunem termenul zilnic după semnul randamentelor intraday

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

  - RV_t^+, RV_t^-: semivariațele pozitivă și negativă ale zilei t
  - $\beta^+ = \beta^-$ readuce HAR, deoarece $\mathrm{RV}_t^+ + \mathrm{RV}_t^- = \mathrm{RV}_t$
- Patton și Sheppard găsesc $\beta^- > \beta^+$
  - Varianța din scăderi prezice mai multă variață viitoare decît varianța din creșteri: o versiune intraday a efectului de levier (Capitolul 5)
- Semnătura poate surprinde un efect înrudit din date zilnice
  - aria Lévy dintre randamente și log RV (secțiunea despre semnături)

Două ferestre JPM reprezentate ca traiectorii (1/2)



- Traiectoria pe 22 de zile are două canale
  - randamentul cumulat de la deschidere la închidere și log RV
  - coarda punctată unește începutul și sfîrșitul, iar zona hașurată dă aria Lévy
- Fereastra calmă (pînă la 15 nov. 2019): traiectoria fluctuează fără direcție
  - log RV se schimbă cu -0,58
  - aria Lévy este -0,011

Două ferestre JPM reprezentate ca traiectorii (2/2)

- Fereastra de stres (pînă la 16 mar. 2020): randamentele intraday scad și apoi recuperează
  - log RV crește cu +4,32 (RV se înmulțește cu 75)
  - aria Lévy este -0,247 (randamente zecimale la calculul ariei, procente pe axă)
- Randamentul intraday net este doar +0,5%
  - crahul s-a produs în mare parte peste noapte, în afara acestui canal
  - nicio adîncime a semnăturii nu poate recupera informația absentă din date

MFM_ch20_signatures

Fundamentele teoretice ale semnăturilor (1/2): unicitate și shuffle

- **Unicitate** Hambly & Lyons (2010): semnătura completă identifică traiectoria
  - $\mathrm{Sig}(X)$ determină X până la o echivalență de tip arbore (adăugarea unor porțiuni parcurse apoi înapoi) și o translație
  - Cu un canal de timp strict crescător nimic nu poate fi parcurs înapoi, deci rămîne doar translația
  - O trunchiere finită Sig^N nu este injectivă: două traiectorii diferite pot avea aceleași niveluri $1..N$
- **Identitatea shuffle**: un produs de termeni ai semnăturii este o sumă de termeni superiori

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

  - Suma se face pe toate intercalările (shuffle) ale celor două cuvinte care păstrează ordinea din fiecare cuvînt
  - Deci un model liniar în $\mathrm{Sig}^N(X)$ conține deja polinoame în incremente

Fundamentele teoretice ale semnăturilor (2/2): universalitate și descreștere

- **Universalitate**: orice funcție continuă de traiectorie este aproximativ liniară în semnătură
  - rezultatul vine din Stone–Weierstrass, prin identitatea shuffle

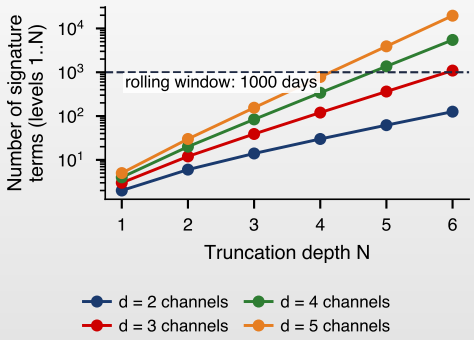
$$\forall \varepsilon > 0 \;\; \exists\, \ell \text{ liniară} : \quad \sup_{X \in K} \left| f(X) - \langle \ell, \mathrm{Sig}(X) \rangle \right| < \varepsilon$$

  - K: o mulțime compactă de traiectorii augmentate cu timpul, cu start comun (în topologia 1-variației); $f : K \to \mathbb{R}$: o funcție continuă de traiectorie
  - $\langle \ell, \mathrm{Sig}(X) \rangle$: o regresie liniară pe un număr finit de termeni ai semnăturii
  - O adîncime suficientă permite aproximarea neliniară, dar nu garantează precizia la adîncimea 3 sau în afara eșantionului
  - Costul este de d^N trăsături, deci este necesară o soluție cu puțini coeficienți nenuli
- **Descreștere factorială**: nivelurile înalte sînt mici
  - $\|X\|_{\text{1-var}}$: lungimea totală a traiectoriei; $\|S^{(k)}\|$: norma euclidiană a nivelului k

