

7504558-why-a-simple-threshold-rule-is-hard-to-beat-volatility-regime-prediction-under-i

SUMMARY: The article investigates how quickly the ability to predict a volatility regime decays when a forecaster's information is delayed by δ days. Using nine years of daily S&P 500 data, the study evaluates four forecasters against a binary target: whether the realized variance five days hence exceeds its trailing 252-day median. A simple threshold rule, "is volatility above its median now," remains above a constant forecast out to three days of delay against the more demanding of the two constants a forecaster could actually have run. Neither the Yang-Zhang estimator nor a walk-forward HAR classifier is distinguishable from the threshold rule at any delay. The article explains that for a median-split target, under approximate joint lognormality, the threshold rule is the Bayes rule among functions of the delayed state alone. The study concludes that the threshold rule is the best cut on the delayed state, and the negative result is attributed to the target, not the models.

METHOD: The study uses daily open, high, low, and close data for the S&P 500 index over nine years. Two volatility proxies are computed over rolling five-day windows: the close-to-close proxy and the Yang-Zhang proxy. The target is defined as whether the close-to-close proxy exceeds its trailing 252-day median. Four forecasters are evaluated, and the forecasters' predictions are compared against the threshold rule. The study avoids look-ahead errors by comparing a delayed observation against a contemporaneous threshold.

KEY FINDINGS:

- The simple threshold rule remains above a constant forecast out to three days of delay.
- Neither the Yang-Zhang estimator nor the walk-forward HAR classifier is distinguishable from the threshold rule at any delay.
- For a median-split target, under approximate joint lognormality, the threshold rule is the Bayes rule among functions of the delayed state alone.

IMPLICATION FOR TRADING: The study's findings suggest that the threshold rule is the best cut on the delayed state, and the negative result is attributed to the target, not the models. Practitioners should consider the threshold rule as a robust baseline for volatility regime prediction under information delay, as the threshold rule outperforms more sophisticated models. However, practitioners should also be aware that the threshold rule does not provide a ceiling, as it is an approximation rather than a benchmark.

Why a Simple Threshold Rule Is Hard to Beat: Volatility Regime Prediction Under Information Delay

Sunny Alex Vellaniakaran

Independent

22 September 2026

Abstract

We ask how quickly the ability to predict a volatility regime decays when the forecaster's information is δ days out of date. On nine years of daily S&P 500 data we evaluate four forecasters against a binary target, namely whether realised variance five days hence exceeds its trailing 252-day median, at delays of zero to 55 trading days. A trivial rule, "is volatility above its median now", stays above a constant forecast out to three days of delay against the more demanding of the two constants a forecaster could actually have run, and a week and a half against the weaker one, with the crossing itself unresolved at the next delay in each case; beyond that the data cannot separate it from a constant. Neither the Yang-Zhang estimator (+2.0 points at zero delay, 95% CI $[-5.0, +8.2]$) nor a walk-forward HAR classifier (-4.0 points, $[-11.5, +2.5]$) is distinguishable from it at any delay; the intervals are wide enough to leave a HAR advantage of up to 12.5 accuracy points unexcluded, so what is shown is a failure to separate and not an equality. We explain that instead of reporting it: for a median-split target, under approximate joint lognormality, the threshold rule is the Bayes rule among functions of the delayed state alone. Measured accuracies sit at 85–106% of that restricted-information Gaussian benchmark, exceeding it at six of ten delays, so Section 5 treats it as an approximation, never a ceiling. The benchmark is the best cut on the delayed state rather than the same-sign rule: the correlation turns negative at the two longest delays, where a same-sign rule is beaten by its own negation and is therefore not a benchmark for anything. The estimated headroom is small relative to sampling uncertainty in this sample, so the negative result is about the target, not the models. Three robustness checks follow, on the loss, the windows and the split. The third breaks the median-split theory: at an upper quantile the Bayes cut becomes z_q/ρ instead of zero, and at the 90th percentile the Bayes rule itself attains only 90.1% against a majority-class 90.0%.

Keywords: volatility regimes; information delay; threshold rule; Bayes benchmark; forecast evaluation; binary targets.

JEL codes: C53, C58, G17, C22.

1. Introduction

A forecaster whose data arrives late is in a different position from one whose data is merely noisy. The question of how much predictive content survives a delay of δ periods is distinct from, though in population equivalent to, the familiar question of multi-step forecasting, and it has an obvious practical reading: how much is timely data worth?

This note reports a small, fully reproducible experiment on that question, and a negative result. The contribution is not a method. It is a measurement, a set of baselines that the measurement is compared against, and an argument for why the measurement could not have come out any other way.

Two things motivated the exercise. First, the volatility-forecasting literature almost never reports what was attainable, only what was attained; Andersen and Bollerslev [1] are the exception, and their

point, that a low R^2 may be a property of the target and no failure of the model, is the point made here in a different setting. Second, claims about model sophistication under delay are usually made without a trivial baseline in view, and trivial baselines on persistent series are unusually strong.

