








Modelling Financial Markets

Seminar 14: Deep Learning and Time-Series Foundation Models — Hands-on



Daniel Traian PELE
Drd. Antoaneta Amza, antoaneta.amza@csie.ase.ro

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

Better forecasts or just more complex models?

- Question of the seminar
  - **do deep-learning and foundation-model forecasts survive correct inference?**
- Begin with A1–A4 and consult their linked prerequisites as needed; A1 and A2 use VaR (Value-at-Risk) and backtesting from lectures 7 and 8
- Deliverables
  - answers to A1–A4; the extension packs are optional
  - one table and one chart for B1–B4
  - the corrected C1 code
  - one interpretation sentence per exercise
- Notebook for today: open the seminar notebook in Colab
  - the notebook sections are named after the exercises (A1, B1, . . .)
- Exercises, their types and their order: the Exercise Map
- Seminar: Prof. Daniel Traian Pele, danpele@ase.ro; Antoaneta Amza, PhD candidate, antoaneta.amza@csie.ase.ro

Data Used

Series	Symbol, field	Calendar	Test period	Used in
S&P 500	`GSPC.INDX`, close	US trading days	04.01.2016–18.09.2026 (2 693)	A1, A3, B1, B3
Bitcoin	`BTC-USD.CC`, close	every day	01.01.2019–18.09.2026 (2 818)	B4
BET	`BET`, close	BVB trading days	04.01.2016–18.09.2026 (2 675)	B4
BET-tracking ETF	`TVBETETF.RO`, adjusted close	BVB trading days	common days with BET	B4
SPY 5-minute bars	`SPY.US`, 5-minute close	US trading hours	11.10.2022–18.09.2026 (979)	B2

- Daily prices of the course data, one series per asset (field used: close / adjusted close); daily log returns in % (Notation (1/2)) on each series' own calendar
- The one-day-ahead forecasts computed for Chapter 14, saved with the chapter: returns (B1, B4), realised variance (B2), VaR and ES (B3); every forecast uses data up to the previous close

Setup: Run and Verify

```
# notebook section "Setup" defines
# load_csv(name): a forecast table
d = load_csv('ch14_returns_sp500.csv')
print(d.shape, d.index[0].date(),
      d.index[-1].date())
print(d[['y', 'hist', 'lstm',
         'chronos2_q0.50']].head(3).round(4))
rv = load_csv('ch14_rv.csv')
print(rv.shape)
```

- Expected output: (2693, 51), 2016-01-04 to 2026-09-18; `rv`: (979, 12)

date	y	hist	lstm	$q_{0.50}$
2016-01-04	−1.5422	0.0265	0.0595	0.0988
2016-01-05	0.2010	0.0263	0.1044	0.0869
2016-01-06	−1.3202	0.0263	0.0587	0.0861

- `y`: the realised return (%); other columns: forecasts made the previous evening
- If your shape or first row differs, stop and check the file name

Notation Used in Every Exercise (1/2): Returns, R^2 and the DM Test

- r_t: daily log return in %, $r_t = 100 \ln(P_t / P_{t-1})$, P_t the close of day t; $\hat{r}_t$: its forecast made the evening before; portfolio loss $L_t = -r_t$; L with a model index (L_t^A, $L_{k,t}$): a forecast loss
- **Out-of-sample R^2** against the zero forecast

$$R^2 = 1 - \frac{\sum_t (r_t - \hat{r}_t)^2}{\sum_t r_t^2}$$

  - $R^2 > 0$: smaller squared errors than "always zero"; it can be negative, unlike the in-sample R^2
- **DM test** (Diebold–Mariano, Diebold & Mariano (1995)): $d_t = L_t^A - L_t^B$, the loss of forecast A minus that of B over T days

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{LRV}}/T}}, \qquad \widehat{\mathrm{LRV}} = \hat{\gamma}_0 + 2\sum_{j=1}^{q} \left(1 - \frac{j}{q+1}\right)\hat{\gamma}_j$$

  - $\bar{d}$: mean of d_t; $\hat{\gamma}_j$: autocovariance of d_t at lag j; $\widehat{\mathrm{LRV}}$: Newey–West long-run variance with q lags
  - approximately $N(0,1)$ under equal accuracy; negative: A better; $|\mathrm{DM}| > 1.96$ rejects at 5%

Notation Used in Every Exercise (2/2): Variance Losses and Tail Risk

- RV: realised variance, the sum of squared 5-minute returns of a day; $\hat{h}_t$: a variance forecast
- **QLIKE loss** (quasi-likelihood, Patton (2011)): 0 only for a perfect forecast

$$L_t = \frac{\mathrm{RV}_t}{\hat{h}_t} - \log \frac{\mathrm{RV}_t}{\hat{h}_t} - 1$$

- **VaR and ES (Expected Shortfall)** at level α (Chapter 7), from the quantiles q_u of r_{t+1}, $u \in (0, 1)$

$$\mathrm{VaR}_\alpha = -q_\alpha, \qquad \mathrm{ES}_\alpha = -\frac{1}{\alpha}\int_0^\alpha q_u \, du$$

  - positive numbers (losses); ES averages all quantiles below q_α, so $\mathrm{ES}_\alpha \geq \mathrm{VaR}_\alpha$
  - breach indicator $I_t = 1\{L_t > \mathrm{VaR}_t\}$; a correct VaR has $P(I_t = 1) = \alpha$
- **Kupiec test** (Kupiec (1995)): x breaches in T days, observed rate $\hat{\pi} = x/T$

$$\mathrm{LR}_{uc} = -2\big[(T - x)\ln(1 - \alpha) + x \ln \alpha - (T - x)\ln(1 - \hat{\pi}) - x \ln \hat{\pi}\big] \sim \chi^2_1$$

  - $\mathrm{LR}_{uc} > 3.84$ rejects correct coverage at 5%, both for too many and for too few breaches

Neural Networks and Gradient Descent

- **A layer** of a neural network

$$h = \varphi(Wx + b), \qquad \tanh(a) = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}}, \qquad \mathrm{ReLU}(a) = \max(a, 0)$$

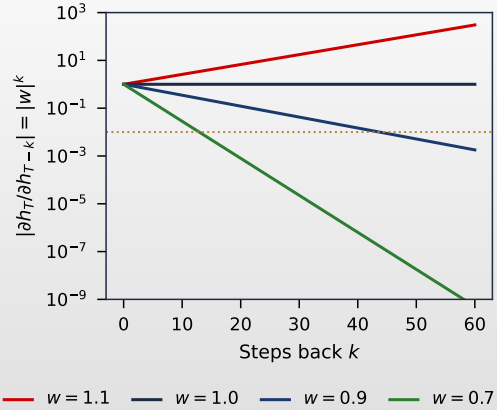
  - x: inputs; W: weight matrix; b: bias vector; φ: activation, applied element by element
  - tanh maps to $(-1, 1)$, with derivative $1 - \tanh^{2}(a) \leq 1$

- **Training**: gradient descent on a loss $\mathcal{L}(\theta)$, e.g. the mean squared error (MSE), θ = all weights

$$\theta_{k+1} = \theta_k - \eta \, \nabla_\theta \mathcal{L}(\theta_k)$$

  - $\eta > 0$: learning rate; Adam rescales each coordinate of the gradient by its recent size
  - **backpropagation** (Rumelhart, Hinton & Williams (1986)): the gradient by the chain rule, layer by layer, from the output back
  - **early stopping**: stop when the loss on a validation period stops falling; the **seed** fixes the random start

Recurrent Networks and Vanishing Gradients



- **Recurrent neural network** (RNN): one state h_t, updated every day

$$h_t = \tanh(Wh_{t-1} + Ux_t + b), \qquad \hat{y}_{t+1} = v^\top h_t + c$$

  - x_t: input; U: its weights; v, c: output weights and intercept; the same W at every step

- **Gradient through time** (Pascanu, Mikolov & Bengio (2013)): one Jacobian per step

$$\frac{\partial h_T}{\partial h_{T-k}} = \prod_{j=0}^{k-1} D_{T-j} W, \qquad D_t = \operatorname{diag}(1 - h_t^2)$$

- Scalar case, $D_t = 1$: the product is w^k (chart)
  - $w = 0.9$: $0.9^{20} = 0.12$, $0.9^{50} = 0.005$; a day 50 steps back hardly moves the weights: the gradient *vanishes*
  - $|w| > 1$: the gradient explodes

Matrix Tools: Spectral Radius, Norm, VAR(1)

- **Spectral radius** of a square matrix W, with eigenvalues λ_i

$$\rho(W) = \max_i |\lambda_i|$$

  - for a triangular matrix the eigenvalues are the diagonal entries

- **Spectral norm**: the largest stretch of a unit vector

$$\|W\|_{\mathbf{2}} = \max_{\|v\|=\mathbf{1}} \|Wv\| = \sqrt{\lambda_{\max}(W^\top W)}, \qquad \|AB\|_{\mathbf{2}} \le \|A\|_{\mathbf{2}} \|B\|_{\mathbf{2}}$$

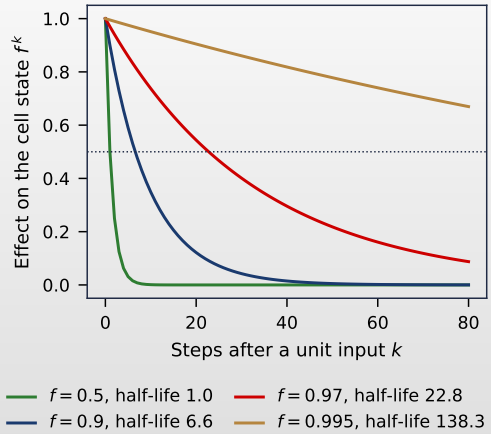
  - $\rho(W) \le \|W\|_{\mathbf{2}}$, with equality for symmetric W

- **VAR(1) model** (Vector AutoRegression of order 1) and its impulse response at horizon k

$$h_t = W h_{t-\mathbf{1}} + \varepsilon_t \;\Rightarrow\; h_t = \sum_{k \ge \mathbf{0}} W^k \varepsilon_{t-k}, \qquad \frac{\partial h_t}{\partial \varepsilon_{t-k}} = W^k$$

  - stable (shocks die out) if and only if $\rho(W) < 1$

The LSTM Cell



- **LSTM** (long short-term memory, Hochreiter & Schmidhuber (1997)): gates f_t, i_t, $o_t \in (0,1)$, $\sigma(a) = 1/(1+e^{-a})$

$$f_t = \sigma(W_f[h_{t-\mathbf{1}}, x_t] + b_f), \quad i_t = \sigma(W_i[h_{t-\mathbf{1}}, x_t] + b_i)$$

$$c_t = f_t \odot c_{t-\mathbf{1}} + i_t \odot g_t, \qquad h_t = o_t \odot \tanh(c_t)$$

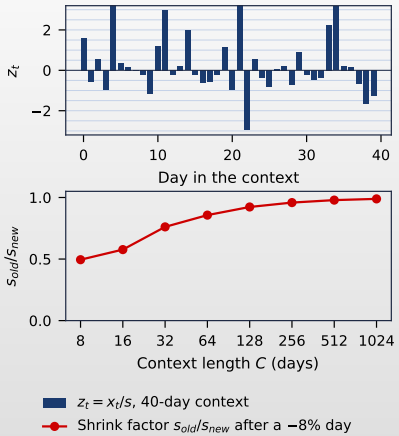
  - f_t: forget gate; i_t: input gate; o_t: output gate; $g_t = \tanh(W_g[h_{t-\mathbf{1}}, x_t] + b_g)$: candidate content
  - c_t: cell state; $\odot$: element-by-element product; $[h_{t-\mathbf{1}}, x_t]$: stacked vector; σ: the sigmoid function