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

  - Trunchierea la o adîncime mică N pierde puțin pentru traiectorii de lungime moderată, motiv pentru care canalele se rescalează (Anexă)

Numărul de trăsături: dimensiunea trunchierii



- Numărul de termeni din nivelurile $1..N$ ale unei traiectorii d-dimensionale
  - o sumă geometrică, deoarece nivelul k are d^k cuvinte
  - $\dim(d, N) = \sum_{k=\mathbf{1}}^{N} d^k = \frac{d^{N+\mathbf{1}} - d}{d - \mathbf{1}}$
- În acest capitol: $3 + 9 + 27 = 39$ de termeni
  - $d = 3$ canale (timp, log RV, randament cumulat), adîncimea $N = 3$
  - Trei sînt constanți (cuvintele doar cu litera timpului), deci 36 variază
- $d = 5$, $N = 4$ ar da 780 de termeni pentru 1 000 de observații
  - selecția LASSO ar deveni instabilă între mulți termeni corelați
- Aici $N = 3$ și $L = 22$ sînt fixate dinainte
  - $N = 2$ servește ca verificare de robustețe

MFM_ch20_signatures

Două invarianțe și punctul de bază

- **Invarianța la translație**: $\mathrm{Sig}(X + v) = \mathrm{Sig}(X)$, deci contează doar incrementele
  - $v \in \mathbb{R}^d$: orice vector constant
  - Semnătura nu reflectă **nivelul** volatilității, deci nivelul intră separat: termenii log-HAR joacă rolul factorilor externi x_τ ai lui Gu et al.
- **Augmentarea cu punct de bază** Morrill et al. (2020)
  - traiectoria pornește din origine, apoi parcurge punctele ferestrei
    $$\hat{X} = (0,\ X_{u_{\mathbf{0}}},\ X_{u_{\mathbf{1}}}, \ldots, X_{u_{L-\mathbf{1}}}), \qquad \mathrm{Sig}(\hat{X}) = \exp_{\otimes}(X_{u_{\mathbf{0}}}) \otimes \mathrm{Sig}(X)$$
  - Nivelul 1 devine $X_{u_{\mathbf{0}}} + (X_{u_{L-\mathbf{1}}} - X_{u_{\mathbf{0}}}) = X_{u_{L-\mathbf{1}}}$: punctul de start nu apare separat, dar termenii de ordin superior depind de nivelul absolut (de exemplu $\ln \mathrm{RV}$), deci semnătura devine sensibilă la nivel
  - Translatarea traiectoriei ca să pornească din 0 nu are acest efect: este ea însăși o translație (Seminarul 20, C2 folosește segmentul adăugat la început)
- **Invarianța la reparametrizare**: $X_{\psi(u)}$, cu ψ crescătoare, are aceeași semnătură
  - fără canal de timp, două evoluții cu aceeași formă spațială, parcurse în ritmuri diferite, au aceeași semnătură
  - canalul de timp le distinge raportîndu-le la un ceas fix; traiectoria augmentată rămîne invariantă la reparametrizarea ei în ansamblu
  - $u_i = i/(L-1)$ duce fiecare fereastră pe $[0, 1]$: durata absolută (22 față de 2 zile) cere timp fizic sau un regresor separat
- În acest capitol: augmentare cu timpul, fără punct de bază
  - deci nucleul compară forme, nu niveluri
  - consecința apare la ponderile nucleului, în secțiunea Rezultate