2. Data

Daily open, high, low and close for the S&P 500 index, 12 September 2017 to 11 September 2026, 2,262 trading days, from a retail export. The series was reordered to ascending and validated: dates strictly increasing, high not below low, high not below either open or close, low not above either, prices strictly positive. No survivorship or revision adjustment is possible with this source, which is a limitation recorded in Section 6.

3. Method

3.1 Volatility proxies

Two proxies are computed over rolling five-day windows. The close-to-close proxy is the mean of squared log returns over the window. For the second, write O_s , H_s , L_s and C_s for the open, high, low and close of session s , and $o_s = \log(O_s / C_{s-1})$, $c_s = \log(C_s / O_s)$, $u_s = \log(H_s / O_s)$, $d_s = \log(L_s / O_s)$ for the overnight, open-to-close, high and low log returns. The Yang-Zhang proxy combines overnight, open-to-close and Rogers-Satchell [7] components,

$$\sigma_{YZ}^2 = \sigma_o^2 + k \sigma_c^2 + (1 - k) \sigma_{RS}^2, \quad k = 0.34 / (1.34 + (n+1)/(n-1)), \quad (1)$$

where σ_o^2 and σ_c^2 are the sample variances of o_s and c_s over the n sessions ending at t , σ_{RS}^2 is the mean over them of $u_s(u_s - c_s) + d_s(d_s - c_s)$, and k is the weight for which the sum is minimum-variance and drift-independent under a Brownian price, with $n = 5$, following Yang and Zhang [8]. The two proxies differ in how they measure, not in what they see. Close-to-close combines the overnight and intraday moves in one return and ignores the high-low range, since $\log(C_s/C_{s-1}) = o_s + c_s$; Yang-Zhang separately identifies the overnight component and adds range information. The comparison therefore tests the value of the richer measurement, not whether overnight moves matter. An earlier version of this note said the close-to-close proxy discarded the overnight gap, which is wrong: it contains it, unseparated.

3.2 Target

Let v_t be the close-to-close proxy and M_t its median over the trailing 252 days, computed using data up to t only. The target is

$$y_t = \mathbf{1}\{v_{t+h} > M_t\}, \quad h = 5. \quad (2)$$

A forecaster who is δ days late knows the proxy only to $t - \delta$, and therefore knows $M_{t-\delta}$, not M_t . All four forecasters predict this same close-to-close-defined label; CC and YZ denote alternative information variables a forecaster may hold, not alternative evaluation targets, so every accuracy in this note is measured against one common event. Comparing a delayed observation against a contemporaneous threshold is a look-ahead error; we avoid it, and measure its size in Section 4.3.

3.3 Forecasters

Four forecasters, each seeing only information dated $t - \delta$ or earlier. *Majority* predicts the class that was more frequent in training. *Persist* predicts $\mathbf{1}\{v_{t-\delta} > M_{t-\delta}\}$, in close-to-close and Yang-Zhang variants. *HAR* is a logistic regression on the daily, weekly and monthly averages of the log proxy, the heterogeneous-autoregressive structure of Corsi [3] cast as a classifier, fitted by iteratively reweighted least squares on a rolling 1,250-day window, refitted every 21 days. Its training pairs are restricted to labels already resolved at $t - \delta$, so no label information crosses the information boundary.

3.4 Evaluation

Walk-forward over 694 test days, 27 November 2023 to 2 September 2026, after the 252-day median warm-up, the 1,250-day training window and the 55-day maximum delay are set aside. Because the five-day targets overlap, ordinary binomial intervals would be too narrow; we use a moving-block bootstrap [6] on the correctness series with block length $8h = 40$ and 2,000 resamples, and the same procedure on paired differences. An earlier version of this note used $2h = 10$, which is the shortest block that covers the overlap itself and nothing beyond it; the companion paper's bandwidth measurement finds the loss differential still carrying autocorrelation out to lag 38, so a block of 10 resamples dependent blocks as though they were independent and returns intervals that are too narrow. Section 6.4 reports what the choice is worth: the block is varied over 10, 20 and 40, and a stationary bootstrap with matched expected block length is run beside it.

4. Results

Table 1 reports out-of-sample accuracy. The base rate in the test window is 0.468, so a constant "no" forecast would have scored 53.2%, but that constant is chosen with hindsight. A forecaster at the start of the window does not know which way the balance will fall, and choosing from the training window gives two answers depending on its shape: a rolling 1,250-day window picks "no" at every refit and also scores 53.2%, while the expanding window behind the majority rule in Table 1 picks the other class and scores 46.8%. Both are honest and they differ by 6.3 points, which is wider than several of the model differences discussed below. An earlier version took 53.2% as the hurdle on the ground that two of the three constructions agree on it, which counts the hindsight rule as a vote it is not entitled to: only two of the three are available to a forecaster at all, and of those two, 53.2% is simply the higher. We therefore treat 53.2% as a *conservative constant-class hurdle* rather than as the constant-forecast level, and report the expanding-window rule in Table 1 so that a reader can see the other. The walk-forward majority rule scores 46.8% because the training-window majority was the opposite class, which is honest walk-forward behaviour and no error. Which of the two is used changes an answer, so both are carried through rather than one being chosen. Section 4.1 reports the delay at which each forecaster stops being distinguishable from a constant under each hurdle; for the threshold rule those delays are $\delta = 3$ and $\delta = 8$, and the gap between them is the price of the choice. The higher hurdle is the harder test of this note's positive claim, that the rule beats a constant while the mark is fresh, and the easier test of its negative claim, that it stops doing so; naming which claim is being served is what keeps the choice from being made twice in opposite directions.

