

Modelling Financial Markets

Chapter 8: Backtesting and Evaluating Risk Forecasts



Daniel Traian PELE

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

Can a Risk Forecast Be Proven Wrong?

> *"The elicitability of a functional opens the possibility to compare competing forecasts and to rank them in terms of their realized scores."*
>
> — *Fissler & Ziegel (2016)*

- GARCH-EVT (GARCH volatility with an extreme value theory tail): the VaR (Value-at-Risk) 1% is breached on 1.15% of days on the S&P 500 (Standard & Poor's) and on 0.97% on the BET (target 1%)
  - is a rate close to 1% enough to call a model good?
- With $T = 250$ days and a true breach rate of 2%, the Kupiec POF test (proportion of failures) rejects only 24% of the time
  - a doubled breach rate goes unrejected in three years out of four
- The dynamic quantile (DQ) test rejects every S&P 500 model, filtered historical simulation (FHS) included
- ES (Expected Shortfall) cannot be ranked on its own, but the pair (VaR, ES) can, with the consistent FZ0 (Fissler–Ziegel loss, zero-homogeneous version) score
  - FHS has the lowest mean FZ0 loss on the S&P 500 and the BET

Central chart: Power of the POF Test

How Can We Tell Whether a Risk Forecast Is Useful?

- **Calibration**
  - is a VaR (Value-at-Risk) or ES (Expected Shortfall) forecast right, day after day?
  - can a backtest tell?
- **Comparison**
  - which of several forecasts is better?
  - is the gap significant?
- **Decision**: what can we conclude when the tail holds only a few days a year?

Worked exercises: Seminar 8, A1 (Kupiec), A2 (Christoffersen), A4 (Z_2), B2 (FZ0 and DM); readings and derivations: ▸ Appendix

What You Should Be Able to Do After This Chapter

- derive the local power of the POF test and the minimiser of the pinball loss
- compute the tests of the number and the timing of VaR breaches, with the size and power of each test at your T
- compute ES backtests: exceedance residuals, Z_2 by simulation and statistics built on the PIT (probability integral transform)
- interpret rankings of (VaR, ES) forecasts by a consistent score, DM tests and the MCS
  - DM: Diebold–Mariano
  - MCS: model confidence set
- interpret what a distribution-free (conformal) VaR does and does not guarantee
- Detailed examples and advanced topics are in Further Topics, before the Appendix
  - the footer button leads there; the Back button returns to the starting slide
- Further Topics and Appendix: reading for familiarity

What Do We Test? Losses and VaR

- **Backtesting**: compare risk forecasts made with the information $\mathcal{F}_{t-1}$ with the outcomes that followed
  - $\mathcal{F}_{t-1}$: everything known at the close of day $t-1$
- Daily log return: $X_t = r_t = \ln(P_t/P_{t-1})$, where P_t is the closing price on day t
- Daily **loss**: $L_t = -r_t$ (a positive value means money lost)
- **VaR**: the loss exceeded with probability α, minus the α-quantile q_α of the return (Chapter 7)
  - $\mathrm{VaR}_\alpha(X) = -\inf\{x : \ P(X \le x) > \alpha\} = -q_\alpha(X)$
  - equivalently, the $(1-\alpha)$-quantile of the loss L_t
  - e.g. VaR 1% = 3%: the loss should exceed 3% on about 1 day in 100

What Do We Test? Expected Shortfall

- **ES**: the average of the quantiles q_u over all tail levels $u \leq \alpha$, i.e. the mean loss beyond VaR
  - $\mathrm{ES}_\alpha(X) = -\frac{1}{\alpha}\int_0^\alpha q_u(X)\,du = E\left[L_t \mid L_t > \mathrm{VaR}_\alpha\right]$ (continuous case)
  - e.g. Student-t losses with $\nu = 4$, standard deviation 1% and $\alpha = 2.5\%$: VaR $= 1.96\%$, ES $= 2.82\%$
  - always $\mathrm{ES} \geq \mathrm{VaR}$
- Why these levels?
  - VaR 1% is the Basel backtesting level (1996)
  - ES 2.5% is the FRTB (Fundamental Review of the Trading Book) capital measure (2019)
- Under the Normal distribution, ES $2.5\% = 2.338\,\sigma$ and VaR $1\% = 2.326\,\sigma$
  - ES 2.5% keeps about the same prudence, but also accounts for the depth of the tail

Breaches and Conditional Calibration

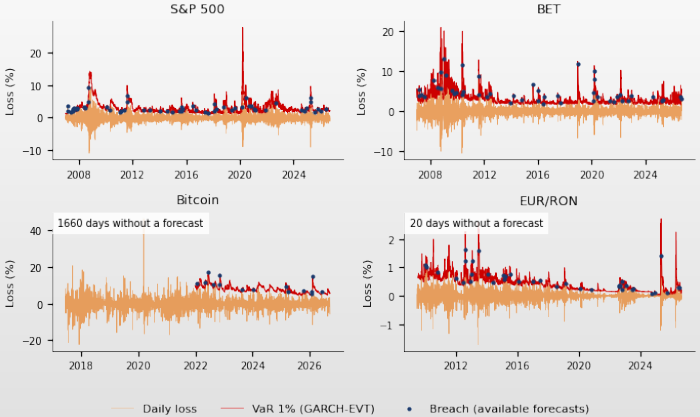
- A **breach** (exception): $I_t = 1\{L_t > \mathrm{VaR}_t\}$, where VaR_t is the forecast made on day $t-1$ for day t
- One-day forecasts: 10-day forecasts would overlap and produce dependent breaches
- **Conditional calibration** (Christoffersen (1998)): $P(I_t = 1 \mid \mathcal{F}_{t-1}) = \alpha$ for all t
  - so the I_t are i.i.d. Bernoulli(α)
- Two testable consequences
  - **Unconditional coverage**: about αT breaches in T days
  - **Independence**: breaches do not cluster

Six Ways to Forecast Tail Losses

Model	One-day VaR and ES from
HS	historical simulation: the empirical quantile of order $1-\alpha$ of the last 1000 losses
Normal	the Normal distribution with the window mean and standard deviation
Student-t	a Student-t fitted to the window (location, scale, degrees of freedom ν)
GARCH-t	GARCH (generalised autoregressive conditional heteroskedasticity, Chapter 5): $\sigma_t^2 = \omega + a\,\varepsilon_{t-1}^2 + b\,\sigma_{t-1}^2$, with $\varepsilon_{t-1} = r_{t-1} - \mu$; a: reaction to yesterday's shock; b: persistence; Student-t shocks
FHS	filtered HS: GARCH volatility $\times$ empirical quantiles of $z_t = \varepsilon_t/\sigma_t$
GARCH-EVT	EVT (extreme value theory) tail: GARCH volatility $\times$ a GPD (generalised Pareto) tail fitted to the largest 10% of standardised losses (100 exceedances out of 1 000) (McNeil & Frey (2000))

- **Unconditional** (HS, Normal, Student-t): one distribution for the whole 1000-day window
  - every day in the window has the same weight, so the forecast reacts slowly to a volatility burst
- **Conditional** or **filtered** (GARCH-t, FHS, GARCH-EVT): a scale σ_t updated daily by GARCH $\times$ a standardised tail

GARCH-EVT VaR in Four Markets (1/2)



MFM_ch8_four_assets ▶ Interactive

GARCH-EVT VaR in Four Markets (2/2)

- Orange: daily losses; red: VaR 1% forecast by GARCH-EVT; blue dots: breaches ($L_t > \mathrm{VaR}_t$)
- The VaR moves with volatility: it doubles or triples in 2008, March 2020 and the 2022 crypto crash
- GARCH-EVT VaR 1% breach rate: 1.15% on the S&P 500 and 0.97% on BET (target: 1%)
- On Bitcoin and EUR (euro)/RON (euro/leu), the VaR line starts after the period of withheld (not computed) GARCH forecasts (Further Topics)
- EUR/RON has a much smaller scale: a managed exchange rate, long calm spells, rare jumps
- Type of evidence: description (it shows what happened in the data, without testing a cause)
- Question for the chapter: is a rate close to 1% enough to call a model good?

MFM_ch8_four_assets

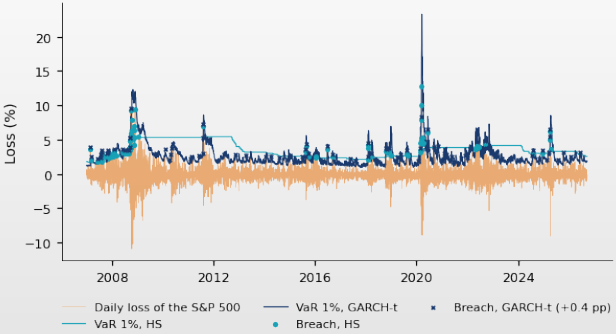
What a Backtest Can and Cannot Identify

- Each backtest checks one implication of conditional calibration
- Non-rejection is not validation
  - it is informative only if the test has power against the relevant alternatives at the available T
- H_0 is a **joint** hypothesis on the whole forecasting procedure
  - distribution, volatility dynamics, estimated parameters, window length and refit rule
- A rejection does not say which component failed
- One outcome per day: evidence accumulates only over many days
- The textbook χ^2 limits treat VaR_t as known
  - the effect of estimation: backtests under estimation risk

Does VaR Have the Right Number of Breaches, at the Right Times?

- Kupiec: the number of breaches
- Christoffersen: first-order memory
- Duration tests: the spacing of breaches
- DQ (Dynamic Quantile): any predictor known yesterday
- For each test: the null, the statistic, its distribution under H_0, and size and power at the available T
- We start with two S&P 500 forecasts with almost the same number of breaches

S&P 500: Two VaR Forecasts, 2007–2026



- Chart: VaR 1% of HS (turquoise) and GARCH-t (blue)
  - daily losses (orange) and each model's breaches (markers)
- HS is a step function: it moves only when extreme days enter or leave the 1 000-day window
- GARCH-t rises from the day after a volatility shock
- Breaches: 83 for HS and 88 for GARCH-t, in 4 959 days
- HS breaches come in bursts (2008, 2020), so **when** they occur must be tested too

MFM_ch8_var_forecasts ▶ Interactive

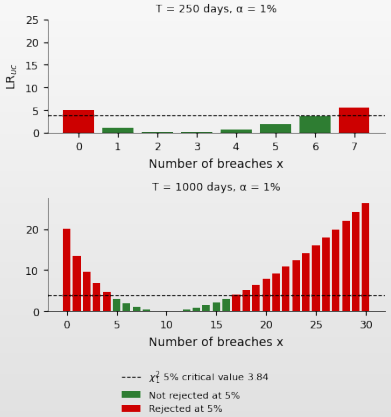
Kupiec (1995): the POF (Proportion of Failures) Test

- What it tests: whether the breach rate equals α (unconditional coverage), whatever their timing
- Data: $x = \sum_t I_t$ breaches in T days; estimated breach rate $\hat{\pi} = x/T$
- H_0: $\pi = \alpha$, against H_1: $\pi \neq \alpha$; π: the true breach probability
- Likelihood ratio statistic (Kupiec (1995)): compares the Binomial likelihood of x under α with that under $\hat{\pi}$
  - $\mathrm{LR}_{uc} = -2 \ln \dfrac{(1-\alpha)^{T-x}\alpha^x}{(1-\hat{\pi})^{T-x}\hat{\pi}^x} \xrightarrow{d} \chi^2_1$
  - $\mathrm{LR}_{uc} \geq 0$ equals 0 when $\hat{\pi} = \alpha$ and grows as $\hat{\pi}$ moves away from α

The POF Test: Decision Rule and an Example

- Reject at 5% if $\mathrm{LR}_{uc} > 3.84$ ($= 1.96^2$, the 95% quantile of χ^2_1)
- The test is two-sided: too few breaches also lead to rejection
- For small T we use the exact Binomial(T, α) distribution of x
- Actual size: the sum of the Binomial probabilities of the rejected counts x
  - with discrete x, no rule has size exactly 5%
- Example: S&P 500, GARCH-t, year 2008: $x = 8$ breaches in $T = 253$ days, against 2.53 expected
- $\mathrm{LR}_{uc} = 7.60$, with $p = 0.0058$ (χ^2_1 approximation)
  - exact two-sided Binomial p-value (doubling the smaller tail): $2P(X \geq 8) = 0.009$
- Both reject at 5%: in 2008 the GARCH-t VaR was breached too often

How Many Breaches Are Acceptable? (1/2)



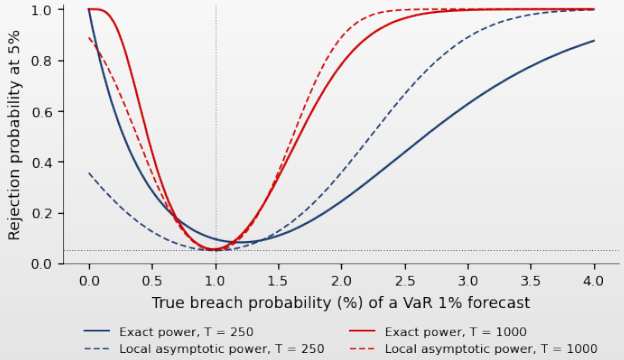
MFM_ch8_kupiec ▶ Interactive

How Many Breaches Are Acceptable? (2/2)

- LR_{uc} against the number of breaches x, one year (top) and four years (bottom) of VaR 1%
- $T = 250$: the test accepts x from 1 to 6; $T = 1000$: from 5 to 16
- Even $x = 0$ is rejected in one year
  - $\mathrm{LR}_{uc} = -500 \ln 0.99 = 5.03$: a VaR never breached is too conservative
- One year holds about 2.5 informative days for VaR 1%
  - use long samples, two levels ($\alpha = 1\%$ and 2.5%) and several tests