Trei ferestre: ponderi și informație

- Distanțele $\delta = (0{,}1;\ 0{,}4;\ 0{,}9)$ și $\gamma = 2$
- Ponderile nenormalizate: $(0{,}819;\ 0{,}449;\ 0{,}165)$, cu suma 1,433
- Ponderile normalizate: $w = (0{,}571;\ 0{,}313;\ 0{,}115)$
- $n_{\text{eff}} = 1/\sum w_i^2 \approx 2{,}28$, față de 3 pentru ponderi egale
- γ acționează ca o temperatură inversă
  - un γ mai mare concentrează ponderile
  - cîștigăm asemănare și pierdem informație

Protocolul de evaluare: varianța HAC

- De ce nu varianța empirică simplă: d_t este autocorelată
  - ținte suprapuse pînă la lagul $h-1$, gruparea volatilității (*volatility clustering*) în pierderi
  - $\hat{\Gamma}_{\mathbf{0}}/T$ subestimează incertitudinea lui $\bar{d}$ (și respinge prea des) cînd suma autocovarianțelor este pozitivă, cazul tipic al țintelor suprapuse
- Estimatorul Newey–West Newey & West (1987), robust la heteroscedasticitate și autocorelare (HAC)

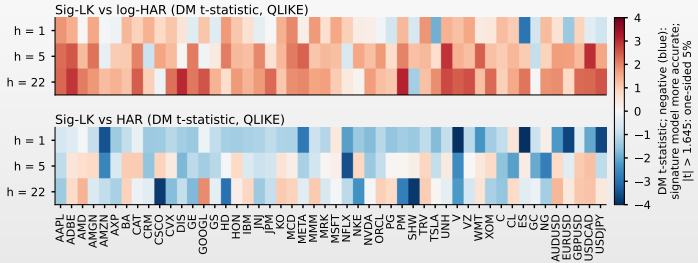
$$\hat{\Omega} = \hat{\Gamma}_{\mathbf{0}} + 2\sum_{j=\mathbf{1}}^{q}\Big(1-\frac{j}{q+1}\Big)\hat{\Gamma}_j, \qquad \hat{\Gamma}_j = \frac{1}{T}\sum_{t=j+\mathbf{1}}^{T}(d_t-\bar{d})(d_{t-j}-\bar{d})$$

  - $\hat{\Gamma}_j$: autocovarianța empirică a lui d_t la lagul j (nu temperatura γ a nucleului)
  - Ponderile Bartlett $1-j/(q+1)$ scad liniar cu lagul și garantează $\hat{\Omega}\geq 0$
  - $q=\max(h,\lceil T^{\mathbf{1/3}}\rceil)$ laguri: cel puțin h, pentru dependența creată de suprapunere ($\lceil\cdot\rceil$: rotunjire în sus)

Două exemple numerice: Holm, BH și MCS

- Holm compară trei valori p ordonate, (0,004; 0,023; 0,08)
  - succesiv cu (0,05/3; 0,05/2; 0,05)
  - primele două trec, a treia nu, deci ne oprim
- BH le compară cu (0,05/3; $2 \cdot 0{,}05/3$; 0,05)
  - alege cel mai mare rang care trece: tot două
- Într-o familie de 50 de teste, aceleași valori p pot să nu mai treacă
  - de aceea familia se fixează înaintea rezultatelor
- MCS cu trei modele A, B și C
  - diferențele standardizate de pierdere: 2,4; −0,5; −0,7
  - valoarea p bootstrap este 0,04, deci la 10% eliminăm A
- Recalculăm testul doar pentru B și C
  - noua valoare p este 0,30, deci ne oprim și rămîn B și C
- Faptul că testul nu le separă nu dovedește că pierderile lor în populație sînt egale
- Acoperirea de 90% este o garanție pe eșantioane repetate
  - nu o probabilitate a posteriori pentru mulțimea obținută

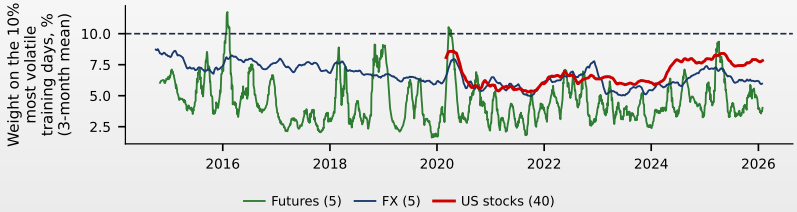
Statisticile DM pentru fiecare activ și orizont