- Constant forget gate f: $c_t = fc_{t-\mathbf{1}} + \text{input}$, an AR(1) (AutoRegressive) process with coefficient f

$$\text{half-life} = \frac{\ln 0.5}{\ln f}$$

  - f near 1: more persistence over a finite horizon; for $f < 1$ the effect still dies out geometrically (chart)

Foundation Models and Chronos Scaling



- Time-series **foundation model**: pretrained once on many series, used **zero-shot**: no estimation on the target series
  - input: the last C values (**context**, here $C = 512$ days); output: a grid of quantiles q_u of the next value
  - point forecast: the median $q_{\mathbf{0.5}}$ or the mean $\int_{\mathbf{0}}^{\mathbf{1}} q_u \, du$
- **Original Chronos** (Ansari et al. (2024)): mean scaling and tokens (numbered bins); Chronos-2 instead uses standardisation and inverse hyperbolic sine

$$s = \frac{1}{C}\sum_{t} |x_t|, \qquad z_t = \frac{x_t}{s}, \qquad j = \operatorname{round}\left(\frac{z_t + 15}{\Delta}\right)$$

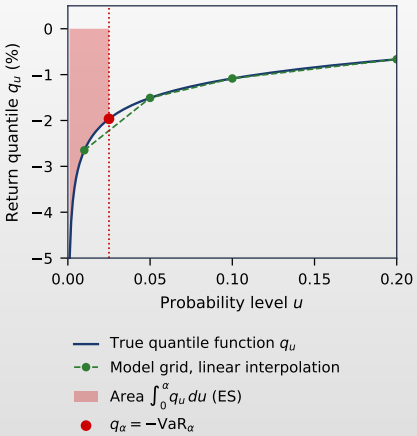
  - $\Delta = 30/4092$: bin width; 4 093 bins on $[-15, 15]$; the bins are modelled like words
  - how a crash M changes the scale s and the other z_t: derive it in A3 (lower panel, simulated)

The Forecasts We Test

Forecast	What it is	Used in
Historical mean, AR(1)	classical benchmarks: mean over an expanding window, AR(1) on a 1 000-day rolling window	B1
LSTM	recurrent network; returns: average over 5 seeds, log RV: average over 3 seeds	B1, B2
Chronos-2	119 million parameters, quantiles 1%, 5%, . . . , 99%; weights released 30.10.2025 (Ansari et al. (2025))	A1, B1, B2, B3
Chronos-Bolt	48 million parameters, quantiles 10%–90%; released 25.11.2024	B1, B2
TimesFM-2.5	231 million parameters, mean and quantiles 10%–90%; released 01.10.2025 (Das et al. (2024))	B1, B2, B3
FHS + foundation-model volatility	hybrid: scale $\hat{\sigma}_t$ = the model's forecast of $\lvert r_t \rvert$, then filtered historical simulation (FHS, McNeil & Frey (2000)) on $r_t/\hat{\sigma}_t$	B3

- Every forecast is made the evening before, from data up to that close; in class you only test the saved forecasts, no network is trained
- **Contamination**: a pretrained model may have seen our series; a test free of weight contamination uses pinned checkpoints (revision and SHA-256 hash) and only days after their release, from 3 November 2025 (about 220 trading days, in A2 and C1; 218 days with RV in B2)
- More in Chapter 14: attention and patches, the full test design, the post-release evidence

VaR and ES from a Quantile Grid



- A model returns quantiles only on a grid, e.g. $u \in \{0.01, 0.05, 0.10, \dots\}$; between grid points, **linear interpolation**

$$q_u \approx q_{u_1} + \frac{u - u_1}{u_2 - u_1}\left(q_{u_2} - q_{u_1}\right), \quad u_1 \le u \le u_2$$

- ES_α of Notation (2/2): the area under the quantile function up to α, divided by α (red area; Student-t with 4 degrees of freedom)
  - below the first grid level the model says nothing: the tail must be assumed (A1, A7)
- Normal benchmark, zero mean, with $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$

$$\frac{\mathrm{ES}_\alpha}{\mathrm{VaR}_\alpha} = \frac{\varphi(z_{1-\alpha})}{\alpha\, z_{1-\alpha}}$$

  - φ, Φ: standard Normal density and distribution function; fatter tails give a larger ratio

Beating the Zero Forecast

- **Block-bootstrap interval** for R^2: resample blocks of 20 consecutive days (circular), recompute R^2, read the 2.5% and 97.5% percentiles
- **Share of correct signs** $\frac{1}{T}\sum_t 1\{\text{sign}\,\hat{r}_t = \text{sign}\,r_t\}$; benchmark: the share of up days, not 50%
- **Nested models**: the historical mean and AR(1) contain the zero forecast as a special case; under H_0 their estimation noise makes DM too conservative
- **Clark–West** (Clark & West (2007)): adjusted loss differential, with $\hat{r}_{0,t}$ the benchmark and $\hat{r}_{1,t}$ the larger model

$$f_t = (r_t - \hat{r}_{0,t})^2 - \left[(r_t - \hat{r}_{1,t})^2 - (\hat{r}_{0,t} - \hat{r}_{1,t})^2\right]$$

  - zero benchmark ($\hat{r}_{0,t} = 0$): $f_t = 2 r_t \hat{r}_{1,t}$
  - one-sided HAC (Heteroskedasticity and Autocorrelation Consistent) t-test of $E[f_t] = 0$; reject if $t > 1.645$ at 5%
- **Many models, one family**: a search over six models finds a "winner" by luck; the tests on the next slide control for it

Many Models at Once: SPA, Romano–Wolf, the Model Confidence Set

- $d_{k,t} = L_{\mathbf{0},t} - L_{k,t}$: gain of model k over the benchmark; $\bar{d}_k$: its mean; $\hat{\omega}_k$: its long-run standard deviation, estimated by bootstrap; $n = T$ is the number of matched forecast days
- **SPA test** (Superior Predictive Ability, Hansen (2005)): $H_{\mathbf{0}}$: no model beats the benchmark, $\max_k E[d_{k,t}] \leq 0$

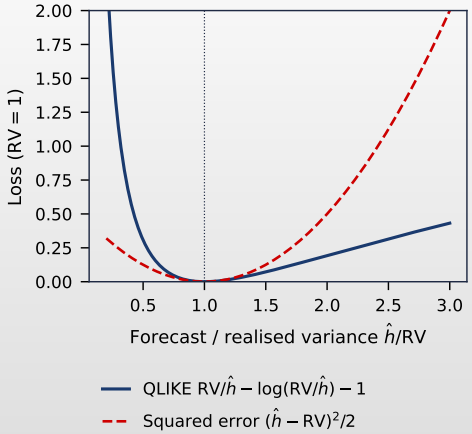
$$T^{SPA} = \max_k \max\left(\frac{\sqrt{n}\,\bar{d}_k}{\hat{\omega}_k},\ 0\right)$$

  - p-value: share of bootstrap replications of the recentred maximum above T^{SPA}; stationary bootstrap, blocks of random length with mean 20 days
- **Romano–Wolf** (Romano & Wolf (2005)): step-down; reject every model above the 95% bootstrap quantile of the maximum, remove them, repeat; one adjusted p-value per model
- **Model confidence set** (MCS, Hansen, Lunde & Nason (2011)): the set of models that contains the best one with 90% probability

$$t_m = \frac{\bar{L}_m - \bar{L}_\cdot}{\widehat{\mathrm{se}}(\bar{L}_m - \bar{L}_\cdot)}, \qquad T_{\max} = \max_m t_m$$

  - $\bar{L}_m$: mean loss of model m; $\bar{L}_\cdot$: average over the surviving models; if the bootstrap p of $T_{\max}$ is below 0.10, drop the worst model and repeat

Volatility Forecasts: HAR, GARCH-t, QLIKE



- **HAR** (heterogeneous autoregressive, Corsi (2009)): daily, weekly and monthly averages of past RV

$$\mathrm{RV}_{t+\mathbf{1}} = b_{\mathbf{0}} + b_{\mathbf{1}}\mathrm{RV}_t + b_{\mathbf{5}}\overline{\mathrm{RV}}_t^{(\mathbf{5})} + b_{\mathbf{22}}\overline{\mathrm{RV}}_t^{(\mathbf{22})} + u_{t+\mathbf{1}}$$

  - $\overline{\mathrm{RV}}_t^{(j)}$: mean of the last j days; log-HAR: the same in $\log \mathrm{RV}$, back-transformed with $\exp(\cdot + \hat{s}^{\mathbf{2}}/2)$

- **GARCH(1,1)**-t (Generalised AutoRegressive Conditional Heteroskedasticity): variance from daily returns only

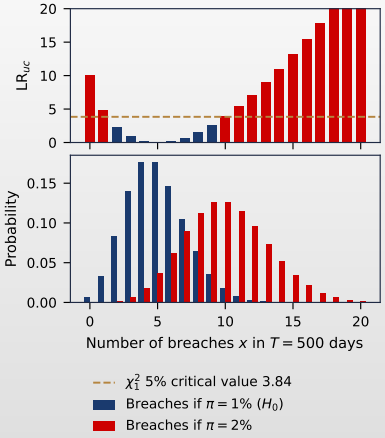
$$\sigma_t^{\mathbf{2}} = \omega + \alpha r_{t-\mathbf{1}}^{\mathbf{2}} + \beta\sigma_{t-\mathbf{1}}^{\mathbf{2}}, \qquad r_t = \sigma_t z_t$$

  - z_t: Student-t, unit variance; $\alpha + \beta < 1$: persistence

- **QLIKE** (Notation (2/2)): penalises under-prediction more than over-prediction; its expected value is minimised by $\hat{h} = E[\mathrm{RV} \mid \text{information}]$ (A10)

- Random walk benchmark: $\hat{h}_{t+\mathbf{1}} = \mathrm{RV}_t$

Size and Power of the Kupiec Test

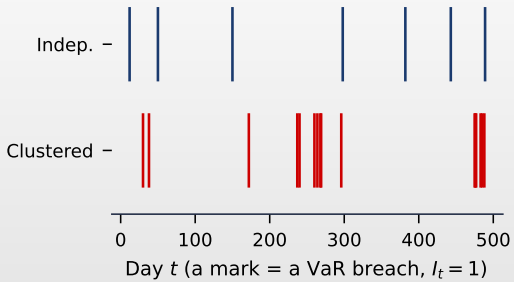


- Breaches of a VaR with true breach probability π: binomial distribution

$$P(x = k) = \binom{T}{k} \pi^k (1 - \pi)^{T-k}, \quad k = 0, \ldots, T$$

  - mean $T\pi$; $T = 500$, $\pi = 1\%$: 5 breaches expected
- **Rejection region**: the x with $\mathrm{LR}_{uc}(x) > 3.84$ (red bars, top)
- **Exact size**: $P_{\pi=\alpha}(x \in \text{region})$; **power**: $P_{\pi}(x \in \text{region})$ for a wrong π
  - x is discrete, so the size is not exactly 5%
- Chart: $T = 500$, $\alpha = 1\%$: reject if $x \leq 1$ or $x \geq 10$; size 7.1%; against $\pi = 2\%$ the power is only 54%
- A non-rejection on a short sample is weak evidence of correct coverage

Clustered Breaches: Christoffersen and DQ



Independent: $x = 7$, $\hat{\pi}_{01} = 0.014$, $\hat{\pi}_{11} = 0.00$

Clustered: $x = 16$, $\hat{\pi}_{01} = 0.027$, $\hat{\pi}_{11} = 0.19$

- A correct VaR breaches with probability α *independently* of yesterday
- **Christoffersen independence test** (Christoffersen (1998)): n_{ij} = number of days with $I_{t-1} = i$, $I_t = j$