MFM_ch8_kupiec

Power of the POF Test



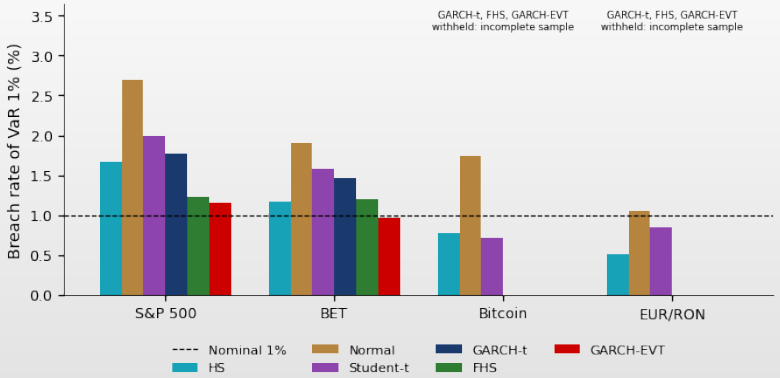
- Axes: true breach rate (horizontal) and rejection probability at 5% (vertical)
- Solid line: exact Binomial calculation; dashed line: local approximation $\chi^2_1(\lambda)$
  - $\lambda = T(\pi - \alpha)^2/(\alpha(1-\alpha))$
- $T = 250$, $\pi = 2\%$: exact power 24%, local power 36%
- So a doubled breach rate goes unrejected in three years out of four
- The rule $\mathrm{LR}_{uc} > 3.84$ has an actual size of 9.5% at $T = 250$; $P(x = 0) = 8.1\%$ alone
- For 50% power: $\pi \approx 2.7\%$ at $T = 250$ and $\pi \approx 1.7\%$ at $T = 1000$

MFM_ch8_kupiec ▶ Interactive

Power of the POF Test: Local Alternatives

- Local alternatives: the true rate approaches α as T grows, $\pi_T = \alpha + h/\sqrt{T}$
  - h: a fixed distance on the $\sqrt{T}$ scale
- Why: with a fixed $\pi \neq \alpha$ the power tends to 1
  - local alternatives keep it between 0 and 1 and make it computable
- $\mathrm{LR}_{uc} = \dfrac{T(\hat{\pi}-\alpha)^2}{\alpha(1-\alpha)} + o_p(1) \xrightarrow{d} \chi^2_1(\lambda), \qquad \lambda = \dfrac{T(\pi-\alpha)^2}{\alpha(1-\alpha)}$
- λ: the noncentrality parameter; it grows with T and with $(\pi-\alpha)^2$
- Power $= P(\chi^2_1(\lambda) > 3.84)$ grows with λ

Breach Rates Across Models and Markets



MFM_ch8_var_forecasts ▶ Interactive

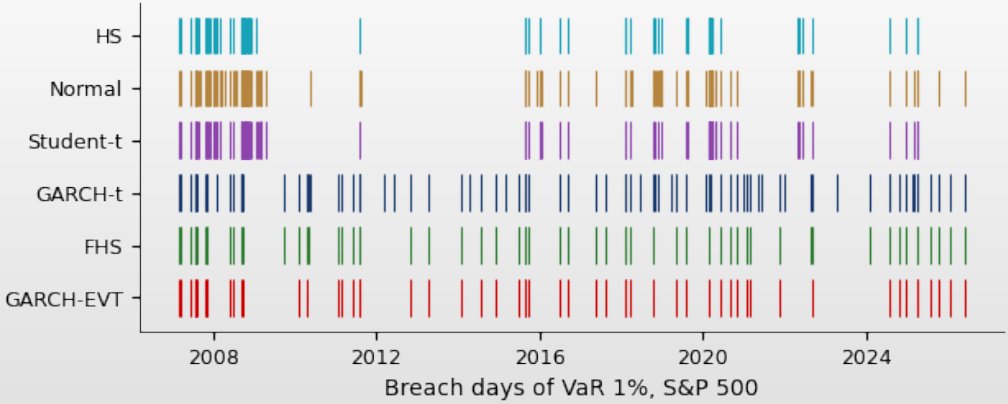
Reading the Breach Rates

- Share of days with $L_t > \mathrm{VaR}_t$ for VaR 1%; the Normal distribution fails except on EUR/RON
  - S&P 500 2.70%, BET 1.90%, Bitcoin 1.74%
- GARCH-t and FHS: same GARCH volatility, different tail (t versus empirical), on S&P 500 and BET
  - GARCH-t 1.77% and 1.46% versus FHS 1.23% and 1.20%
  - evidence against the symmetric t tail on the loss side
- HS is too conservative on EUR/RON (0.51%): an old crisis in the window keeps its VaR high

Christoffersen (1998): Testing Independence

- What it tests: whether today's breach depends on yesterday's
  - the Kupiec test ignores the order of breaches
- Model I_t as a first-order Markov chain with $\pi_{ij} = P(I_t = j \mid I_{t-1} = i)$
- n_{ij}: transitions from state i yesterday to state j today
  - $\hat{\pi}_{01} = \frac{n_{01}}{n_{00}+n_{01}}$, $\hat{\pi}_{11} = \frac{n_{11}}{n_{10}+n_{11}}$
- H_0: $\pi_{01} = \pi_{11} = \pi$ (no memory), with $\hat{\pi} = (n_{01} + n_{11})/\sum n_{ij}$
- $\mathrm{LR}_{ind} = -2 \ln \dfrac{(1-\hat{\pi})^{n_{00}+n_{10}} \hat{\pi}^{n_{01}+n_{11}}}{(1-\hat{\pi}_{01})^{n_{00}} \hat{\pi}_{01}^{n_{01}} (1-\hat{\pi}_{11})^{n_{10}} \hat{\pi}_{11}^{n_{11}}} \xrightarrow{d} \chi^2_1$
- Numerator: one breach probability; denominator: probabilities that depend on yesterday's state
  - reject at 5% above 3.84
- Conditional coverage: $\mathrm{LR}_{cc} = \mathrm{LR}_{uc} + \mathrm{LR}_{ind} \xrightarrow{d} \chi^2_2$ tests both properties (5% critical value: 5.99)

When Do Breaches Happen? (1/2)



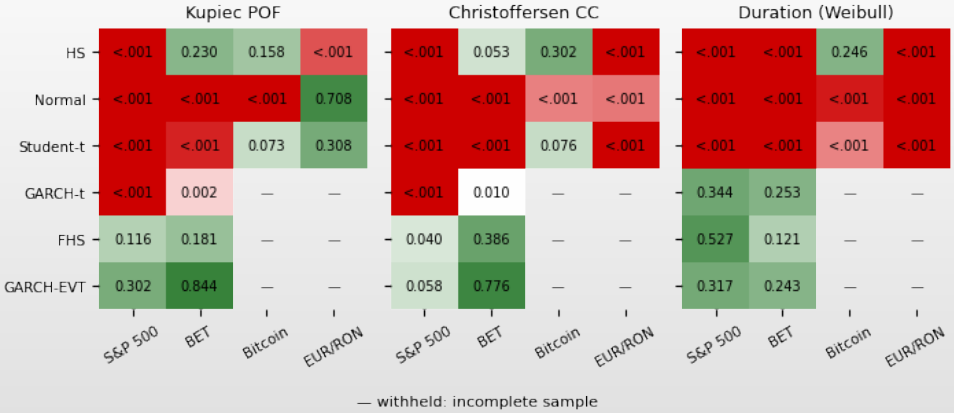
MFM_ch8_coverage_tests ▶ Interactive

When Do Breaches Happen? (2/2)

- Each tick marks one breach of VaR 1% on the S&P 500; one row per model
- HS, Normal and Student-t: bursts in 2008, 2015–2016, 2018, 2020, 2022
- Filtered (GARCH) models: no multi-month bursts, yet some back-to-back breaches
- A subsample test can miss clustering: HS, 2007–2010, $n_{00} = 928$, $n_{01} = 38$, $n_{10} = 38$, $n_{11} = 3$
- $\hat{\pi}_{11} = 0.0732$ versus $\hat{\pi}_{01} = 0.0393$, yet $\mathrm{LR}_{ind} = 0.95$, $p = 0.33$
- The reason: $\hat{\pi}_{11}$ rests on only 3 back-to-back pairs

MFM_ch8_coverage_tests

Summary of VaR Backtests (1/2)



MFM_ch8_coverage_tests ▶ Interactive

Summary of VaR Backtests (2/2)

- The chart: p-values of the Kupiec POF, conditional coverage (CC) and duration (Weibull) tests, VaR 1%
  - red: rejection, green: no rejection; "—": withheld GARCH forecast
- The DQ test (not in the chart; asymptotic χ^2_6 p-values) rejects every S&P 500 model
  - including FHS ($p < 0.001$) and GARCH-EVT ($p < 0.001$)
  - on BET it does not reject FHS ($p = 0.200$) or GARCH-EVT ($p = 0.279$)
  - hits are predictable from the VaR level and lagged hits, a failure that counting breaches cannot detect
- Different tests examine different failures: read the pattern, not one p-value
  - the unconditional models fail the duration test almost everywhere
  - GARCH-EVT is not rejected by conditional coverage or duration tests on BET
- Type of evidence: comparison of procedures (it shows which method works better on the same data, not why)

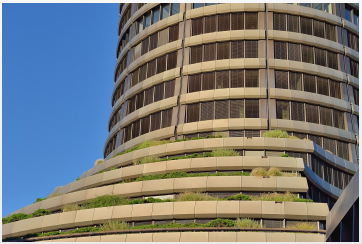
MFM_ch8_coverage_tests

How Does the Supervisor Test? The Basel Traffic Light

Exceptions	$P(X \leq x)$	Zone
0	8.11%	green
2	54.32%	green
4	89.22%	green
5	95.88%	yellow
7	99.60%	yellow
9	99.97%	yellow
10	99.99%	red

- Banks with internal models compare VaR 1% daily with the actual P&L of the last 250 days (BCBS (1996))
  - P&L: profit and loss
- What it tests: only the number of exceptions, not their size or timing
- $X \sim \text{Binomial}(250, 0.01)$: the exceptions of a correct VaR 1% in 250 days
- $P(X \leq x) = \sum_{k=0}^{x} \binom{250}{k} 0.01^k \, 0.99^{250-k}$: the second column of the table
- Yellow starts where $P(X \leq x) \geq 95\%$, red where $P(X \leq x) \geq 99.99\%$
- Why a count rule: simple, automatic, hard to game

Backtesting Rules in 1996 and 2019

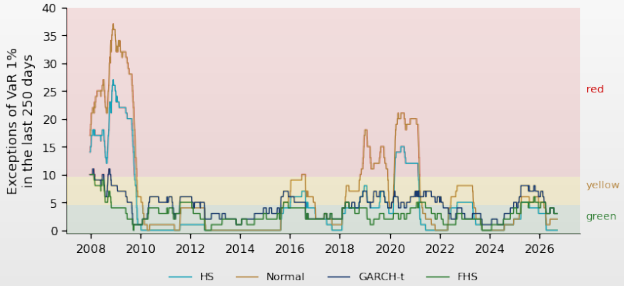


BIS tower, Basel: seat of the BCBS secretariat

Photo: Taxiarchos228 (2014); Free Art License; Wikimedia Commons

- BCBS (1996): capital on the 10-day VaR 1%, with multiplier 3
- In 1996, the multiplier rises by 0.40–0.85 in the yellow zone and by 1 in the red zone
- BCBS (2019) (FRTB): capital on ES 2.5%, but backtesting stays on the VaR 1% of the last 250 days
- FRTB multiplier, by zone
  - 1.50 in green (0–4 exceptions)
  - 1.70–1.92 in yellow (5–9)
  - 2.00 in red (10 or more)
- A trading desk moves to the standardised approach if, in 12 months, it has
  - more than 12 exceptions of VaR 1%
  - or more than 30 of VaR 2.5%
- IMA (internal models approach) approval also requires a P&L attribution test
- The Basel backtesting rule counts VaR exceptions; ES is not tested directly
  - that ES cannot be ranked alone (non-elicitability) does not prevent ES backtesting (Acerbi & Szekely (2014), ES section)
  - it concerns the comparison of forecasts (elicitability section)

The Traffic Light Through Time



- VaR 1% exceptions in the last 250 days, S&P 500; bands: green, yellow, red zones
- Share of days in red: Normal 22.3%, HS 14.7%, GARCH-t 2.0%, FHS 0.8%
- GARCH-t is yellow 46.5% of the time
- Peak: Normal 37 and HS 27 exceptions, in the 250 days ending 01.12.2008
- Overlapping windows: the penalty lasts a year after a crisis, when capital is already strained
- Type of evidence: comparison of procedures

MFM_ch8_traffic_light ▶ Interactive

FRTB Desk Check on the Last 250 Days (Limits 12 and 30)

Model	S&P 500	BET	Bitcoin	EUR/RON
HS	0 / 2	4 / 7	3 / 4	1 / 5
Normal	2 / 2	6 / 9	3 / 5	1 / 4
Student-t	0 / 2	4 / 9	1 / 4	5 / 8
GARCH-t	3 / 6	4 / 9	3 / 5	4 / 5
FHS	3 / 4	4 / 6	4 / 5	3 / 4
GARCH-EVT	3 / 3	2 / 7	4 / 5	3 / 4

- Cells: exceptions of VaR 1% / VaR 2.5% over the last 250 days of each series, up to 18.09.2026
- On these days, GARCH forecasts exist on all four markets
- All models sit comfortably within the limits 12 and 30
- Illustrative only: FRTB applies the limits to desk P&L (actual and hypothetical), not to index returns
- The last year was calm: a green zone today says little about the next crisis

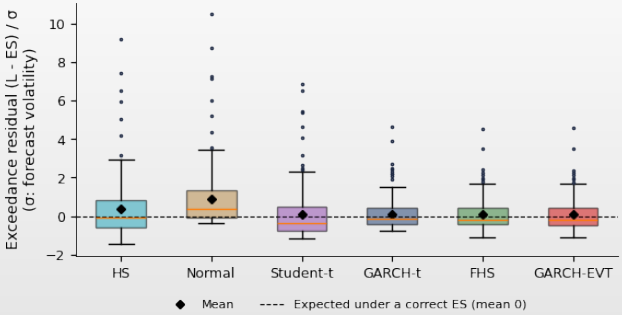
Is ES Forecast Correctly? What Can Be Tested

