

Why GAS-t Fails on Pre-Institutional Bitcoin: A Robust Deficit with a Loss-Conditional Diagnosis

Yi-Hao Lai

Department of Finance, Da-Yeh University, Changhua, Taiwan

Abstract

We document a previously unreported reversal in Generalized Autoregressive Score (GAS) volatility models with Student-t innovations applied to Bitcoin: over the pre-institutional period (January 2017 to December 2020, $n = 1,441$ out-of-sample days), GAS-t produces 11.79% worse quasi-likelihood (QLIKE) loss than the GJR-Normal benchmark (Diebold-Mariano-HLN $t = -4.40$, $p = 1.2 \times 10^{-5}$). Two natural hypotheses—that score-driven GAS dynamics fail, or that Student-t innovations fail—are typically conflated in the GAS literature. We resolve this via a factorial five-model decomposition (GJR-N, GJR-t, GAS-N, GAS-t, GJR-N-standardized) which shows that GAS-Normal *recovers* to near-parity with GJR-Normal (DM-HLN = -1.97 , $p = 0.049$: marginal at the conventional 5% level and far below the $|t| > 3$ bar we apply throughout) while GAS-t and GJR-t both reverse, pointing at the Student-t innovation rather than at the GAS dynamics. We report the limits of that diagnosis rather than only its point estimate. The evidence against the dynamics is the weaker half, because the dynamics contrast does not reach the $|t| > 3$ bar. The GJR-Normal benchmark was estimated from a single start

Email address: yhlai@mail.dyu.edu.tw (Yi-Hao Lai)

and the GAS cells from 100, but in each of the 23 Period 1 refit windows the single start reaches the best objective of 100 starts to within 10^{-6} , so the difference in search effort does not move the benchmark’s fitted objective. The deficit itself is robust: it survives all three Patton (2011) losses we score it under ($t = -4.40, -5.32, -6.37$), a ± 60 -day shift of the period boundary, and a leave-one-year-out jackknife. The attribution is not: the GAS-Normal versus GAS-t contrast is $+3.30$ under QLIKE but changes sign and loses significance under MSE (-1.01) and under $b = -1$ (-0.31), and in a leave-one-year-out jackknife it keeps its sign in every fold but clears $|t| > 3$ in only one. A Generalized Error Distribution innovation—also fat-tailed—recovers to parity, which narrows the culprit from heavy tails in general to the Student-t score’s specific down-weighting of extremes. A Markov-Switching GAS-t extension rescues GAS-t to statistical parity with GJR-Normal under both estimation protocols we report (DM-HLN vs single-state GAS-t = $+4.60$ in the original estimation and $+3.35$ under a common 100-start re-estimation; vs GJR-Normal $+0.27$ and -1.36 , neither significant), at a cost of ten free parameters against four—parity, not superiority. In sample, both the Akaike and the Bayesian information criteria prefer the switching specification in each of the six Period 1 estimation windows of the switching model. Because neither protocol leaves a statistically detectable residual gap, a regime-mixture explanation of the Period 1 reversal is not rejected by this test, and we do not claim a structural one. The reversal disappears in the FTX-Luna evaluation window (January–December 2023), where GAS-t is 1.09% better than the benchmark; in the spot-ETF evaluation window (January–April 2026) the point estimate is still 4.10% adverse but rests on

100 evaluation days and is statistically indistinguishable from zero. Institutional flows plausibly altered Bitcoin’s return distribution. We provide practitioner guidance on when GAS-t is and is not appropriate for crypto volatility forecasting.

Keywords: Bitcoin, GAS models, Student-t innovation, volatility forecasting, negative result, Markov-switching, regime change

JEL: C22, C53, C58, G17

1. Introduction

Bitcoin’s daily turnover on regulated derivatives venues now rivals that of mid-cap equity index futures, and a growing share of pension and endowment portfolios carries explicit cryptoasset volatility exposure through spot exchange-traded funds approved in January 2024. Aggregate open interest on CME Bitcoin futures has expanded by roughly two orders of magnitude since the contract was launched at the end of 2017, and combined notional on spot, perpetual, and quarterly futures markets routinely exceeds USD 50 billion per day. Volatility forecasting for Bitcoin is therefore no longer a niche exercise: it feeds margin engines at centralised clearinghouses, value-at-risk models used by bank treasuries with crypto desks, and increasingly the prudential stress tests that supervisors apply to institutions holding digital asset exposure. The question of which conditional-variance specification best captures Bitcoin’s tail behaviour has, accordingly, moved from an academic curiosity to an operational concern with direct implications for capital allocation, margining, and systemic risk monitoring.

Within the volatility-modelling toolkit available to applied researchers,

the Generalized Autoregressive Score (GAS) framework of [Creal et al. \(2013\)](#) has emerged as the dominant score-driven workhorse. Under a Student-t observation density—hereafter GAS-t—the score function automatically down-weights extreme observations, which in principle should suit assets with heavy tails ([Harvey, 2013](#)). The empirical record points the same way for score-driven dynamics: across 289 cryptocurrencies, [Catania and Grassi \(2017\)](#) find that a score-driven filter attains a higher in-sample likelihood than a generalized autoregressive conditional heteroskedasticity (GARCH) filter of equal parameter count under the same skewed heavy-tailed density, and [Lucas and Zhang \(2016\)](#) show that a score-driven filter with a Student-t forecasting density forecasts Value-at-Risk as well as or better than earlier exponentially weighted methods on exchange rates and individual equities. The combination of a heavy-tailed innovation distribution with a score-driven variance recursion has become, for many practitioners, the default first choice when fat-tailed financial data are encountered.

This paper documents a robust counter-finding and decomposes its source. Estimating GAS-t on Bitcoin daily returns from January 2015 through April 2026, using a rolling in-sample window of up to 750 daily observations with re-fits every 63 observations and one-step-ahead out-of-sample forecasts, we find that over the pre-institutional sub-period (January 2017 to December 2020, $n = 1,441$ OOS days) GAS-t produces a QLIKE loss (the quasi-likelihood loss of [Patton, 2011](#)) of 2.2276 against 1.9926 for a vanilla GJR-Normal benchmark, the Gaussian asymmetric GARCH of [Glosten et al. \(1993\)](#), hereafter GJR-GARCH—an 11.79% deterioration. The corresponding Diebold-Mariano-HLN statistic of -4.40 ($p = 1.2 \times 10^{-5}$)

clears the $|t| > 3$ bar that we impose throughout as a multiple-testing safeguard, adopted by analogy with the cross-sectional threshold of [Harvey et al. \(2016\)](#) rather than derived from it (Section 3). The reversal is not a marginal effect or a sensitivity-analysis curiosity: a specification that the theory of score-driven filtering recommends for fat-tailed data is decisively beaten on Bitcoin, an asset whose extreme kurtosis and frequent jumps are well documented ([Liu and Tsyvinski, 2021](#)).

The puzzle has two natural resolutions, which existing studies typically conflate. The first attributes the failure to the GAS dynamics themselves: perhaps the score-driven updating rule misreads Bitcoin’s idiosyncratic information arrival process, and a more conventional GARCH-type recursion is required. Under this view, the score function—which weighs incoming observations by their relative likelihood under the assumed density—becomes a liability when the assumed density misrepresents the true tail, dampening the variance response to genuine tail events. The second attributes the failure to the Student-t innovation distribution itself: perhaps Bitcoin’s pre-institutional return distribution, while heavy-tailed, departs from the t-shape in ways that cause the t-likelihood to misallocate variance to the tail. The two hypotheses point to opposite remedies—abandon the GAS class versus retain GAS but change the innovation—and the practitioner literature offers no clean diagnostic to separate them. A paper that finds GAS-t losing on a fat-tailed asset typically stops at the observation, without telling the reader whether the appropriate response is to switch dynamics, switch densities, or both.

Our approach addresses this conflation directly. First, we split the sample

along an institutional-structure criterion that is exogenous to the volatility data: a pre-institutional period (2017–2020, before regulated futures, custodial infrastructure, and large institutional inflows took hold), an FTX/Luna era (2023, post-crash recovery), and a spot-ETF regime evaluated through a preliminary 2026Q1 OOS window after the rolling warm-up is consumed. Second, we run a five-model factorial design that crosses the innovation distribution (Normal versus Student-t) with the variance dynamics (GJR-GARCH versus GAS), producing the four corner specifications GJR-N, GJR-t, GAS-N, and GAS-t, plus a standardised-Normal control GJR-N-std that holds dynamics fixed while perturbing only the innovation scaling. The factorial isolates the marginal contribution of each modelling choice. The GJR-Normal benchmark and its standardised control were fitted from a single fixed start and the other three cells from 100 starts. In every Period 1 refit window, however, the single-start objective matches the best of 100 starts to within 10^{-6} in negative log-likelihood, so the difference in search effort leaves the benchmark’s fitted objective unchanged there (Section 3.5). Third, we extend the analysis to a two-state Markov-switching GAS-t specification in which each state carries its own variance recursion, in the manner of [Haas et al. \(2004\)](#). The specification avoids the path-dependence problem of regime-switching GARCH recursions and tests whether the reversal is an unmodelled-regime artefact.