- Statisticile t Diebold–Mariano ale diferenței de QLIKE, cîte o coloană pentru fiecare activ
  - 40 de acțiuni, 5 futures, 5 perechi FX
  - albastru: Sig-LK este mai precis; roșu: mai puțin precis
- Față de log-HAR (sus): culori palide la $h = 1$, iar la $h = 5, 22$ predomină roșul
  - excepțiile albastre: ES, TSLA (Tesla) și EURUSD
  - cinci active semnificativ roșii înaintea corecției
- Față de HAR (jos): albastru la $h = 1$ pentru că Sig-LK moștenește ținta logaritmică

MFM_ch20_case_study

Ponderea perioadelor de stres în nucleu (1/2)



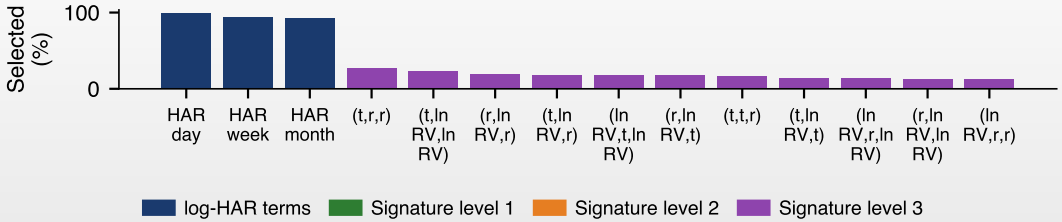
- Ponderea pe care Sig-LK ($h = 1$) o atribuie celor mai volatile 10% dintre zile
  - zilele din fereastra de antrenare
  - mediată pe clasă și pe intervale de trei luni
- Pentru acțiuni, ponderea rămîne sub cele 10% date de ponderarea egală
  - la fiecare reestimare, media este 6,7%
  - variază între 4,4% și 9,9% (maximul pe 17 mar. 2020)

Ponderea perioadelor de stres în nucleu (2/2)

- FX rămîne sub 10% în tot eșantionul
- Futures depășesc 10% (media pe trei luni) doar în două episoade scurte
  - începutul lui 2016 (pînă la 11,7%)
  - martie–aprilie 2020
- Mecanismul din spatele rezultatului nul pentru H2
  - ponderile nu implementează „învățăm din crizele trecute"

MFM_ch20_case_study

Variabilele selectate de LASSO (1/2)



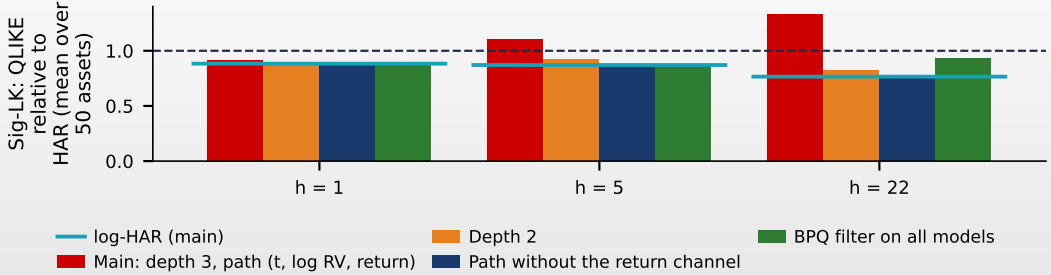
- Proporția reestimărilor în care un regresor face parte din suportul ales prin BIC pentru Sig-LK
  - pe toate activele și orizonturile
  - etichete: HAR day/week/month = cei trei termeni log-HAR, t = timp, ln RV = $\ln \mathrm{RV}$, r = randament cumulat
- LASSO redescoperă HAR
  - termenii log-HAR zilnic, săptămînal și lunar sînt păstrați în 99%, 94% și, respectiv, 93% dintre reestimări
- Suportul mediu are 8,7 regresori pentru Sig-L și 6,7 pentru Sig-LK