- ES is the mean loss beyond VaR: the information comes only from about αT tail days
- S&P 500, $\alpha = 2.5\%$: about 124 days out of 4 959; in one year, 6.25 days
- Over one year, the standard error of the tail mean is of order $1/\sqrt{6.25} = 0.40$ standard deviations of the tail losses
- Three families of tests
  - exceedance residuals (McNeil–Frey)
  - the Z_1, Z_2 statistics (Acerbi–Szekely)
  - PIT-based statistics (Du–Escanciano, multinomial tests)
- All need more than the ES forecast: the VaR, the volatility or the whole forecast distribution
  - a test on ES alone: the case study

McNeil and Frey (2000): Exceedance Residuals

- What it tests: whether the mean loss on breach days is the one forecast by ES, on the volatility scale
- Exceedance residual: $e_t = (L_t - \mathrm{ES}_t)/\sigma_t$ on days with $L_t > \mathrm{VaR}_t$ (VaR 2.5%)
  - $\sigma_t > 0$: the volatility forecast at $t-1$
- If VaR **and** ES are right: $E[e_t \mid L_t > \mathrm{VaR}_t, \mathcal{F}_{t-1}] = 0$
  - a wrong VaR changes the conditioning event
- H_0: $E[e_t] = 0$ against H_1: $E[e_t] > 0$ (ES underestimated)
- Statistic: $t = \bar{e}/(s_e/\sqrt{N})$, with $\bar{e}$ and s_e the mean and standard deviation of the N residuals
- p-value: i.i.d. bootstrap of the centred residuals (5 000 replications) (McNeil & Frey (2000))
  - because the residuals are few and heavy-tailed
- The bootstrap assumes independent residuals
  - with dependent residuals, use a block bootstrap

Exceedance Residuals on the S&P 500



- Distribution of $(L_t - \mathrm{ES}_t)/\sigma_t$ on breach days of VaR 2.5%; diamonds: means
- Means: Normal 0.875, HS 0.363, Student-t 0.082, GARCH-t 0.100, FHS 0.104, GARCH-EVT 0.068
- Bootstrap p
  - Normal <0.001, HS <0.001, Student-t 0.173, GARCH-t 0.025
  - FHS 0.072, GARCH-EVT 0.157
- Normal: tail mean too low by almost one volatility unit
  - the extreme residuals, up to 10 volatilities, matter most for capital

MFM_ch8_es_backtests ▶ Interactive

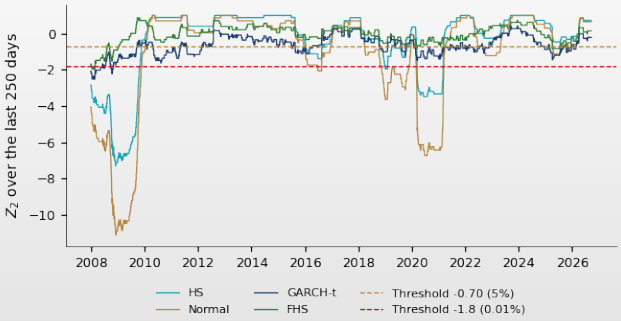
Acerbi and Szekely (2014): Testing ES Directly

- Breach indicator of VaR 2.5%: $I_t = 1\{L_t > \mathrm{VaR}_t^{\alpha}\}$, $\alpha = 0.025$
  - $N_T = \sum_t I_t$: the number of breaches in T days
  - ES_t: the ES forecast for day t
- $Z_1 = 1 - \frac{1}{N_T} \sum_t \frac{L_t I_t}{\mathrm{ES}_t}$, conditional on breaches
  - is the average tail loss as large as forecast?
- $Z_2 = 1 - \frac{1}{T\alpha} \sum_t \frac{L_t I_t}{\mathrm{ES}_t}$: tests frequency and size together
- Under H_0 (correct forecast distribution) $E[Z_2] = 0$; underestimated risk gives $Z_2 < 0$ (Acerbi & Szekely (2014))
- For $T = 250$ and many distributions, the 5% and 0.01% critical values are close to -0.70 and -1.8
  - reference thresholds for an ES traffic light

Getting p-Values by Simulation

- The null distribution of Z_2 depends on the forecast distributions F_t, so critical values are simulated
  - F_t: the distribution function of L_t forecast on day $t-1$
- Step 1, repeated $M = 2\,000$ times: for each day draw $U_t \sim \text{Uniform}(0,1)$
  - set $L_t^* = F_t^{-1}(U_t)$ (F_t^{-1}: the quantile function)
- Step 2: compute Z_2^* from the simulated losses, with the same VaR and ES forecasts
  - a calibration with the forecast distributions held fixed and independent U_t: estimation risk and GARCH refitting are ignored
- p-value = share of $Z_2^* \leq$ the observed Z_2; with $M = 2\,000$ the smallest nonzero p-value is 0.0005; no more extreme simulated value means $p < 0.0005$, not $p = 0$
- Only tail draws matter: for HS and FHS we keep the 25 largest losses of each day's window ($25 = 2.5\% \times 1\,000$)
- S&P 500, full sample: $Z_2 = -0.664$ for GARCH-t ($p <$0.001) and -0.079 for FHS ($p = 0.209$)
- The 5% critical value at $T = 4\,959$ is about -0.150, closer to 0 than -0.70 at $T = 250$

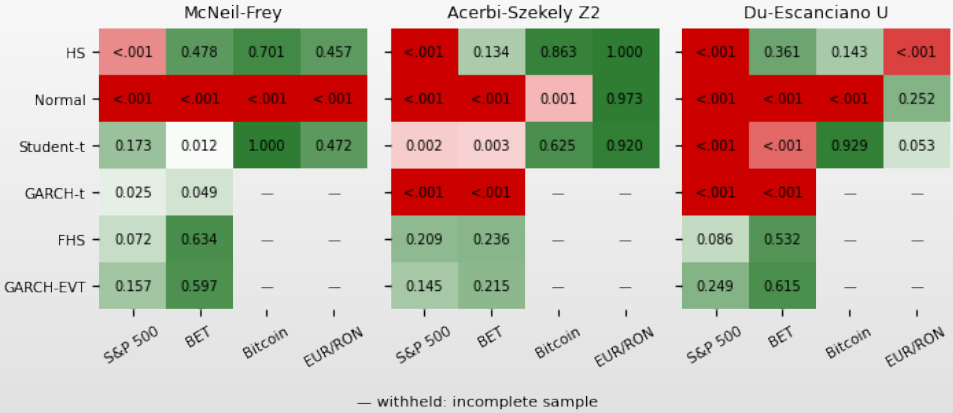
Z_2 Over Rolling 250-Day Windows



- Z_2 for ES 2.5% on the S&P 500, recomputed daily on the last 250 days
  - horizontal lines: the thresholds -0.70 and -1.8
- Minimum: Normal -11.13 (window ending 04.12.2008), HS -7.30 (01.12.2008)
- Minimum: GARCH-t -2.52 (05.02.2008), FHS -1.97 (17.01.2008)
- Windows with $Z_2 < -1.8$: Normal 21.9%, HS 15.8%, GARCH-t 3.0%, FHS 0.7%
- GARCH-t is below -0.70 in 44.3% of windows
- Daily checking is repeated testing: false alarms accumulate (sequential monitoring, ▸ Appendix)

MFM_ch8_es_backtests ▸ Interactive

Summary of ES Backtests (1/2)



MFM_ch8_es_backtests ▶ Interactive

Summary of ES Backtests (2/2)

- At the 5% significance level, FHS and GARCH-EVT are not rejected on S&P 500 and BET
  - neither by McNeil–Frey, nor by Z_2, nor by Du–Escanciano U, taken separately
- Different nulls and inputs, different decisions
  - Student-t on the S&P 500 has McNeil–Frey $p = 0.173$ but Z_2 $p = 0.002$
- These are calibration tests, not rankings
- GARCH-t is rejected by Z_2 on S&P 500 ($p < 0.001$) and BET ($p < 0.001$)
- Read U with its sign: on EUR/RON, HS has $U = -3.68$ ($p < 0.001$)
  - a rejection because losses stay **shallower** than forecast
  - this matches Z_2 ($p = 1.000$) and its VaR 1% breach rate of 0.51%
- Type of evidence: comparison of procedures

MFM_ch8_es_backtests

How Do We Rank Forecasts? Consistent Scores

- A backtest asks whether a model is **acceptable**: several models can pass, or all can fail
- Comparative backtesting asks which model is **better** (Nolde & Ziegel (2017))
- Ranking needs a **scoring function** $S(x, L)$
  - $S(x, L)$: the penalty for reporting x when L occurs
  - compare $\bar{S} = \frac{1}{T}\sum_t S(x_t, L_t)$, lower = better
- S is **consistent** for $\theta(F)$, and θ is **elicitable**, if $\theta(F) = \arg\min_x E_F[S(x, L)]$ (Gneiting (2011))
- $\theta(F)$: the target functional (e.g. VaR) of the loss distribution F; E_F: expectation when $L \sim F$
- Ranking needs the true value to be the **unique** minimiser (strict consistency)

The Pinball Loss and ES

- VaR is elicitable, with the pinball loss $S(v, L) = (1\{L > v\} - \alpha)(L - v)$, where v is the VaR forecast
- If $E|L| < \infty$ and the density $f > 0$ at the quantile:
  $$\frac{\partial}{\partial v} E_F[S(v, L)] = F(v) - (1 - \alpha) = 0 \;\Rightarrow\; v = F^{-1}(1 - \alpha) = \mathrm{VaR}_\alpha$$
- The second derivative is $f(v) > 0$, so the minimum is unique
- ES alone is **not** elicitable: no scoring function of the reported ES e alone is strictly consistent for ES on the relevant class of distributions (Gneiting (2011))
- The reason: the level sets of ES are not convex under mixtures of distributions (Appendix)
- Hence ES is ranked jointly with VaR (next slide)

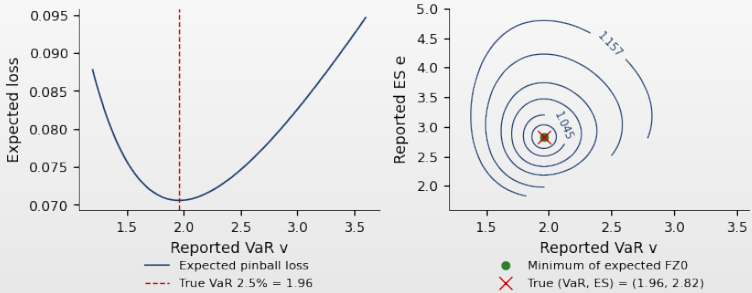
VaR and ES Together: Fissler and Ziegel (2016)

- The pair (VaR, ES) is jointly elicitable: Fissler & Ziegel (2016) characterise the FZ (Fissler–Ziegel, family of scoring functions for VaR and ES) family of consistent scoring functions
- FZ0: the member of the FZ family with zero-homogeneous score differences (Patton, Ziegel & Chen (2019)), for positive losses, $0 < v \leq e$ and $E|L| < \infty$
- $\mathrm{FZ0}(v, e; L) = \dfrac{1\{L > v\}(L - v)}{\alpha\, e} + \dfrac{v}{e} + \ln e - 1$
- v, e: the VaR and ES forecasts (positive losses); L: the realised loss
- The first term penalises breaches, scaled by ES
- The last terms penalise reporting a large ES
- Lower mean FZ0 = better joint VaR and ES forecast
  - $\mathrm{FZ0}(cv, ce; cL) = \mathrm{FZ0}(v, e; L) + \ln c$: differences between models do not depend on the scale of losses

The FZ0 Loss: a One-Day Example

- $\alpha = 2.5\%$, forecasts $v = 3\%$ and $e = 3.9\%$, losses in percent
- Loss $L = 4.5\%$ (breach): $\mathrm{FZ0} = 15.385 + 1.130 = 16.515$
- Loss $L = 1\%$ (no breach): only $v/e + \ln e - 1 = 0.769 + 1.361 - 1 = 1.130$ remains
- In decimals ($v = 0.03$, $e = 0.039$) only $\ln e$ changes
  - each value drops by $\ln 100 = 4.605$
  - to -3.475 without and 11.910 with the breach
- This is why mean FZ0 losses on decimal returns are negative (S&P 500, FHS: -3.551)
  - differences between models are unchanged

Elicitability in Pictures



- Student-t losses with $\nu = 4$ and unit variance, $\alpha = 2.5\%$
- Left: the expected pinball loss is lowest at the true VaR 2.5%, 1.96
- Right: the expected FZ0 loss is lowest at (1.96, 2.82)
  - the true (VaR, ES) up to the grid
- A forecaster who knows the truth has no incentive to report anything else: this is what makes ranking by average score meaningful

MFM_ch8_elicitability ▶ Interactive

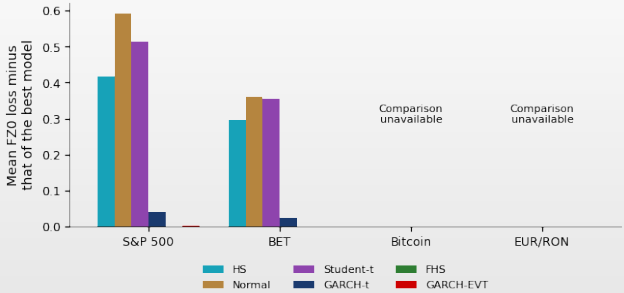
Is the Difference Between Two Forecasts Significant?