Four findings emerge. In the pre-institutional sub-period, both Student-t models reverse against the GJR-Normal benchmark: GJR-t records DM-HLN of -3.35 and GAS-t records -4.40 , while GAS-N recovers to near-parity (-1.97 , $p = 0.049$: marginal at the conventional 5% level and far

below the $|t| > 3$ bar we apply throughout) and GJR-N-std is indistinguishable from the baseline (-0.06). The direct innovation contrast within the GAS class—GAS-N versus GAS-t—yields DM-HLN of $+3.30$ in favour of the Normal specification, locating the failure at the innovation distribution rather than at the score-driven dynamics — but only under QLIKE, since the same contrast changes sign and loses significance under two other losses in the [Patton \(2011\)](#) family. It also keeps its sign, but not its clearance of the $|t| > 3$ bar, when single years are removed. The deficit is robust and the attribution is conditional, and [Section 5](#) keeps the two separate throughout. The Markov-switching extension delivers a substantial DM-HLN improvement over single-state GAS-t ($+4.60$ in the original estimation, $+3.35$ under a common 100-start re-estimation) and reaches statistical parity with the GJR-Normal benchmark under either protocol (DM-HLN = $+0.27$ and -1.36 , neither significant). Because neither protocol leaves a statistically detectable residual deficit, the test does not reject a regime-mixture account of the Period 1 reversal. The in-sample likelihoods point the same way: in each of the six Period 1 Markov-switching estimation windows, both the Akaike and the Bayesian information criteria (AIC and BIC) prefer the switching specification to GJR-Normal. In the FTX/Luna and spot-ETF sub-periods, all five specifications fall within a narrow band ($|\text{DM-HLN}| < 1.1$), but the two windows say different things. Over the FTX/Luna window GAS-t is 1.09% better than the benchmark, a genuine absence of deficit. Over the spot-ETF window its point estimate is still 4.10% adverse, on only 100 evaluation days, so that window neither confirms nor rules out a continuing deficit.

This paper makes three contributions to the volatility-forecasting literature. First, we document a regime-specific negative result for GAS-t on Bitcoin and reconcile it with the broader GAS literature. The full-sample reversal that motivates the analysis is driven by the pre-institutional sub-period; once that sub-period is separated, the FTX/Luna window shows no GAS-t deficit and the spot-ETF window is too short to decide. This narrows an apparent anomaly: the theoretical case for Student-t innovations on heavy-tailed data is not refuted in general, but it does not extend to a market microstructure regime—retail-dominated Bitcoin before regulated derivatives had deepened—in which, on our evidence, the Student-t innovation is the component that fails. Nor does the finding contradict [Catania and Grassi \(2017\)](#): their score-driven advantage on an early cryptocurrency sample concerns the dynamics under a fixed skewed heavy-tailed density, not the Student-t innovation that our factorial isolates.

Second, we locate the reversal via factorial decomposition. Holding the score-driven dynamics and the 100-start estimation protocol fixed, we show that under QLIKE the Student-t innovation carries the deficit, and we report the three findings that bound how far that diagnosis travels: it does not survive a change of loss function, it clears the $|t| > 3$ bar in only one of four leave-one-year-out folds although its sign survives all four, and a Generalized Error Distribution innovation with comparably heavy tails recovers to parity. The companion finding that the dynamics are not the culprit is weaker. GAS-Normal recovers to statistical parity with GJR-Normal while both GAS-t and GJR-t reverse against it, but the dynamics contrast itself does not reach the $|t| > 3$ bar. The GJR-Normal benchmark was fitted from a single start

and GAS-Normal from 100; the single start reaches the best objective of 100 starts in every one of the 23 Period 1 refit windows, so the difference in search effort does not account for the contrast at the level of the fitted objective. This is a sharper diagnosis than the stylised observation that “GAS-t loses on Bitcoin”: it tells the practitioner which component of the specification to revisit, with the evidence for retaining the other stated at its actual strength. To our knowledge, no prior paper has separated these two channels in the cryptoasset volatility literature.

Third, we test a Markov-switching rescue, report that it succeeds on loss, and draw the inference that actually follows. The switching specification improves on single-state GAS-t by a DM-HLN margin above our $|t| > 3$ bar under both estimation protocols we report (+4.60 and +3.35) and leaves no statistically detectable gap to GJR-Normal under either, which is the outcome that the decision rule we state before reporting the test (Section 6) assigns to a regime-confounding explanation rather than to a structural one. We therefore do not claim that regime-confounding is ruled out; an earlier version of this paper did, on a reading of the same numbers we withdraw in Section 6. What the rescue establishes out of sample is a cost: the forecast accuracy GJR-Normal reaches with four coefficients requires ten in the switching specification. In sample the comparison runs the other way, because the switching specification fits each of the six Period 1 Markov-switching estimation windows better than GJR-Normal by a margin that both AIC and BIC favour. Parsimony is therefore an out-of-sample argument only. The finding has both methodological and substantive implications: methodologically, it warns against stating a two-branch test and then reading the result

onto the branch the argument needs; substantively, it leaves the Student-t mis-calibration on pre-institutional Bitcoin diagnosed as to its locus—the innovation distribution, under QLIKE—but open as to whether that mis-calibration is a property of the tail itself or of a regime mixture the pooled density cannot represent.

The remainder of the paper proceeds as follows. Section 2 reviews the relevant literature on GAS modelling, cryptoasset volatility forecasting, Markov-switching extensions, and negative-result methodology in financial forecasting. Section 3 describes the Bitcoin daily return data, defines the three institutional-structure sub-periods, specifies the five competing models and the Markov-switching extension, and details the rolling-window estimation and Diebold-Mariano-HLN evaluation protocol. Section 4 presents the cross-period reversal decomposition. Section 5 reports the factorial diagnosis of the root cause. Section 6 examines the Markov-switching rescue. Section 7 discusses why the reversal is pre-institutional and what that implies for cryptoasset volatility modelling more broadly. Section 8 reports robustness checks covering lookahead audits, optimiser stability, alternative period cuts, alternative loss functions, and a leave-one-year-out jackknife; the cross-asset replication on Ethereum and Binance Coin was specified but did not complete, and no cross-asset claim is made anywhere in this paper. Section 9 concludes.

2. Literature Review

This section organises the literature into five thematic strands that jointly motivate our research design. The first three strands establish the method-

ological building blocks—score-driven volatility dynamics, the comparative evidence between GAS-t and GJR-GARCH on financial returns, and the empirical state of Bitcoin volatility modelling. The fourth strand summarises the Markov-switching extensions that we use to test whether a regime-aware specification can rescue the GAS-t mechanism on pre-institutional Bitcoin. The fifth strand situates our negative finding within the broader credibility movement in finance forecasting, which has increasingly demanded that anomalous results be reconciled with prior literature through rigorous decomposition rather than discarded as outliers. We close the section by articulating the specific gap that motivates the factorial-plus-rescue research design developed in Section 3.

2.1. The Generalised Autoregressive Score Framework

The Generalised Autoregressive Score (GAS) family was introduced by [Creal et al. \(2013\)](#) as a unifying observation-driven approach in which the conditional density is updated through the scaled score of the predictive likelihood. [Blasques et al. \(2015\)](#) show that the score-driven update is information-theoretically optimal in a local sense: it reduces, in expectation, the Kullback–Leibler divergence between the model-implied density and the unknown data-generating process at each step, which gives GAS specifications an a priori appeal whenever the chosen innovation distribution is a reasonable approximation of the true conditional distribution. [Harvey \(2013\)](#) developed the parallel Dynamic Conditional Score (DCS) framework, paying particular attention to specifications with heavy-tailed innovations such as the Student-t and showing analytically that the score-driven update down-weights extreme observations rather than letting them dominate the

volatility recursion, in contrast to standard GARCH innovations. [Blasques et al. \(2022\)](#) provided the asymptotic theory and the conditions for stationarity, ergodicity, and consistency of maximum-likelihood estimation for the GAS class, and [Blasques et al. \(2016\)](#) supplied the complementary in-sample confidence and out-of-sample forecast bands for the time-varying parameters such a model produces. Together, these works made GAS-t the natural workhorse for series in which heavy tails and time-varying volatility coexist, and they motivate a theoretical prior that GAS-t should weakly dominate GJR-Normal on a heavy-tailed asset such as Bitcoin. Our negative finding directly challenges this prior, and the factorial design in Section 3 is constructed precisely to identify which component of the GAS-t apparatus—the score-driven recursion or the Student-t innovation assumption—is responsible for the breakdown.

2.2. Comparative Evidence on GAS-t versus GJR-GARCH

Direct out-of-sample evidence on GAS-t against a GJR-Normal benchmark is thinner than the theoretical case. [Lucas and Zhang \(2016\)](#) introduce a score-driven exponentially weighted moving average that updates volatility and higher moments through the score of a skewed Student-t forecasting density, and show that it forecasts Value-at-Risk as well as or better than earlier methods of the exponentially weighted family on six exchange rates and six individual equities; GJR-GARCH is not among their benchmarks. [Catania and Grassi \(2017\)](#), discussed in the next subsection, compare score-driven and GARCH filters under a common density. Neither study, and none we are aware of, establishes as an empirical regularity that a Student-t GAS specification beats a GJR-Normal benchmark out of sample on heavy-tailed

assets. The prior our result runs against is therefore the theoretical one set out in the previous subsection, together with the practice of treating GAS-t as the default for fat-tailed data, rather than a documented horse-race regularity. Our paper reports a case in which that prior fails decisively, on pre-institutional Bitcoin, and locates the failure at the Student-t innovation under QLIKE, while GAS-Normal recovers to parity with GJR-Normal.