Variabilele selectate de LASSO (2/2)

- Cel mai frecvent termen al semnăturii este (t,r,r) (în 27% dintre reestimări)
  - toți termenii frecvenți sînt de nivelul 3
- Fiind cubici în incremente, într-un crah ajung mult în afara intervalului de antrenare
  - și sînt amplificați de $\exp(\cdot)$

MFM_ch20_case_study

Robustețea Sig-LK



- Sig-LK relativ la HAR (media pe 50 de active)
  - în specificația principală
  - în două variante preînregistrate
  - cu filtrul BPQ (abaterea declarată)
- Linia de referință marchează log-HAR, iar o bară sub 1 înseamnă o prognoză mai precisă decît HAR

MFM_ch20_case_study

Robustețea Sig-LK: variantele

- Cu adîncimea 2, rapoartele sînt 0,922 la $h = 5$ și 0,820 la $h = 22$
  - față de 1,104 și 1,334 în specificația principală
- Fără canalul randamentelor: 0,878 la $h = 5$ și 0,770 la $h = 22$
  - ablațiile sugerează că eșecurile la orizont lung sînt legate de termenii care conțin canalul randamentelor
- Adîncimea 2 le reduce și ea: ablațiile sugerează un rol al termenilor de nivelul 3
  - eliminarea unor termeni schimbă însă și selecția, distanțele nucleului și ponderile
- La $h = 5$ și $h = 22$, varianta fără randamente se apropie de log-HAR (0,871, 0,765)
  - adîncimea 2 rămîne inferioară
- Ambele variante reduc instabilitatea, fără a confirma H1

Trăsături și scalare

◂ Parcurs ◂ Descreștere ◂ Nucleu

- Regresorii în ziua τ pentru modelele cu semnături: 3 termeni log-HAR + 39 de termeni ai semnăturii
  - $\xi_\tau = \left[\ln \mathrm{RV}_\tau,\ \ln \mathrm{RV}_\tau^{(w)},\ \ln \mathrm{RV}_\tau^{(m)},\ \mathrm{Sig}^3(X_{\tau-21:\tau})\right]$, traiectoria X = (timp, $\ln \mathrm{RV}$, r cumulat), 22 de puncte
- Fiecare canal q (altul decît timpul) se înmulțește cu $\kappa_q = 1/(s_q\sqrt{L-1})$
  - astfel încît deplasarea lui totală tipică pe fereastră este aproximativ 1
  - s_q: abaterea standard a incrementelor zilnice ale canalului q, de la începutul ferestrei de antrenare pînă la originea t ($W + h$ zile, W: lungimea ferestrei de antrenare)
  - $L - 1 = 21$: numărul de increment dintr-o traiectorie, folosit doar în factorul $\sqrt{L-1}$
  - Timpul nu se rescalează (increment constante) și ia valori de la 0 la 1
- Înaintea LASSO fiecare coloană este standardizată cu momentele ponderate ale ferestrei de antrenare
  - coloanele constante se elimină
  - coloane constante: cuvintele formate doar din litera timpului
- Variabilele brute din τ folosesc date pînă la τ
  - scalele și standardizarea folosesc date pînă la originea t

Estimare: alegerea lui λ fără deplasarea de anticipare

- În fiecare fereastră de antrenare: întreaga traiectorie de regularizare LASSO ponderată, prin LARS
  - Pentru fiecare suport de pe traiectorie: reestimare OLS ponderată (pasul 2 din Gu et al.)
- Alegem suportul care minimizează un criteriu de tip BIC
  - cu mărimea efectivă a eșantionului Kish
  - o euristică a acestui studiu
  - $\mathrm{BIC}(S) = n_{\mathrm{eff}} \ln \hat{s}_S^2 + (|S|+1) \ln n_{\mathrm{eff}}, \quad n_{\mathrm{eff}} = 1/\sum_\tau w_\tau^2$
  - S: un suport (mulțimea regresorilor păstrați), cu mărimea $|S|$ ($+1$ pentru constantă); $\hat{s}_S^2$: varianța reziduală ponderată a reestimării pe S
- Validarea încrucișată ar cere 50 de active $\times$ 3 orizonturi $\times$ sute de reestimări
  - în timp ce BIC folosește doar fereastra
  - Mărimea Kish corectează ponderile inegale, nu dependența serială (ținte suprapuse, $h > 1$), deci criteriul nu este un BIC bazat pe verosimilitate pentru aceste date