- The mean score ranks forecasts, but small gaps can be sampling noise
- Diebold–Mariano: do two forecasts have the same mean loss?
- Giacomini–White: can yesterday's information predict the loss difference?
- Nolde–Ziegel: is the internal model better than a standard model?
- MCS: which set contains the best models with a given probability?
- All use the FZ0 loss of the pair (VaR 2.5%, ES 2.5%)

The Diebold–Mariano Test on FZ0 Losses

- What it tests: whether two forecasts have the same mean loss
- $d_t = \mathrm{FZ0}_{A,t} - \mathrm{FZ0}_{B,t}$; H_0: $E[d_t] = 0$ (equal predictive accuracy), Diebold & Mariano (1995)
- $\mathrm{DM} = \dfrac{\bar{d}}{\sqrt{\widehat{\mathrm{LRV}}(d_t)/T}} \xrightarrow{d} N(0,1)$
- $\widehat{\mathrm{LRV}} = \hat{\gamma}_0 + 2\sum_{k=1}^{q}\left(1 - \frac{k}{q+1}\right)\hat{\gamma}_k$: the long-run variance with Newey–West weights (Newey & West (1987))
  - $\hat{\gamma}_k$: autocovariance of d_t at lag k
- Why: tail losses cluster, so d_t is autocorrelated
- Lag rule: $q = \lfloor 4(T/100)^{2/9} \rfloor = 9$ for $T = 4\,959$
- $\mathrm{DM} > 0$: model A has the larger mean loss (worse); $|\mathrm{DM}| > 1.96$: significant at 5%

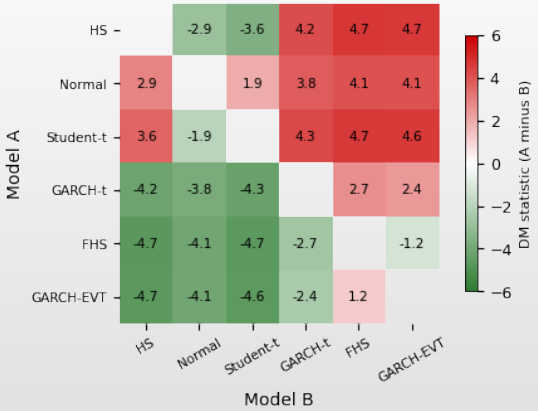
Average FZ0 Loss by Model



- Mean FZ0 loss of the (VaR 2.5%, ES 2.5%) forecasts minus that of the best model in each market
- On S&P 500 and BET, all six models are compared on the same evaluation days
- Lowest FZ0 loss: FHS on S&P 500 and BET
- The mean FZ0 loss of GARCH-t exceeds that of FHS by 0.040 on the S&P 500
- The size of a gap does not tell whether the data separate the models: we test it

MFM_ch8_comparative ▶ Interactive

Pairwise Diebold–Mariano Tests, S&P 500 (1/2)



MFM_ch8_comparative ▶ Interactive

Pairwise Diebold–Mariano Tests, S&P 500 (2/2)

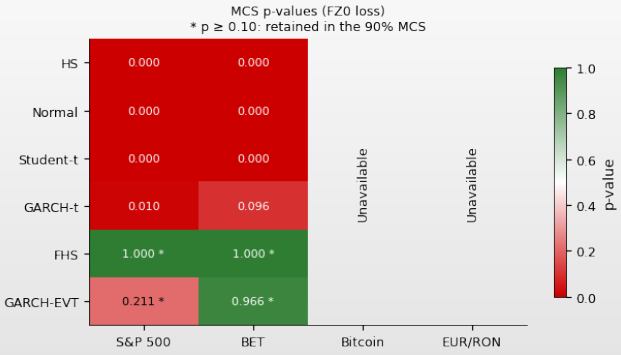
- DM of model A (row) minus B (column) on FZ0 losses; red: A worse, green: A better
- Against FHS
  - HS $\mathrm{DM} = 4.69$, GARCH-t $\mathrm{DM} = 2.66$: both significantly worse
  - GARCH-EVT $\mathrm{DM} = 1.24$: indistinguishable
  - but the conditional Giacomini–White test rejects it against FHS ($p = 0.035$, Further Topics)
- Normal minus HS: $\mathrm{DM} = 2.88$: a slow empirical quantile still beats a thin tail
- Fifteen pairwise tests: a multiple-testing problem
  - which models survive **jointly**? (model confidence set, MCS)

MFM_ch8_comparative

The Model Confidence Set: the Idea

- MCS (Hansen, Lunde & Nason (2011)): the set that contains the best models with asymptotic probability at least $1 - \alpha_{MCS}$
- What it corrects: the 15 pairwise DM tests do not control the family-wise error
  - the MCS compares all models jointly
- The guarantee needs the method's regularity conditions
  - stationary, weakly dependent loss differences
- We use $\alpha_{MCS} = 10\%$: keep models with MCS p-value ≥ 0.10
- A large MCS is an informative result: the data do not separate the models

Models Retained in the MCS



- MCS p-values on the FZ0 loss of the pair (VaR 2.5%, ES 2.5%)
  - models with $p \geq 0.10$ form the 90% MCS
- S&P 500 and BET: the MCS contains FHS and GARCH-EVT
- GARCH-t is excluded on the S&P 500 ($p = 0.010$) and, narrowly, on BET ($p = 0.096$)
- HS, Normal and Student-t: $p \leq 0.001$ on both markets
- Bitcoin and EUR/RON: no MCS, because GARCH forecasts are withheld (slide on withheld forecasts)
- Type of evidence: comparison of procedures

MFM_ch8_comparative ▶ Interactive

Can Coverage Be Guaranteed Without Distributional Assumptions?

- The models so far assume a distribution (Normal, t, GPD) or a stable past (HS)
- **Conformal prediction** (Lei et al. (2018), Angelopoulos & Bates (2021)) turns any forecast into a quantile with a coverage guarantee
- Score: $s_t = L_t/\hat{\sigma}_t$, where $\hat{\sigma}_t$ is the standard deviation of the previous 250 returns
- Calibration set: the last $n = 250$ scores (one trading year), sorted $s_{(1)} \leq \cdots \leq s_{(n)}$
- Split conformal VaR: $\mathrm{VaR}_t = \hat{\sigma}_t\, s_{(k)}$, with $k = \lceil (n+1)(1-\alpha) \rceil$ ($\lceil x \rceil$: the smallest integer $\geq x$)
- For $n = 250$ and $\alpha = 0.01$: $k = 249$, the second-largest score

Split Conformal: the Guarantee and Its Limits

- The guarantee, if the $n+1$ scores are **exchangeable**: $1-\alpha \leq P(L_t \leq \mathrm{VaR}_t) \leq 1-\alpha+\frac{1}{n+1}$ (no ties)
  - exchangeable: the joint law is invariant to permutations
- Here the guarantee is only approximate (extensions: Appendix)
- Volatility clustering breaks exchangeability
- $\hat{\sigma}_t$ comes from overlapping rolling windows, so the scores are not independent
- Genuine split conformal fits the score once, on a separate sample

Adaptive Conformal Inference (ACI)

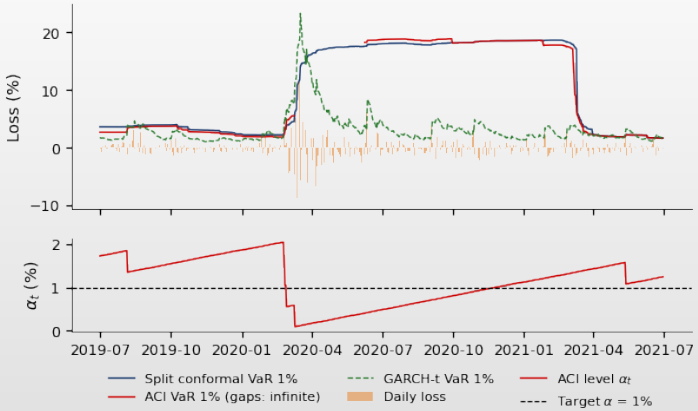


Emmanuel Candès, co-author of adaptive conformal inference

Photo: Cmichel67 (2026); CC BY-SA 4.0; Wikimedia Commons

- Gibbs & Candès (2021), ACI (Adaptive Conformal Inference): the working level is updated after every day
  - $\alpha_{t+\mathbf{1}} = \alpha_t + \gamma\,(\alpha - \mathrm{err}_t)$, $\mathrm{err}_t = 1\{L_t > \mathrm{VaR}_t\}$
  - α_t: the conformal quantile level on day t; $\gamma > 0$: step size; $\alpha_{\mathbf{1}}$: starting level
- After a breach α_t falls (VaR rises); after calm days it creeps back up
- We use $\gamma = 0.005$: a small step keeps the level stable
  - a large step reacts faster but leads more often to infinite VaR
- The guarantee, for any sequence: $\left|\frac{1}{T}\sum_t \mathrm{err}_t - \alpha\right| \le \frac{\max(\alpha_{\mathbf{1}}, 1-\alpha_{\mathbf{1}}) + \gamma}{\gamma T}$
  - a long-run **average**, not every day or every crisis
- If $\alpha_t < 1/(n+1)$ the required quantile does not exist: VaR $= +\infty$
- If $\alpha_t \ge 1$ the threshold is $-\infty$ and a breach is certain; the guarantee needs both conventions

Conformal VaR Through COVID-19 (1/2)



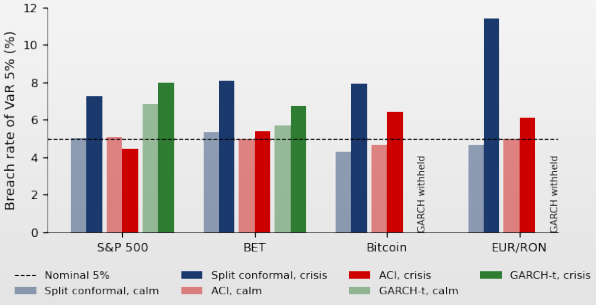
MFM_ch8_conformal ▶ Interactive

Conformal VaR Through COVID-19 (2/2)

- S&P 500, July 2019–June 2021, VaR 1%
  - top: split conformal, ACI and GARCH-t
  - bottom: the ACI level α_t
- From 20.02 to 30.04.2020: 50 days, with 0.5 expected breaches
- Observed breaches: split conformal 6, ACI 4 (8% of days)
- ACI set VaR $= \infty$ on 37 of these days: no new breach, but an infinite VaR carries no information
- The split conformal VaR 1% stays around 18% for a year
  - $\hat{\sigma}_t$ and $s_{(\mathbf{249})}$ use the last 250 days, so they keep the crash until March 2021
- GARCH-t adapts within days: after calm days, the weight of a past shock in $\sigma_t^{\mathbf{2}}$ falls by the factor b per day
  - with the parameters estimated on the S&P 500, 1990–2026: $b = 0.892$, half-life of about 6 days
  - the variance forecast reverts on average at rate $a + b = 0.994$: half-life of about 124 days

MFM_ch8_conformal

Coverage in Calm and Crisis Periods



- Chart: VaR 5% breach rate by market and method, on calm and crisis days
- Crisis day, defined in advance: the 20-day volatility exceeds the 90th percentile of the previous 1 000 days
- Why VaR 5%: 713 S&P 500 crisis days give about 36 expected breaches, against 7 for VaR 1%
- Split conformal, all days: S&P 500 5.34%, BET 5.63%, Bitcoin 4.58%, EUR/RON 5.06%
- Split conformal, crisis days: S&P 500 7.29%, BET 8.08%, Bitcoin 7.95%, EUR/RON 11.43%
- ACI, crisis days: S&P 500 4.49%, BET 5.38%
  - ▶ for VaR 1%, this requires an infinite VaR on 13.4% of S&P 500 evaluation days
- Marginal coverage is not crisis coverage
- Type of evidence: comparison of procedures

MFM_ch8_conformal ▶ Interactive

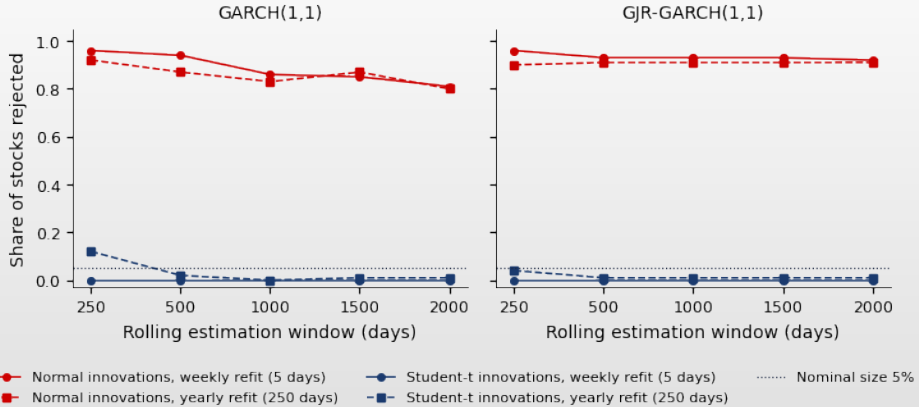
Case Study (1/4): Can ES Be Backtested Using Only ES Forecasts?

- Bayer & Dimitriadis (2022): can ES forecasts be backtested with only the observed returns and these forecasts?
- Relevance: FRTB sets capital from ES 2.5% (the regulator's view)
  - the ES tests above also need VaR, the volatility or the forecast distribution, which supervisors may not receive
  - rejecting such a joint test does not say which forecast is wrong
- The idea: a Mincer–Zarnowitz regression of returns on the ES forecasts
- H_0: intercept 0 and slope 1 (a correct forecast needs no correction)
- The regression can be estimated because (VaR, ES) is jointly elicitable (Fissler & Ziegel (2016))
  - a joint quantile–ES regression (Patton, Ziegel & Chen (2019))

Case Study (2/4): the ES Regression

- Return convention: $\hat{e}_t = -\widehat{\mathrm{ES}}_t < 0$ is the ES forecast on the return scale, with $\tau = \alpha = 2.5\%$
- ES equation: $r_t = \gamma_1 + \gamma_2 \hat{e}_t + u_t^e$, with $\mathrm{ES}_\tau(u_t^e \mid \mathcal{F}_{t-1}) = 0$, so $\gamma_1 + \gamma_2 \hat{e}_t$ is the true conditional ES
- γ_1: intercept; γ_2: slope; correct forecasts: $\gamma_1 = 0$, $\gamma_2 = 1$
- ESR (expected shortfall regression) also needs a quantile equation, because ES is estimated jointly with VaR
- **Auxiliary ESR**: quantile equation on $(1, \hat{v}_t)$, so it also needs VaR forecasts $\hat{v}_t$
- **Strict ESR**: quantile equation on $(1, \hat{e}_t)$, so it needs ES forecasts only
- **Intercept ESR**: $r_t - \hat{e}_t = \gamma_1 + u_t^e$ (slope fixed at 1)
  - one-sided t-test of H_0: $\gamma_1 \geq 0$ against $\gamma_1 < 0$ (ES underestimated)