2.3. Bitcoin Volatility Modelling

A separate strand has focused on volatility dynamics in cryptocurrency markets. [Catania and Grassi \(2017\)](#) were among the first to apply score-driven specifications to cryptocurrencies. Across 289 cryptocurrencies, over a sample running from 2013 to mid-2017, they find that a score-driven filter under a Generalized Hyperbolic Skew Student-t density attains a higher in-sample likelihood than a GARCH filter of equal parameter count under the same density, with an average gain near 5% and a gain of 2.8% for Bitcoin, the smallest of their four headline series. Their design holds the density fixed and varies the dynamics, so it speaks to the score-driven recursion rather than to the Student-t innovation, and it is consistent with our finding that the score-driven dynamics are not where the pre-institutional deficit lies. What neither their study nor any other we cite establishes is a Student-t innovation advantage on early Bitcoin. [Yi et al. \(2018\)](#) studied volatility connectedness across cryptocurrencies and documented a high degree of common variation, which motivated a cross-asset replication on Ethereum and Binance Coin that did not run to completion (Section 8). [Liu and Tsyvinski \(2021\)](#) provided a comprehensive treatment of cryptocurrency risk and return, emphasising that Bitcoin exhibits extreme kurtosis and frequent jumps that are

not well explained by exposure to traditional risk factors. [Klein et al. \(2018\)](#) analysed Bitcoin against gold and other store-of-value assets using BEKK and DCC specifications, concluding that Bitcoin’s volatility dynamics differ structurally from those of mature assets in ways that complicate direct transplantation of equity-tested models. These contributions establish that the pre-institutional Bitcoin return distribution is unusual along multiple dimensions and create a *prima facie* case that off-the-shelf GAS-t specifications may not transfer cleanly from equities to early-era cryptocurrencies.

2.4. Markov-Switching Volatility Models

The Markov-switching volatility literature provides the natural framework for testing whether a single-regime GAS-t specification is misspecified due to unmodelled regime variation rather than a fundamental mismatch between the Student-t innovation and the underlying return distribution. [Gray \(1996\)](#) avoided the path-dependence problem by collapsing the lagged state variances into a probability-weighted average. [Klaassen \(2002\)](#) refined that collapse so that, among other things, multi-period-ahead forecasting becomes a recursive procedure, and found that the resulting regime-switching GARCH delivers significantly better out-of-sample volatility forecasts than single-regime GARCH for three U.S. dollar exchange rates. [Haas et al. \(2004\)](#) removed the collapse altogether: each regime carries its own variance recursion driven by the observed data, so the likelihood is tractable without any path integration. The MS-GAS-t rescue test in Section 6 adapts the Haas construction to score-driven dynamics (Section 3.4); it avoids path-dependence and allows a like-for-like comparison against the single-state GAS-t. [Hansen et al. \(2011\)](#) developed the Model Confidence Set, which

retains the set of models that cannot be distinguished from the best under a given loss, and [Hansen et al. \(2003\)](#) applied an earlier version of the procedure to a set of 55 volatility models. We do not run the Model Confidence Set: our comparisons are pairwise DM-HLN tests of each cell against a fixed benchmark and between the factorial cells (Section 3.6). [Marcucci \(2005\)](#) applied Markov-switching GARCH to stock market volatility and demonstrated that regime-switching extensions can substantially improve forecast accuracy when underlying volatility exhibits structural breaks, which is the leading alternative explanation for our pre-institutional negative result and which the MS-GAS-t test in Section 6 is designed to falsify.

2.5. Negative Results in Finance Forecasting

A growing body of work has argued that the credibility of empirical finance depends on the willingness to publish, scrutinise, and decompose negative findings. [Welch and Goyal \(2008\)](#) showed that out-of-sample evidence for many widely cited equity premium predictors collapses once the analysis is conducted with care, and they advocated for forecasting evaluations grounded in genuine real-time information sets. [Harvey et al. \(2016\)](#) and [Harvey \(2017\)](#) extended this credibility critique to the cross-section of expected returns and to factor models more broadly, recommending substantially higher t-statistic thresholds for new claims and emphasising the importance of distinguishing data-mined patterns from structural findings. Our paper sits squarely within this tradition: we document a previously unreported reversal of GAS-t on pre-institutional Bitcoin, apply a $|\text{DM-HLN}| > 3$ bar adopted by analogy with [Harvey et al. \(2016\)](#) (Section 3) as a sub-period stability gate, and use a factorial decomposition to locate the source of the

deficit rather than reporting the negative result in isolation.

2.6. Gap and Contribution Positioning

Despite the depth of each of the strands surveyed above, no prior paper decomposes a Bitcoin GAS-t reversal into innovation versus dynamics components, nor tests a Markov-switching GAS-t rescue as a falsification device for the regime-confounding hypothesis. The cryptocurrency volatility literature has either reported score-driven advantages that concern the dynamics under a fixed density (Catania and Grassi, 2017) or applied alternative specifications without isolating the role of the innovation distribution (Klein et al., 2018; Liu and Tsyvinski, 2021). The GAS literature has made the theoretical case for Student-t score-driven filters on heavy-tailed series (Creal et al., 2013; Harvey, 2013) but has not addressed the question of why the same specification might fail on a structurally distinct heavy-tailed asset. The Markov-switching literature has provided the tools for regime-aware estimation but has rarely been combined with score-driven volatility models in the cryptocurrency context. The factorial-plus-rescue design developed in Section 3 fills this gap by crossing the innovation distribution with the score-driven recursion within a single estimation framework and by stress-testing the resulting diagnosis against a Markov-switching extension.

3. Data and Methodology

3.1. Data

We use daily closing prices for Bitcoin (BTC-USD) from 1 January 2015 through 15 April 2026, retrieved from Yahoo Finance via the `yfinance`

Python API. We request the series with `auto_adjust=False` and compute returns from the `Adj Close` column the vendor then returns. The return series used in every estimate is pinned: the archived run reads a hashed snapshot of the 4,121 daily returns stored in the experiment record rather than re-querying the vendor. Each of the three sub-periods is estimated independently and re-consumes its own rolling in-sample warm-up from its own start date (750 observations in Periods 1 and 2, 735 in Period 3; Section 3.2), rather than sharing a single warm-up at the head of the full sample, so the three out-of-sample windows total 1,886 days ($1,441 + 345 + 100$) and 2,235 raw observations never enter any evaluation window. Period 2 illustrates what this costs: its sub-period begins on 1 January 2021, but its evaluable window opens only on 21 January 2023, so the Terra/Luna and FTX collapses of 2022 lie inside its training data and not inside anything this paper evaluates. Returns are simple percentage returns, $r_t = 100 (P_t/P_{t-1} - 1)$, and the realized variance proxy follows the squared-return convention $\sigma_t^{2,\text{proxy}} = r_t^2$, which the proxy-robustness result of Patton (2011) licenses for daily-frequency loss-function evaluation even though it is noisier than an intraday realized-volatility measure. The pinned snapshot lives in the experiment record rather than in the paper directory, so a replicator must take it from there; re-querying the vendor API would return a series that has since moved.

3.2. Three-Period Split

Rather than partitioning the OOS window by calendar year, equal-length thirds, or data-driven changepoint detection, we segment the sample by *institutional structure*. The motivation is mechanistic: Bitcoin’s return-generating process is hypothesized to be qualitatively different before institu-

tional custody, derivatives liquidity and spot-ETF arbitrage became routine. Data-driven segmentation (e.g., Bai-Perron breakpoint tests on the return variance) would mechanically locate breaks at the largest realized-volatility excursions—typically the March 2020 COVID shock, the May 2021 China-ban crash, and the November 2022 FTX collapse—but those are *consequences* of the prevailing market structure, not its causes, and using them as cut-points would conflate idiosyncratic drawdowns with structural regime change. Anchoring the partition to externally observable institutional events (custody platform launches, futures-ETF approval, spot-ETF approval) avoids look-ahead snooping in the partitioning step itself and yields period boundaries that are taken from institutional history and not optimized against forecast performance.

The three OOS sub-samples are:

- **Period 1 — Pre-institutional (2017-01-21 → 2020-12-31, $n_{\text{OOS}} = 1,441$ days).** No US spot ETF, no major regulated institutional custody at scale, CME futures only from December 2017 and no CME options on them until early 2020. Retail-dominated order flow, fragmented exchange microstructure, and persistent basis dislocation between spot and futures.
- **Period 2 — FTX-Luna recovery (2023-01-21 → 2023-12-31, $n_{\text{OOS}} = 345$ days).** Post-Terra/Luna and post-FTX collapse, institutional rebuild phase: surviving exchanges re-collateralize, derivatives open interest recovers, but no spot ETF yet. Order flow is partially institutional and increasingly tied to traditional risk-on/risk-off cycles.

- **Period 3 — Spot-ETF era (OOS window 2026-01-05 → 2026-04-14, $n_{\text{OOS}} = 100$ days, preliminary).** The period’s calendar anchor is the 10 January 2024 SEC approval of US-listed spot Bitcoin ETFs (BlackRock IBIT, Fidelity FBTC, and others); after the rolling estimation warm-up, the evaluable out-of-sample window falls in 2026Q1, i.e., two years into the institutionalized regime, by which point creation-redemption arbitrage had fully linked Bitcoin’s price to large-asset-manager flow and traditional brokerage rails. This window therefore measures the mature spot-ETF regime rather than the immediate post-approval transition. The window length is set by the design, not by data availability. The warm-up is the shorter of 750 observations and the sub-period length minus 100, floored at 500, so any sub-period of at most 850 observations is evaluated on exactly 100 days; this sub-period holds 835 observations, which gives a 735-observation warm-up and 100 evaluation days. At the 500-observation floor the same sub-period would have yielded 335 evaluation days. We did not run that alternative, and Period 3’s small sample is a consequence of the window rule rather than of the calendar. Conclusions for Period 3 are reported but not used to support the paper’s primary diagnostic claim, which is established on Period 1.