Table 1. Out-of-sample accuracy (%) by delay, with 95% moving-block bootstrap intervals.

δ	majority	persist-CC	persist-YZ	har-CC	har-YZ
0	46.83	65.85 [57.3, 73.1]	67.87 [59.1, 73.6]	61.82 [53.0, 68.7]	68.59 [61.5, 73.3]
1	46.83	65.56 [57.8, 72.6]	66.71 [58.9, 72.2]	61.53 [52.7, 68.2]	67.44 [60.1, 72.2]
2	46.83	64.41 [57.1, 70.7]	65.56 [57.5, 71.2]	61.24 [52.4, 68.4]	66.57 [59.7, 72.0]

δ	majority	persist-CC	persist-YZ	har-CC	har-YZ
3	46.83	62.10 [54.3, 68.0]	63.26 [55.2, 68.7]	60.81 [52.3, 67.9]	64.84 [57.3, 70.0]
5	46.83	60.23 [52.6, 66.9]	60.95 [52.9, 66.6]	58.50 [49.7, 66.1]	60.95 [52.6, 67.0]
8	46.83	56.34 [49.3, 63.5]	57.20 [49.1, 63.8]	54.76 [45.1, 62.8]	57.20 [48.3, 64.6]
13	46.83	52.88 [45.7, 60.4]	54.90 [47.0, 61.7]	52.74 [43.4, 61.1]	53.75 [45.2, 60.7]
21	46.83	49.86 [42.1, 57.3]	50.00 [41.9, 57.6]	54.03 [44.5, 62.0]	51.87 [44.2, 58.4]
34	46.83	50.43 [42.7, 56.9]	49.86 [42.5, 58.4]	50.43 [41.2, 57.9]	47.84 [39.8, 55.2]
55	46.83	48.27 [40.3, 55.5]	44.24 [36.6, 51.4]	46.54 [36.2, 54.9]	46.25 [37.2, 53.7]

Every column falls with delay and they fall at different rates, which is the whole content of the table. The two persistence rules and the HAR classifier all start near 66–69% at zero delay and converge on the majority rule's 46.83% by three weeks; by $\delta = 21$ no forecaster's interval excludes either constant-class hurdle. The intervals are wide, roughly seven accuracy points either side, so what the table establishes is the decay, not an ordering among the four. The comparisons that need an ordering are made on paired differences in Table 2, where the common days cancel and the intervals narrow.

4.1 The decay

Accuracy falls from about 68% at zero delay to the constant-forecast level by roughly two weeks. Against the conservative 53.2% hurdle the threshold rule's interval lies entirely above it through $\delta = 3$; against the 46.8% expanding-window rule it does so through $\delta = 8$. Each of those crossings is stated at the last delay that clears in every bootstrap seed, which is not the same as the last delay that clears in the seed one happens to draw. At $\delta = 5$ the lower endpoint sits within a quarter of a point of the higher hurdle and clears it in 1 of 20 seeds; at $\delta = 13$ it clears the lower hurdle in 18 of 20. Those two delays are unresolved by this sample and are reported as such rather than assigned to a side. The other three forecasters shift in the same direction: persist-CC and har-YZ run through $\delta = 3$ against the higher hurdle and $\delta = 8$ against the lower, and har-CC clears the higher one at no delay firmly, clearing even $\delta = 0$ in only 5 of 20 seeds, while clearing the lower through $\delta = 5$. Which constant is used therefore changes the answer for all four, which is why this note quotes both rather than calling either the constant-forecast level. Covering a hurdle is a failure to distinguish, not a demonstration of equality; what the table licenses is the statement that beyond these delays the data cannot separate the forecaster from a constant, not that the two are equal. The apparent non-monotonicity between $\delta = 13$ and $\delta = 21$ is within noise: those two accuracies differ by less than one point against a bootstrap standard error of roughly five on the difference.

4.2 Neither refinement helps

Table 2. Paired accuracy differences (percentage points) with 95% block-bootstrap intervals. No interval excludes zero.