Case Study (3/4): Tails Matter More Than Refits



MFM_ch8_esr_case ▶ Interactive

Case Study (4/4): Tails Matter More Than Refits

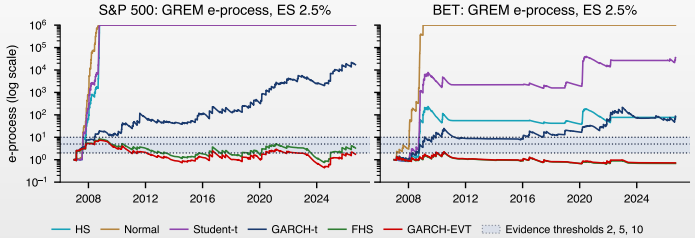
- Source: Bayer & Dimitriadis (2022), Table 3: the 200 largest S&P 500 stocks, 2010–2019
- Axes: estimation window, 250–2 000 days (horizontal)
  - vertical: share of stocks whose ES 2.5% is rejected by the one-sided 5% intercept ESR
- Solid line: refit every 5 days; dashed line: every 250 days (Table 3 also has 21, 62 and 125 days)
- GJR-GARCH (Glosten–Jagannathan–Runkle): GARCH with an extra response to negative shocks
- Normal innovations: the rejected share is 0.80–0.96 in all 50 cells
  - 5 windows $\times$ 5 refit frequencies $\times$ GARCH/GJR-GARCH
- Student-t innovations: at most 0.12 with 250-day windows and at most 0.01 with windows of at least 1 000 days
- Heavy tails and a long window matter most; the refit frequency matters only for short windows
- Type of evidence: comparison of procedures
  - the authors note that the classical covariance over-rejects in small samples

MFM_ch8_esr_case

E-Backtesting: Monitoring ES Continuously

- Wang, Wang & Ziegel (2026), *Management Science*: an ES backtest that can be checked every day without losing its validity
- The problem with the ES tests above: a p-value is valid for a sample fixed in advance
  - a supervisor who repeats the test every day and stops at the first rejection has a type I error above 5%
- **E-value**: a nonnegative statistic with mean at most 1 under H_0
  - a large value is evidence against H_0, as a small p-value is
- The paper's claims
  - the test needs only the (VaR, ES) forecasts and the realised losses, not the model that produced them
  - validity needs no asymptotic approximation and holds at any stopping time (anytime-valid)
  - for a fixed loss and VaR, a higher, more prudent ES gives a smaller e-statistic e_t
- The paper's empirical result: NASDAQ (National Association of Securities Dealers Automated Quotations) Composite, ES 2.5% from AR(1)-GARCH(1,1), from 3 Jan 2005
  - Normal innovations, GREM (growth rate with a mixture) e-process (the accumulated evidence against H_0, defined below): threshold 2 is reached after 540 trading days, threshold 10 after 713
  - skewed-t innovations with the ES raised by 10%: no threshold reached by 31 Dec 2021

E-Backtesting: ES 2.5% E-Processes on the S&P 500 and BET



- Horizontal: the evaluation days, Jan 2007 – Sep 2026 (S&P 500: 4 959 days; BET: 4 937 days)
  - vertical: the GREM e-process of each model, on a log scale; it starts at 1 (no evidence)
- Band: thresholds 2, 5 and 10; reaching $1/\delta$ rejects the forecasts with a type I error of at most δ (50%, 20%, 10%)
- Above 10^6, the top of the axis: on the S&P 500, Normal, Student-t and HS (4 Sep 2008, 6 Oct 2008, 15 Oct 2008); on BET, Normal (7 Jan 2009)
- FHS and GARCH-EVT stay below 10 on the S&P 500 and below 2.3 on BET

MFM_ch8_ebacktest ▶ Interactive

E-Backtesting: the E-Statistic and the Course Replication

- The e-statistic for ES 2.5%: $e_t = \dfrac{(L_t - v_t)_+}{0.025\,(s_t - v_t)}$, with $(x)_+ = \max(x, 0)$
  - L_t: the loss on day t; v_t: the VaR 2.5% forecast; s_t: the ES 2.5% forecast (paper: z_t, r_t)
  - H_0: for every t, $v_t = \mathrm{VaR}_t$, $s_t \geq \mathrm{ES}_t$ and $s_t > v_t$, with forecasts known at $t-1$
  - under H_0 the conditional mean of e_t is at most 1; a wrong VaR can push it above 1 even with the correct ES
- The e-process: $M_t = \prod_{s \leq t}(1 - \lambda_s + \lambda_s e_s)$, with $\lambda_s \in [0,\ 0.5]$ chosen from the data up to $s-1$ only
  - under H_0, M_t is a nonnegative supermartingale, and Ville's inequality gives $P(\sup_t M_t \geq 1/\delta) \leq \delta$
- λ_s maximises the log growth of M_t (the paper's Taylor approximation)
  - GREE (growth rate with empirical e-statistics): from the e-values of the last 500 days
  - GREL (growth rate with empirical losses): from the losses of the last 500 days, at today's forecast
  - GREM (growth rate with a mixture): the average of the two e-processes, the authors' recommended choice
- The course replication, against the paper
  - the chapter's forecasts: six models, rolling window of 1,000 days (paper: AR(1)-GARCH(1,1), 500 days)
  - S&P 500 and BET, 2007–2026 (paper: NASDAQ Composite, 2005–2021); Bitcoin and EUR/RON have withheld GARCH forecasts
  - GREE starts in Jan 2007 with no earlier e-values; the paper had five years of forecasts before 2005

E-Backtesting: Evidence from 2007 On, Except for FHS and GARCH-EVT

Model	S&P 500: max M_t	first day with $M_t \geq 10$	BET: max M_t	first day with $M_t \geq 10$
HS	$1.5 \cdot 10^{9}$	19 Oct 2007	237	10 Oct 2008
Normal	$1.5 \cdot 10^{21}$	9 Aug 2007	$9.3 \cdot 10^{10}$	21 Jan 2008
Student-t	$4.7 \cdot 10^{11}$	15 Aug 2007	$4.1 \cdot 10^{4}$	6 Oct 2008
GARCH-t	$2.2 \cdot 10^{4}$	6 Jun 2008	218	17 Feb 2009
FHS	8.55	—	2.16	—
GARCH-EVT	7.47	—	2.28	—

- S&P 500: Normal reaches threshold 2 on 26 Jul 2007 and threshold 10 on 9 Aug 2007; HS on 19 Oct 2007; GARCH-t on 6 Jun 2008
  - the mean of e_t: 1.89 for GARCH-t and 1.33 for FHS (expected value for a correct ES: 1)
- FHS and GARCH-EVT on the S&P 500: peaks of 8.55 and 7.47 (29 Sep 2008); on BET: 2.16 and 2.28
- HS on BET: Z_2 does not reject on the whole sample ($p = 0.134$), but the e-process reaches 10 on 10 Oct 2008 (peak 237)
- ES raised by 10%: FHS and GARCH-EVT on the S&P 500 stay below 5 (peaks 3.38 and 3.46); GARCH-t on BET stays below 2 (1.79)
  - an empirical result: the bets λ_s are estimated from the past, so the GREM e-process is not guaranteed to be monotone in the ES
- Type of evidence: a test of the ES conditional on a correct VaR 2.5%, valid on any day it is read
  - it does not rank the models: that is the job of comparative backtesting and the MCS
- Limits: after a prudent spell, M_t falls and needs time to grow again; thresholds 2 and 5 carry a large type I error (50%, 20%)

An Open Question: Crisis-Conditional Coverage

- Distribution-free VaR guarantees are **marginal**
  - split conformal needs exchangeability
  - ACI controls only the long-run average
  - section "Conformal Prediction for Tail Risk"
- Formal question: is the crisis breach rate still α?
  - with a crisis flag $C_t \in \mathcal{F}_{t-1}$ known ex ante
  - $C_t = 1$ on a crisis day, 0 otherwise
- H_0: $P(L_t > \mathrm{VaR}_t \mid C_t = 1) = \alpha$ against H_1: $P(L_t > \mathrm{VaR}_t \mid C_t = 1) \neq \alpha$
- Tested through the moment $E[C_t(I_t - \alpha)] = 0$, with I_t the breach on day t
  - HAC (heteroskedasticity and autocorrelation consistent) inference or a block bootstrap, because crisis breaches cluster; then the Holm correction
- Stakes: capital is needed exactly on the days when marginal guarantees say least

The Research Loop with an AI Assistant

- The AI assistant: an LLM such as Claude, ChatGPT or Gemini
  - AI: artificial intelligence
  - LLM: large language model
- Literature, with DOIs checked on Crossref
- Replicate a known number first
- Three rival explanations: stale scale, changing tail shape, regime-selection artefact
- Pre-register the tests
- Ask the assistant for a hostile referee report (the prompts: Appendix)

Mini-Case: Does a GARCH Scale Repair Crisis Coverage?

- Alternative score: the standardised loss $s_t = (L_t + \mu_t)/\sigma_t$, with μ_t, σ_t the GARCH-t forecasts of the mean and volatility
- H_0: $E[C_t(I_t - 0.05)] = 0$, i.e. the crisis breach rate of VaR 5% is 5%
  - t test with a HAC (Newey–West) variance, because crisis breaches cluster
  - Holm correction (a step-down adjustment of the p-values) within each experiment: four markets, and the markets with complete GARCH forecasts (n/a: withheld GARCH forecasts)

Market	Split conformal, full calendar			GARCH scale, common calendar		
	Crisis days	Rate	Holm p	Crisis days	Rate	Holm p
S&P 500	713	7.29%	0.187	613	2.77%	0.003
BET	520	8.08%	0.187	506	3.95%	0.233
Bitcoin	264	7.95%	0.187	n/a	n/a	n/a
EUR/RON	245	11.43%	0.119	n/a	n/a	n/a

ai_discovery_case.py

Mini-Case: Reading the Result

- Two calendars: split conformal on the full calendar of the conformal section
  - the GARCH scale needs 250 calibration days: the S&P 500 has 4 709 days, 613 of them crisis days
  - on these days, the split conformal crisis rate is 7.01%
- Split conformal: crisis rates above 5% on every market, from 7.29% to 11.43%
  - with HAC inference and the Holm correction, no rejection at 5% (smallest Holm p: 0.119, EUR/RON)
- GARCH scale on the S&P 500: 2.77% in crises, below 5% (Holm p 0.003), so too conservative in crises
- GARCH scale on BET: 3.95%, the gap to 5% is not significant (Holm p 0.233)
- Non-rejection does not prove crisis coverage: crisis samples are small and HAC inference is asymptotic
- Type of evidence: comparison of procedures

The Practical Decision



National Bank of Romania, Bucharest

Photo: Britchi Mirela (2018); CC BY-SA 4.0; Wikimedia Commons

- **What I choose**: FHS as the main model, GARCH-EVT as the alternative, revisited as evidence accumulates
  - FHS has the lowest mean FZ0 on the S&P 500; on BET, FHS and GARCH-EVT tie to three decimals
- **Under which conditions**
  - FHS and GARCH-EVT pass Z_2 at 5% and stay in the 90% MCS on the S&P 500 and BET (full scorecard: Appendix)
  - The evidence comes from long samples: one year of data has low power
- **What result would make me drop it**
  - A Z_2 rejection at 5% or an exit from the 90% MCS at evaluation dates fixed in advance, adjusted for repetition
  - In daily monitoring: the ES e-process above 20, i.e. the 5% level
  - A rejection of the crisis moment $E[C_t(I_t - \alpha)] = 0$ (HAC inference, Holm)
  - Already against it: DQ rejects every S&P 500 model, FHS included
- A Romanian portfolio (BET and EUR/RON): Seminar 8, C2

Check Your Understanding, then Chapter 9

Retrieval questions

- Why does the rule "$\mathrm{LR}_{uc} > 3.84$" reject a correct VaR 1% in about one year in ten?
- Why can a model with the right number of breaches still fail?
- Why can ES be backtested but not ranked on its own?
- What does ACI guarantee?
- What does ACI not guarantee?

Chapter 9: realised volatility

- Measure daily variance from intraday returns
- Evaluate volatility forecasts with robust scores such as QLIKE (quasi-likelihood loss) and the MCS

Numeric Self-Checks

1. Kupiec: $x = 3$ breaches of VaR 1% in $T = 500$ days; compute $\hat{\pi}$ and LR_{uc}, then decide at 5%
2. FZ0: one day, $\alpha = 5\%$, $v = 2\%$, $e = 2.5\%$
   - compute the score for $L = 3\%$ and for $L = 1\%$, in percent and in decimals
3. Local power: $T = 1000$, $\alpha = 1\%$, true $\pi = 2\%$
   - compute λ and the power, then compare with the dashed curve of the power chart
4. Z_1 and Z_2: $T = 250$, $\alpha = 2.5\%$, five breaches with $L_t/\mathrm{ES}_t = 1.1,\ 0.9,\ 1.3,\ 1.0,\ 1.2$
   - compute both and explain their signs

Answers with intermediate steps: Appendix. Worked exercises on the course data: Seminar 8, A1, A4, B2

Which Data? Four Markets

- S&P 500: a developed, liquid equity market
- BET (Bucharest Stock Exchange index): an emerging, less liquid market
- Bitcoin: trades every day and has very heavy tails
- EUR/RON: the BNR (National Bank of Romania) reference rate, a managed float
- Why these: four different volatility and tail regimes, on which the same tests can behave differently
- S&P 500, BET and Bitcoin: daily closes
  - each series keeps its own trading calendar

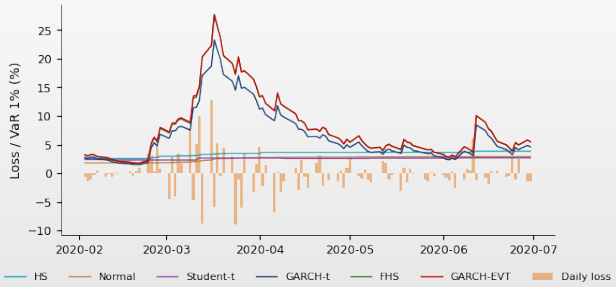
How Are Forecasts Evaluated? The Rolling Window

