

Scale Before Shape: A Distribution-Free Diagnostic for Plug-In VaR Failure*

Yi-Hao Lai[†]

September 6, 2026

Abstract

A plug-in Value-at-Risk built on a volatility forecast can fail coverage tests for two very different reasons: the forecast is right about shape but wrong about scale, or it is wrong about the shape of the tail itself. We give a distribution-free diagnostic that separates the two. For a plug-in VaR at level α , the implied scale $c_{\text{emp}}(\alpha)$ is the $(1 - \alpha)$ empirical quantile of the standardized exceedance ratio $u_t = r_t/\text{VaR}_t$; a scale failure appears as $c_{\text{emp}} \neq 1$ at both levels with $\Delta c = c_{\text{emp}}(1\%) - c_{\text{emp}}(5\%)$ indistinguishable from zero, while a shape failure appears as a significant Δc . Inference uses a paired moving-block bootstrap and assumes no parametric innovation law. Applied to a HAR-RV plug-in VaR on 0050.TW, the diagnostic classifies the failure as scale, and a scale correction estimated from the same information set moves the affected specification back inside the calibrated region. We also report a cautionary construction result. The forecast-accuracy-versus-tail-adequacy divergence that motivated this line was an artifact of scoring a point-forecast loss against a realized-variance proxy different from the return series on which the VaR was evaluated; on the aligned target the loss ranking is not merely weaker but

*All errors are my own. The return series is 0050.TW daily closing prices; the realized-variance series is built from TAIFEX TX futures tick data. A self-contained replication package—the frozen input series together with a single script that regenerates every number reported in this note—accompanies this submission and will be deposited in a public repository upon acceptance.

[†]Department of Finance, Da-Yeh University. Corresponding author. Email: yhlai@mail.dyu.edu.tw

reverses sign and is statistically indistinguishable from zero. The practical rule is ordered: check target alignment, then scale, then shape.

Keywords: Value-at-Risk; backtesting; realized volatility; HAR; forecast evaluation; proxy consistency

JEL: C52, C53, G17

1 The claim under test and where it came from

A familiar working rule in applied volatility forecasting is that the specification with the best out-of-sample point-forecast loss is the one to plug into a risk measure. The rule is convenient and it is not obviously wrong: the same conditional variance enters both objects.

An apparent counterexample motivated this note. In our own earlier work on the Taiwan market, a HAR-RV specification won a Diebold–Mariano comparison on QLIKE decisively while over-violating its own 1% VaR, and a GJR-GARCH specification that lost the QLIKE comparison was tail-adequate. Read at face value, that pattern says central forecast skill and tail adequacy are separate properties and the folk rule is unsafe.

This note reports that the counterexample dissolves under a pre-registered gating test, and that what survives the dissolution is more useful than the original claim. [Section 2](#) gives the construction error. [Section 3](#) reports the tail pattern that remains real after it is removed. [Sections 4](#) and [5](#) give the diagnostic and apply it. [Section 6](#) states the ordered rule and its limits.

2 Why the counterexample dissolves

The QLIKE comparison and the VaR backtest were not evaluated against the same object. QLIKE was scored against a TAIFEX TX five-minute realized-variance proxy, while the VaR was evaluated against 0050.TW close-to-close returns. [Patton \(2011\)](#) shows that the ranking of volatility forecasts under a loss function is only robust to proxy noise for a restricted class of losses, and that the target must be the quantity the forecast is meant to predict. Scoring a loss on one target and adjudicating adequacy on another is outside that guarantee.

Rebuilding the comparison so that both legs address the same target removes the divergence. On the aligned target—0050.TW squared returns—the Diebold–Mariano statistic is $t = 1.4694$ with $p = 0.142$ on $n = 436$ paired observations, and the better model is GJR rather than HAR. The sign has reversed relative to the original comparison

and the magnitude is not distinguishable from zero. Against the pre-registered guardrail of $|t| > 3$ —a deliberately conservative house threshold motivated by the multiple-testing argument of [Harvey et al. \(2016\)](#) rather than by any Diebold–Mariano critical value—the leg does not pass.