Înapoi

Perioadele de evaluare

Clasa	Evaluarea începe	Se termină	Zile pe activ (mediana)
Acțiuni	27 ian. 2020	29 ian. 2026	1 499
Futures	25 sept. 2014	29 ian. 2026	2 919
FX	27 aug. 2014	29 ian. 2026	2 952

- 265 929 de prognoze pentru fiecare model, pe cele trei orizonturi
  - 89 893 ($h = 1$), 89 293 ($h = 5$), 86 743 ($h = 22$)
- Acțiunile încep în 2015, deci evaluarea lor începe în ianuarie 2020
  - crahul din pandemia COVID-19 este în eșantionul fiecărei acțiuni
  - Futures și FX încep în 2009 și sînt evaluate din 2014

înapoi

Temperaturile alese ale nucleului

	Sig-LK			log-HAR-K		
Clasa	$h = 1$	$h = 5$	$h = 22$	$h = 1$	$h = 5$	$h = 22$
Acțiuni	0,5	0,5	0,5	0,5	0,5	1,0
Futures	2,0	1,0	0,5	2,0	1,0	4,0
FX	0,5	2,0	0,5	4,0	2,0	4,0

- Valoarea lui c din grila $\{0{,}5; 1; 2; 4\}$
  - aleasă pe blocul de validare al fiecărei clase și al fiecărui orizont

înapoi

Metrica articolului pe datele VOLARE

Mediana erorilor medii pe active $\lvert \mathrm{RV} - F \rvert / \mathrm{RV}$ (%)	$h = 1$	$h = 5$	$h = 22$
HAR	57,2	48,9	51,7
log-HAR	50,4	41,9	38,7
HARQ	55,0	49,0	50,7
SHAR	57,9	49,2	51,8
log-HAR-K	49,7	41,0	38,7
Sig-L	50,1	42,6	42,6
Sig-LK	50,6	43,6	43,7

- Mediana pe active a erorii relative medii: aproximativ 40–60%
  - față de 1–2% pentru tarifele de transport la Gu et al.
- RV zilnică este o țintă zgomotoasă, cu cozi groase
  - modelele evaluate au erori relative mult mai mari decît cele raportate pentru tarifele de transport
  - aceasta nu stabilește o limită inferioară a erorii posibile
  - Numărul prognozelor Sig-LK care depășesc de zece ori maximul în eșantion: 8 la $h = 1$, 29 la $h = 5$ (concentrate pe 10 active) și 36 la $h = 22$
  - Log-HAR nu are nicio astfel de prognoză

Înapoi

Filtrul „insanity”, revăzut

QLIKE relativ la HAR, filtrul BPQ pe toate modelele	$h = 1$	$h = 5$	$h = 22$
HAR	1,000	1,000	1,000
log-HAR	0,763	0,654	0,778
HARQ	0,995	0,912	1,014
SHAR	1,014	1,003	1,000
log-HAR-K	0,772	0,659	0,779
Sig-L	0,871	0,819	0,904
Sig-LK	0,890	0,867	0,930

- Filtrul îmbunătățește HARQ față de HAR, așa cum au urmărit Bollerslev, Patton și Quaedvlieg
  - 0,995 la $h = 1$ în loc de 1,140
- Sig-LK relativ la log-HAR, ambele filtrate: 1,238 la $h = 1$, 1,424 la $h = 5$ și 1,209 la $h = 22$
- Cu filtru, distanța crește la $h = 1, 5$ și scade la $h = 22$
- Filtrul se aplică modelelor cu semnături mult mai des decît lui log-HAR
  - prognoze înlocuite la $h = 1$ (din 89 893): HAR 180, HARQ 140, Sig-LK 143, log-HAR 29