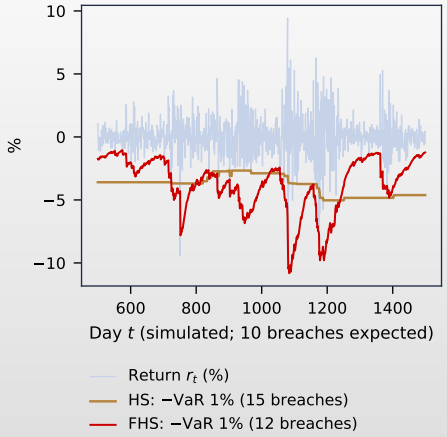
$$\hat{\pi}_{01} = \frac{n_{01}}{n_{00} + n_{01}}, \qquad \hat{\pi}_{11} = \frac{n_{11}}{n_{10} + n_{11}}$$

  - LR (Likelihood Ratio) test of $\pi_{01} = \pi_{11}$, χ^2_1; clustered breaches give $\hat{\pi}_{11} \gg \hat{\pi}_{01}$ (chart)
  - conditional coverage jointly tests the breach rate and independence: $\mathrm{LR}_{cc} = \mathrm{LR}_{uc} + \mathrm{LR}_{ind} \sim \chi^2_2$ under the null
- **Dynamic quantile test** (DQ, Engle & Manganelli (2004)): regress the hit $I_t - \alpha$ on a constant, lagged hits and VaR_t

$$\mathrm{DQ} = \frac{\hat{\beta}^\top X^\top X \hat{\beta}}{\alpha(1-\alpha)} \sim \chi^2_k$$

  - X: the k regressors; $\hat{\beta}$: OLS (Ordinary Least Squares) coefficients; a correct VaR makes every coefficient zero

Historical Simulation, FHS and the FZ0 Loss



- **Historical simulation** (HS): VaR = minus the empirical α-quantile of the last W returns
- **Filtered HS** (FHS, McNeil & Frey (2000)): standardise, take the quantile, rescale with today's volatility

$$\mathrm{VaR}_{\alpha,t} = -\hat{\sigma}_t \, \hat{q}_\alpha \left(\left\{ \frac{r_s}{\hat{\sigma}_s} \right\}_{s=t-W}^{t-1} \right)$$

  - $\hat{\sigma}_t$: GARCH or foundation-model volatility; FHS reacts to volatility, HS does not (chart, simulated)
- **FZ0 (Fissler–Ziegel loss, zero-homogeneous version) loss** (Fissler–Ziegel, zero-homogeneous version, Patton, Ziegel & Chen (2019)) of a pair (VaR v, ES e), loss L, level α

$$\mathrm{FZ0} = \frac{1\{L > v\}(L - v)}{\alpha e} + \frac{v}{e} + \log e - 1$$

  - lower is better; compares models by DM; it penalises both missed tails and needlessly large VaR and ES

Tools for the Extension Packs

- **PIT** (Probability Integral Transform) (Diebold, Gunther & Tay (1998)): $u_t = \hat{F}_t(r_t)$, the predictive distribution function at the outcome; i.i.d. uniform if the forecast is correct
  - Berkowitz (Berkowitz (2001)): $z_t = \Phi^{-1}(u_t) \sim N(0,1)$; LR test of $\mu = 0$, $\sigma = 1$, χ^2_2
- **Quantile score** QS (Quantile Score, pinball loss) of a quantile forecast q_u, averaged over the grid levels u_k

$$\mathrm{QS}_t = \frac{2}{K} \sum_{k=1}^{K} \left(1\{r_t < q_{u_k}\} - u_k\right)\left(q_{u_k} - r_t\right) \geq 0$$

- **Peaks over threshold** (A7): excesses $L - u$ over a high threshold u follow a generalised Pareto distribution (GPD)

$$P(L - u \leq y \mid L > u) = 1 - \left(1 + \frac{\xi y}{\beta}\right)^{-1/\xi}$$

  - ξ: shape (larger: heavier tail); $\beta > 0$: scale
- **Conformal recalibration** (Gibbs & Candès (2021)): shift a quantile by the empirical quantile of recent errors; the adaptive version updates the level $\alpha_{t+1} = \alpha_t + \gamma(\alpha - I_t)$

Exercise Map

Exercise	Question	Type	Model
A1	how do we read VaR and ES from a quantile grid?	Solved	–
A2	how much power has a 220-day backtest?	Solved	–
A3	how does a crash move the Chronos scale?	Solved	–
A4	why do RNN gradients vanish?	Solved	–
B1	can any model beat zero for S&P 500 returns?	Solved	–
B2	is a foundation model better than log-HAR for volatility?	Solved	–
B3	do foundation-model VaR forecasts pass the backtests?	Solved	–
C1	what is wrong in an AI answer?	Proposed	A2, B3
B4	Bitcoin and BET: is there any signal? (at home)	Proposed	B1
Extension packs	A5–A10, B5–B11, C2, C3 (optional, after the seminar)	both	on each task

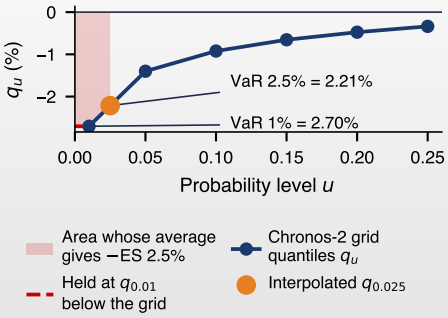
- **[Solved]**: full solution in the slides and in the notebook, a worked model; **[Proposed]**: derive or reproduce the result yourself
- A: derivations on paper; B: inference on the saved forecasts, in the notebook, each ending with an interpretation question; C: open questions and a critique of an artificial intelligence (AI) answer

A1 [Solved]: VaR 1% and ES 2.5% from a Quantile Grid — Task

- **Question**: how do we turn the quantile grid of a foundation model into VaR and ES?
- **Data**: the last zero-shot Chronos-2 forecast for the S&P 500 (Standard & Poor's), with quantiles for 18.09.2026 from the context of 512 returns up to 17.09.2026: $q_{0.01} = -2.704\%$, $q_{0.05} = -1.400\%$; grid levels 1%, 5%, 10%, . . . , 99%
- **Tasks**
  1. Compute VaR 1% with the convention $\mathrm{VaR}_\alpha = -q_\alpha$.
  2. Interpolate $q_{0.025}$ linearly between $q_{0.01}$ and $q_{0.05}$ and compute VaR 2.5%.
  3. Compute ES 2.5%, holding $q_u = q_{0.01}$ for $u < 0.01$ and using the linear q_u on $[0.01, 0.025]$.
- **Report**: VaR 1%, VaR 2.5%, ES 2.5% and the ratio ES 2.5% / VaR 2.5%
- **Interpretation**: compare the ratio with its value for a zero-mean Normal distribution at the same level: what does the grid miss?

Back Exercise list

A1 [Solved]: Solution



1. VaR 1% $= -q_{0.01} = 2.704\%$.
2. $q_{0.025} = q_{0.01} + \frac{0.015}{0.04}(q_{0.05} - q_{0.01}) = -2.704 + 0.375\,(-1.400 + 2.704) = -2.215\%$, so VaR 2.5% $= 2.215\%$.
3. Split the integral (red area): $\int_0^{0.01} q_u\,du = 0.01\,q_{0.01}$ and $\int_{0.01}^{0.025} q_u\,du = 0.015\,(q_{0.01} + q_{0.025})/2$; ES 2.5% $= -\frac{1}{0.025}[\ldots] = 2.557\%$.

⊡ Interpretation: ES/VaR = 1.15, below the Normal value $\varphi(1.96)/(0.025 \cdot 1.96) = 1.19$

- φ: standard Normal density
- The grid cuts the tail below 1%, so this ES is a tail-truncated approximation

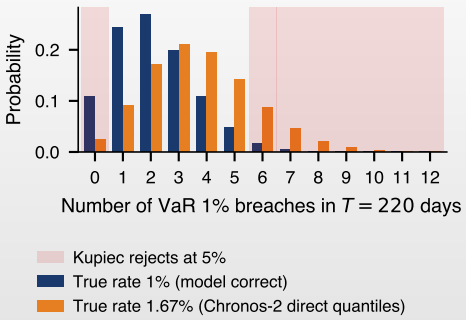
MFM_ch14_seminar

A2 [Solved]: The Exact Power of the Kupiec Test — Task

- **Question**: can 220 days after the release of the weights tell a correct VaR 1% from one that is breached 1.67% of the time?
- **Data**: $T = 220$ days (3.11.2025–18.09.2026); breaches $x \sim \mathrm{Bin}(T, \pi)$ (Size and Power of the Kupiec Test); two-sided Kupiec test of Notation (2/2) at 5%; 1.67% is the full-sample breach rate of raw Chronos-2 on the S&P 500 (B3)
- **Tasks**
  1. Compute $\mathrm{LR}_{uc}(x)$ for $x = 0, 1, 2, \ldots$ with $\alpha = 1\%$ and find the rejection region.
  2. Compute the exact size, the probability of rejecting when $\pi = 1\%$.
  3. Compute the power, the probability of rejecting when $\pi = 1.67\%$.
- **Report**: the rejection region, the size and the power
- **Interpretation**: what does a non-rejection over 220 days tell a risk manager?

Back Exercise list

A2 [Solved]: Solution



1. $x = 0$: $\mathrm{LR} = -2 \cdot 220 \ln 0.99 = 4.42 > 3.84$, reject (too few breaches); $x = 5$: LR $= 2.65$; $x = 6$: LR $= 4.51$; rejection region $x = 0$ or $x \geq 6$.
2. Size $= P_{1\%}(x = 0) + P_{1\%}(x \geq 6) = 0.134$, not 0.05: the rule uses the asymptotic χ^2_1 cut-off; a rejection region calibrated exactly on the binomial would keep the size below 5%.
3. Exact power of this rule $= 0.9833^{220} + P_{1.67\%}(x \geq 6) = 0.025 + 0.165 = 0.189$; expected breaches 2.2 under H_0, 3.7 under the alternative.

⊡ Interpretation: the two distributions overlap almost entirely

- Four times in five a model with 67% too many breaches passes
- So "not rejected" is no evidence of correct coverage

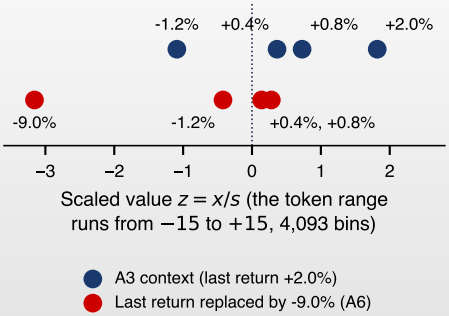
MFM_ch14_seminar

A3 [Solved]: Mean Scaling in Chronos: Who Moves the Scale? — Task

- **Question**: how much does one new day change the scale that Chronos divides the context by?
- **Data**: the Chronos scale s, scaled values z and bins j of Foundation Models and Chronos Scaling; a four-day context $x = (0.8, -1.2, 0.4, 2.0)$ (%)
- **Tasks**
  1. The window moves one day: a new value M enters and x_{out} leaves; derive s_{new} as a function of s_{old}, M, x_{out} and C.
  2. Derive the factor by which the scaled value z of every unchanged day changes.
  3. For the four-day context compute s, the scaled values z and the bin indices j.
- **Report**: the formula of steps 1–2 and a table of x, z, j
- **Interpretation**: how does the effect of one crash on the scale depend on C?