- Rolling window of 1 000 observations (about four years of equity trading)
  - enough data for GARCH and 100 exceedances for the GPD tail
- Each forecast uses data up to the previous day only (no look-ahead bias)
- GARCH, Student-t and EVT (Extreme Value Theory) parameters are refitted every 20 days (about one trading month)
  - the GARCH volatility is updated daily
- Evaluation starts on 01.01.2007 or, for a shorter series, after its first 1 000 observations
  - last day: 18.09.2026
- S&P 500 and BET (from 2007): 4 959 and 4 937 evaluation days
- EUR/RON (rate from July 2005, so evaluated from 2009): 4 352 days
- Bitcoin (data from September 2014, so evaluated from 2017): 3 384 days

GARCH Forecasts Withheld on Bitcoin and EUR/RON

- A GARCH(1,1)-t fit is **admissible** if the optimiser converged and $\omega > 0$, $a, b \geq 0$, $\nu > 2$, $a + b < 1$
  - with a numerical margin of 10^{-8}
  - i.e. the unconditional variance $\omega/(1 - a - b)$ is finite
- At an inadmissible refit we keep the last admissible parameters
  - the volatility is still filtered on current data
- With no earlier admissible fit, the GARCH-t, FHS and GARCH-EVT forecasts are **withheld**
  - neither computed nor imputed
- Bitcoin: the first 1 660 of 3 384 evaluation days (the first 83 refits all have $a + b \geq 1 - 10^{-8}$)
- EUR/RON: the first 20 of 4 352 evaluation days (the first refit)
- Full-sample GARCH results and the MCS are reported only on the S&P 500 and BET
  - the MCS needs all six models
  - "n/a" in tables, "—" in charts

March 2020: The Stress Test Nobody Planned



- S&P 500, February–June 2020: daily losses (bars) and the VaR 1% of the six models
- Largest loss: 12.77%, on 16.03.2020
- A correct VaR 1% has 1.04 expected breaches in these 104 days
- Breaches, unconditional models: HS 12, Normal 17, Student-t 17
- Breaches, conditional models: GARCH-t 3, FHS 2, GARCH-EVT 2
- A time-varying scale mattered more than the tail: Student-t (heavy tail, fixed scale) has 17 breaches

MFM_ch8_var_forecasts

Local Power by Hand

- $\chi^2_1(\lambda)$ is the law of $(Z+\sqrt{\lambda})^2$ with $Z \sim N(0,1)$, and $3.84 = 1.96^2$, so
  - power $= P(|Z+\sqrt{\lambda}| > 1.96) = \Phi(\sqrt{\lambda} - 1.96) + \Phi(-\sqrt{\lambda} - 1.96)$, Φ: the $N(0,1)$ distribution function
- $T = 250$, $\pi = 2\%$: $\lambda = 2.525$, $\sqrt{\lambda} = 1.589$
  - power $= \Phi(-0.371) + \Phi(-3.549) = 0.355 + 0.0002 = 0.356$: the 36% of the dashed curve
- The exact power is lower, 24%: the test accepts x from 1 to 6, an event of probability 75.7% when $\pi = 2\%$
  - with 2.5 expected breaches a year, the count is too discrete and too skewed for the Normal approximation behind $\chi^2_1(\lambda)$

Independence on the Full S&P 500 Sample

- p-values of LR_{ind}, VaR 1%: HS, Normal and Student-t <0.001
  - GARCH-EVT 0.031; FHS 0.046; GARCH-t 0.095
- Independence is rejected at 5% for every model except GARCH-t
- After a Holm correction for the six tests, only HS, Normal and Student-t stay rejected
- GARCH-EVT: a breach follows a breach with probability 0.053, against 0.011 after a calm day
- The joint CC test does not reject GARCH-EVT ($p = 0.058$): its correct coverage part dilutes the signal

MFM_ch8_coverage_tests

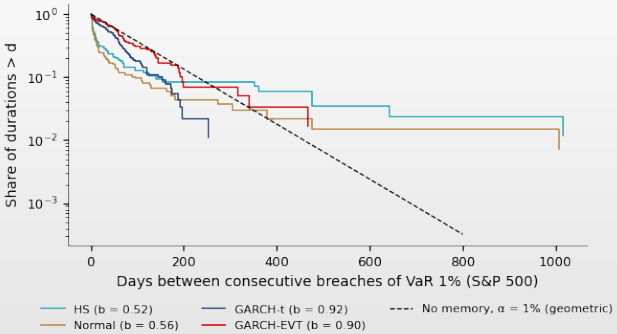
Duration-Based Tests

- What it tests: whether the gaps between breaches are those of a memoryless process
  - it uses every gap, not only back-to-back pairs
- Duration D_i: the number of days between breach $i-1$ and breach i
- For a correct VaR, D_i is geometric with mean $1/\alpha$ (100 days for $\alpha = 1\%$)
- Christoffersen & Pelletier (2004): approximate the discrete durations by the Weibull density $f(D; a, b) = a^b \, b \, D^{b-1} e^{-(aD)^b}$
- $a > 0$: rate (inverse scale); $b > 0$: shape (not the GARCH a, b)
- $b = 1$: the exponential, memoryless case
- $b < 1$: short durations are too frequent, i.e. breaches cluster

The Duration Test: the Statistic

- H_0: $b = 1$ with a free, against H_1: $b \neq 1$
- $\mathrm{LR}_{dur} = 2\left[\ell(\hat{a}, \hat{b}) - \ell(\tilde{a}, 1)\right] \xrightarrow{d} \chi^2_1$, where $\tilde{a}$ maximises the likelihood with $b = 1$
- $\ell(a, b) = \sum_i [C_i \ln S(D_i) + (1 - C_i) \ln f(D_i)]$, with $S(D) = e^{-(aD)^b}$ the survival function
- $C_i = 1$ for the first and last durations, which are censored
  - the first does not start at a breach, the last does not end at one
- The test does not check the mean $1/\alpha$: coverage is left to the Kupiec test
- At $T \geq 1\,000$ it detects clustering better than the conditional coverage test
  - at $T = 250$ it does not (the Monte Carlo at the end of the section)

How Long Between Breaches?



- Axes: duration d in days (horizontal) and the share of durations longer than d (vertical)
  - the vertical axis is on a log scale
- Dotted line: memoryless (geometric), $\alpha = 1\%$
- HS: $b = 0.52$, $p < 0.001$; median duration 7 days
- Normal: $b = 0.56$
- GARCH-t: $b = 0.92$, $p = 0.344$; median duration 45 days
- GARCH-EVT: $b = 0.90$, $p = 0.317$
- The long HS tail: years without breaches after a crisis leaves a high VaR in the window

MFM_ch8_coverage_tests

Regression-Based Backtests: the DQ Test

- What it tests: whether breaches can be predicted from any information known yesterday
  - not only from yesterday's breach
- Under conditional calibration $\mathrm{Hit}_t = I_t - \alpha$ is a martingale difference
  - $E[\mathrm{Hit}_t \mid \mathcal{F}_{t-1}] = 0$, $\mathrm{Var}(\mathrm{Hit}_t \mid \mathcal{F}_{t-1}) = \alpha(1-\alpha)$
- Regress $\mathrm{Hit}_t = X_t\beta + u_t$ with $X_t = (1, \mathrm{Hit}_{t-1}, \ldots, \mathrm{Hit}_{t-4}, \mathrm{VaR}_t)$ and test $\beta = 0$ (Engle & Manganelli (2004))
- $\mathrm{DQ} = \dfrac{\hat{\beta}' X' X \hat{\beta}}{\alpha(1-\alpha)} = \dfrac{\mathrm{Hit}' X (X'X)^{-1} X' \mathrm{Hit}}{\alpha(1-\alpha)} \xrightarrow{d} \chi^2_6$
- X: the $T \times 6$ regressor matrix; Hit: the vector of centred hits
  - $\hat{\beta}$: the OLS (ordinary least squares) estimator
  - u_t: the error
- DQ: the explained sum of squares, scaled by the null variance

The DQ Test: Special Cases and Size

- With $X_t = 1$: the score form of the Kupiec test, $T(\hat{\pi} - \alpha)^2/(\alpha(1-\alpha))$
- With $X_t = (1, \mathrm{Hit}_{t-1})$: a 2-degree-of-freedom analogue of LR_{cc}
  - asymptotically equivalent but not numerically equal
- Testing only the lag coefficient corresponds to LR_{ind}
- What it adds: power against memory beyond lag 1 and against breaches that depend on the VaR level
- Limitation: the χ^2_6 approximation over-rejects a correct model
  - the Monte Carlo at the end of the section
- The DQ p-values in this chapter are asymptotic; critical values have to be simulated
  - under the data-generating process of the Monte Carlo experiment, the 95% quantile of DQ at $T = 4\,959$ is 14.2, against 12.59 for χ^2_6
  - FHS and GARCH-EVT on the S&P 500: simulated frequencies of 0.6% and 0.5%, against asymptotic p-values of 0.014% and 0.011%
  - a benchmark under this teaching process, not definitive p-values for the empirical data

Size and Power of VaR Backtests: a Monte Carlo

- DGP (data generating process, the simulated truth): GARCH(1,1)-t fitted to the S&P 500 1990–2026
  - $a = 0.102$, $b = 0.892$, $\nu = 6.1$
  - VaR 1%; 1000 paths; rejection rates (%) at 5%

Forecaster	T	POF	Cond. cov.	Duration	DQ
Correct (size)	250	**9.7**	1.1	15.0	8.6
	1000	6.4	3.3	8.0	11.0
	5000	4.9	3.4	4.6	**8.8**
Normal-GARCH (power)	250	**10.6**	4.8	14.8	18.8
	1000	30.2	19.9	8.2	30.2
	5000	83.7	77.2	7.3	70.7
HS, 1000 days (power)	250	51.5	17.0	**12.0**	32.5
	1000	46.4	46.9	68.3	73.8
	5000	57.1	86.6	100.0	99.5

MFM_ch8_size_power

Size and Power of VaR Backtests: How to Read the Table

- Correct: true parameters and σ_t; the rejection rate is the size (target: 5%)
- Normal-GARCH: true σ_t, Normal tail; the rejection rate is power
- HS, 1 000 days: no volatility dynamics; the rejection rate is power
- An undefined duration p-value counts as non-rejection
  - undefined: fewer than 3 durations, in 28% of correct samples at $T = 250$
- Monte Carlo standard error: 0.7 percentage points at 5% and 1.6 at 50%
- In the bold POF cell (9.7%) the standard error is 0.9: the size is clearly above 5%

Size and Power of VaR Backtests: What the Table Shows

- Within a year, a Normal tail on the correct volatility is nearly invisible
  - POF rejects 10.6% of the time, about its size (9.7%)
- HS clustering is rarely detected by the duration test at $T = 250$ (12.0%)
- At $T = 1\,000$ the duration test (68.3%) and DQ (73.8%) detect HS far better than POF (46.4%)
- The conditional coverage test is conservative: size 1.1% at $T = 250$
- DQ over-rejects a correct model even at $T = 5\,000$ (8.8%): its critical values should be simulated
- At $T = 250$ no test has both correct size and high power

Backtests Under Estimation Risk

- Forecasts use $\hat{\theta}$ estimated on R days and the test uses P out-of-sample days
  - the hits $\hat{I}_t = 1\{L_t > \mathrm{VaR}_t(\hat{\theta})\}$ are not Bernoulli(α) even under a correct model
- The correction depends on the forecasting scheme (fixed, rolling, recursive) and grows with P/R (Escanciano & Olmo (2010))
  - expansion in the Appendix
- Fixed scheme, 400 GARCH(1,1)-t paths, $R = 1000$
  - the fit is admissible on 324 of them (simulation results)
- This chapter uses a rolling window ($R = 1000$, refitted every 20 days)
  - a fixed-scheme experiment does not measure the size of these rolling-window tests
- Remedy: reproduce the estimation inside the calibration
  - e.g. a parametric bootstrap that re-estimates the model on every simulated path
  - Seminar 8, B9, also discusses the selection of admissible paths

Du and Escanciano (2017): Cumulative Violations

- What it tests: how deep into the tail losses go, not only whether they breach VaR
  - it uses the forecast distribution through the PIT
- PIT: $u_t = F_t(L_t)$, uniform on $(0,1)$ if the forecast distribution F_t is right
- Cumulative violation $H_t = \frac{1}{\alpha}\left(u_t - (1-\alpha)\right) 1\{u_t > 1-\alpha\}$
  - under H_0: $E[H_t] = \alpha/2$, $\mathrm{Var}(H_t) = \alpha(1/3 - \alpha/4)$
- Unconditional test: $U = \sqrt{T}\left(\bar{H} - \alpha/2\right)/\sqrt{\alpha(1/3-\alpha/4)} \xrightarrow{d} N(0,1)$, with $\bar{H}$ the sample mean of H_t
- The sign of U
  - $U > 0$: losses deeper than forecast (risk underestimated)
  - $U < 0$: shallower (risk overestimated)
- The heat map in all ES tests at a glance gives the two-sided $p = 2\Phi(-|U|)$
  - the one-sided 5% test rejects for $U > 1.645$
- Conditional test (Du & Escanciano (2017)): the Box–Pierce statistic $C_5 = T\sum_{j=1}^{5}\hat{\rho}_j^2 \xrightarrow{d} \chi^2_5$
  - $\hat{\rho}_j$: the autocorrelations of H_t

Multinomial VaR Backtests: an Implicit ES Test

- ES 2.5% averages the quantiles beyond VaR 2.5%: test several VaR levels at once (Kratz, Lok & McNeil (2018))
- $N = 4$ levels $\alpha_j = 2.5\%\,(1 - (j-1)/4)$: 2.5%, 1.875%, 1.25% and 0.625%
  - by its PIT u_t each day falls in one of 5 cells
- Pearson statistic: $S = \sum_{j=0}^{4} (O_j - Tp_j)^2/(Tp_j) \xrightarrow{d} \chi^2_4$, with O_j the observed and Tp_j the expected days in cell j
- Cell probabilities: $p_0 = 97.5\%$, $p_1 = \cdots = p_4 = 0.625\%$
  - a large S flags counts far from the expected ones

p-value	S&P 500	BET	Bitcoin	EUR/RON
HS	<0.001	0.725	0.564	0.004
Normal	<0.001	<0.001	<0.001	<0.001
Student-t	<0.001	0.024	0.015	0.172
GARCH-t	<0.001	<0.001	n/a	n/a
FHS	0.003	0.862	n/a	n/a
GARCH-EVT	0.378	0.853	n/a	n/a