înapoi

Verificările cercetătorului

- **Metoda este cea din articol**: nucleul, ponderile, reestimarea, ca în ecuațiile citate
  - Test: fiecare componentă a metodei indică ecuația, tabelul sau secțiunea articolului din care provine, iar deciziile nepublicate sînt marcate ca alegeri ale studiului
- **Fără deplasarea de anticipare** (*look-ahead bias*): λ, scalele, c și trăsăturile folosesc doar date disponibile la origine
  - Test: dacă modificați orice valoare de după origine (de exemplu, înmulțiți RV viitoare cu 10), prognoza făcută la origine nu trebuie să se schimbe
- **Reperul izolează ideea**: ablația 2×2 este raportată, nu doar modelul complet față de HAR
- **Funcțiile de pierdere și testele sînt aplicate corect**: $\mathrm{QLIKE}(\mathrm{RV}, \mathrm{RV}) = 0$ și $\mathrm{QLIKE}(\mathrm{RV}, F) > 0$ în rest
  - Varianța DM este HAC, cu lățimea de bandă (*bandwidth*) de cel puțin h

Înapoi

Bucla descoperirii, cererile (1/2)

- **Citim**: extragem exact ce a făcut articolul
  - „Din Gu et al. (2024, KDD), listează fiecare ecuație a estimatorului, fiecare hiperparametru și valoarea raportată și reperul fiecărui tabel. Marchează ce nu este publicat.”
  - Verificați fiecare element în PDF: numerele ecuațiilor, ale tabelelor, nucleul folosit efectiv
- **Stabilim corespondența**: ce se transferă, ce trebuie adaptat
  - „Ținta noastră este varianța realizată zilnică. Pentru fiecare componentă a metodei, spune dacă se transferă, ce trebuie schimbat și ce reper o izolează.”
- **Preînregistrăm**: ipoteze, grilă, teste, corecție
  - fixate într-un document datat și semnat de echipă înaintea primului rezultat

Înapoi

Bucla descoperirii, cererile (2/2)

- **Construim și testăm piesele**: semnăturile față de identități, nu față de altă bibliotecă
  - „Scrie teste unitare pentru o semnătură de adîncime 3: nivelul 1 egal cu incrementul, partea simetrică a nivelului 2 egală cu jumătate din produsul exterior, identitatea lui Chen pe o traiectorie împărțită.”
- **Rulăm grila, apoi o supunem unor teste de rezistență**
  - „Iată toate cele 600 de rulări (50 de active × 3 orizonturi × 4 specificații: principală, adîncimea 2, fără canalul randamentelor, filtrul BPQ). Ce concluzii rămîn valabile după corecția Holm pe 50 de active? Ce activ sau ce subperioadă ar da un titlu înșelător?”
- **Explicăm rezultatul nul**: identificăm mecanismul, nu o justificare

Înapoi

Acronime folosite în acest material (1/3)

- **AAPL**: Apple Inc. (Nasdaq ticker) — *Apple Inc., simbolul bursier Nasdaq*
- **ACM**: Association for Computing Machinery — *Asociația pentru Mașini de Calcul*
- **AI**: Artificial Intelligence — *inteligență artificială*
- **AR**: AutoRegressive model — *model autoregresiv*
- **ARIMA**: AutoRegressive Integrated Moving Average (model) — *model autoregresiv integrat cu medie mobilă*
- **AUD**: Australian Dollar (ISO 4217 code) — *dolarul australian (cod ISO 4217)*
- **BH**: Benjamini–Hochberg (procedure) — *procedura Benjamini–Hochberg*
- **BIC**: Bayesian Information Criterion — *criteriul informațional bayesian*
- **BPQ**: Bollerslev–Patton–Quaedvlieg (insanity filter of the HARQ paper) — *Bollerslev–Patton–Quaedvlieg (filtrul din articolul HARQ)*
- **CAD**: Canadian Dollar (ISO 4217 code) — *dolarul canadian (cod ISO 4217)*
- **CL**: Crude oil (WTI light sweet) futures, NYMEX ticker — *contractul futures pe petrol (WTI), simbolul NYMEX*
- **COVID-19**: COronaVIrus Disease 2019 — *boala provocată de coronavirus, 2019*
- **DM**: Diebold–Mariano (test) — *testul Diebold–Mariano de comparare a prognozelor*
- **ES**: E-mini S&P 500 futures (CME ticker) — *contractul futures E-mini S&P 500 (simbolul CME)*
- **ETF**: Exchange-Traded Fund — *fond tranzacționat la bursă*
- **EURUSD**: Euro / US dollar exchange rate (FX pair code) — *cursul euro / dolar american (codul perechii FX)*