3.3. *Five Competing Models*

To locate the source of the Period 1 GAS-t reversal, we estimate a 2×2 factorial of innovation distribution (Normal vs Student-t) crossed with conditional-variance dynamics (GJR-GARCH vs score-driven GAS), plus one standardized-Normal control (Table 1):

Table 1: The five competing specifications (2×2 factorial plus control).

Model	Innovation	Dynamics
M1 GJR-N	Normal	GJR-GARCH(1,1)
M2 GJR-t	Student-t	GJR-GARCH(1,1)
M3 GAS-t	Student-t	GAS score-driven
M4 GAS-N	Normal	GAS score-driven
M5 GJR-N-std	Standardized Normal	GJR-GARCH(1,1)

As a model design the factorial is orthogonal in its 2×2 core. As estimated, the cells differ in search protocol (Section 3.5). The M4-vs-M3 contrast varies the *innovation* while holding both the GAS dynamics and the 100-start protocol fixed. The M4-vs-M1 contrast varies the *dynamics* holding Normal innovations fixed, and it also moves from a 100-start to a single-start protocol. The M2-vs-M1 contrast, a parallel innovation effect within GJR dynamics, carries the same protocol difference. Section 3.5 shows that in Period 1 this protocol difference leaves the fitted objectives unchanged. M5 (a Normal-innovation GJR specification with returns demeaned and scaled to unit unconditional variance) serves as a control for any artifact of variance-targeting normalization rather than the innovation distribution per se.

The GJR-GARCH(1,1) conditional variance follows Glosten et al. (1993):

$$h_t = \omega + \alpha r_{t-1}^2 + \gamma r_{t-1}^2 \mathbf{1}\{r_{t-1} < 0\} + \beta h_{t-1}, \quad (1)$$

with positivity and stationarity constraints $\omega > 0$, $\alpha, \beta, \gamma \geq 0$, and $\alpha + \beta + \frac{1}{2}\gamma < 1$. The score-driven GAS specification of Creal et al. (2013) replaces the squared-innovation forcing term with the scaled score of the conditional log-likelihood. For a Student-t innovation with ν degrees of freedom, the

time-varying log-variance $f_t = \ln h_t$ obeys

$$f_{t+1} = \omega + A S(\nu) \nabla_t + B f_t, \quad (2)$$

with

$$\nabla_t = \frac{1}{2} \left[\frac{(\nu + 1) r_t^2}{(\nu - 2) h_t + r_t^2} - 1 \right], \quad S(\nu) = \frac{2\nu}{(\nu + 3)(\nu - 2)},$$

where ∇_t is the score of the unit-variance Student-t log-density with respect to f_t . The scale factor $S(\nu)$ is the one the archived code applies. It is not the inverse Fisher information for f_t . Under the model, $r_t^2/[(\nu - 2)h_t + r_t^2]$ follows a $\text{Beta}(\frac{1}{2}, \frac{\nu}{2})$ distribution, so $\mathbb{E}[\nabla_t] = 0$ and $\mathbb{E}[\nabla_t^2] = \nu/[2(\nu + 3)]$; the inverse Fisher information is therefore $2(\nu + 3)/\nu$. In a single-state model with A free, the product $A S(\nu)$ is a reparameterisation of the score coefficient, and the choice of scale leaves the forecasts unchanged unless A reaches a bound. The admissible sets are $A \in [10^{-6}, 2]$ and $\nu \in [2.1, 102]$. The M3 parameter vectors are not archived, so we cannot report how often either parameter sits at a bound. The Normal-innovation GAS analogue (M4) uses the inverse-Fisher-scaled Gaussian score, $f_{t+1} = \omega + A s_t^N + B f_t$ with $s_t^N = r_t^2/h_t - 1$. As $\nu \rightarrow \infty$, $2\nabla_t$ tends to s_t^N , so the direction of the Student-t update nests the Gaussian one. The coded scale $S(\nu)$ tends to zero in the same limit, so the nesting holds for the score and not for the estimated coefficient: the M4-vs-M3 contrast compares two separately estimated recursions, not a restriction of one. What differs between them is the Student-t tail-discounting term $(\nu - 2)h_t/[(\nu - 2)h_t + r_t^2]$, which down-weights large $|r_t|$ events, and the contrast asks whether that term improves or degrades one-step-ahead variance forecasts on pre-institutional Bitcoin.

3.4. Markov-Switching GAS-t

To test whether a regime-switching extension can rescue GAS-t on Period 1, we estimate a two-state Markov-switching GAS-t (MS-GAS-t) that adapts the construction of [Haas et al. \(2004\)](#) to score-driven dynamics. Each state keeps its own log-variance path, and each path is updated by that state's own score of the observed return; the state probabilities enter only the likelihood filter and the forecast mixture. Nothing is collapsed. The collapsing recursions of [Gray \(1996\)](#) and [Klaassen \(2002\)](#) instead replace each lagged state variance with a probability-weighted average before the update. All three constructions avoid the path-dependence problem that [Cai \(1994\)](#) and [Hamilton and Susmel \(1994\)](#) raised for regime-switching models with GARCH-type variance recursions, which renders exact likelihood evaluation intractable. Klaassen's recursion approaches the uncollapsed one only as the regimes become perfectly persistent ($p_{00}, p_{11} \rightarrow 1$), and our fitted transition probabilities are well away from that limit. States in a two-state mixture are identified only up to relabelling, so the persistences depend on how a state is followed across refits, and we report two conventions. Followed by index, under which no parameter other than the tail parameter changes sides between the two states in any refit, the persistences are $p_{00} \in [0.72, 0.85]$ and $p_{11} \in [0.25, 0.57]$. Followed instead by the tail parameter, the ordering differs in the first refit, where state 0 carries the larger ν , and the own-state persistence ranges over $[0.42, 0.85]$ for the fatter-tailed state and $[0.25, 0.75]$ for the thinner-tailed one. Both conventions give weak persistence in at least one state. At these persistences the uncollapsed and collapsing filters are different estimators in our sample, and replication should follow the archived

code rather than a textbook collapsing recursion. The fits also sit on their parameter bounds in most refits. The score coefficient of one state reaches its upper bound of 2.0 in five of the six Period 1 refits; that state is the thinner-tailed one in four of those five and the fatter-tailed one in the first. The persistence coefficient of the other state reaches its upper bound of 0.999 in three refits. The MS results of Section 6 are therefore estimated at the constraint set rather than at an interior optimum in most refits. The bound on the score coefficient also couples it to the tail parameter. The coefficient multiplies $S(\nu_j)\nabla_{j,t}$, and $S(\nu_j)$ falls as ν_j rises; see Eq. (2). At the bound, a larger ν_j therefore lowers the largest variance response the state can make, so the likelihood trades tail thickness against responsiveness. In a state at the bound the fitted ν_j does not measure tail thickness alone, and Section 6 draws no inference from its level.

The archived recursion is as follows. Let $S_t \in \{1, 2\}$ denote the latent regime, with transition matrix $P = [p_{ij}]$ where $p_{ij} = \Pr(S_t = j \mid S_{t-1} = i)$, and let $\xi_{t|t-1}(j) = \Pr(S_t = j \mid \mathcal{F}_{t-1})$ denote the one-step-ahead state probability delivered by the Hamilton (1989) filter. Each state carries its own log-variance path, which is updated at every t by that state's own score of the realised return, whichever state the filter infers:

$$f_{j,t+1} = \omega_j + A_j u_{j,t} + B_j f_{j,t}, \quad j \in \{1, 2\}, \quad (3)$$

where $u_{j,t} = S(\nu_j)\nabla_{j,t}$ is the scaled Student-t score of Eq. (2), evaluated at variance $\exp(f_{j,t})$ and tail parameter ν_j . The one-step-ahead forecast is the probability-weighted mixture $\hat{h}_{t+1|t} = \sum_j \xi_{t+1|t}(j) \exp(f_{j,t+1})$. The collapsing recursion of Klaassen (2002) would differ by replacing each lagged

state variance with its expectation conditional on the current state, computed from $\xi_{t-1|t-1}$ and P , before every update. The sample log-likelihood is the standard Hamilton-filter form

$$\ln L = \sum_{t=1}^T \ln \left[\sum_{j=1}^2 \xi_{t|t-1}(j) g(r_t \mid S_t = j; \theta_j) \right], \quad (4)$$

where $g(\cdot)$ is the standardized Student-t density with state-specific degrees of freedom ν_j . The parameter vector has ten elements: an intercept ω_j , a score coefficient A_j , a persistence coefficient B_j and a tail parameter ν_j for each state, plus the two transition probabilities. We initialize state probabilities at the ergodic distribution implied by P and both log-variances at the log sample variance of the training window.