Back Exercise list

A3 [Solved]: Solution



1. The sum of absolute values loses $|x_{\text{out}}|$ and gains $|M|$; dividing by C: $s_{\text{new}} = s_{\text{old}} + (|M| - |x_{\text{out}}|)/C$.
2. Unchanged day x: $z_{\text{new}} = x/s_{\text{new}} = z_{\text{old}} \cdot s_{\text{old}}/s_{\text{new}}$.
3. $s = (0.8 + 1.2 + 0.4 + 2.0)/4 = 1.10$; $z = (+0.727; -1.091; +0.364; +1.818)$; $\Delta = 0.00733$; $j = (2145, 1897, 2096, 2294)$, near the centre $j = 2046$ of $z = 0$.

⊡ Interpretation: a crash moves s by about $|M|/C$ (red: the last day replaced by −9%, all other days shrink); a 512-day context hardly reacts

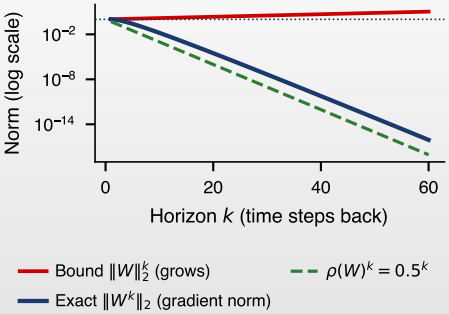
MFM_ch14_seminar

A4 [Solved]: Vanishing Gradients as VAR Impulse Responses — Task

- **Questions**
  - why does the gradient of a simple recurrent network vanish?
  - which number tells us how fast?
- **Data**: the RNN of Recurrent Networks and Vanishing Gradients with two hidden units, $W = \begin{pmatrix} 0.5 & 0.8 \\ 0 & 0.5 \end{pmatrix}$, linearised at $h^* = 0$ ($b = 0$, $x = 0$); no data file
- **Tasks**
  1. Derive the Jacobian $\partial h_T / \partial h_{T-k}$ with the chain rule and prove that its spectral norm is at most $\|W\|_2^k$.
  2. Show that the linearised network is a VAR(1) in h_t and that the Jacobian equals its impulse response at horizon k.
  3. Compute the spectral radius $\rho(W)$, the norm $\|W\|_2$, the bound $\|W\|_2^k$ and the exact $\|W^k\|_2$ for $k = 5, 20, 50$.
- **Report**: the two derivations and a table of the numbers of step 3
- **Interpretation**: is the norm bound or the spectral radius the right criterion for vanishing gradients?

Back Exercise list

A4 [Solved]: Solution



1. One step: $\partial h_t/\partial h_{t-\mathbf{1}} = D_t W$, $D_t = \mathrm{diag}(1 - h_t^{\mathbf{2}})$; over k steps $\prod_j D_{T-j} W$; since $\|D_t\|_{\mathbf{2}} \leq 1$ and norms are submultiplicative, the product is at most $\|W\|_{\mathbf{2}}^k$.
2. At $h^* = 0$, $D_t = I$: $h_t \approx W h_{t-\mathbf{1}} + U x_t$, a VAR(1); from $h_T = \sum_k W^k U x_{T-k}$, $\partial h_T/\partial h_{T-k} = W^k$ is the impulse response.
3. W is triangular: $\rho(W) = 0.50$; $\|W\|_{\mathbf{2}} = 1.040$, so the bound $1.040^{\mathbf{20}} = 2.20$ grows; exactly $W^k = 0.5^k \begin{pmatrix} 1 & 1.6k \\ 0 & 1 \end{pmatrix}$: $\|W^{\mathbf{5}}\|_{\mathbf{2}} = 0.254$, $\|W^{\mathbf{20}}\|_{\mathbf{2}} = 3.1 \times 10^{-\mathbf{5}}$, $\|W^{\mathbf{50}}\|_{\mathbf{2}} = 7.1 \times 10^{-\mathbf{14}}$.

⊡ Interpretation: the norm bound is sufficient, not necessary

- The gradient vanishes like $k\,0.5^k$ because $\rho(W) < 1$, the VAR stability criterion
- Gelfand: $\|W^k\|^{\mathbf{1}/k} \to \rho(W)$

MFM_ch14_seminar

B1 [Solved]: Can Anything Beat Zero for the S&P 500? — Task

- **Question**: does any of six models forecast daily S&P 500 returns better than the zero forecast, once we test correctly?
- **Data**: `ch14_returns_sp500.csv`: realised return `y` and the forecasts `hist` (historical mean), `ar1` (1,000-day window), `lstm` (5 seeds), `chronos2_point`, `bolt_point`, `timesfm_point` (medians); 2 693 days from 04.01.2016
- **Tasks**
  1. Compute the out-of-sample R^2 of each model and a 95% circular block-bootstrap interval (blocks of 20 days, 1,000 draws).
  2. Test each model against zero with the DM test on squared errors (two-sided, 5%) and compute the share of correct signs.
  3. Apply the Clark–West test (Clark & West (2007)) to the historical mean and AR(1), which nest the zero forecast.
  4. Apply SPA (Hansen (2005)) and Romano–Wolf (Romano & Wolf (2005)) to the six models (stationary bootstrap, mean block 20 days).
- **Report**: a chart of R^2 with intervals, a table (R^2, interval, DM [p], correct signs) and the Clark–West, SPA and Romano–Wolf p-values
- **Interpretation**
  - is an R^2 of about 1% meaningful here?
  - why is 55% correct signs not evidence of skill?
- **Question for the class**: what do you think: does the LSTM beat the zero forecast?

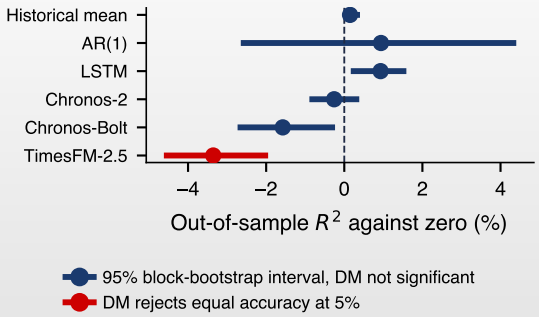
Back Exercise list

B1 [Solved]: Solution (1/3) — Step by Step

1. Align forecasts and returns on the 2 693 test days; losses $e^2_{0,t} = r_t^2$ (zero) and $e^2_{1,t} = (r_t - \hat{r}_t)^2$ (model); $R^2 = 1 - \sum e_1^2 / \sum e_0^2$. Bootstrap: draw blocks of 20 consecutive days with wrap-around, recompute R^2 1,000 times, take the 2.5% and 97.5% percentiles; LSTM: R^2 +0.93%, interval [+0.17; +1.59].
2. DM: $d_t = e^2_{1,t} - e^2_{0,t}$, Newey–West variance with $\lfloor 4(T/100)^{2/9} \rfloor$ lags; LSTM: DM $= -1.70$, $p = 0.090 > 0.05$: equal accuracy is not rejected, although the bootstrap interval excludes zero. LSTM correct signs: 53.5%, below the 54.6% of up days.
3. Clark–West with the zero benchmark (Beating the Zero Forecast): a one-sided HAC t-test of $E[r_t \hat{r}_t] = 0$ (Clark & West (2006)); one-sided p-values, DM vs Clark–West (CW): historical mean 0.048 vs 0.007; AR(1) 0.320 vs 0.055.
4. SPA: bootstrap the maximum of the recentred studentised gains; $p = 0.130$ (H_0: no model beats zero); Romano–Wolf removes rejected models step by step: historical mean 0.164, LSTM 0.161.

⊡ **Answer**: no, after the correction for six models: SPA $p = 0.130$, Romano–Wolf $p = 0.161$

B1 [Solved]: Solution (2/3) — Chart



- Dots: R^2; bars: 95% block-bootstrap intervals; red: DM rejects equal accuracy with zero at 5%
- TimesFM-2.5 is significantly worse than zero; AR(1) has a very wide interval, driven by 2020
- Interpretation: CW detects the small drift of the historical mean
  - After the six-model search (SPA, Romano–Wolf) nothing beats zero

MFM_ch14_seminar

B1 [Solved]: Solution (3/3) — Table

Model	R^2 (%)	95% CI	DM [p]	Correct sign (%)
Historical mean	+0.15	[−0.02; +0.40]	−1.67 [0.095]	54.6
AR(1)	+0.94	[−2.65; +4.40]	−0.47 [0.639]	54.1
LSTM	+0.93	[+0.17; +1.59]	−1.70 [0.090]	53.5
Chronos-2	−0.26	[−0.89; +0.38]	+0.81 [0.418]	54.7
Chronos-Bolt	−1.57	[−2.73; −0.24]	+1.76 [0.078]	54.6
TimesFM-2.5	−3.35	[−4.62; −1.95]	+3.23 [0.001]	53.6

- The index rose on 54.6% of days: "always up" already gets that share of signs right
- LSTM: the bootstrap interval excludes zero, yet DM $p = 0.090$: both concern zero forecasting gain, but differ in finite samples: a percentile interval for a ratio with 20-day resampling blocks versus HAC inference on the mean loss difference with 8 lags. Neither survives the six-model search

MFM_ch14_sem_partB

B2 [Solved]: Realised Volatility: Better than log-HAR? — Task

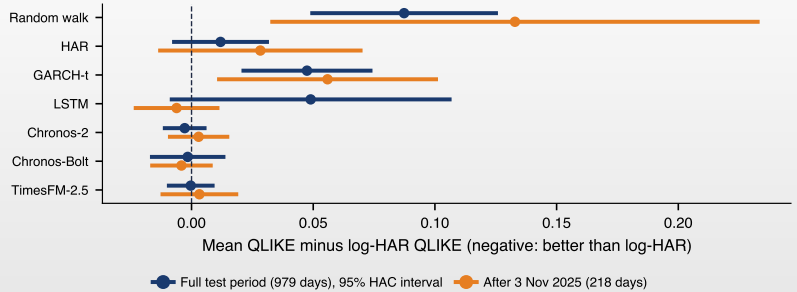
- **Question**: are zero-shot foundation models better than log-HAR at forecasting SPY (SPDR S&P 500 ETF Trust) realised variance?
- **Data**: `ch14_rv.csv`: `RV` (sum of 78 squared 5-minute returns on the 09:30–16:00 New York grid; the first return from the open and close of the 09:30 bar; full-grid days only) and forecasts `RW`, `HAR`, `logHAR` (Corsi (2009); 500-day window), `GARCH-t`, `LSTM`, `Chronos-2`, `Chronos-Bolt`, `TimesFM-2.5` (grid mean); 979 days
- **Tasks**
  1. Compute the daily QLIKE loss of each model and its mean over the 979 days.
  2. Test each model against log-HAR with DM on the QLIKE differences (two-sided, 5%) and plot the mean differences with 95% HAC intervals.
  3. Compute the 90% model confidence set (MCS, Hansen, Lunde & Nason (2011); block bootstrap, blocks of 10 days).
  4. Repeat steps 1–2 on the 218 days after 3 November 2025, when all three sets of weights were public.
- **Report**: the chart and a table: QLIKE, DM [p], MCS p, QLIKE and DM after the release
- **Interpretation**
  - what does membership of the MCS tell you?
  - what does it not tell you?