δ	har-CC – persist-CC	har-YZ – persist-YZ	persist-YZ – persist-CC
0	–4.03 [–11.5, +2.5]	+0.72 [–4.9, +6.3]	+2.02 [–5.0, +8.2]
3	–1.30 [–7.8, +4.9]	+1.59 [–3.2, +6.8]	+1.15 [–4.6, +5.6]
5	–1.73 [–8.5, +3.9]	+0.00 [–4.8, +4.8]	+0.72 [–4.9, +5.0]
8	–1.59 [–8.9, +3.3]	+0.00 [–5.6, +5.9]	+0.86 [–5.3, +5.2]
13	–0.14 [–7.9, +6.3]	–1.15 [–7.8, +4.6]	+2.02 [–3.5, +5.9]

δ	har-CC – persist-CC	har-YZ – persist-YZ	persist-YZ – persist-CC
21	+4.18 [−4.8, +12.5]	+1.87 [−3.9, +6.5]	+0.14 [−4.0, +4.5]

The Yang-Zhang proxy is worth about two accuracy points at zero delay and less thereafter, never distinguishable from zero. The HAR classifier is worse than the threshold rule at short delays and not distinguishable from it at long ones. That last phrase is doing less work than “no better” would, and deliberately: no interval excluding zero is a failure to distinguish, never a demonstration of equality, and the bound that failure carries is in the intervals themselves. The least favourable of them, har-CC minus persist-CC at $\delta = 21$, leaves a HAR advantage of up to 12.5 accuracy points unexcluded. So what this table rules out is a large and consistent gain from either refinement, not a modest one, and Section 7 records the width that bound reflects. On this target, over this sample, neither the richer measurement nor the multi-horizon regression structure earns its place at any delay we can resolve.

4.3 The look-ahead leak is negligible

Comparing the delayed proxy against M_t rather than $M_{t-\delta}$, the error described in Section 3.2, changes accuracy by at most 0.43 points at any delay where the rule has skill, and at $\delta = 34$ it lowers accuracy by 1.73. The leak is real and should be avoided, but it explains nothing.

5. Why the threshold rule is a strong benchmark

A negative result is only interesting with a mechanism attached. Here there is one, and it is close to trivial once stated.

Write the two decision variables in log form relative to their own thresholds: $a = \log(v_{t-\delta} / M_{t-\delta})$ for the forecaster, $b = \log(v_{t+h} / M_t)$ for the outcome. The target is $\mathbf{1}\{b > 0\}$ and the threshold rule is $\mathbf{1}\{a > 0\}$. If (a, b) are approximately jointly normal with correlation ρ and near-zero medians, then $P(b > 0 | a)$ is monotone in a and crosses one half at $a = 0$, and the threshold rule is the Bayes rule *among functions of a alone*. The accuracy it then attains is

$$A^*(\delta) = 1/2 + \arcsin \rho(\delta) / \pi. \quad (3)$$

Two things must be said about that object before Table 3 is read, because an earlier draft said neither and called it a ceiling.

It is not a ceiling for the models compared against it. The HAR classifier sees three features, the current delayed state and its five- and twenty-two-day means, and making the sign rule optimal over that larger set needs a sufficiency condition, $P(b > 0 | a, w, m) = P(b > 0 | a)$, which bivariate normality of (a, b) does not supply. We tested it: a logistic on all three features beats a logistic on a alone at none of the ten delays at $|z| > 1.96$, so the condition is not rejected here, and the extra lags are not where the excess over A^* comes from.

What does fail is the centring assumption, and our own data rejects it. Estimating $P(b > 0 | a)$ non-parametrically on deciles of a gives a function that is not monotone and that crosses one half at $a = +0.27$ instead of at zero. The rule's threshold is in the wrong place. That is why a logistic free to choose its own threshold beats the sign rule at several short delays, and why measured accuracy exceeds the benchmark at six of the ten delays below. A quantity the process can exceed is an approximation and not a bound, so we call A^* a *restricted-information Gaussian benchmark* and claim nothing stronger for it.

Table 3 evaluates it on the same test window.

Table 3. Attainable accuracy under the normal approximation, against measured accuracy. Two Gaussian columns are shown and only the second is a benchmark. “Same-sign A^* ” is $1/2 + \arcsin(\rho)/\pi$, the accuracy of the rule this note actually runs; where $\rho < 0$ it falls below one half, which means the rule is beaten by its own negation and cannot bound anything. “Best cut” is the Bayes accuracy implied by the fitted non-centred Gaussian model, maximised over thresholds on the delayed state. It is at least one half at every delay, which the same-sign column is not, so the last column is taken against it. “Attainable” here means attainable by the Bayes classifier *under that fitted approximation*, not a bound on what any forecaster could achieve in the data-generating process: measured accuracy exceeds it at six of the ten delays, which is itself evidence that the approximation is one.

δ	$\rho(a, b)$	same-sign A^*	Gaussian benchmark, best cut (fitted model)	persist- YZ	har- YZ	% of benchmark
0	0.4502	64.87	64.99	67.87	68.59	104.4
1	0.4039	63.23	63.37	66.71	67.44	105.3
2	0.3595	61.71	61.86	65.56	66.57	106.0
3	0.3179	60.30	60.48	63.26	64.84	104.6
5	0.2539	58.17	58.40	60.95	60.95	104.4

δ	$\rho(a, b)$	same-sign A*	Gaussian benchmark, best cut (fitted model)	persist- YZ	har- YZ	% of benchmark
8	0.1879	56.02	56.33	57.20	57.20	101.5
13	0.1712	55.48	55.82	54.90	53.75	98.3
21	0.1741	55.57	55.91	50.00	51.87	89.4
34	-0.0418	48.67	52.38	49.86	47.84	95.2
55	-0.1202	46.17	52.27	44.24	46.25	84.6

The measured accuracies sit between 90% and 106% of the benchmark. An earlier draft explained the excess above 100% by saying that log realised variance has heavier tails than the normal approximation allows. That was a guess. It was never measured, and it points at the wrong assumption: tails are a statement about the shape of each margin, while equation (3) is an *orthant* probability and so depends on where the margins sit relative to a threshold fixed at zero. Neither a nor b is centred at zero here, because both are log ratios to a *trailing* median, and a trailing median of a persistent right-skewed series is not its current median. So the non-centred version is the one this data needs.