3.5. Estimation

All five competing models are estimated by maximum likelihood (MLE) over a rolling in-sample window and re-fitted every 63 observations, producing one-step-ahead variance forecasts $\hat{h}_{t+1|t}$ for every OOS day; MS-GAS-t is re-fitted every 252 observations to reflect its greater computational cost. The window length is the shorter of 750 observations and the sub-period length minus 100, floored at 500; it is 750 in Periods 1 and 2 and 735 in Period 3 (Section 3.2), and the window rolls at a fixed length throughout. The 750-observation window is calibrated to span at least one complete BTC volatility cycle in pre-institutional data. Because Bitcoin trades every day, 63 observations are nine calendar weeks and 252 are about eight months; both cadences were chosen for computational cost, and we attach no institutional interpretation to them.

The estimation protocol differs across cells, and we state it cell by cell because it bears on the factorial. M2 GJR-t, M3 GAS-t and M4 GAS-N are estimated from 100 seeded starts per refit—the fixed base start plus random draws over the admissible parameter bounds—and the start with the highest log-likelihood is retained. M1 GJR-Normal and M5 GJR-N-std are estimated from a single fixed start. MS-GAS-t is estimated from two starts, a fixed vector and its state-swapped copy, with an iteration cap of 200 against 500 for the other cells, and every refit after the first starts from the previous refit’s estimates. Our design standard is 100 starts per fit on every cell. The run reported here meets it for three of the five cells and not for the GJR-Normal cells or the Markov-switching extension.

For the GJR-Normal cells the shortfall is measured. The robustness battery of Section 8 re-fits every Period 1 refit window of every cell from 100 starts, and the first of those starts is the production start. Across all 23 windows, the production single-start objective exceeds the best of the 100 by at most 7.2×10^{-7} in negative log-likelihood for M1 and 4.1×10^{-7} for M5. Both gaps are below 10^{-6} , the smallest objective difference the estimation code’s own basin diagnostic tabulates. At the level of the fitted objective, the single start therefore reached the best basin that 100 starts found, in every Period 1 window. The GJR-t cell records the same comparison within the production run: across its 23 Period 1 refit windows, no 100-start fit improves on the single fixed start by more than 0.5 in negative log-likelihood, the threshold the code uses to count a change of basin; the largest improvement is 6.3×10^{-6} , and 2 of the 23 windows improve by more than 10^{-6} . Equal objectives do not prove equal parameter vectors, so a later re-estimation puts M1, M2, M5

and MS-GAS-t on the same seeded, model-independent 100-start schedule as M3 and M4, removes the warm-start chain between MS-GAS-t refits, and re-derives the forecasts in all three periods. For the single-state comparisons the DM-HLN statistics move by at most 0.21 (Period 2) and no significance call changes. The Markov-switching comparisons move more, and Section 6 reports both sets side by side (Table 3). The re-estimation was specified after the results in this paper were known, so it is a non-blind sensitivity check rather than a replacement estimate, and its lowest objective among the 100 scheduled starts is not a claim of global optimality.

The experiment record archives the cross-start log-likelihood dispersion but not a runner-up basin comparison, so we make no claim that any retained mode is separated from its nearest competitor by more than numerical tolerance; Section 5 reports what the dispersion diagnostic does and does not establish, cell by cell. The record also stores no per-refit parameter vectors for M1–M5; only the MS-GAS-t parameters are archived. Optimization uses `scipy.optimize.minimize` with the L-BFGS-B algorithm and finite-difference gradients; no standard errors are computed, and none are reported.

Forecasts are constructed in a strictly lookahead-safe manner: at each OOS day t , the variance forecast $\hat{h}_{t|t-1}$ is built only from information dated $t - 1$ or earlier, and the realized loss QLIKE_t pairs $\hat{h}_{t|t-1}$ with r_t^2 . In code the safeguard is an index check rather than a series shift: at every refit, in both the single-state and the Markov-switching paths, the training window is the block of observations ending immediately before the forecast index, and the code asserts that the window’s start plus its length equals that index. An independent code review on 2026-04-17 returned a conditional pass on

lookahead safety, with a verification follow-up still outstanding; we report it at that status. All random initializations use seeds drawn from a fixed master seed of 42 for reproducibility.

3.6. Evaluation

Predictive accuracy is assessed using the QLIKE loss of [Patton \(2011\)](#),

$$\text{QLIKE}_t = \frac{r_t^2}{\hat{h}_t} - \ln \frac{r_t^2}{\hat{h}_t} - 1, \quad (5)$$

which is robust to noise in the squared-return proxy in the sense that ranking is consistent under unbiased volatility proxies. Pairwise model comparisons use the [Diebold and Mariano \(1995\)](#) test with the [Harvey et al. \(1997\)](#) small-sample correction (DM-HLN), implemented with a Newey-West ([Newey and West, 1987](#)) HAC variance estimator with Bartlett weights $1 - k/(L + 1)$ and lag truncation $L = \lfloor T^{1/3} \rfloor$, where T is the number of loss differentials. As coded, the lag-0 variance uses the divisor $T - 1$, each lag- k autocovariance is the mean of its $T - k$ cross-products about the full-sample mean, the one-step horizon gives the correction factor $\sqrt{(T - 1)/T}$, and two-sided p -values are taken from Student's t with $T - 1$ degrees of freedom. QLIKE is undefined on a day with $r_t^2 = 0$; one such Period 1 day is dropped, so the Period 1 tests use 1,440 of the 1,441 evaluation days. We impose $|t_{\text{DM-HLN}}| > 3$ as our own threshold for stable sub-period inference. The bar is borrowed by analogy rather than by citation: [Harvey et al. \(2016\)](#) derive $t > 3$ as a multiple-testing correction for newly claimed factors in the cross-section of expected returns, not for pairwise forecast-loss comparisons, and we are aware of no result that transfers their calculation to this setting. We adopt it because

the same multiple-testing pressure applies to a five-model factorial screened across three sub-periods, and we apply it symmetrically—to the findings that support our argument and to those that do not—so that it cannot function as a filter chosen after the fact. Alongside it we report Spearman rank correlations between forecast variance and the squared return as a non-parametric complement to QLIKE-based testing (Section 4). Sub-period DM-HLN statistics are reported separately for Periods 1, 2 and 3 to support the cross-period reversal-decomposition argument developed in Section 4.

4. Cross-Period Reversal Decomposition

The empirical strategy proceeds in two stages. Before any attempt to diagnose the source of the GAS-t reversal documented on the full Bitcoin sample, we must first establish where, in calendar time, the reversal actually lives. Conflating periods is the dominant failure mode in the cryptocurrency volatility literature: pooled-sample comparisons routinely report a single Diebold-Mariano statistic spanning regimes whose return distributions and institutional microstructure have been radically reshaped by the introduction of CME futures (2017), the FTX-Luna failures (2022–2023), and the U.S. spot ETF approvals (January 2024). A pooled negative result could in principle arise from a uniformly poor specification, from a single regime in which the specification breaks down, or from offsetting wins and losses across regimes whose direction happens to cancel. These three possibilities are observationally distinct and carry different methodological implications, with different recommendations for the practitioner who needs to choose a volatility model for a forward-looking forecast. Cross-period decomposition

is therefore not a robustness exercise appended after the main analysis; it is the precondition for any factorial diagnosis to be interpretable. Sections 5 and 6 condition on the period in which the deficit is concentrated and would have no defensible target without the partition established here. We follow the convention common to negative-result methodology papers (Welch and Goyal, 2008; Harvey et al., 2016) of leading with the partition that disciplines all subsequent inference, rather than presenting an averaged result and treating the partition as a robustness check.

The three-period partition described in Section 3 is chosen from institutional history: the cut points correspond to identifiable structural events in the Bitcoin market microstructure rather than to data-driven changepoint detection. The pre-institutional period ends on 31 December 2020, a calendar cut on a gradual process: the first corporate-treasury purchases of Bitcoin fall inside it, in the second half of 2020, while the large institutional inflows and the spot-ETF approvals come later; the FTX-Luna sub-period is defined to contain the major centralized exchange and stablecoin failures of 2022, although its 750-day warm-up means that those failures inform estimation while the evaluated window opens afterwards, on 21 January 2023; and the spot-ETF regime is evaluated through a preliminary 2026Q1 OOS window that likewise arrives only after that sub-period’s own warm-up has elapsed. Fixing the cut points from institutional history rather than from the data avoids the in-sample search bias that would inflate the apparent significance of any cross-period contrast. We have no archived record that dates the choice of cut points before the data were examined, so the reader should weigh the partition as a stated design choice rather than as a pre-

registered one. The approach aligns the analysis with the structural-break testing tradition of [Andrews \(1993\)](#) and [Bai and Perron \(1998\)](#), albeit with the cut points taken from institutional history rather than estimated.

Table 2 reports the headline Diebold-Mariano-HLN test statistics for models M2–M5 against the M1 GJR-Normal benchmark across the three institutionally motivated periods. The signs follow the standard QLIKE convention: a negative statistic indicates that the alternative model produces a higher (worse) QLIKE loss than the benchmark. We report conventional two-sided 5% significance for the pairwise DM-HLN comparisons and treat our more conservative $|t| > 3$ bar ([Section 3](#)) as a robustness tier for especially strong sub-period evidence.