We state plainly what this does to the original thesis. Without a forecast-loss win on a common target there is no divergence to explain, so the orthogonality claim is withdrawn rather than weakened. The gating experiment recorded this verdict in its own artifact as H2_REJECTED before any of the present text was written.

Three properties of the inference deserve statement rather than assumption, and we state them as what the code does rather than as what the literature offers. The comparison is non-nested, so the [Diebold and Mariano \(1995\)](#) statistic is well defined on the loss differential. Its standard error is a canonical Bartlett HAC approximation with the usual bandwidth rule. It carries no [Harvey et al. \(1997\)](#) small-sample correction, and the asymptotics are not adjusted for the estimation error in the generated forecasts in the manner of [West \(1996\)](#); we claim neither refinement, and either would have to be added before this leg could support a conclusion sharper than the one drawn here. Since the leg is used to withdraw a claim rather than to establish one, both omissions work against the reading we adopt. The first-order autocorrelation of the loss differential is 0.0159, so the HAC adjustment is doing little work here in any case.

3 What remains: the tail pattern

The tail-side asymmetry survives the rebuild. Over the out-of-sample window—2023-03-01 to 2024-12-31, $n = 450$ trading days, refit every 63 days—the HAR plug-in with a Cornish–Fisher tail fails the 1% trinity of Kupiec unconditional coverage, Christoffersen independence and the Basel traffic light, while the GJR plug-in with the same tail treatment passes it.

Table [1](#) reports the backtest at both α levels. HAR+CF realizes 14 violations in 450 days, a rate of 3.11% against an expected 1%, with Kupiec $p = 0.0003$ and a yellow traffic

light; its independence test passes ($p = 0.4467$), so the failure is one of level rather than clustering. GJR+CF realizes 2 violations in 450 days (0.44%), with Kupiec $p = 0.1833$ and a green light.

Table 1: Out-of-sample VaR backtest. 0050.TW close-to-close returns, 2023-03-01 to 2024-12-31, $n = 450$ trading days, parameters refit every 63 days. *Viol.* is the realized number of exceedances and *Rate* the same figure in percent, against an expected α . *Kupiec* is the unconditional-coverage p -value and *Indep.* the Christoffersen independence p -value; the joint conditional-coverage statistic is stored separately in the artifact and is not what this column reports. The light at 5% is marked $\dagger$ because it is a custom α -scaled extension: the Basel traffic light is defined at 1% only.

α	Model	Viol.	Days	Rate (%)	Kupiec p	Indep. p	Light
1%	HAR+CF	14	450	3.11	0.0003	0.4467	yellow
1%	GJR+CF	2	450	0.44	0.1833	1.0000	green
5%	HAR+CF	40	450	8.89	0.0006	0.1886	yellow †
5%	GJR+CF	21	450	4.67	0.7429	0.9850	green †

The denominator and the source both matter here and we give both. The 2/450 figure is this experiment’s own count on its rebuilt realized-variance series. A predecessor study on its own series recorded 3/450 for the same cell, and the two constructions are reconciled in a replication bridge in which 11 of 14 cells match exactly (78.6%), with GJR+CF one of the three that differ by a single violation. We also avoid any uniqueness framing: in the predecessor study a second specification reached the same trinity outcome, so “the only specification that passes” would be false.

Three scale- and shape-corrected HAR variants were estimable, and none of them recovers the full trinity at both α levels; the count of variants rescuing HAR is zero. This is the fact against which Section 5 must be read.