Acronime folosite în acest material (2/3)

- **FDR**: False Discovery Rate — *rata descoperirilor false*
- **FWER**: Family-Wise Error Rate — *probabilitatea de cel puțin o eroare de tip I în familia de teste*
- **FX**: Foreign eXchange — *piața valutară*
- **GARCH**: Generalised AutoRegressive Conditional Heteroskedasticity — *heteroscedasticitate condiționată autoregresivă generalizată*
- **GC**: Gold futures, COMEX ticker — *contractul futures pe aur, simbolul COMEX*
- **GW**: Giacomini–White (test) — *testul Giacomini–White*
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors) — *erori standard robuste la heteroscedasticitate și autocorelație*
- **HAR**: Heterogeneous AutoRegressive (model) — *model autoregresiv heterogen*
- **HARQ**: HAR model with Realised Quarticity — *model HAR cu cvarticitatea realizată*
- **JPM**: JPMorgan Chase & Co. (NYSE ticker) — *JPMorgan Chase & Co. (simbolul bursier NYSE)*
- **KDD**: Knowledge Discovery and Data Mining (ACM SIGKDD conference) — *descoperirea cunoașterii și data mining (conferința ACM SIGKDD)*
- **LARS**: Least-Angle RegreSsion (algorithm) — *algoritmul regresiei unghiului minim*
- **LASSO**: Least Absolute Shrinkage and Selection Operator — *regresie penalizată L1 cu selecția variabilelor*
- **MCS**: Model Confidence Set — *mulțimea de încredere a modelelor*
- **MSE**: Mean Squared Error — *eroarea pătratică medie*
- **NG**: Natural Gas futures, NYMEX ticker — *contractul futures pe gaze naturale, simbolul NYMEX*

Acronime folosite în acest material (3/3)

- **NVDA**: NVIDIA Corporation (Nasdaq ticker) — *NVIDIA Corporation, simbolul bursier Nasdaq*
- **NW**: Newey–West (standard errors) — *erori standard Newey–West*
- **NYSE**: New York Stock Exchange — *Bursa din New York*
- **OLS**: Ordinary Least Squares — *metoda celor mai mici pătrate*
- **PDE**: Partial Differential Equation — *ecuație cu derivate parțiale*
- **PRDS**: Positive Regression Dependence on a Subset (of p-values) — *dependența pozitivă de regresie pe o submulțime (a p-value-urilor)*
- **QLIKE**: Quasi-LIKElihood (loss function) — *funcția de pierdere de tip cvasi-verosimilitate*
- **RQ**: Realised Quarticity — *cvarticitatea realizată*
- **RV**: Realised Variance — *varianța realizată*
- **SHAR**: Semivariance HAR model — *model HAR cu semivarianțe*
- **SUA**: Statele Unite ale Americii
- **TSLA**: Tesla, Inc. (Nasdaq ticker) — *Tesla, Inc., simbolul bursier Nasdaq*
- **USDJPY**: US dollar / Japanese yen exchange rate (FX pair code) — *cursul dolar american / yen japonez (codul perechii FX)*
- **VOLARE**: VOLatility Archive for Realized Estimates (database) — *arhiva de estimatori realizați ai volatilității (bază de date)*
- **WTI**: West Texas Intermediate (crude oil benchmark) — *țițeiul de referință West Texas Intermediate*
- **XOM**: Exxon Mobil Corporation (NYSE ticker) — *Exxon Mobil Corporation, simbolul bursier NYSE*

Bibliografie (1/3)

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

Bibliografie (2/3)

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

Bibliografie (3/3)

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