Table 2: Cross-Period DM-HLN Test Statistics versus GJR-Normal (M1)

Comparison	Period 1 (pre-institutional, $n = 1,441$)	Period 2 (FTX-Luna, $n = 345$)	Period 3 (spot-ETF, $n = 100$)
M2 GJR-t vs M1	− 3.35	−0.26	+0.79
M3 GAS-t vs M1	− 4.40	+0.53	−0.80
M4 GAS-N vs M1	− 1.97	+0.25	−1.00
M5 GJR-N-std vs M1	−0.06	−0.19	+0.75

Notes: Negative statistics indicate that the alternative model produces a higher (worse) QLIKE loss than the M1 GJR-Normal benchmark. Bold entries denote conventional two-sided 5% significance; the stricter $|t| > 3$ bar we impose throughout—our own multiple-testing safeguard, adopted by analogy with the cross-sectional threshold of [Harvey, Liu and Zhu \(2016\)](#) rather than derived from it—is discussed separately as a robustness tier. The archived run’s own screen flags significance at $|t| > 2$, so a replicator reading those flags will disagree with the bolding in this table for the M4-versus-M1 Period 1 cell (−1.97, $p = 0.049$). The Diebold-Mariano test uses the Harvey-Leybourne-Newbold (1997) small-sample correction (DM-HLN) with a Newey-West HAC variance estimator. Sample sizes refer to out-of-sample one-step-ahead daily forecasts.

The Period 1 (pre-institutional) column is the empirical core of the paper. Three observations organize the discussion. First, both Student-t specifica-

tions reverse decisively against the Normal benchmark: M2 GJR-t loses with $\text{DM-HLN} = -3.35$ and M3 GAS-t loses with $\text{DM-HLN} = -4.40$, both well past the $|t| > 3$ bar and corresponding to two-sided p-values of 8.1×10^{-4} and 1.2×10^{-5} respectively. The QLIKE point estimates corroborate the test: M3 GAS-t produces an average QLIKE of 2.2276 against the benchmark 1.9926, an 11.79% deterioration that is substantial in absolute and relative terms. Second, neither Normal-innovation specification reverses at the bar this paper applies: M4 GAS-N delivers $\text{DM-HLN} = -1.97$ with a two-sided p-value of approximately 0.049, which clears the conventional 5% level by a hair but sits far below the $|t| > 3$ bar, and the placebo standardized-Normal M5 returns $\text{DM-HLN} = -0.06$, statistically indistinguishable from the benchmark. The placebo establishes that the Normal-Normal comparison is properly calibrated and that any reversal must come from the genuine structural variation across cells, not from numerical artifacts of the QLIKE loss or the DM-HLN correction. Third, both t-innovation reversals (-3.35 for GJR-t, -4.40 for GAS-t) lie further from zero than the Normal innovation result (-1.97) on the same side of it, with the GAS-t deterioration the most severe. The ordering pre-stages the factorial diagnosis: Student-t innovations are doing damaging work in Period 1, and the damage is amplified, not generated, by GAS dynamics. The full QLIKE bar chart by model and period is provided in Figure 1, and Figure 2 presents the same evidence as a DM-HLN heatmap that makes the regime-specificity visually unmistakable.

The rank correlations do not reproduce the QLIKE ordering. In Period 1 the Spearman correlation between forecast variance and the squared return is 0.35 for both Student-t models (GJR-t 0.35, GAS-t 0.35), 0.33 for GJR-

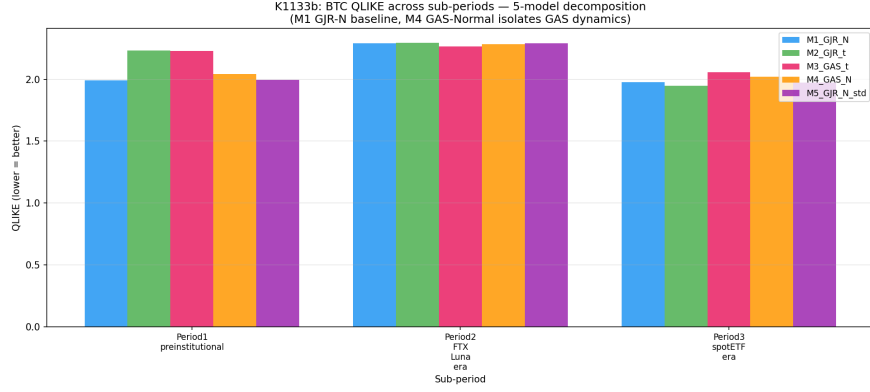


Figure 1: Average QLIKE loss by model and period. Bars group the five competing specifications (M1–M5) within each of the three institutional-structure sub-periods. The GAS-t (M3) bar exceeds the GJR-Normal (M1) benchmark in Periods 1 and 3 and falls below it in Period 2; only the Period 1 gap is statistically significant, and the Period 3 gap rests on 100 evaluation days. Rendered from the 2026-09-06 run; the plotted QLIKE values differ from those behind Table 2 only for M2, by about 10^{-5} .

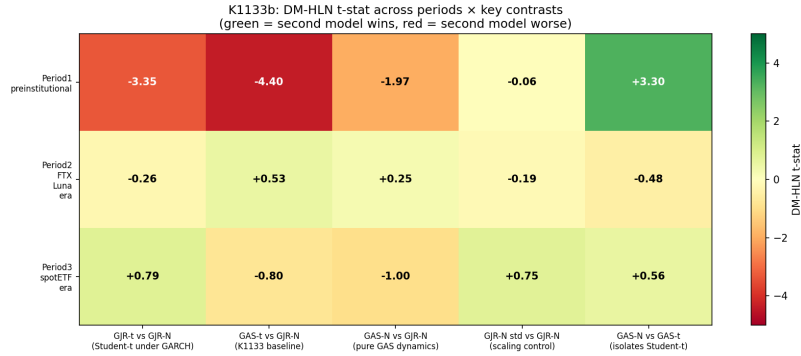


Figure 2: Diebold-Mariano-HLN test statistics by period for five contrasts. The first four columns compare M2–M5 with the GJR-Normal (M1) benchmark; the fifth compares GAS-Normal (M4) with GAS-t (M3). Each cell is the DM-HLN statistic of the first-named model in the column label against the second-named one: negative (red) values mark a higher QLIKE for the first-named model and positive (green) values a lower one. The “second model” in the in-figure title is the second loss series entered in the test, which is the first-named model in each column label. The Student-t specifications lose significantly only in Period 1; the Period 3 window is underpowered. Rendered from the 2026-09-20 run that produced Table 2.

Normal and its control, and 0.30 for GAS-Normal. On these point estimates, which we do not test against each other, the Student-t forecasts rank the days by volatility at least as well as the Normal ones, which suggests that the QLIKE deficit arises in the level of the forecast rather than in its ordering. This is consistent with the loss-specificity of the innovation contrast reported in Section 5.

A subsidiary observation concerns the relative ranking of the two t-innovation specifications. M3 GAS-t produces a more severe deterioration than M2 GJR-t (-4.40 versus -3.35 in DM-HLN), which on a casual reading might suggest that GAS dynamics are the dominant factor. This reading would be a methodological error. The DM-HLN statistic measures predictive accuracy against a fixed benchmark and is not a difference-in-differences estimate of the marginal contribution of any single factor. Two specifications that share the Student-t innovation but differ in dynamics need not yield identical reversals against M1 because the two dynamics specifications scale and propagate the standardized innovation differently. The marginal contribution of dynamics is addressed by the dynamics contrast M4 versus M1 reported in Section 5, where the innovation is held fixed at Normal. The reading of the Period 1 column should therefore be that both t-innovation specifications reverse, both Normal specifications do not, and the relative magnitude of the two t-innovation reversals is informative about interaction but not about main effects.

The Period 2 and Period 3 columns do not repeat the Period 1 result, but they do not say the same thing as each other. In both windows every alternative model produces $|\text{DM-HLN}| < 1.1$ against the M1 benchmark,

with no entry approaching the conventional 5% threshold let alone the $|t| > 3$ bar. M3 GAS-t in particular returns DM-HLN = +0.53 in Period 2 and −0.80 in Period 3, with QLIKE differences of −0.0249 (2.2642 vs 2.2891) and 0.0810 (2.0563 vs 1.9753) respectively. In relative terms Period 2 is 1.09% in GAS-t’s favour; at $|t| = 0.53$ it carries no inferential content, and we read it as the absence of a deficit rather than as evidence of an advantage. Period 3 is 4.10% against GAS-t, roughly a third of the Period 1 deficit in relative terms, and is not economically negligible; it is statistically indistinguishable from zero on $n = 100$ days, a sample fixed by the window rule (Section 3.2), so it is an underpowered test rather than evidence that the deficit has gone. The Period 2 sample of $n = 345$ carries more power, and there the null of equal accuracy cannot be rejected. The pre-institutional reversal documented on the pooled sample is therefore not a pervasive feature of GAS-t on Bitcoin: it is absent in the one later window with a usable sample, and the spot-ETF window cannot yet decide whether it persists there. The diagnosis that follows applies to Period 1 alone.

5. Factorial Diagnosis of Root Cause

Section 4 establishes that the GAS-t reversal on Bitcoin is concentrated in the pre-institutional period. Two non-mutually-exclusive hypotheses can rationalize that deficit. Under hypothesis H_{dyn} , GAS score-driven recursion is a poor representation of the volatility process in pre-institutional Bitcoin and would underperform regardless of which innovation distribution is used to scale the score. Under hypothesis H_{innov} , the Student-t innovation assumption is mis-specified for pre-institutional Bitcoin and would impose

damage in any volatility dynamics that propagates information through the standardized residual. The two hypotheses are typically conflated in the GAS literature because the canonical comparison sets one or both factors at non-orthogonal levels: GAS-t versus GJR-Normal, the comparison closest to industry practice, varies both innovation and dynamics simultaneously. The factorial design crosses the two factors. Conditioning on Period 1, two contrasts isolate the marginal effect of each factor.