Back Exercise list

B2 [Solved]: Solution (1/3) — Step by Step

1. Daily loss $L_{t,m} = \mathrm{RV}_t/\hat{h}_{t,m} - \log(\mathrm{RV}_t/\hat{h}_{t,m}) - 1$; e.g. on 11.10.2022, RV = 1.578 and the Chronos-2 forecast 1.484 (both daily squared percentage points): $L = 1.064 - \log 1.064 - 1 = 0.0020$; means: Chronos-2 0.223, log-HAR 0.226.
2. $d_t = L_{t,m} - L_{t,\log\mathrm{HAR}}$; Chronos-2: $\bar{d} = -0.0028$, DM $= -0.61$, so the HAC standard error is $\bar{d}/\mathrm{DM}$ and the 95% interval $[-0.0118;\ +0.0062]$ contains 0 ($p = 0.539$).
3. MCS: the elimination loop of Many Models at Once: SPA, Romano–Wolf, the Model Confidence Set with QLIKE losses; a model's MCS p is the maximum of all elimination p-values through its elimination, ensuring nested retained model sets.
4. After 3.11.2025: Chronos-2 $\bar{d} = +0.0029$, interval $[-0.0097;\ +0.0155]$: wider, because 218 days carry less information.

B2 [Solved]: Solution (2/3) — Chart



- Dots: mean QLIKE difference to log-HAR; bars: 95% HAC intervals; blue: full period, orange: after 3.11.2025
- Chronos-2 has the lowest QLIKE, but every foundation-model interval contains 0
  - Random walk and GARCH-t are worse and leave the MCS
- Interpretation: MCS membership means "not distinguishable at 90%", not "equally good"

MFM_ch14_seminar

B2 [Solved]: Solution (3/3) — Table

	Full period (979 days)			After 3.11.2025 (218 days)	
Model	QLIKE	DM [p]	MCS p	QLIKE	DM [p]
Random walk	0.313	+4.44 [<0.001]	0.005	0.356	+2.59 [0.010]
HAR	0.238	+1.17 [0.241]	0.33	0.251	+1.32 [0.187]
log-HAR	0.226	–	0.95	0.223	–
GARCH-t	0.273	+3.45 [<0.001]	0.03	0.279	+2.41 [0.016]
LSTM	0.275	+1.66 [0.097]	0.29	0.217	−0.68 [0.494]
Chronos-2	0.223	−0.61 [0.539]	1.00	0.226	+0.45 [0.650]
Chronos-Bolt	0.224	−0.20 [0.842]	0.95	0.219	−0.63 [0.527]
TimesFM-2.5	0.225	−0.07 [0.948]	0.95	0.226	+0.40 [0.692]

- DM against log-HAR; positive: the model is worse than log-HAR
- After the release the LSTM has the lowest QLIKE (0.217), but only the random walk and GARCH-t differ significantly from log-HAR
- HAR in levels over-forecasts; GARCH-t reacts to daily returns, not to intraday information

MFM_ch14_sem_partB

B3 [Solved]: Backtesting the Tail Forecasts on the S&P 500 — Task

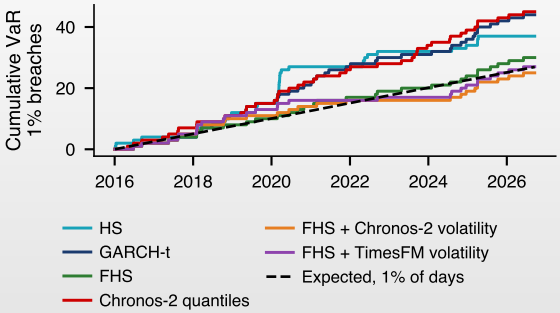
- **Question**: do raw Chronos-2 quantiles, or FHS with a foundation-model volatility, give a usable VaR 1% and ES 2.5%?
- **Data**: `ch14_risk_sp500.csv`: return `y` and, for six models, columns `<model>|VaR1`, `|VaR2.5`, `|ES2.5`: HS, GARCH-t, FHS (1,000-day window, refit every 20 days), `C2-raw` (raw Chronos-2 quantiles), `C2-FHS` and `TFM-FHS` (FHS with Chronos-2 or TimesFM volatility); 2 693 days
- **Tasks**
  1. Count the VaR 1% breaches of each model and plot their cumulative number against the expected 1% line.
  2. Apply the Kupiec test and the Christoffersen independence test (Christoffersen (1998), χ^2_1) at 5%.
  3. Compute the mean FZ0 loss (Patton, Ziegel & Chen (2019)) of (VaR 2.5%, ES 2.5%) and DM tests against FHS.
  4. For FHS and the Chronos-2 hybrid, compute the DQ test (Engle & Manganelli (2004)): regress $I_t - 0.01$ on a constant, four lagged breaches and VaR_t.
- **Report**: the chart and one table: breaches, rate, Kupiec p, independence p, FZ0, DM against FHS
- **Interpretation**: why does the hybrid pass the coverage test and still lose to FHS?

Back Exercise list

B3 [Solved]: Solution (1/3) — Step by Step

1. Breaches $x = \sum_t I_t$, rate $\hat{\pi} = x/T$; raw Chronos-2: $x = 45$, $\hat{\pi} = 1.67\%$ against 26.9 expected: LR_{uc} gives $p = 0.001$, correct coverage rejected.
2. Independence: count the transitions n_{ij} of $I_{t-1} \to I_t$ and test $\pi_{01} = \pi_{11}$ (Clustered Breaches: Christoffersen and DQ); the hybrid: $p = 0.021$, breaches cluster.
3. FZ0 of Historical Simulation, FHS and the FZ0 Loss with v = VaR 2.5%, e = ES 2.5%, $\alpha = 0.025$; hybrid 1.084 vs FHS 0.999, DM +2.65.
4. DQ $= \hat{\beta}^\top X^\top X \hat{\beta}/(0.01 \cdot 0.99) \sim \chi^2_6$, X: the six regressors, $\hat{\beta}$: their OLS coefficients: hybrid $p = 0.001$, FHS $p = 0.021$.

B3 [Solved]: Solution (2/3) — Chart



- Cumulative VaR 1% breaches of the six models; black dashed: 1% of days
- Raw Chronos-2 and GARCH-t run above the expected line (too narrow)
  - The hybrids follow it on average, but in steps around 2020 and 2022 (clusters)
- Interpretation: the hybrid takes its tail from the data through FHS, so coverage passes
  - Its forecast volatility is noisy: sometimes too low (clustered breaches, independence $p = 0.021$), sometimes too high (higher FZ0)

MFM_ch14_seminar

B3 [Solved]: Solution (3/3) — Table

Model	Breaches	Rate (%)	p Kupiec	p indep.	FZ0	DM vs FHS
HS	37	1.37	0.065	0.001	1.309	+2.46
GARCH-t	44	1.63	0.002	0.205	1.026	+1.34
FHS	30	1.11	0.559	0.045	0.999	–
Chronos-2 quantiles	45	1.67	0.001	0.223	1.147	+3.16
FHS + Chronos-2 vol.	25	0.93	0.705	0.021	1.084	+2.65
FHS + TimesFM vol.	27	1.00	0.989	0.029	1.059	+1.89

- Raw Chronos-2: 1.67% breaches, Kupiec $p = 0.001$: its quantiles are too narrow; hybrid: 0.93%, $p = 0.705$
- Right on average, but clustered (independence $p = 0.021$, DQ $p = 0.001$) and an FZ0 loss significantly above FHS (DM +2.65)

MFM_ch14_sem_partB

C1 [Proposed]: Critique an AI Answer — the Answer

- A student asks an AI assistant: "Backtest zero-shot Chronos-2 one-day-ahead VaR 1% for the S&P 500 on the last 1 500 days, with a 512-day context"
- The assistant answers with this code (spx: daily closes, imports omitted):

```
pipe = BaseChronosPipeline.from_pretrained('amazon/chronos-2')
r = 100 * np.log(spx).diff().dropna().to_numpy()      # daily returns, %
test = np.arange(len(r) - 1500, len(r))
q01 = np.empty(len(test))
for i, t in enumerate(test):                          # forecast of r[t]
    ctx = torch.tensor(r[t - 511:t + 1], dtype=torch.float32)   # 512 days
    q, _ = pipe.predict_quantiles([ctx[None, :]], prediction_length=1,
                                  quantile_levels=[0.01])
    q01[i] = q[0][0, 0, 0].item()
x = int((r[test] < q01).sum())                        # breaches
print(f"VaR 1% = {q01.mean():.2f}%, breaches = {x}")
```

- "Output: VaR 1% = −2.41%, breaches = 5, i.e. 0.33% against 15 expected: the model is conservative, so it passes the Kupiec test easily"

Back Exercise list

C1 [Proposed]: Critique an AI Answer — Task

- **Model**: A2 and B3 [Solved]
- **Question**: which three errors occur in the code and conclusion (discuss training contamination separately)?
- **Data**: the code and output of the previous slide; S&P 500 daily closes, last 1 500 days, zero-shot Chronos-2, 512-day context
- **Tasks**
  1. For each error, state what is wrong and why it matters.
  2. Say how you would detect it: an index check, the VaR convention of Chapter 7, or the test statistic.
  3. Give the corrected line or conclusion, and compute the Kupiec statistic and its p-value for 5 breaches in 1 500 days.
  4. Log the three errors in `AI_ERRORS.md`: what was asked, what the AI answered, what was wrong, how you found out, the correct version.
- **Report**: the three errors with their detection and correction, the Kupiec statistic, your log entry
- **Interpretation**
  - is "too few breaches" good news for a risk manager?
  - is it good news for a regulator?
  - do the last 1 500 days give a clean test of a model published in autumn 2025?

B4: Data and Comparison Direction

- Use the close column of BET.csv and adjusted close of TVBETETF.RO.csv. Align price dates first, then form returns.
- The one-sided alternative is that the candidate LSTM has lower expected forecast loss than the zero forecast, the B1 benchmark. Produce a B1-style model table for Bitcoin and a separate one for BET (Bucharest Exchange Trading, index).
- A statistical signal alone does not identify net profitability. For a trading claim specify a position rule and turnover, compute its break-even cost, and compare it with an explicitly assumed cost rather than declaring the signal tradable from the p-value.

B4 [Proposed]: Bitcoin and BET: Is There Any Signal? — Task

- **Model**: B1 [Solved]
- **Questions**
  - do the B1 results hold for Bitcoin and for the Romanian BET index?
  - does any BET signal survive stale prices?
- **Data**: `ch14_returns_btc.csv` (2 818 calendar days from 01.01.2019) and `ch14_returns_bet.csv` (2 675 days from 04.01.2016), same columns as B1; the BET-tracking ETF `TVBETETF.RO` (adjusted close)
- **Tasks**
  1. Repeat steps 1–2 of B1 for both markets and plot the R^2 intervals side by side.
  2. For the BET LSTM, compute the one-sided DM and Clark–West p-values, Romano–Wolf and SPA within the six BET models, and the Holm-adjusted p over the 18 one-sided DM tests (six models $\times$ three markets).
  3. Compute the BET lag-1 autocorrelation and the correlation of the LSTM forecast with r_{t-1} (Lo & MacKinlay (1990)).
  4. Join the BET and ETF prices on common days, then apply the LSTM signal to the ETF return of the same day and of the next day (R^2, Clark–West p).
- **Report**: the two-panel chart, the two tables and the numbers of steps 2–4
- **Interpretation**
  - which market has predictable daily returns?
  - is the signal tradable after costs (Chapter 2)?