4 The diagnostic

Let $\text{VaR}_t(\alpha)$ be a plug-in one-day-ahead Value-at-Risk and r_t the realized return. Define the exceedance ratio

$$u_t = \frac{r_t}{\text{VaR}_t(\alpha)},$$

and let the *implied scale* $c_{\text{emp}}(\alpha)$ be the $(1-\alpha)$ empirical quantile of u_t over the evaluation window. If the plug-in were correctly calibrated at α , the $(1-\alpha)$ quantile of u_t would be 1 by construction. Deviations from 1 therefore measure how much the reported risk number would have to be multiplied by to achieve nominal coverage *ex post*.

The diagnostic uses two levels. Write $\Delta c = c_{\text{emp}}(1\%) - c_{\text{emp}}(5\%)$. Two failure channels separate:

- **Scale.** $c_{\text{emp}} \neq 1$ at both levels, with Δc indistinguishable from zero. The reported risk is uniformly too small (or too large) by a common factor; the shape of the standardized tail is right.
- **Shape.** Δc significantly different from zero. The misstatement grows or shrinks as one moves further into the tail, which a single multiplicative factor cannot fix.

Inference is by paired moving-block bootstrap on u_t , which preserves the dependence in the exceedance ratio and requires no assumption about the innovation law. This is the sense in which the diagnostic is distribution-free: it never asks what the standardized residuals are distributed as, only where their empirical quantiles sit relative to one.

Under the primary run the two failing HAR cells classify the same way. HAR+CF has $c_{\text{emp}}(1\%) = 1.401$ with 95% CI $[1.021, 2.099]$ and $\Delta c = 0.131$ with CI $[-0.250, 0.680]$; HAR+HistSim has $c_{\text{emp}}(1\%) = 1.300$ with $\Delta c = 0.114$, CI $[-0.220, 0.682]$. Both are classified SCALE: the level is wrong, the shape is not detectably so. Table 2 collects the estimates.

We note the diagnostic’s own limit at the point where it matters. The Δc confidence intervals are wide, so “no detectable shape failure” is a statement about power at $n = 450$ and not a demonstration that the shape is correct. A wide interval containing zero and a tight interval containing zero support different conclusions, and only the second would license the stronger reading.

Table 2: Implied scale and its two-level difference, with paired moving-block bootstrap 95% confidence intervals. $c_{\text{emp}}(\alpha)$ is the $(1 - \alpha)$ empirical quantile of the exceedance ratio $u_t = r_t/\text{VaR}_t(\alpha)$, so a correctly calibrated plug-in gives $c_{\text{emp}} = 1$; $\Delta c = c_{\text{emp}}(1\%) - c_{\text{emp}}(5\%)$. A cell is classified SCALE when the c_{emp} interval excludes 1 while the Δc interval contains zero, and calibrated when the c_{emp} interval contains 1. HAR-a is the scale-corrected variant of Section 5; the intervals are wide enough that “no detectable shape failure” is a statement about power at $n = 450$.

Cell	$c_{\text{emp}}(1\%)$			Δc			Class
	est.	CI lo	CI hi	est.	CI lo	CI hi	
HAR+CF	1.401	1.021	2.099	0.131	−0.250	0.680	SCALE
HAR+HistSim	1.300	0.945	2.001	0.114	−0.220	0.682	SCALE
HAR-a+CF	1.165	0.855	1.741	0.108	−0.207	0.559	calibrated

5 Applying it

The classification points at a specific remedy: if the failure is scale, a single multiplicative correction estimated from the information set available at each refit should move the cell back toward calibration. Variant HAR-a implements exactly that, with the scale factor re-estimated at each refit; across refits the estimate averages $s = 1.203$, ranging from 1.185 to 1.225.

After correction, HAR-a+CF has $c_{\text{emp}}(1\%) = 1.165$, CI $[0.855, 1.741]$, and $c_{\text{emp}}(5\%) = 1.057$, CI $[0.936, 1.370]$. Both intervals contain 1, and the cell is classified calibrated.

Three qualifications keep this from being read as more than it is, and each is a claim we deliberately do not make.