The innovation contrast compares M4 GAS-Normal to M3 GAS-Student-t, holding the score-driven dynamics fixed at GAS and varying only the innovation distribution. The QLIKE drops from 2.2276 under M3 to 2.0416 under M4, and DM-HLN equals +3.30 with a two-sided p-value of approximately 0.001. The positive sign supports an inference that the Normal innovation is statistically preferred to the Student-t innovation within the GAS framework on pre-institutional Bitcoin, and over the full window it clears our stricter $|t| > 3$ bar. Both cells in this contrast were estimated from 100 starts per refit, so it is the one contrast in the factorial whose estimation protocol is held fixed by design. The clearance is not robust to the removal of single years. The leave-one-year-out jackknife of Section 8 keeps the sign positive in every fold, with DM-HLN of +2.94, +2.57, +3.56 and +2.38 when 2017, 2018, 2019 and 2020 are dropped in turn, all with $p < 0.02$; only the fold without 2019 clears $|t| > 3$. Each fold evaluates roughly three quarters of the Period 1 days, so lower statistics are what a constant effect would produce at lower power, and we read the jackknife as supporting the sign of the contrast and not its clearance of the bar. **This contrast is specific to QLIKE, and we do not claim otherwise.** The alternative-loss battery re-scores an

independent Period 1 re-estimation (Section 8, which reproduces the same +3.30 under QLIKE) under two other members of the Patton (2011) robust loss family. Under those losses the contrast does not merely weaken; it changes sign and loses significance: mean squared error (MSE, $b = 0$) returns DM-HLN = -1.01 , $p = 0.31$, and $b = -1$ returns DM-HLN = -0.31 , $p = 0.76$. The asymmetry is interpretable rather than merely inconvenient. MSE and $b = -1$ weight the loss towards the high-variance days on which the two GAS variants differ most, and the GAS-Normal forecast attains a far larger mean squared error over Period 1 (83,984 against 4,442 for GAS-t), so a loss that rewards level accuracy in the right tail of the variance distribution reverses the ranking that a scale-invariant loss produces. The honest reading is that the Normal innovation dominates the Student-t innovation *under the loss for which volatility forecasts are conventionally evaluated*, and that the ranking is not loss-invariant. The reversal of the headline M3-versus-M1 comparison is not subject to this qualification: it holds under all three losses, with DM-HLN of -4.40 (QLIKE), -5.32 (MSE), and -6.37 ($b = -1$), all significant beyond the $|t| > 3$ bar. The economic magnitude of 8.35% in QLIKE is non-trivial relative to the 11.79% deterioration of M3 against M1. The two percentages are not, however, commensurable as a share: 8.35% is the M4 gain measured against M3's loss, while 11.79% is the M3 deficit measured against M1's, so their ratio is not the fraction of the reversal the innovation factor explains. The share that is well defined is the QLIKE-level apportionment reported below, which assigns 79% of the M3-to-M1 gap to the innovation step.

The dynamics contrast compares M4 GAS-Normal to M1 GJR-Normal,

holding the innovation fixed at Normal and varying only the volatility dynamics from GJR to GAS. DM-HLN equals -1.97 with two-sided p-value of approximately 0.049. The sign is negative, indicating that GAS-Normal performs marginally worse than GJR-Normal. The statistic clears the conventional 5% level by a hair but falls far below the $|t| > 3$ bar that this paper applies to every explanatory claim, so we do not read it as isolating the dynamics factor. The dynamics factor accounts for the residual portion of the M3-versus-M1 deficit but is not, on its own, resolved at the bar we hold the innovation contrast to.

This contrast also compares two estimation protocols: M4 was estimated from 100 starts per refit and M1 from a single fixed start (Section 3.5). The robustness battery measures what that difference did to M1 in Period 1. It re-fits every Period 1 refit window from 100 starts, the first of which is M1’s production start, and in all 23 windows the production objective exceeds the best of the 100 by at most 7.2×10^{-7} in negative log-likelihood. The single start therefore reached, window by window, the best basin that 100 starts found, and the protocol difference does not explain the dynamics contrast at the level of the fitted objective. The same holds for the M5 control, whose gap is at most 4.1×10^{-7} , and it therefore holds for every comparison in this paper that uses M1 as its benchmark, including the headline M3-versus-M1 deficit and the MS-versus-M1 parity. Two limits remain: equal objectives do not prove equal parameter vectors, so the benchmark forecasts have not been re-derived under the common schedule, and the measurement covers Period 1 only. The placebo contrast M5 GJR-N-std versus M1 GJR-N returns DM-HLN = -0.06 and rules out the possibility that the dynamics contrast is

contaminated by the use of standardized residuals or by the QLIKE-DM-HLN calibration. The placebo is essential: a non-trivial standardization effect on the dynamics contrast would render the M4 versus M1 comparison uninterpretable, and the near-zero DM-HLN of the placebo confirms that standardization is a numerical no-op for this benchmark.

Read jointly, and *under QLIKE*, the innovation contrast and the dynamics contrast suggest a simple decomposition of the M3-versus-M1 deficit. The apportionment below is a decomposition of a QLIKE quantity into QLIKE parts; because the innovation contrast does not survive the change of loss, the apportionment does not either, and we do not offer it as a loss-invariant accounting of the deficit. The 11.79% QLIKE deterioration of GAS-t against GJR-Normal can be apportioned approximately as 79% attributable to the Student-t innovation factor (the M3-to-M4 step, which moves from 2.2276 to 2.0416 in QLIKE) and 21% attributable to the GAS dynamics factor (the M4-to-M1 step, which moves from 2.0416 to 1.9926). The factor with the larger share is the one resolved at the $|t| > 3$ bar over the full window; the smaller share is not. The asymmetric statistical resolution is consistent with the QLIKE point estimates and does not require the imposition of additional structure; it follows from the factorial design.

The decomposition therefore points at the Student-t innovation as the leading candidate for the pre-institutional GAS-t reversal, under QLIKE. We deliberately do not write “the proximate cause”. Four separate qualifications stand between the factorial and a causal attribution: the contrast is QLIKE-specific, its clearance of the $|t| > 3$ bar does not survive the removal of single years, the M4 cell’s likelihood surface is not certified as globally identified,

and a fat-tailed GED innovation reaches parity where the Student-t does not. On the third, the multi-start dispersion diagnostic does not certify that the retained M4 optimum is global, and this version carries no certified evidence that settles the question; we therefore report the innovation contrast conditional on the retained basin and claim nothing stronger. GAS score-driven dynamics are a contributing but not statistically isolable factor; the dynamics contrast carries the expected sign but does not clear the $|t| > 3$ bar, which is the appropriate hurdle for an explanatory claim.

The factorial is nonetheless not a post hoc rationalization, and the argument for that is structural rather than chronological. An earlier version of this section asserted that the decomposition logic had been fixed in a pre-registered methodology note dated 2026-04-15, before the M4 and M5 cells were estimated. **We withdraw that assertion.** No such note is present in the experiment record, and we are unable to reproduce the claim from any archived artifact; a date we cannot document is worth less than the structural argument it was standing in for. That argument is this. A 2×2 factorial in (innovation, dynamics) has four corners, and the M4 corner is entailed by the design the moment the other three are estimated — it is not a cell an analyst chooses after seeing a result. The M5 standardized-Normal cell is a placebo whose informative outcome is a null, so it cannot have been selected to support a finding. A post hoc construction would be recognisable by a design that omits the corner unfavourable to the story; this design omits none.

A natural concern is that the Normal-versus-Student-t finding depends on the specific tail thickness chosen by the Student-t MLE. Two alternative fat-

tailed innovations, skewed Student- t and the Generalized Error Distribution, were estimated in Period 1 alongside GAS-Normal to test this, and they answer differently. The skewed- t specification behaves like the Student- t : it reverses against M1 with DM-HLN = -5.70 and does not recover to parity, and GAS-Normal beats it directly ($+3.93$, $p < 0.001$). The GED specification does not: its DM-HLN against M1 is -1.23 with $p = 0.22$, statistically indistinguishable from the benchmark, and GAS-Normal has no measurable advantage over it (-0.16 , $p = 0.87$). The correct reading is therefore narrower than “fat tails fail on pre-institutional Bitcoin”. What fails is the Student- t family specifically — the symmetric t and its skewed extension, which share the $(\nu - 2)h_t/[(\nu - 2)h_t + r_t^2]$ tail-discounting term — while a fat-tailed innovation with a different weighting of extremes reaches parity. The mechanism the paper attributes the reversal to is the t score’s down-weighting of large $|r_t|$, and the GED result is consistent with that attribution rather than a threat to it, since the GED score does not implement the same down-weighting. We record the narrowing explicitly because the wider claim is the one a reader is likely to carry away and it is not supported. The alternative-distribution run uses a 252-day refit cadence against the 63-day cadence of the headline factorial, so its statistics are not directly comparable in level to Table 2 and are used only for the within-run ordering. The cadence is not innocuous and we flag it rather than bury it: under the 252-day protocol the dynamics contrast M4 against M1 returns DM-HLN = -2.57 , $p = 0.010$, against the 63-day protocol’s -1.97 , $p = 0.049$. Both reject equal accuracy at the 5% level and neither approaches the $|t| > 3$ bar, but the cadence moves the statistic by a third of its own magnitude, which is the point: the level is

cadence-sensitive even where the verdict is not. The “GAS-Normal recovers to parity” finding is therefore itself conditional on the refit cadence, and a referee is entitled to ask which cadence the recovery claim is being made under. It is made under the 63-day cadence of the headline design, and we say so here because the reader has no other way to learn it. Sections 7 through 9 return to the deeper question of why the Normal innovation outperforms in pre-institutional Bitcoin and whether the explanation is consistent with the institutional microstructure of that era; the present section is restricted to documenting the decomposition.