Back Exercise list

Team Project Step

- One asset not covered in class (a BVB stock, a crypto-asset, an exchange rate or a commodity)
- Its volatility or VaR 1%: one foundation model zero-shot against a GARCH benchmark; context ends the day before the forecast; post-release period reported separately
- DM test and, for VaR, coverage and independence tests
- A Quantlet folder (code, Metainfo, charts); AI use in `AI_USE.md`, caught errors in `AI_ERRORS.md`; extension packs are optional
- Assessment criteria: question, data and conventions, method, uncertainty (intervals, robustness), limitations

Extension Packs: What Each One Needs

Pack	Exercises	Prerequisites
1. Recurrent memory	A5	A4
2. Tokens and quantile grids	A6, A7	A1, A3
3. Tail risk and calibration	B5, A8, B6 [Solved], B7	A1, B3
4. Power	A9, B8	A2, B2
5. Volatility forecasts	A10 [Solved], B9, B10	B2
6. Deep-learning practice	B11 [Solved]	B1
7. Projects	C2, C3	B3; the lecture case study

- Do the packs in order within a pack; each exercise names its worked model; the notebook section carries the exercise name
- C2 and C3 are possible starts for the team project

A5 [Proposed]: The LSTM Cell State as an AR(1) — Task

- **Model**: A4 [Solved]
- **Questions**
  - how long does an LSTM with a constant forget gate remember?
  - can it reproduce long memory?
- **Data**: cell $c_t = fc_{t-1} + ig_t$, constant $f = \sigma(b_f)$ and i, i.i.d. g_t with mean 0 and variance σ_g^2 (distinct from the sigmoid function σ); ACF (AutoCorrelation Function) of S&P 500 $|r_t|$ since 2000: 0.289 (lag 1), 0.132 (lag 50), 0.059 (lag 150)
- **Tasks**
  1. Use the stationary solution (initialize c_0 from its invariant distribution). Show that c_t is a stationary AR(1) for $0 < f < 1$ and derive $\mathrm{Var}(c_t)$ and $\mathrm{Corr}(c_t, c_{t-k})$.
  2. Compute the half-life $\ln 0.5 / \ln f$ for $b_f = 1$ and for $f = 0.99$, the bias that gives $f = 0.99$, and f^{250} for $b_f = 5$.
  3. Fit $\rho_k = \rho_1 f^{k-1}$ with $f^{49} = \rho_{50}/\rho_1$ and predict ρ_{150}.
- **Report**: the derivations and the numbers of steps 2–3
- **Interpretation**
  - why can no fixed $f < 1$ reproduce long memory?
  - what does the network need instead?

Back Exercise list

A6 [Proposed]: A Crash Day in the Context — Task

- **Model**: A3 [Solved]
- **Question**: does a crash make ordinary days look calm to Chronos?
- **Data**: the scaling of A3; S&P 500 on 16 March 2020: the return -12.77% enters a 512-day context and a day of 0.51% leaves; before the crash $s_{\text{old}} = 0.741\%$
- **Tasks**
  1. Replace the last value of the A3 context by -9.0 and recompute s, z, j and the factor $s_{\text{old}}/s_{\text{new}}$.
  2. Apply the formula of A3 to the S&P 500 case and compute s_{new} and the factor.
  3. Compute the scaled value of an ordinary 0.4% day before and after the crash, in both cases.
- **Report**: s, the factor and z of the 0.4% day in both cases
- **Interpretation**: does a 512-day context make ordinary days look calm after a crash?

Back Exercise list

Quantile Levels and Loss Thresholds

- In step 1 of A7, u is the probability that indexes q_u.
- In steps 2–3, u is a loss threshold in percentage points, and N_u is the number of losses above it.
- Data for the tail fit: `GSPC.INDX.csv`, column `close`. Compute $r_t = 100 \ln(P_t/P_{t-1})$ without rounding; use the last 1000 returns through the stated date and set $L = -r$.
- Use the linearly interpolated empirical 95th percentile (`numpy.quantile`, `method="linear"`); fit the strictly positive excesses with `scipy.stats.genpareto.fit`, fixing `floc=0`.
- The threshold is a return-unit quantity, not the probability 0.95. For example, a 2% loss threshold and a 3% loss produce an excess of one percentage point.

A7 [Proposed]: Clamping the Grid Tail; a GPD Tail — Task

- **Model**: A1 [Solved]
- **Questions**
  - in which direction does the grid bias ES?
  - what does a fitted heavy tail say?
- **Data**: the grid of A1; S&P 500 losses $L = -r$ over the last 1,000 days to 18.09.2026; peaks over threshold: $u =$ the 95% quantile of L, excesses $L - u \sim \mathrm{GPD}(\xi, \beta)$ of Tools for the Extension Packs, $\xi < 1$ (McNeil & Frey (2000)); n: sample size, N_u: number of excesses
- **Tasks**
  1. Prove that holding $q_u = q_{0.01}$ for $u < 0.01$ never raises ES 2.5%, and show that the linear interpolation on $[0.01, 0.025]$ can err either way.
  2. Derive $\mathrm{ES}_\alpha = \frac{\mathrm{VaR}_\alpha}{1-\xi} + \frac{\beta - \xi u}{1-\xi}$ from the GPD mean excess function.
  3. Fit ξ, β by maximum likelihood and compute $\mathrm{VaR}_\alpha = u + \frac{\beta}{\xi}[(n\alpha/N_u)^{-\xi} - 1]$ for 1% and 2.5%, and ES 2.5%.
- **Report**: u, N_u, $\hat{\xi}$, $\hat{\beta}$, VaR, ES and ES/VaR
- **Interpretation**: why is ES/VaR larger here than in A1?

Back Exercise list

B5 [Proposed]: Bitcoin, and VaR 10% on All Three Markets — Task

- **Model**: B3 [Solved]
- **Questions**
  - do the B3 conclusions hold for Bitcoin?
  - are foundation-model quantiles too narrow even at 10%?
- **Data**: `ch14_risk_btc.csv` (2 818 days from 01.01.2019); columns `<model>|VaR10` of HS, FHS and the raw Chronos-2, Chronos-Bolt and TimesFM-2.5 quantiles in `ch14_risk_<market>.csv`
- **Tasks**
  1. Repeat steps 1–3 of B3 for Bitcoin.
  2. For VaR 10%, compute the breach rates and the Kupiec test with $\alpha = 10\%$ for the five models in the three markets.
  3. Plot the rates against the 95% range of a correct model.
- **Report**: the Bitcoin table, the chart and a 3×5 table of rates with Kupiec p
- **Interpretation**: is the narrowness of foundation-model quantiles specific to the 1% level?

Back Exercise list

Tail Risk of a Normal Distribution with a Nonzero Mean

- For a predictive return $r = m + \sigma Z$, with Z standard Normal, the loss is $L = -r$. A positive m shifts the loss distribution left.
- Let $z_{1-\alpha} = \Phi^{-1}(1-\alpha) > 0$. Then $\mathrm{VaR}_\alpha = -m + \sigma z_{1-\alpha}$ and $\mathrm{ES}_\alpha = -m + \sigma\varphi(z_{1-\alpha})/\alpha$.
- The ES formula follows by averaging the upper Normal tail: the integral of $z\varphi(z)$ above $z_{1-\alpha}$ is $\varphi(z_{1-\alpha})$; divide by the tail probability α.
- In A8, the median supplies m and the gap between the median and first decile supplies σ. Use B5 for the empirical breach-rate comparison.

A8 [Proposed]: A "VaR 1%" from Deciles — Task

- **Model**: A1 [Solved]; after B5
- **Question**: what does a model that returns only deciles report when asked for VaR 1%?
- **Data**: Chronos-Bolt (small), same day and context as A1: $q_{0.1} = -0.691\%$, $q_{0.5} = 0.071\%$
- **Tasks**
  1. State the value returned when VaR 1% is requested from a decile grid clamped at its lowest level.
  2. Assume a Normal predictive distribution with median $q_{0.5}$ and $\hat{\sigma} = (q_{0.5} - q_{0.1})/z_{0.9}$, $z_{0.9} = 1.2816$ the 90% standard Normal quantile; compute VaR 1% and ES 2.5%.
  3. Compare with Chronos-2 in A1 and with the VaR 10% breach rates of B5.
- **Report**: the three numbers and the comparison
- **Interpretation**: which test would settle whether the Normal extrapolation works?

Back Exercise list

B6 [Solved]: Calibration of the Whole Predictive Distribution — Task

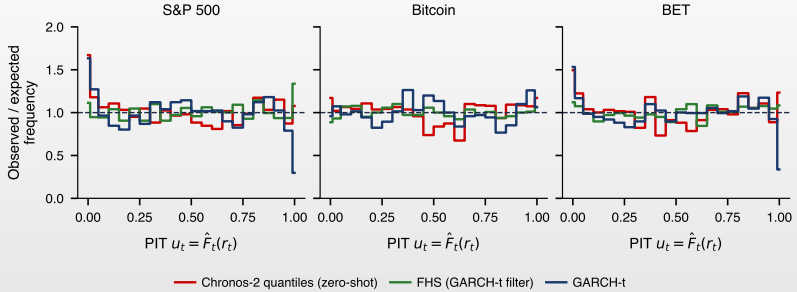
- **Question**: is the whole Chronos-2 predictive distribution calibrated, or only its centre?
- **Data**: S&P 500, 2 693 days; the raw Chronos-2 grid of 21 quantiles (1%, 5%, . . . , 99%) in `ch14_returns_sp500.csv`; FHS and GARCH-t evaluated on the same grid
- **Tasks**
  1. Compute the PIT u_t of Tools for the Extension Packs by linear interpolation, censored outside $[q_{1\%}, q_{99\%}]$, and apply Pearson's χ^2 to the counts in the 22 intervals.
  2. Apply the Berkowitz LR test (Berkowitz (2001)) of $\mu = 0$, $\sigma = 1$ to $z_t = \Phi^{-1}(u_t)$ with a censored Normal likelihood (χ^2_2).
  3. Compute the quantile score QS_t of Tools for the Extension Packs and DM of Chronos-2 minus FHS, overall, in the centre and in the tails.
- **Report**: the PIT histogram and a table: tail shares, $\hat{\mu}/\hat{\sigma}$, Berkowitz p, χ^2 p, QS
- **Interpretation**
  - is the failure in the centre or in the tails?
  - why censor instead of extrapolating?

Back Exercise list

B6 [Solved]: Solution (1/3) — Step by Step

1. If $q_{1\%,t} \le r_t \le q_{99\%,t}$, u_t interpolates the levels between the two neighbouring quantiles; otherwise record only "below" or "above"; $\chi^2 = \sum_{j=1}^{22}(O_j - Tp_j)^2/(Tp_j)$, O_j = observed count and p_j = width of interval j.
2. $\ell(\mu,\sigma) = \sum_{\text{inside}} \log \frac{1}{\sigma}\varphi\Big(\frac{z_t-\mu}{\sigma}\Big) + n_L \log \Phi\Big(\frac{c_L-\mu}{\sigma}\Big) + n_H \log\Big[1-\Phi\Big(\frac{c_H-\mu}{\sigma}\Big)\Big]$, $c_L = \Phi^{-1}(0.01)$, $c_H = \Phi^{-1}(0.99)$; φ, Φ: standard Normal density and distribution function; n_L, n_H: days below and above the grid; $\mathrm{LR} = 2[\ell(\hat\mu,\hat\sigma) - \ell(0,1)]$.
3. QS on all 21 levels, on the 17 central levels and on the 4 tail levels; DM with Newey–West variance.