- n/a: withheld GARCH result (slide on withheld forecasts)

Multinomial Cell Counts: S&P 500, GARCH-t

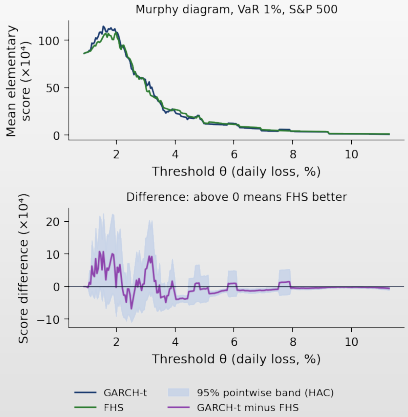
PIT interval	Observed days	Expected days
[0, 0.975]	4761	4 835.02
(0.975, 0.98125]	39	30.99
(0.98125, 0.9875]	46	30.99
(0.9875, 0.99375]	52	30.99
(0.99375, 1]	61	30.99

- The first cell lies outside the upper 2.5% loss tail
- Each of the four tail cells contains more days than expected under calibration

Is the Ranking Robust to the Score? Elementary Scores

- A quantile has infinitely many consistent scores: $S(v, L) = (1\{L \leq v\} - \tau)(g(v) - g(L))$, with g increasing and $\tau = 1 - \alpha$
- Each is a mixture of **elementary scores** (Ehm, Gneiting, Jordan & Krüger (2016))
  - $S_\theta(v, L) = (1\{L < v\} - \tau)\,(1\{\theta < v\} - 1\{\theta < L\})$
  - $\theta \in \mathbb{R}$: a threshold on the loss scale
- S_θ penalises a forecast only when v and L fall on opposite sides of θ
- The pinball loss is $\int S_\theta \, d\theta$
- A is preferred to B under **every** consistent score if and only if $\bar{S}_\theta(A) \leq \bar{S}_\theta(B)$ for all θ
  - $\bar{S}_\theta$: mean elementary score over T days
- **Murphy diagram**: $\bar{S}_\theta$ against θ
- If the curves cross, the ranking depends on the chosen score

Murphy Diagram: GARCH-t versus FHS (1/2)



MFM_ch8_comparative

Murphy Diagram: GARCH-t versus FHS (2/2)

- S&P 500, VaR 1%, 2007–2026; θ from 0.9% to 11.3%
  - pinball ($\times 10^4$): GARCH-t 3.852, FHS 3.861, a tie under the usual score
- Below $\theta = 5\%$ the curves cross
  - FHS better on 47% of thresholds, band containing zero on 86%
  - above 5% GARCH-t lower on 89%: no dominance
- The 95% bands use HAC standard errors and are pointwise, not a simultaneous test of dominance
  - HAC: heteroskedasticity- and autocorrelation-consistent
- State the score before looking at the results

MFM_ch8_comparative

Giacomini–White: Comparing Forecasting Methods

- DM targets the population accuracy of **models**
- With rolling 1 000-day windows and refits we compare **forecasting methods**
  - estimation error does not vanish as T grows
- Giacomini & White (2006): H_0: $E[d_t \mid \mathcal{F}_{t-1}] = 0$ for $d_t = \mathrm{FZ0}_{A,t} - \mathrm{FZ0}_{B,t}$
- The test is valid for nested models too, with finite rolling windows (not with recursive ones)
- What it adds: it tests whether the loss difference can be predicted from information known at $t-1$
  - not only whether the mean gap is zero

Giacomini–White: Statistic and Results

- Test function $h_{t-1} = (1, d_{t-1})'$, known at $t-1$; $Z_t = h_{t-1} d_t$
- $\mathrm{GW} = T\,\bar{Z}'\hat{\Omega}^{-1}\bar{Z} \xrightarrow{d} \chi^2_2$, $\hat{\Omega} = T^{-1}\sum_t Z_t Z_t'$ (one-step forecasts)
- $\bar{Z}$: the mean of Z_t; 2 degrees of freedom, one per instrument; $h_{t-1} = 1$ gives the DM form
- S&P 500, against FHS: GARCH-EVT $\mathrm{GW} = 6.71$ ($p = 0.035$), GARCH-t 7.60 ($p = 0.022$), HS 53.16 ($p < 0.001$)
- GARCH-EVT, tied under DM, is rejected conditionally
  - equal conditional predictive ability is rejected for the instruments h_{t-1}
- The rejection alone does not show when FHS scores lower
- Which instrument drives the rejection shows only when each is tested separately

Three-Zone Comparative Backtesting (Nolde and Ziegel)

- Compare the internal model A with a standard model S (here HS) by mean FZ0 loss (Nolde & Ziegel (2017))
  - with two one-sided 5% tests
- H_0^-: A is at least as good as S; H_0^+: A is at most as good
- **Green**: reject H_0^+ ($\mathrm{DM} < -1.645$); **red**: reject H_0^- ($\mathrm{DM} > 1.645$); **yellow**: neither
- E.g. $\mathrm{DM} = -1.8$ is green, although $|\mathrm{DM}| < 1.96$
- The burden of proof flips: the bank must show its model is better, not the supervisor that it is worse

DM (A minus HS)	S&P 500	BET	Bitcoin	EUR/RON
GARCH-t	−4.21	−5.25	n/a	n/a
FHS	−4.69	−5.53	n/a	n/a
GARCH-EVT	−4.67	−5.57	n/a	n/a

- Colour gives the statistical zone; n/a: withheld GARCH result (slide on withheld forecasts)

The Model Confidence Set: the Statistic

- For model i in the current set $\mathcal{M}$: $d_{i\cdot} = \bar{L}_i - \frac{1}{|\mathcal{M}|}\sum_{j\in\mathcal{M}} \bar{L}_j$
  - its mean loss relative to the set average
- $t_i = d_{i\cdot}/\widehat{\mathrm{se}}(d_{i\cdot})$ and $T_{\max} = \max_{i\in\mathcal{M}} t_i$; H_0: all models in $\mathcal{M}$ have the same expected loss
- $\widehat{\mathrm{se}}$ and the null distribution of $T_{\max}$ come from a circular block bootstrap with 10-day blocks
  - with 1 000 replications
- If H_0 is rejected, drop the model with the largest t_i and repeat
- MCS p-value of a model: the largest p-value of the tests up to its elimination
- With 1 000 replications the smallest resolvable p-value is about 0.001: a "0.000" in the chart means $p \leq 0.001$

Case Study (1/5): What Is Estimated, Under Which Assumptions

- M-estimator: the parameters minimise the sample mean of an FZ0-type loss
- Asymptotic theory under misspecification (the paper, Theorem 1.4)
  - R packages esback and esreg
- Assumption 1, for strict and intercept ESR
  - the quantile equation is right only if VaR is an affine function of ES
  - e.g. pure scale models
- The covariance estimator used here estimates some components under a location-scale assumption
  - the sandwich covariance under misspecification does not, in general, need this assumption
- Sandwich covariance: $A^{-1}BA^{-1}$; A: curvature of the mean loss; B: variance of its gradient
- Location-scale model: $r_t = \mu_t + \sigma_t z_t$ with i.i.d. standardised shocks z_t

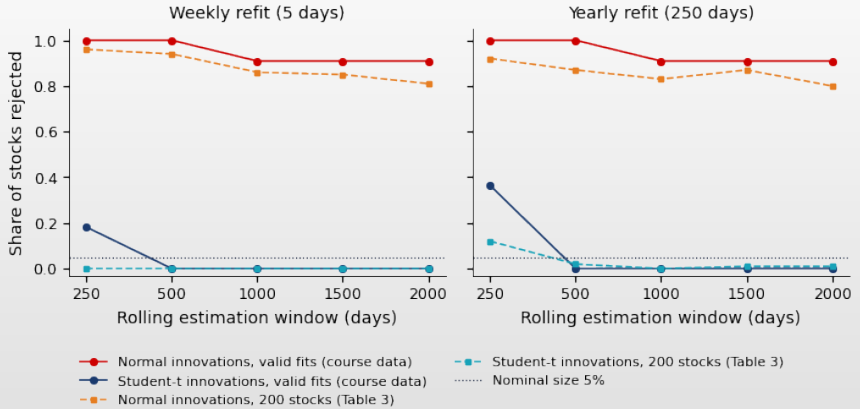
Case Study (2/5): the Paper's Monte Carlo Design

- The paper, Sec. 2: 10 000 replications, T from 250 to 5 000
- EGARCH-t (exponential GARCH, Student-t shocks), calibrated to the S&P 500
- AR(1)-GARCH(1,1): first-order autoregressive mean, GARCH variance
- Two Student-t GAS (generalised autoregressive score) models
- Benchmarks: McNeil & Frey (2000) and the calibration tests of Nolde & Ziegel (2017)

Case Study (3/5): Size and Power in the Paper's Simulations

- Actual size at nominal 5%, EGARCH-t, $T = 250$ and $T = 1\,000$ (Table 1)
  - strict ESR, robust covariance: 0.09 and 0.05
  - strict ESR, classical covariance: 0.24 and 0.11
  - simple conditional calibration test: 0.29 and 0.14
- Power (Fig. 3): exceedance residual tests barely react to two errors
  - wrong volatility dynamics or a wrong level $\tilde{\tau} \neq \tau$
  - ESR detects both
- A bank could report ES at a less extreme level without being rejected by the older tests

Case Study (4/5): Table 3 Replicated on the Course Data



MFM_ch8_esr_case

Case Study (5/5): Table 3 Replicated on the Course Data

- Solid line: 11 large S&P 500 stocks, January 2010 – August 2019 (2 432 days); dashed line: Bayer & Dimitriadis (2022), Table 3
- GARCH(1,1) with constant mean; 25 window–refit combinations (the 5×5 grid)
  - intercept ESR with esback
- Normal innovations: the rejected share is 0.82–1.00 (paper: 0.80–0.96)
- Student-t, windows of at least 1 000 days: no stock rejected (paper: at most 0.01)
- Student-t, 250-day windows: 0.09–0.55 (paper: 0.00–0.12)
- With 11 stocks each rejection adds $1/11 \approx 0.09$, and the standard error of a share reaches $\sqrt{0.25/11} \approx 0.15$, an i.i.d. benchmark: the stocks share market shocks
- The central result replicates
  - the gap at the 250-day window is large even against this standard error

MFM_ch8_esr_case

Further Reading and Resources

- Fissler & Ziegel (2016): VaR and ES are jointly elicitable, the basis of the FZ0 loss
- Nolde & Ziegel (2017): comparative backtests and the traffic light for VaR and ES
- Acerbi & Szekely (2014): model-free ES backtests, including Z_2
- Entropy-based VaR and ES forecasts: Pele, Lazar & Dufour (2017), Pele & Mazurencu-Marinescu-Pele (2019), Pele, Lazar & Mazurencu-Marinescu-Pele (2019); when two ES forecasts can be told apart: Pele & Mazurencu-Marinescu-Pele (2026)
- Textbooks: Franke, Härdle & Hafner (2019) (VaR and backtesting chapters) and Härdle, Chen & Overbeck (2017)
- Quantinar: Machine learning in Financial Risk
- Code and data: github.com/danpele/MFM/tree/main/Quantlets/Ch_08
  - S&P 500, BET, Bitcoin: daily closes from the course data
  - EUR/RON: BNR reference rate

Back

Estimation Risk, the Fixed-Scheme Expansion

- Fixed scheme (Escanciano & Olmo (2010)): $\frac{1}{\sqrt{P}} \sum_t (\hat{I}_t - \alpha) = \frac{1}{\sqrt{P}} \sum_t (I_t - \alpha) + \sqrt{P/R}\, D' \sqrt{R}(\hat{\theta} - \theta_0) + o_p(1)$
- First term: hits centred at the true parameters θ_0
  - second: the effect of the estimation error $\hat{\theta} - \theta_0$
- $D = E\left[\partial P(L_t > \mathrm{VaR}_t(\theta) \mid \mathcal{F}_{t-1})/\partial\theta\right]$ at θ_0
  - the sensitivity of the breach probability to the parameters
- The estimation term adds variance of order P/R to $\sqrt{P}(\hat{\pi} - \alpha)$ and vanishes only if $P/R \to 0$
- So the χ^2 critical values are too small when P/R is large
- Assumptions: smooth $\mathrm{VaR}_t(\theta)$, $\sqrt{R}$-consistent $\hat{\theta}$, stationarity
  - rolling and recursive schemes change the form of the correction
- Remedies: the corrected variance or subsampling of Escanciano & Olmo (2010)
  - or a parametric bootstrap that refits the model on every simulated path

Back

Estimation Risk: Simulation Results

- Fixed scheme: 400 paths, $R = 1000$, Kupiec test at 5%
- Fits on the 400 paths
  - 324 admissible
  - 75 numerically unidentified
  - 1 fails the finite-variance admissibility condition
- We report only the true-parameter rates, on all 400 paths
  - a rate on the 324 admissible paths would be conditional on estimation success
- True parameters: 10.0% at $P = 250$, 5.2% at $P = 1000$, 4.0% at $P = 5000$
- Monte Carlo standard error at a 5% rejection probability: about 1.1 percentage points

Back

Osband's Principle and Identification Functions

- Osband's principle (Fissler & Ziegel (2016)): consistent scores are built from an identification function V
  - V has $E_F[V(x, L)] = 0$ only at the true value
- VaR: $V(v, L) = 1\{L \leq v\} - (1 - \alpha)$; the pinball loss is its integral in v
- (VaR, ES): $V_1 = \alpha - 1\{L > v\}$ and $V_2 = e - v - \frac{1}{\alpha} 1\{L > v\}(L - v)$
  - their expectations vanish only at the true pair
- The level sets of ES are not convex under mixtures
  - so no identification function of e alone exists (Seminar 8, A8)
- FZ0 is the member of the FZ family whose differences are invariant to the scale of the losses (Patton, Ziegel & Chen (2019))

Back

Sequential Monitoring of Risk Forecasts

- A regulator checks the forecasts every day
  - repeating a fixed-sample test (e.g. rolling Z_2) inflates the probability of at least one false alarm
- Hoga & Demetrescu (2023): monitoring statistics for VaR and ES with boundary functions
  - the boundary functions keep the probability of any false alarm over the whole horizon below the nominal level
- Compared with the traffic light: evidence accumulates from the start of monitoring
  - instead of a single 250-day snapshot
- The ES regression backtest (Bayer & Dimitriadis (2022)) of the case study can be monitored the same way
  - both address the one-shot design of the traffic light