It costs an integral rather than an argument. For (a, b) bivariate normal, write $\alpha = -\mu_a/\sigma_a$ and $\beta = -\mu_b/\sigma_b$. The accuracy of the sign rule is then

$$A(\delta) = 1 - \Phi(\alpha) - \Phi(\beta) + 2 \Phi_2(\alpha, \beta; \rho), \quad (3')$$

which collapses to (3) when $\alpha = \beta = 0$, since $\Phi_2(0, 0; \rho) = 1/4 + \arcsin \rho / 2\pi$. The Bayes rule among functions of a alone is then a cut at some c rather than at zero, and its accuracy is (3') maximised over c . Both are computed alongside the centred figure in the script behind this section.

The result does not rescue the earlier explanation and does not replace it with a better one, so we report it as it came out. Dropping the centring assumption moves the benchmark by **0.16 points** on average, while measured accuracy sits **2.88 points** above the centred figure at the delays out to eight days and 3.05 above the non-centred one. The centring correction is the right correction and it is not the explanation. What remains is the reason we call A* a benchmark and not a ceiling: a quantity the process exceeds by three points on a 694-day sample, where the intervals are roughly seven points either way, is an approximation whose error we have not attributed rather than a bound the models are pressing against. The binding assumption is the normal approximation; which part of it binds, we do not know.

One figure from the same calculation is worth carrying because it decides what to do about the mis-centred threshold. Moving from the zero cut to the Bayes cut is worth at most **0.52 accuracy points** at any delay out to three weeks, rising to 6.18 points at $\delta = 55$ where the correlation has decayed far enough that the optimal rule is close to never predicting the event at all. That last figure is a statement about the base rate rather than about timing. So the note keeps the sign rule, which has no fitted threshold to defend out of sample, and reports the benchmark beside it.

This is the substance of the negative result. Estimated headroom is small relative to sampling uncertainty, and neither refinement we evaluated exploits it reliably, so their failure to improve on the threshold rule is not evidence that those tools are weak. It is a property of the target: a median-split label discards exactly the magnitude information that a better estimator or a richer regression would have supplied. A study wanting to discriminate between volatility models should not use one.

6. Robustness: three things the design fixed by fiat

Three choices in Sections 3 and 4 were made once and never varied: accuracy as the loss, 252 days as the median window with 1,250 as the training window, and the median itself as the split. Each was reasonable and none was tested. This section tests all three. Two of the answers support the note; the third says the question cannot be settled on nine years of data, and says so instead of reporting a number.

6.1 Accuracy is a coarse loss, so use sharper ones

Accuracy sees only which side of a half a forecast fell on. Two sharper measures are available. AUC asks whether a forecaster orders the days correctly and is invariant to any monotone transform, so it can score a rule that has a margin behind it even when that rule emits no probability: the persistence rules have one, the decision variable $a = \log(v_{t-\delta} / M_{t-\delta})$ whose sign is the rule. The Brier and log scores measure the quality of a probability and are strictly proper. The threshold rules do not provide non-degenerate probability forecasts: treating their hard labels as probabilities of zero or one makes the Brier score equivalent to the error rate and makes the log score infinite after any error, so neither gives a useful probability-calibration comparison. The objection is usefulness rather than computability.

Table 4. The same forecasters under three measures. AUC(margin) ranks by the continuous decision variable a forecaster holds; AUC(label) ranks by the 0/1 it emits, and for a binary forecast equals balanced accuracy. The base-rate row is the recursive base rate, the probability a forecaster could actually have quoted; its AUC is below one half because a nearly constant series is ranked by its own slow drift, and it should not be read as a chance benchmark, which is 0.500.

δ	forecaster	accuracy	AUC (margin)	AUC (label)	Brier	log
0	persist-YZ	67.87%	0.735	0.681	n/a	n/a
0	har-YZ	68.59%	0.705	0.674	0.2195	0.6357
5	persist-YZ	60.95%	0.660	0.611	n/a	n/a
5	har-YZ	60.95%	0.628	0.598	0.2470	0.6997
21	persist-YZ	50.00%	0.535	0.501	n/a	n/a
21	har-YZ	51.87%	0.508	0.507	0.2849	0.7890
21	recursive base rate	42.94%	0.336	0.409	0.2513	0.6958

The objection is answered in the opposite direction to the one it anticipated. Under accuracy the two forecasters are inseparable, as Section 4.2 reported. Under AUC they separate at five of ten delays, and in the threshold rule's favour every time, never the HAR classifier's. Every gap reported here is the HAR classifier minus the threshold rule, so a negative number is the threshold rule ahead; at $\delta = 0$ the gap is -0.0299 with a block-bootstrap interval of $[-0.0590, -0.0026]$. The separation holds at $\delta = 0, 1, 2, 3$ and 5 and fades after: from $\delta = 8$ onward the interval covers zero, while the point estimate keeps the same sign at all ten delays. The raw margin orders the days better than the fitted logistic's probability does, so the claim of parity in Section 4.2 is, if anything, too generous to the refinement.