B6 [Solved]: Solution (2/3) — Chart



- Observed over expected frequency in the 22 grid intervals; 1 = perfect calibration; spikes at the ends: returns outside $[q_{1\%}, q_{99\%}]$
- Interpretation: Chronos-2 is worse than FHS both in the centre and in the tails, most in the tails; censoring avoids inventing a tail the grid does not have

MFM_ch14_calibration

B6 [Solved]: Solution (3/3) — Table

Model	Below $q_{1\%}$ (%)	Above $q_{99\%}$ (%)	$\hat{\mu}$ / $\hat{\sigma}$	Berkowitz p	χ^2 p	QS
Chronos-2	1.67	1.08	-0.035 / 1.030	0.019	0.004	0.5101
FHS	1.11	1.34	$+0.014$ / 0.990	0.599	0.747	0.5062
GARCH-t	1.63	0.30	-0.041 / 0.976	0.023	<0.001	0.5063

- Chronos-2: too many returns below $q_{1\%}$, $\hat{\sigma} > 1$, both tests reject; FHS passes; QS Chronos-2 minus FHS: DM $+3.06$ (centre $+2.51$, tails $+3.65$)

MFM_ch14_calibration

Which Calibration Quantile Does B7 Use?

- At each forecast origin, sort only the previous 250 scores E_t (defined in B7). The current return enters the calibration set only after its forecast has been evaluated.
- Start the adaptive level at the target level. Select order statistic $k = \lceil 251(1 - \alpha_t) \rceil$, counting from 1, without interpolating.
- If the index exceeds 250 or the adaptive level is nonpositive, use a correction of $+\infty$; if the index is below 1 or the level is at least one, use $-\infty$. Do not clip the adaptive level.
- For a target of 1%, the initial index is 249: a library's default interpolated quantile would implement a different rule. The update of B7 is applied after observing the breach.

B7 [Proposed]: Conformal Recalibration of Chronos-2 — Task

- **Model**: B3 and B6 [Solved]
- **Question**: can conformal recalibration make raw Chronos-2 quantiles a usable VaR 1% without a GARCH filter?
- **Data**: the raw Chronos-2 quantiles of B3 in the three markets; evaluation on the days after the first 250 calibration days
- **Tasks**
  1. Correct the 1% and 2.5% quantiles by adaptive conformal inference (Gibbs & Candès (2021)): scores $E_t = \hat{q}_{\alpha,t} - r_t$ on the last 250 days, corrected quantile $\hat{q}_{\alpha,t} - \hat{Q}_{1-\alpha_t}(E)$, update $\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t)$, $\gamma = 0.001$; $\hat{q}_{\alpha,t}$: the raw quantile; $\hat{Q}_{1-\alpha_t}(E)$: empirical $(1-\alpha_t)$ quantile of the scores; $\mathrm{err}_t = 1$ on a breach of the corrected quantile.
  2. Repeat with $\gamma = 0$ (rolling calibration at a fixed level); shift ES 2.5% by the VaR 2.5% correction.
  3. Backtest FHS, raw Chronos-2 and both conformal versions: breaches, conditional coverage, DQ, FZ0 and DM against FHS.
- **Report**: a chart of breaches by method and market and one table
- **Interpretation**: why can a model with correct coverage still lose the FZ0 comparison?

Back Exercise list

A9 [Proposed]: How Long Must the Post-Release Window Be? — Task

- **Model**: A2 [Solved]
- **Question**: how many days after the release does a VaR 1% backtest need to detect a 1.67% or 2% breach rate?
- **Data**: the two-sided Kupiec test of A2 at 5%; true breach rate $\pi = 1.67\%$ or 2%
- **Tasks**
  1. Compute the exact power against $\pi = 1.67\%$ for $T = 1{,}000$ and $T = 2{,}693$.
  2. Find the smallest T with power of at least 80% and the T from which it stays above 80%.
  3. Repeat for $\pi = 2\%$: the power at $T = 220$ and the T for 80% power.
- **Report**: a chart of power against T and the numbers
- **Interpretation**
  - why is power not monotone in T?
  - how many years does a clean backtest need?

Back Exercise list

B8: The Variance of the Pre/Post-Release Difference

- Estimate separate long-run variances before and after release. For planning, treat the two periods as independent blocks; do not pool their variances.
- With long-run variances $v_{\mathrm{pre}}, v_{\mathrm{post}}$, the variance of the difference is approximated by $v_{\mathrm{pre}}/n_{\mathrm{pre}} + v_{\mathrm{post}}/n_{\mathrm{post}}$.
- For the limit in step 3, use the same quantile sum $z_{0.975} + z_{0.8}$ as for the minimum detectable effect (MDE).

B8 [Proposed]: A Formal Pre/Post-Release Test — Task

- **Model**: B2 and A2 [Solved]
- **Question**: does the advantage of Chronos-2 over log-HAR shrink after its weights were published, as memorisation would predict?
- **Data**: `ch14_rv.csv` (B2); cutoff 3 November 2025, the first Monday after all three sets of weights were public (Chronos-2: 30.10.2025); 761 days before, 218 after
- **Tasks**
  1. Compute d_t = QLIKE(Chronos-2) − QLIKE(log-HAR) and its means before and after the cutoff.
  2. Regress d_t on a constant and $1\{t \geq 3.11.2025\}$ with Newey–West errors and test $\beta_1 = 0$ at 5%.
  3. Compute the minimum detectable effect $(z_{0.975} + z_{0.8})\,\mathrm{se}_{HAC}(\hat{\beta}_1)$ (z_p: standard Normal quantiles, 1.96 and 0.84: a 5% two-sided test with 80% power) and its limit when the post window grows without bound.
- **Report**: a chart of d_t with both means and their intervals; the means, $\hat{\beta}_1$, t, p, MDE
- **Interpretation**: does a non-significant $\hat{\beta}_1$ show that Chronos-2 did not memorise the test period?

Back Exercise list

A Quantile Grid Is Not a Complete Distribution

- For the mean computed from the quantile grid, extend the quantile levels to 0 and 1 by holding the outermost available quantiles constant. This fixes the unobserved tails; it is an assumption, not information from the model.
- Apply the target transformation to every quantile first, then integrate those transformed values over the quantile levels with the trapezoidal rule. For log-volatility targets, exponentiating an average log forecast would be a different calculation.
- Equal averaging over the supplied quantiles gives each grid point equal weight. It generally differs from integration over probability levels, especially on an uneven grid; report the chosen tail and quadrature rules.

A10 [Solved]: Which Functional Does QLIKE Elicit? — Task

- **Questions**
  - which point forecast of the variance does QLIKE reward?
  - what does using the median cost?
- **Data**: QLIKE $L(y, h)$ of Notation (2/2), with $y = \mathrm{RV}$, h the forecast; Chronos-2 on $\log \mathrm{RV}$ returns quantiles q_u; average ratio of $\exp(q_{0.5})$ to $\int_0^1 \exp(q_u)\, du$ over the 979 days of B2: 0.8095
- **Tasks**
  1. Minimise $E[L(Y, h) \mid \mathcal{F}]$ over $h > 0$ and show that the minimiser is $E[Y \mid \mathcal{F}]$.
  2. Assume $\log Y \mid \mathcal{F} \sim N(\mu, s^2)$; derive the median, the mean, their ratio and the expected QLIKE excess of the median forecast.
  3. From the Chronos-2 ratio infer an effective common s and the QLIKE excess.
- **Report**: the derivations, s and the excess
- **Interpretation**: why is exp(median) of log RV a biased volatility forecast?

Back Exercise list

A10 [Solved]: Solution

1. $E[L] = E[Y]/h - E[\log Y] + \log h - 1$; $\partial_h E[L] = -E[Y]/h^2 + 1/h = 0 \Rightarrow h^* = E[Y \mid \mathcal{F}]$; the second derivative $1/h^{*2} > 0$: QLIKE is consistent for the mean (Gneiting (2011)).
2. Median e^{μ}, mean $m = e^{\mu + s^2/2}$, ratio $e^{-s^2/2}$; with $h = me^{-s^2/2}$: $E[L(h)] - E[L(m)] = e^{s^2/2} - 1 - s^2/2 > 0$.
3. $s = \sqrt{-2 \ln 0.8095} = 0.650$; excess $1/\text{ratio} - 1 + \ln(\text{ratio}) = 0.0240$; observed: QLIKE 0.249 with the median against 0.223 with the mean.

⊡ Interpretation: monotone maps preserve quantiles, so exp(median of $\log \mathrm{RV}$) is the median of RV, about 20% below its mean; QLIKE rewards the mean

B9 [Proposed]: Median or Mean? — Task

- **Model**: A10 and B2 [Solved]
- **Questions**
  - how much does the choice of functional (median or mean) change the QLIKE of a foundation model?
  - how much does the context length change it?
- **Data**: `ch14_rv.csv`: columns `Chronos-2`, `Chronos-2|median`, `TimesFM-2.5`, `TimesFM-2.5|median`; `ch14_ctx.csv`: Chronos-2 with contexts of 32, 64, 128, 256 and 512 days on the same 979 days
- **Tasks**
  1. For both models compute the QLIKE and the ratio mean forecast / mean RV of the median and of the grid mean.
  2. Test median against mean for Chronos-2 with DM on the QLIKE differences.
  3. Compute the Chronos-2 QLIKE for each context length and compare with log-HAR.
- **Report**: a two-panel chart, the table of step 1, the DM statistic
- **Interpretation**: what does the context effect say about the memory of volatility?

Back Exercise list

B10: Training History and Evaluation Days

- `ch14_rv.csv` contains evaluation forecasts, not the full history required to fit them. The earlier SPY five-minute prices start on 12 October 2020 in the intraday source file.
- Build RV separately within each day from squared log price changes, with returns expressed in percent; use the 09:30–15:55 grid (78 bars, New York time), the first return from the open and close of the 09:30 bar, and full-grid days only. This gives daily variance in percent squared.
- Use that history for the up to 500 usable regression rows and the 22-day lookback; before 500 rows are available, use all rows after the first 22 observations. Evaluate only on the dates in the forecast file, and compare all models on identical dates.

B10 [Proposed]: Leakage Through a Look-Ahead Regressor — Task

- **Model**: B2 [Solved]
- **Questions**
  - how large a gain can a one-day look-ahead create?
  - could an honest model match it?
- **Data**: SPY $\log \mathrm{RV}$ as in B2; log-HAR on a rolling 500-day window with regressors $\log \mathrm{RV}_{t-1}$, $\frac{1}{5}\sum_{j=1}^{5} \log \mathrm{RV}_{t-j}$, $\frac{1}{22}\sum_{j=1}^{22} \log \mathrm{RV}_{t-j}$; forecast $\exp(\hat{y}_t + \hat{s}^2/2)$
- **Tasks**
  1. Build a wrong version whose 5- and 22-day averages run over $j = 0, \ldots, 4$ and $j = 0, \ldots, 21$, so they include the forecast day t.
  2. Compute the QLIKE of the correct and the wrong log-HAR and of Chronos-2 on the same 979 days.
  3. Test wrong against correct with DM and write a leak test for your own code.
- **Report**: the three QLIKE values, the relative gain and the DM test
- **Interpretation**: where can the same error hide in a deep-learning pipeline?