Back

Conformal Prediction Beyond Exchangeability

- Barber, Candès, Ramdas & Tibshirani (2023): coverage without exchangeability
  - the coverage gap is bounded by weighted total-variation distances
  - between the score vector and the same vector with the test score swapped with calibration score i
- How large these distances are under volatility clustering is an empirical question
- Gibbs, Cherian & Candès (2025): exact coverage conditional on a pre-specified class of covariate shifts is attainable
  - which class captures a financial crisis is open
- Choosing the ACI step γ: Zaffran et al. (2022) aggregate ACI experts over a grid of γ values

Back

Case Study, Further Key Results

Finding	Source	Reading
Strict and Auxiliary ESR sizes differ by at most 0.01 for every process and sample size, including AR(1)-GARCH with $\phi = 0.5$	Table 1	In these designs, dropping the VaR forecasts barely changes the size; power is compared separately (Fig. 2–3)
Student-t, 250-day window: the 0.12 share comes from GARCH with a yearly refit; from 500 days GJR-GARCH-t and GARCH-t differ by at most 0.01	Table 3	Leverage in the variance equation barely changes the rejection shares once the window is long

Back

Case Study, Limitations and Extensions

- Limitation stated by the authors: auxiliary and strict ESR are two-sided only (the paper, Sec. 1.3)
- Limitation stated by the authors: the classical covariance over-rejects in small samples
- Replication (Seminar 8, C3): apply strict ESR (two-sided) and intercept ESR (one-sided)
  - to the ES 2.5% forecasts of the six models on the four markets
  - compare with Z_2 and McNeil–Frey
- Extension: replicate Table 3 with GJR-GARCH, more stocks and the period after August 2019
- Open question: the size of the two-sided intercept ESR at $T = 250$ (Table 1, EGARCH-t)
  - 0.04 with the robust covariance
  - 0.15 with the classical covariance
- The one-sided size is reported only for $T = 2\,500$, 0.02 (Table 2): what is it over one year?
- Can a bootstrap of the t-statistic hold it at 5% without losing the power of Fig. 4?

Back

Scorecard Across Four Markets

- Each cell: breach rate of VaR 1% in % / Z_2 p-value / MCS p-value (FZ0)
  - $\leq$0.001: at the resolution of the 1 000 bootstrap replications

Model	S&P 500	BET	Bitcoin	EUR/RON
HS	1.67 / <0.001 / $\leq$0.001	1.17 / 0.134 / $\leq$0.001	0.77 / 0.863 / n/a	0.51 / 1.000 / n/a
Normal	2.70 / <0.001 / $\leq$0.001	1.90 / <0.001 / $\leq$0.001	1.74 / 0.001 / n/a	1.06 / 0.973 / n/a
Student-t	2.00 / 0.002 / $\leq$0.001	1.58 / 0.003 / $\leq$0.001	0.71 / 0.625 / n/a	0.85 / 0.920 / n/a
GARCH-t	1.77 / <0.001 / 0.010	1.46 / <0.001 / 0.096	n/a / n/a / n/a	n/a / n/a / n/a
FHS	1.23 / 0.209 / 1.000	1.20 / 0.236 / 1.000	n/a / n/a / n/a	n/a / n/a / n/a
GARCH-EVT	1.15 / 0.145 / 0.211	0.97 / 0.215 / 0.966	n/a / n/a / n/a	n/a / n/a / n/a

- n/a: withheld GARCH forecast, hence no MCS (slide on withheld forecasts)

Back

Answers to the Numeric Self-Checks (1/2)

1. $\hat{\pi} = 0.006$; $\ln L(\alpha) = 497 \ln 0.99 + 3 \ln 0.01 = -18.81$, $\ln L(\hat{\pi}) = 497 \ln 0.994 + 3 \ln 0.006 = -18.34$
   - $\mathrm{LR}_{uc} = 2(-18.34 - (-18.81)) = 0.94 < 3.84$, $p = 0.33$: not rejected; 5 breaches were expected
2. $L = 3\%$: $\frac{3-2}{0.05 \cdot 2.5} + \frac{2}{2.5} + \ln 2.5 - 1 = 8 + 0.8 + 0.916 - 1 = 8.716$; $L = 1\%$: $0.8 + 0.916 - 1 = 0.716$
   - in decimals only $\ln e$ changes, by $\ln 0.01 = -4.605$: 4.111 and -3.889; the difference between the two days is the same

Self-checks

Answers to the Numeric Self-Checks (2/2)

3. $\lambda = 1000 \cdot 0.01^{\mathbf{2}}/(0.01 \cdot 0.99) = 10.10$, $\sqrt{\lambda} = 3.178$
   - power $= \Phi(1.218) + \Phi(-5.138) = 0.888$ (the second term is below $10^{-\mathbf{6}}$): the 88.8% of the dashed $T = 1000$ curve; exact power 78%
4. $\sum_t L_t I_t / \mathrm{ES}_t = 5.5$, $N_T = 5$, $T\alpha = 6.25$
   - $Z_{\mathbf{1}} = 1 - 5.5/5 = -0.10$: on breach days losses are on average 10% deeper than the ES forecast
   - $Z_{\mathbf{2}} = 1 - 5.5/6.25 = 0.12 > 0$: fewer breaches than the 6.25 expected offset that depth; far from the 5% threshold -0.70

Self-checks

The Discovery Loop (1/3): Prompts

- Prompts (tools: Semantic Scholar, Elicit, Google Scholar; check every DOI on Crossref)
  - `List peer-reviewed work since 2019 on conformal prediction under serial dependence, with DOI, exact guarantee and assumptions.`
  - `Give three rival explanations for crisis over-breaches of split conformal VaR and what each predicts for breach rates by regime.`
  - `Act as a hostile referee: list the strongest reasons this crisis effect is an artefact and the test that would settle each.`

Back

The Discovery Loop (2/3): Checks the Researcher Runs

- **Look-ahead bias**
  - change the observations after day $t-1$ and confirm that VaR_t, the scale and the crisis flag do not move
- **Dependent breaches**
  - Christoffersen and duration tests, or a block bootstrap of the hit sequence
- **Power**
  - 613 crisis days give 6.1 expected breaches of VaR 1%
  - a true 2% rate is detected with probability 57% by an exact two-sided Binomial test at 5% (i.i.d. breaches assumed)

Back

The Discovery Loop (3/3): Further Checks

- **Multiplicity**: 4 markets $\times$ 2 levels $\times$ 3 methods $=$ 24 tests
  - Holm controls the FWER (family-wise error rate)
  - Benjamini–Hochberg controls the FDR (false discovery rate) under independence or positive dependence (PRDS: positive regression dependency on a subset)
  - under arbitrary dependence: Benjamini–Yekutieli
- **Misstated theorems**
  - e.g. "ACI guarantees conditional coverage"
- **Off-by-one quantiles**
  - $k = \lceil (n+1)(1-\alpha) \rceil$ is not the default of `np.quantile`

Back

Project Seed, Crisis-Conditional Coverage

- Question: can a distribution-free VaR 1% and ES 2.5% forecast achieve crisis-conditional coverage?
- Question: does this coverage hold in all four markets under a pre-registered test?
- Replicate first
  - split conformal VaR 5% crisis rates (S&P 500 7.29%, EUR/RON 11.43%)
  - the ACI share of infinite VaR 1% (13.4% of S&P 500 evaluation days)
- Deliverable: a pre-registration (hypotheses, tests, corrections, power) dated before the first run
- Deliverable: a candidate method, e.g. Gibbs, Cherian & Candès (2025) with the volatility regime as conditioning class
  - VaR tests by regime
  - ES tested with the ES regression backtest (Bayer & Dimitriadis (2022)) or Z_2
- Deliverable: code and a two-page report with every specification tried
  - a referee-style critique of your own result
- Log every prompt that shaped a design choice in `AI_USE.md`
  - every error the assistant made in `AI_ERRORS.md`: github.com/danpele/MFM/tree/main/project

Back

Acronyms used in this material (1/3)

- **ACI**: Adaptive Conformal Inference
- **AI**: Artificial Intelligence
- **AR**: AutoRegressive (model), as in AR(1)-GARCH
- **BCBS**: Basel Committee on Banking Supervision
- **BET**: Bucharest Exchange Trading (index)
- **BIS**: Bank for International Settlements
- **BNR**: Banca Națională a României — *the National Bank of Romania*
- **CC**: Conditional Coverage (test)
- **COVID-19**: COronaVIrus Disease 2019
- **DGP**: Data-Generating Process
- **DM**: Diebold–Mariano (test)
- **DOI**: Digital Object Identifier
- **DQ**: Dynamic Quantile (test)
- **EGARCH**: Exponential GARCH
- **ES**: Expected Shortfall
- **ESR**: Expected Shortfall Regression (backtest)

Acronyms used in this material (2/3)

- **EUR**: Euro
- **EVT**: Extreme Value Theory
- **FDR**: False Discovery Rate
- **FHS**: Filtered Historical Simulation
- **FRTB**: Fundamental Review of the Trading Book
- **FWER**: Family-Wise Error Rate
- **FZ**: Fissler–Ziegel (family of scoring functions for VaR and ES)
- **FZ0**: Fissler–Ziegel loss with zero-homogeneous score differences
- **GARCH**: Generalised AutoRegressive Conditional Heteroskedasticity
- **GARCH-EVT**: GARCH filter combined with Extreme Value Theory for the tail
- **GAS**: Generalised Autoregressive Score (model)
- **GJR-GARCH**: Glosten–Jagannathan–Runkle GARCH
- **GPD**: Generalised Pareto Distribution
- **GREE**: Growth Rate with Empirical e-statistics (e-backtest)
- **GREL**: Growth Rate with Empirical Losses (e-backtest)
- **GREM**: Growth Rate with a Mixture (of GREE and GREL)
- **GW**: Giacomini–White (test of conditional predictive ability)
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors)
- **HS**: Historical Simulation

Acronyms used in this material (3/3)

- **IMA**: Internal Models Approach
- **LLM**: Large Language Model
- **LR**: Likelihood Ratio
- **LRV**: Long-Run Variance
- **MCS**: Model Confidence Set
- **NASDAQ**: National Association of Securities Dealers Automated Quotations
- **OLS**: Ordinary Least Squares
- **P&L**: Profit and Loss
- **PIT**: Probability Integral Transform
- **POF**: Proportion Of Failures (test)
- **QLIKE**: Quasi-LIKElihood (loss function)
- **RON**: Romanian leu
- **S&P 500**: Standard & Poor's 500 (index)
- **VaR**: Value-at-Risk

References (1/4)

- Acerbi, C., Szekely, B. (2014). Backtesting expected shortfall. MSCI Inc., December 2014.
- Angelopoulos, A.N., Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511.
- Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J. (2023). Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2).
- Basel Committee on Banking Supervision (1996). Supervisory framework for the use of "backtesting" in conjunction with the internal models approach to market risk capital requirements. Bank for International Settlements.
- Basel Committee on Banking Supervision (2019). Minimum capital requirements for market risk (rev. February 2019). Bank for International Settlements.
- Bayer, S., Dimitriadis, T. (2022). Regression-based expected shortfall backtesting. *Journal of Financial Econometrics*, 20(3), 437–471.
- Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.
- Christoffersen, P., Pelletier, D. (2004). Backtesting value-at-risk: a duration-based approach. *Journal of Financial Econometrics*, 2(1), 84–108.
- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

References (2/4)

- Du, Z., Escanciano, J.C. (2017). Backtesting expected shortfall: accounting for tail risk. *Management Science*, 63(4), 940–958.
- Ehm, W., Gneiting, T., Jordan, A., Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society, Series B*, 78(3), 505–562.
- Engle, R.F., Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4), 367–381.
- Escanciano, J.C., Olmo, J. (2010). Backtesting parametric value-at-risk with estimation risk. *Journal of Business & Economic Statistics*, 28(1), 36–51.
- Fissler, T., Ziegel, J.F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4).
- 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.
- Gibbs, I., Candès, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34; arXiv:2106.00170.
- Gibbs, I., Cherian, J.J., Candès, E.J. (2025). Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society, Series B*, 87(4), 1100–1126.

References (3/4)

- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.
- Hansen, P.R., Lunde, A., Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Härdle, W.K., Chen, C.Y.-H., Overbeck, L. (eds.) (2017). Applied Quantitative Finance, 3rd ed. Springer.
- Hoga, Y., Demetrescu, M. (2023). Monitoring value-at-risk and expected shortfall forecasts. *Management Science*, 69(5), 2954–2971.
- Kratz, M., Lok, Y.H., McNeil, A.J. (2018). Multinomial VaR backtests: a simple implicit approach to backtesting expected shortfall. *Journal of Banking & Finance*, 88, 393–407.
- Kupiec, P.H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73–84.
- Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094–1111.
- 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.
- Newey, W.K., West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.

References (4/4)

- Nolde, N., Ziegel, J.F. (2017). Elicitability and backtesting: perspectives for banking regulation. *Annals of Applied Statistics*, 11(4).
- 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.
- Pele, D.T., Lazar, E., Dufour, A. (2017). Information entropy and measures of market risk. *Entropy*, 19(5), 226.
- Pele, D.T., Lazar, E., Mazurencu-Marinescu-Pele, M. (2019). Modeling expected shortfall using tail entropy. *Entropy*, 21(12), 1204.
- Pele, D.T., Mazurencu-Marinescu-Pele, M. (2019). Using high-frequency entropy to forecast Bitcoin's daily value at risk. *Entropy*, 21(2), 102.
- Pele, D.T., Mazurencu-Marinescu-Pele, M. (2026). Finite-sample precision limits for expected shortfall forecast comparisons. *Mathematics*, 14(13), 2316.
- Wang, Q., Wang, R., Ziegel, J. (2026). E-backtesting. *Management Science*, 72(6), 4952–4973.
- Zaffran, M., Féron, O., Goude, Y., Josse, J., Dieuleveut, A. (2022). Adaptive conformal predictions for time series. *Proceedings of the 39th International Conference on Machine Learning*; arXiv:2202.07282.