Two further readings are worth stating because accuracy cannot reach either. The first is what discretising costs: persist-YZ's margin out-ranks its own emitted label at nine of ten delays, by as much as 0.055 of AUC at $\delta = 0$. A forecaster who reports the sign throws that away, and Table 1 cannot see it. The second is less comfortable for the note. Beyond about three weeks the HAR classifier's *probability* is worse than the recursive base rate on Brier score, significantly so at $\delta = 34$ and

55, while its accuracy at those delays still looks respectable. A user who needs a probability instead of a label should read the Brier column, and the Brier column says the model has stopped being useful earlier than the accuracy column suggests.

6.2 The two windows

The median window is not a free parameter of the forecaster; it defines the target, since $y_t = 1$ exactly when the proxy five days on exceeds M_t . Changing it changes the question. A 63-day median asks whether volatility will exceed its recent quarter's typical level and a 504-day median asks about a two-year level, and those have different base rates and different intrinsic difficulty. Accuracies are therefore not comparable across windows. What is comparable is the contest within each window, and that is what the note's claim is about. Every cell is scored on one common 476-day window, whatever the most demanding cell can afford, so no difference between cells is a difference in sample.

Table 5. Accuracy of the HAR classifier minus the threshold rule, in percentage points, across the window grid, with 95% block-bootstrap intervals. Positive would favour the HAR classifier. Every cell on the same 476 test days.

median	training	$\delta = 0$	$\delta = 3$	$\delta = 5$	$\delta = 13$	$\delta = 21$
63	1250	-5.25	-5.88	-7.14	-0.63	+3.36
126	1250	-7.35	-6.93	-8.19	-2.94	-2.31
252	500	-4.41	-6.09	-5.46	-6.30	-2.94
252	1250	-1.89	-2.52	-2.94	-5.46	-1.47
504	1250	+0.63	-0.84	-2.31	-2.10	-2.31

Across the full grid of four median windows and two training windows at five delays, forty cells in all, the HAR classifier significantly beats the threshold rule in none, and the threshold rule significantly beats the HAR classifier in fifteen. The note's negative result is not a property of 252 days: it survives a fourfold change in the median window and a 2.5-fold change in the training window, and the direction never reverses.

One reading of the same grid does not survive, and we report it because an earlier draft of this section would have got it wrong. The delay at which the rule stops beating a constant is later for short median windows, five days at 63 and 126 against zero days at 252 and 504, which looks like a finding about window length. It is partly a loss of power. Re-running the note's own cell on the largest window it can afford, 728 days instead of the common 476, moves that boundary back out from zero to three days. Shorter median windows also give a target closer to a 50/50 split, which is easier to beat by construction. The ranking across windows stands; the absolute boundary in Section 4.1 is specific to the note's own sample and should not be re-read off this table.

6.3 Off the median: a tail target

Volatility regimes are asymmetric, and a median split does not ask about spikes. The natural extension is to move the target to an upper quantile, and doing so breaks both halves of Section 5 in ways worth being explicit about.

The rule breaks first. For approximately joint normal a and b with correlation ρ , the Bayes rule against the target $1\{b > z_q\}$ predicts one when $P(b > z_q | a)$ exceeds a half. With standardised variables $E[b | a] = \rho a$, so that condition is $\rho a > z_q$, and it becomes a cut on a at

$$c^*(q) = z_q / \rho, \quad z_q = \Phi^{-1}(q) \quad (4)$$

The direction of that cut depends on the sign of ρ , and this note is a case where the sign changes: the rule predicts the event when $a > c^*(q)$ if $\rho > 0$, and when $a < c^*(q)$ if $\rho < 0$, because dividing $\rho a > z_q$ by a negative ρ reverses the inequality. Table 3 shows ρ turning negative at $\delta = 34$ and $\delta = 55$, so the qualification is not decorative.

At $q = 0.5$ this is zero and the note's rule is the Bayes rule, which is Section 5's claim. Away from the median it is not: at $q = 0.9$, with $\rho = 0.52$ at zero delay, the optimal cut is 2.46 standard deviations. Carrying "is volatility above its median now" to a tail target therefore tests a misspecified rule, and it would lose for a reason that has nothing to do with the question. The benchmark breaks second: with the two thresholds no longer both at zero there is no arcsin form, and the benchmark has to be integrated numerically. That integrator reproduces $\frac{1}{2} + \arcsin(\rho)/\pi$ at $q = 0.5$ to a largest absolute difference of 1.41e-07, which is the check that licenses using it elsewhere.