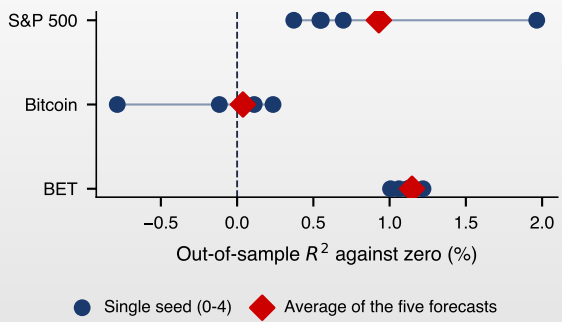
Back Exercise list

B11 [Solved]: The Seed Lottery — Task

- **Question**: how much does the random seed alone change the out-of-sample R^2 of an LSTM?
- **Data**: the return LSTM of Chapter 14 (last 20 standardised returns, 16 hidden units, MSE, Adam, early stopping); five networks identical except the seed, columns `lstm_s0`, ..., `lstm_s4` and their average `lstm` in `ch14_returns_<market>.csv`
- **Tasks**
  1. Compute the out-of-sample R^2 against zero of each seed in the three markets.
  2. Compute the R^2 of the average of the five forecasts.
  3. Compare the spread across seeds with the size of R^2.
- **Report**: a chart and a 3×6 table of R^2 (%)
- **Interpretation**
  - what would a single run on Bitcoin have concluded?
  - why is averaging seeds a form of regularisation?

Back Exercise list

B11 [Solved]: Solution (1/2) — Chart



- Blue: single seeds; red: the average of the five forecasts
- Interpretation: Bitcoin seeds from -0.79% to $+0.24\%$
  - One run could "show" skill or its absence; on the BET all five agree
  - Averaging reduces variance, like bagging

MFM_ch14_seminar

B11 [Solved]: Solution (2/2) — Table

R^2 (%)	Seed 0	1	2	3	4	Average
S&P 500	+0.55	+0.37	+0.70	+0.54	+1.96	+0.93
Bitcoin	+0.11	+0.03	−0.79	−0.12	+0.24	+0.04
BET	+1.06	+1.01	+1.12	+1.22	+1.12	+1.15

1. $R_s^2 = 1 - \sum_t (r_t - \hat{r}_{t,s})^2 / \sum_t r_t^2$ for each seed: columns 0–4
2. The average of the five forecasts, $\bar{r}_t = \frac{1}{5} \sum_s \hat{r}_{t,s}$: column Average; by convexity its squared error is no greater than the average error of the seeds
3. Spread across seeds: on Bitcoin from −0.79% to +0.24%, far above the average R^2 (+0.04%); on the S&P 500 from +0.37% to +1.96%; on the BET from +1.01% to +1.22%, small next to $R^2 = +1.15\%$

MFM_ch14_sem_partB

C2 [Proposed]: Can a Zero-Shot Foundation Model Replace GARCH for the BET? — Task

- **Model**: B3 [Solved]; project extension: fine-tune a foundation model on Romanian stocks
- **Question**: a risk manager proposes Chronos-2 for the daily VaR 1% and ES 2.5% of a BET position: accept, reject, or keep as a challenger?
- **Data**: `ch14_risk_bet.csv`, 2 675 days from 04.01.2016; post-release window from 3.11.2025 (217 days)
- **Tasks**
  1. Backtest VaR 1% of the six models of B3: breaches, Kupiec, independence and DQ.
  2. Compare (VaR 2.5%, ES 2.5%) with FZ0, DM against FHS and the 90% MCS.
  3. Classify the first ten consecutive non-overlapping 250-day blocks into Basel zones, leaving the final incomplete block out (green ≤ 4, yellow 5–9, red ≥ 10 breaches) and repeat step 1 after the release.
  4. Optional live project: pre-register a starting date and raw Chronos-2 with 512 prior daily log returns. Record the next-day VaR before that trading day begins, for 20 trading days, then backtest.
- **Report**: a one-page recommendation with the tables of steps 1–3
- **Interpretation**
  - which evidence would make you switch?
  - when would you switch?

Back Exercise list

C3: Data, Feature Count and Computation

- Use GSPC.INDX daily closes for the monthly sum of squared returns and SPY adjusted closes for monthly total returns. The longer price history is distinct from the 2016-onward forecast evaluation file.
- P counts individual columns, so 12,000 features use 6,000 frequency draws, each producing one sine and one cosine column. Use the same draws across the nested feature counts.
- With only 12 observations per estimation window, compute ridge through the 12 by 12 Gram matrix rather than a 12,000 by 12,000 inverse. Begin with a few draws to check dimensions, then run the requested 1,000 draws.

C3: Monthly Targets and Missing Observations

- Construct lags and the next-month target on a complete monthly index, leaving missing values unfilled. Compute the trailing 12-month volatility before dropping incomplete rows.
- The October 2025 CPI (Consumer Price Index) was never published. In the vintage of 30 November 2025 the latest inflation is still September's; in December 2025 November inflation cannot be computed and is not imputed. The final test has 354 targets, February 1997–August 2026, excluding January 2026.
- The regression window contains 12 complete observations, potentially spanning more than 12 calendar months. The target still remains exactly the next calendar month. Use the same 1000 random-feature draws, seed 0, for all model comparisons.

C3 [Proposed]: The Virtue of Complexity on Course Data (1/2) — Task

- **Model**: B1 for the evaluation workflow; here $R^2 = 1 - \mathrm{Var}(\text{error})/\mathrm{Var}(R)$ (footnote 40 of the paper cited below)
- **Question**: does the virtue of complexity (Kelly, Malamud & Zhou (2024)) appear on the course data?
- **Data**
  - Monthly SPY total returns from month-end adjusted closes, from 1993, minus the previous month's `TB3MS`/1200 (FRED, Federal Reserve Economic Data)
  - Seven predictors of Welch & Goyal (2008) dated t: `tbl` = `TB3MS`/100, `lty` = `GS10`/100, `tms` = `lty` − `tbl`, `dfy` = (`BAA` − `AAA`)/100
  - `infl` (monthly growth of `CPIAUCSL` as known at the end of month t: the ALFRED (ArchivaL Federal Reserve Economic Data) vintage of that day, without later revisions), `svar` (sum of squared daily S&P 500 returns), the excess return
- **Tasks**
  1. Build the monthly data and standardise as in Section V.A: returns by their trailing 12-month uncentred standard deviation, predictors by an expanding-window standard deviation, after 36 months of warm-up.
  2. Generate 12 000 random Fourier features with $\gamma = 2$ (eq. 20); for P from 2 to 12 000 and $\log_{10} z = -3, \ldots, 3$, estimate ridge without intercept on rolling windows of 12 complete observations and form the timing return $\hat{\beta}' S_t R_{t+1}$; average over 1 000 draws.
  3. Estimate the linear kitchen sink on the seven predictors with $z = 0^+$ and $z = 10^3$ on the same windows.
- **Report**: next slide

C3 [Proposed]: The Virtue of Complexity on Course Data (2/2) — Report and Questions

- **Notation**: $R_{t+\mathbf{1}}$: next month's excess return; G_t: the seven predictors
  - S_t: the P random Fourier features $\sin(\gamma\, \omega_i' G_t)$, $\cos(\gamma\, \omega_i' G_t)$, $\omega_i \sim N(0, I)$ drawn once
  - ridge $\hat{\beta}(z) = (zI + \frac{\mathbf{1}}{T} \sum_t S_t S_t')^{-\mathbf{1}} \frac{\mathbf{1}}{T} \sum_t S_t R_{t+\mathbf{1}}$, z: penalty, $T = 12$: window; timing return $\hat{\beta}' S_t R_{t+\mathbf{1}}$; complexity $c = P/T$
- **Report**
  - The curves of $\|\hat{\beta}\|^{\mathbf{2}}$, $R^{\mathbf{2}}$ and Sharpe ratio against $c = P/T$ (Figures 7–8)
  - A Table I-style table for the two linear models and RFF (Random Fourier Features) with $c = 1\,000$, $z = 10^{\mathbf{3}}$
  - Its columns: $R^{\mathbf{2}}$, Sharpe ratio with t, information ratios against SPY and the linear model with t, maximum monthly loss (in standard deviations), skewness
- **Interpretation**
  - Your kitchen sink has $c = 7/12$ against $15/12$ in the paper: what does the ridgeless curve of Figure 7 predict for its $R^{\mathbf{2}}$?
  - Can the Sharpe ratio be positive while $R^{\mathbf{2}} < 0$? Use the split of the error into a correlation and a scale part (footnote 5)
  - With about 30 years, what is the power of a two-sided 5% test for an information ratio of 0.3 ($t \approx$ ratio $\times \sqrt{\text{years}}$)?
- **Project extension**: replace the random features by the embeddings of a pretrained foundation model and repeat step 2

Acronyms used in this material (1/2)

- **ACF**: AutoCorrelation Function
- **ACI**: Adaptive Conformal Inference
- **AI**: Artificial Intelligence
- **AR**: AutoRegressive (model)
- **BET**: Bucharest Exchange Trading (index)
- **BVB**: Bursa de Valori București — *the Bucharest Stock Exchange*
- **CI**: Confidence Interval
- **CPI**: Consumer Price Index
- **CW**: Clark–West (test for nested models)
- **DM**: Diebold–Mariano (test)
- **DQ**: Dynamic Quantile (test)
- **ES**: Expected Shortfall
- **ETF**: Exchange-Traded Fund
- **FHS**: Filtered Historical Simulation
- **FRED**: Federal Reserve Economic Data
- **FZ0**: Fissler–Ziegel loss, zero-homogeneous version
- **GARCH**: Generalised AutoRegressive Conditional Heteroskedasticity

Acronyms used in this material (2/2)

- **GPD**: Generalised Pareto Distribution
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors)
- **HAR**: Heterogeneous AutoRegressive (model)
- **HS**: Historical Simulation
- **LR**: Likelihood Ratio (test)
- **LSTM**: Long Short-Term Memory (network)
- **MCS**: Model Confidence Set
- **MDE**: Minimum Detectable Effect
- **MSE**: Mean Squared Error
- **OLS**: Ordinary Least Squares
- **PIT**: Probability Integral Transform
- **QLIKE**: Quasi-LIKElihood (loss function)
- **QS**: Quantile Score
- **RFF**: Random Fourier Features
- **RNN**: Recurrent Neural Network
- **RV**: Realised Variance
- **SPA**: Superior Predictive Ability (test)
- **US**: United States
- **VAR**: Vector AutoRegression

References (1/3)

- Ansari, A.F., Shchur, O., Küken, J., Auer, A., et al. (2025). Chronos-2: from univariate to universal forecasting. arXiv:2510.15821.
- Ansari, A.F., Stella, L., Turkmen, C., Zhang, X., et al. (2024). Chronos: learning the language of time series. *Transactions on Machine Learning Research*; arXiv:2403.07815.
- Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4), 465–474.
- Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.
- Clark, T.E., West, K.D. (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135(1–2), 155–186.
- Clark, T.E., West, K.D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196.
- Das, A., Kong, W., Sen, R., Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. *ICML*, PMLR 235; arXiv:2310.10688.
- Diebold, F.X., Gunther, T.A., Tay, A.S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4), 863–883.

References (2/3)

- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Engle, R.F., Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4), 367–381.
- Gibbs, I., Candès, E. (2021). Adaptive conformal inference under distribution shift. *NeurIPS*; arXiv:2106.00170.
- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.
- Hansen, P.R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4), 365–380.
- Hansen, P.R., Lunde, A., Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Kelly, B., Malamud, S., Zhou, K. (2024). The virtue of complexity in return prediction. *Journal of Finance*, 79(1), 459–503.
- Kupiec, P.H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73–84.

References (3/3)

- Lo, A.W., MacKinlay, A.C. (1990). An econometric analysis of nonsynchronous trading. *Journal of Econometrics*, 45(1–2), 181–211.
- McNeil, A.J., Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3–4), 271–300.
- Pascanu, R., Mikolov, T., Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *ICML*; arXiv:1211.5063.
- Patton, A.J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.
- Patton, A.J., Ziegel, J.F., Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, 211(2), 388–413.
- Romano, J.P., Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237–1282.
- Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.
- Welch, I., Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4), 1455–1508.