First, *reclassification is not rescue*. The rescue criterion set before the experiment required the full trinity at both α levels, and no variant met it (Section 3). What the correction demonstrates is that the diagnostic’s own classification responds to the intervention the diagnostic recommends. That is evidence the diagnostic measures what it claims to measure; it is not evidence that the corrected model is adequate for capital purposes.

Second, HAR+Normal was already classified calibrated before any correction, so its post-correction status carries no information about the remedy and we do not cite it as support.

Third, HAR+HistSim’s implied scale is numerically unchanged by the correction: $c_{\text{emp}}(1\%)$ is 1.30007 both before and after. equivariant to a multiplicative rescaling of the conditional standard deviation, not a failure of the correction. The experiment records an explicit equivariance check whose maximum absolute deviation is at machine precision (1.4×10^{-17}), and we report the invariance as a sanity check that the implementation behaves as the theory requires.

6 A practical rule, and what the diagnostic cannot do

The ordering is the contribution as much as the statistic is. Before attributing a coverage failure to the tail:

1. **Check target alignment.** A point-forecast loss and a VaR backtest must address the same return object. Our own divergence survived four experiments before a gating design caught it, which is the reason we report the construction error rather than quietly correcting it.
2. **Check scale.** Compute c_{emp} at two levels. If $c_{\text{emp}} \neq 1$ and Δc is not distinguishable from zero, the cheap fix is a multiplicative correction and no re-specification of the tail is warranted by the evidence.
3. **Only then consider shape.** A significant Δc is what licenses a change of innovation law or tail model.

Four limits bound what we have shown. The evidence is one market and one window, $n = 450$; whether the scale-before-shape ordering generalizes needs a second series, and we do not claim it does. The Δc intervals are wide, so shape failures of moderate size would not be detected here. The Basel traffic light we report at 5% is a custom α -scaled extension and not canonical: Basel’s rule is defined at 1% only, and we label it wherever it appears. Finally, the diagnostic addresses calibration of a reported quantile; it says nothing about whether the risk measure is coherent, which is a separate axis.

7 Data, design and controls

The return series is 0050.TW close-to-close, dividend-adjusted. The realized-variance series is built from TAIFEX all-contract TX tick data: for each day the maximum-volume active contract is selected and one continuous five-minute path is formed over the window $(13:30(D-1), 13:30(D)]$, with no cross-contract returns. The series spans 2,191 days and includes 108 roll days, with a mean of 222.3 five-minute returns per day.

Out-of-sample evaluation runs from 2023-03-01 to 2024-12-31, $n = 450$, refitting every 63 days, with seed 20260712; GJR estimation uses 120 multistarts and RGL 40.

Four controls are reported because each addresses a way this design could have been wrong. A session-alignment check confirms that evening ticks are stamped to the previous calendar day in all 40 files examined, so the overnight session precedes rather than follows its trading day. A look-ahead audit perturbs every observation at or after each forecast origin by a factor of ten and requires origin forecasts and correction parameters to return bit-identical; all probes pass. A boundary audit of the realized-variance window passes on all 60 days checked. The GJR likelihood is multimodal—119 distinct basins across 120 starts at the first refit—so the reported fit is the best of many rather than a single optimization, and the sensitivity of the resulting conditional standard deviation to the starting draw is reported alongside it.

8 Conclusion

A reported divergence between forecast accuracy and tail adequacy should be treated as a hypothesis about construction before it is treated as a finding about markets. In our case it was construction: a loss scored against one target and a backtest adjudicated against another. What remains after that correction is a tail-side pattern with a diagnosable cause, and a distribution-free statistic that says which cause. The statistic is cheap, it requires no distributional assumption, and it comes with an ordering—alignment, then scale, then shape—that tells a practitioner which fix to try first and, equally usefully, which fixes the evidence does not yet justify.

References

- Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196.
- Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Harvey, D., Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), 281–291.
- Harvey, C. R., Liu, Y. and Zhu, H. (2016). . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68.
- Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73–84.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.
- West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5), 1067–1084.