Table 6. The Gaussian benchmark off the median, at $\delta = 0$ with $\rho = 0.521$. "Majority" is the accuracy of never predicting the event. The benchmark is the accuracy of the Bayes rule; the last column is the accuracy the note's median rule would attain against the same target.

q	z_q	$c^*(q)$	majority	benchmark	note's rule
0.50	0.000	0.00	50.0%	67.45%	67.43%
0.65	0.385	0.74	65.0%	71.06%	65.37%
0.75	0.674	1.29	75.0%	77.13%	62.42%
0.90	1.282	2.46	90.0%	90.11%	55.33%

The bottom row is the result, and it is a theoretical one that no amount of data would change. At the 90th percentile the Bayes rule itself attains 90.11% against a majority-class 90.00%. Even a forecaster who knows ρ exactly and applies the optimal cut beats "never predict a spike" by a ninth of a percentage point. Accuracy at a tail target is not a demanding measure that the threshold rule might fail; it is a measure on which no rule can demonstrate anything, and a paper reporting a 90% accuracy against a spike target would be reporting the base rate.

Measured on balanced accuracy instead, the mean of sensitivity and specificity, with chance at 0.500 whatever the base rate, the two rules change places depending on the objective. At fifteen of twenty quantile-delay cells the rule that maximises raw accuracy is not the rule that maximises balanced accuracy: the fitted cut is right more often by almost never calling a spike, while the note's median rule catches spikes at the cost of false alarms. Neither is better; they answer different questions, which is the loss-function objection arriving in a different place from where it was aimed.

What we cannot do is say which rule predicts spikes best. On 728 test days the 90th-percentile target has 104 positive days and, with five-day overlap, on the order of ten independent episodes. Of fifteen pairwise comparisons at that quantile, the balanced-accuracy intervals separate one. The tail target is predictable: at $q = 0.9$ and $\delta = 0$ the median rule's balanced accuracy is 0.678 with an interval of $[0.608, 0.749]$, clear of chance. But this sample cannot rank the forecasters against it. Nine years does not contain enough independent volatility spikes, and no change of estimator fixes a shortage of events.

6.4 The block length, and a bootstrap that does not need one

Every interval in this note comes from a moving-block bootstrap, and a moving-block bootstrap has one free parameter that no data choose for it. Too short and neighbouring blocks are resampled as

though they were independent when they are not, which narrows every interval and manufactures significance; too long and too few distinct blocks remain, which widens them and buries it. Neither error announces itself. An earlier version of this note also described a block of $2h = 10$ while producing every table at $8h = 40$, so the question of what the choice is worth had to be answered rather than noted.

Table 7 answers it. The block is varied over 10, 20 and 40, the range any reader would defend: 10 covers the overlap in the five-day target and nothing beyond it, 40 is the length the companion paper's bandwidth measurement implies once the loss differential is found to carry autocorrelation out to lag 38. Beside each, the stationary bootstrap of Politis and Romano (1994) is run with its expected block length matched to the same number. That is not a fourth length but a different object: block lengths are drawn from a geometric law rather than fixed, so the resampled series is stationary and the answer does not turn on one integer. Matching the means isolates the shape of the block distribution from its size.

Table 7. Block-length and bootstrap-family sensitivity, 694 test days, ten delays. “mb” is the moving-block bootstrap at the stated fixed block; “sb” is the stationary bootstrap of Politis and Romano with the same *expected* block. Both columns count the delays at which the paired har-YZ minus persist-YZ difference has a 95% interval excluding zero. Accuracy uses 2,000 draws; AUC uses 800, because AUC is recomputed inside every draw rather than averaged from per-day contributions. The note's own tables use mb 40.

scheme	block	accuracy: delays separating	AUC: delays separating	which delays, on AUC	all favour the threshold rule
mb	10	0 of 10	7 of 10	0, 1, 2, 3, 5, 8, 13	yes
sb	10	0 of 10	6 of 10	0, 1, 2, 3, 5, 13	yes
mb	20	0 of 10	5 of 10	0, 1, 2, 3, 5	yes
sb	20	0 of 10	5 of 10	0, 1, 2, 3, 5	yes
mb	40	0 of 10	5 of 10	0, 1, 2, 3, 5	yes
sb	40	0 of 10	5 of 10	0, 1, 2, 3, 5	yes

The two claims this note rests on behave differently, and the difference is the useful part. Parity on accuracy is untouched: at every one of the ten delays, under all three block lengths and under both bootstrap families, the paired accuracy difference covers zero. That claim is about the data. The AUC separation is not untouched. It stands at seven of ten delays with a block of 10, six under the stationary bootstrap at the same mean, and five at blocks of 20 and 40 under both families. The count falls as the block grows, which is the direction a spurious count falls in, and it has converged by 20: the delays that drop out, $\delta = 8$ and 13, are the two whose intervals were closest to zero to begin with. We therefore report five, the count at the block this note uses, and record here that a shorter block would have given more.

What does not move at all is the direction. Across the six schemes there are thirty-three significant AUC cells and all thirty-three favour the threshold rule; not one favours the HAR classifier at any block length under either bootstrap. The separation's size is a function of the resampling choice and its sign is not, which is the distinction a sensitivity table exists to draw.

7. Limitations

The test window is 694 days and the intervals are wide; nothing here separates differences smaller than about five accuracy points. A single index over a single nine-year window cannot speak to other

markets or other regimes. The evaluation period is narrower still: it begins in November 2023 and contains no major macro transition, so the objection that a rolling median fails to adjust across a structural regime change cannot be tested here at all. That is a shortage of sample rather than of method. The companion paper, whose test window runs from May 2007 and contains 2008, 2020 and 2022, does test it, and finds the cross-sectional result strengthens in stress instead of breaking down. The price series is a retail export with no point-in-time guarantee, so any revision to the underlying index is silently inherited. The Gaussian benchmark of Section 5 is a normal approximation, and Table 3 shows it is not exact; Section 6.3 gives its generalisation away from the median, which has no closed form and is integrated numerically. The delay axis is applied to an information set assumed empty after $t - \delta$; a real late forecaster would still observe contemporaneous prices, so these numbers are a lower bound on what such a forecaster could do. Accuracy on a binary target is a coarse loss. Section 6.1 answers that within this design: AUC scores the ranking, and the Brier and log scores evaluate probability forecasts of the binary event, which is what they are proper for. QLIKE and the proxy-robust variance losses of Patton [5] are the appropriate measures for continuous volatility forecasts rather than for this one, and the companion paper uses them there. Finally, Section 6.3 leaves a question open instead of answering it: this sample contains too few independent volatility spikes to rank forecasters against a tail target, and that is a shortage of events and not of method. What would repair that is worth naming, because the obvious remedy is the wrong one. Sampling the same nine years more finely measures the same five-day target more precisely; it does not create independent tail episodes, whose number is set by the calendar. Redefining the target at an intraday horizon would create them, but it would be a different question. The episode count grows in only two honest ways: a longer panel, decades and not years, which for range-based estimators means data most researchers cannot obtain; or a wider one, since spikes accumulate across markets as well as through time, so a cross-section reaches a given episode count in a fraction of the calendar, and the companion paper's eight markets over nineteen years is the shape of design this question needs. A third route avoids the shortage without solving it: score forecasters against a target that PRICES the tail instead of waiting for it, since an option surface quotes a tail every day and a daily score needs no spike to arrive. Each of the three is a different paper, and none of them is this one.

8. Reproducibility

Every number in this note was produced by seven scripts, all run on the data described in Section 2 with a fixed seed. They ship in the repository named in the main paper, whose README states which script produces which result; `run_all.py --check` runs them all and diffs each line against stored reference output. One of the seven imports another instead of restating it, so the accuracy column of Table 3 and the accuracy columns of Table 1 cannot drift apart. A separate verifier checks every row of Table 3 in this note against that reference output and fails if any cell disagrees.

That check exists because of what it found. Until late in the writing, Table 3 came from a scratch script that read a price file from the author's desktop and compared the benchmark against accuracy figures pasted in as literals from an earlier run. Nobody else could have run it, and had `lab02` changed, the pasted column would have gone on agreeing with a number that no longer existed. Every figure in the table survived being recomputed properly, which is the good outcome and was not the guaranteed one.

Disclosure of AI use

During preparation of this note, the author used Claude (Anthropic) as an assistive tool for drafting and revising text and research code.

The research question, mathematical development, empirical design, data construction, analysis, interpretation of results, and decisions regarding the note's claims were the author's own. The author independently specified and executed all analyses; reviewed, tested, and validated the code and outputs; and takes sole responsibility for the note's contents, including all methods, results, interpretations, and reported numbers.

No AI system is an author of this work and no AI system held authorial responsibility for any part of it.

References

1. Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review* 39(4), 885–905.
2. Bollerslev, T., Patton, A. J. and Quaedvlieg, R. (2016). Exploiting the errors: a simple approach for improved volatility forecasting. *Journal of Econometrics* 192(1), 1–18.
3. Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7(2), 174–196.
4. Hansen, P. R. and Lunde, A. (2014). Estimating the persistence and the autocorrelation function of a time series that is measured with error. *Econometric Theory* 30(1), 60–93.
5. Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160(1), 246–256.
6. Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* 17(3), 1217–1241.
7. Rogers, L. C. G. and Satchell, S. E. (1991). Estimating variance from high, low and closing prices. *Annals of Applied Probability* 1(4), 504–512.
8. Yang, D. and Zhang, Q. (2000). Drift-independent volatility estimation based on high, low, open, and close prices. *Journal of Business* 73(3), 477–491.
9. Li, Z. and Zakamulin, V. (2020). The term structure of volatility predictability. *International Journal of Forecasting* 36(2), 723–737.
10. Bucchieri, G. and Corsi, F. (2021). HARK the SHARK: realized volatility modeling with measurement errors and nonlinear dependencies. *Journal of Financial Econometrics* 19(4), 614–649.