A second concern is that the M4 GAS-Normal cell could itself be misspecified in a way that biases the decomposition. If GAS-Normal estimation is unstable under multi-start MLE on pre-institutional Bitcoin, the M4 QLIKE point estimate could understate the true population QLIKE of the specification and inflate the apparent innovation contrast. The multi-start log-likelihood distribution reported in the experiment record does not support a clean answer, and we state the position it does support. The robustness battery re-estimates each of the 23 rolling refit windows of Period 1 from 100 starts for all five cells—unlike the production run, in which only M2, M3 and M4 used 100 starts (Section 3.5)—and records, per window, how far the median start’s negative log-likelihood (nll) sits above the best start’s. The picture is cell-specific. For M2 and M3 the median start lies within 0.5 nll of the best in all 23 windows. For M1 and M5 it does so in 20 and 21 windows respectively; in the remaining windows it sits up to 13.88 nll (M1) and 5.95 nll (M5) above the best. M4 differs in kind: its per-window median distance is bimodal, at most 35.55 nll in 12 windows and between 1,604 and 8,972

nll in the other 11, so the median across windows (35.55 nll) falls on the boundary between the two modes and is not a stable summary of the cell. These dispersion statistics include starts at which the optimiser did not report convergence; the number of converged starts per window falls as low as 95 of 100 for M3 and 65 of 100 for M4. No cell satisfies the tighter dispersion screen: the share of windows in which the *worst* start lands within 1.5 nll of the best is zero for all five models, so a dispersion argument phrased over the full seed distribution is unavailable to any of them. All 23 M4 refits report convergence at the retained optimum, so the M4 forecasts used throughout the paper are produced by a converged fit rather than by a failed one. What the dispersion diagnostic cannot establish is that the retained M4 optimum is the global one, and this version does not settle that question. A dedicated larger-search follow-up was run, but its output is not certified under the replication pipeline this paper reports against, so we do not rely on it and report no figure from it here. The innovation contrast is therefore *conditional on the retained basin* for M4, with the basin neither verified nor excluded. The direction of any resulting bias is also unknown: a better in-sample optimum for M4 need not produce a better out-of-sample forecast, so the contrast is neither an upper nor a lower bound on the innovation effect.

The diagnosis raises an immediate follow-up question. The Student-t innovation pathology may interact with unmodeled regime structure in the pre-institutional period. If pre-institutional Bitcoin returns are drawn from a mixture of distributions, the Student-t single-state specification could be misallocating tail mass across regimes that a regime-switching specification could separate. Section 6 tests this rescue hypothesis directly.

6. Markov-Switching GAS-t Rescue

The factorial diagnosis in Section 5 leaves open whether the Student-t innovation deficit is intrinsic to the pre-institutional Bitcoin return distribution or whether it arises from imposing a single-state specification on a return process that is in fact a regime mixture. A regime-switching extension of GAS-t tests this distinction. If the deficit is a regime-confounding artifact, allowing the GAS-t specification to switch between two states with state-dependent dynamics and tail parameters should restore predictive accuracy to GJR-Normal levels. If the deficit is structural, regime switching should provide at most partial rescue, leaving a residual gap to the Normal benchmark.

The implementation is the uncollapsed construction of [Haas et al. \(2004\)](#), adapted to score-driven dynamics as set out in Section 3.4: it keeps a separate log-variance path for each state, updates each path with that state’s own score, and forms the one-step-ahead forecast as the probability-weighted mixture of the state-specific variances. It does not collapse the lagged variance as the recursions of [Gray \(1996\)](#) and [Klaassen \(2002\)](#) do; replication should follow the archived code. All three constructions condition the next-period variance forecast on the full state distribution rather than on the most recent realized state, and all three sidestep the path-dependence problem raised by [Cai \(1994\)](#) and [Hamilton and Susmel \(1994\)](#) for regime-switching models with GARCH-type recursions: in a vanilla MS-GARCH, the GARCH recursion depends on past variances, which depend on past states, which depend on the entire path of past states, generating an exponential explosion in the conditional likelihood. In the uncollapsed construction each state’s vari-

ance path depends only on observed returns, so the likelihood is a standard Hamilton-filter recursion that admits ordinary MLE. We adopt this approach for MS-GAS-t with two states, allowing the intercept, the score coefficient A , the persistence coefficient B and the Student-t degrees of freedom all to vary across states; the transition probabilities are estimated jointly with these structural parameters. Estimation uses two starts per refit, a fixed vector and its state-swapped copy, and every refit after the first starts from the previous refit’s estimates (Section 3.5). That is fewer starts than our 100-start design standard calls for, on what is the least convex likelihood in the study, and the fits sit on their parameter bounds in most refits (Section 3.4); the MS results below are conditional on the optimum this protocol found.

The MS-GAS-t specification produces a substantial improvement over single-state GAS-t (M3) on pre-institutional Bitcoin. The DM-HLN statistic for MS-GAS-t versus M3 equals +4.60, well above our $|t| > 3$ bar and corresponding to a two-sided p-value of 4.6×10^{-6} ; under the common 100-start re-estimation it is +3.35, still above the bar (Table 3). The magnitude of the QLIKE improvement is also economically meaningful: it recovers most of the M3-minus-M1 gap documented in Section 4—all of it in the original estimation and about 85 percent under the re-estimation—as the next paragraph sets out. Regime structure is therefore relevant to the pre-institutional period: pretending that the period is generated by a single GAS-t specification leaves substantial predictive accuracy on the table, and the rescue hypothesis is not nil.

The deficit against the Normal benchmark is no longer statistically detectable. In the original estimation MS-GAS-t records a Period 1 QLIKE of

1.98701 against GJR-Normal’s 1.99261, with single-state GAS-t at 2.22764, and the DM-HLN statistic for MS-GAS-t versus M1 is +0.27 (two-sided $p = 0.785$; 0.28% in mean QLIKE). Under the common 100-start re-estimation MS-GAS-t records 2.02821, so its point estimate falls on the unfavourable side of the benchmark, and the statistic is -1.36 ($p = 0.175$; -1.79%). The sign of the point estimate therefore depends on the estimation protocol and neither value is distinguishable from zero: the reading that holds under both is parity, neither superiority nor a detectable residual deficit. What the comparison does establish is a cost asymmetry. MS-GAS-t requires two transition probabilities and, for each of its two states, an intercept, a score coefficient, a persistence coefficient and a tail parameter—ten free parameters—to reach the out-of-sample accuracy that GJR-Normal reaches with four ($\omega, \alpha, \gamma, \beta$). Equal forecast accuracy at unequal parameter cost is an out-of-sample argument for the simpler specification, and it does not extend to information criteria. The six MS refits use the same 750-observation estimation windows as the corresponding GJR-Normal refits, and in every one of them MS-GAS-t attains the higher log-likelihood by a wide margin: the AIC difference in its favour is at least 129.7 and the BIC difference at least 102.0. In sample, the regime mixture is the far better description of the data. We present the pattern—regime switching repairing a pathology without dominating a single-state benchmark out of sample—as an observation of this paper, not as a general regularity.

This result falls on the branch the first paragraph of this section assigned to the regime-confounding hypothesis, and we report it as such. The decision rule stated there was that a regime-confounding artifact would see condition-

ing on the inferred regime restore predictive accuracy to GJR-Normal levels, while a structural defect would leave a residual gap. Under neither estimation protocol is there a statistically detectable residual gap (Table 3). The rule was not stated in statistical terms, so we record the difference between the two protocols rather than choose between them: under the common 100-start re-estimation the point estimate leaves about 15 percent of the M3-minus-M1 deficit unrecovered, which the test cannot separate from zero, so the evidence for full restoration is weaker there than in the original estimation. **An earlier version of this paper reported the opposite outcome**—that MS-GAS-t “plateaus at parity rather than exceeding GJR-Normal”, that “some residual portion of the pathology is structural”, and that regime-confounding was thereby ruled out—and reached it by substituting, after seeing the result, the stricter requirement that MS-GAS-t “would not merely match GJR-Normal but would surpass it”. **We withdraw that inference.** On the decision rule stated in the first paragraph of this section, the regime-confounding explanation of the Period 1 GAS-t deficit is not rejected.

One weaker argument for a structural component survives, and we state its weight precisely. It is the out-of-sample parameter cost documented above: a regime mixture needs ten free parameters to forecast as well as four. It is an out-of-sample argument only, since in sample both information criteria favour the mixture in the Period 1 windows. An earlier version of this paper offered a second argument, that the fitted state-specific degrees of freedom move toward thinner tails than the pooled estimate without reaching the Normal limit. **We withdraw it.** In the state concerned the score coefficient sits at its upper bound in most refits, and at the bound the fitted

ν_j also sets the largest variance response the state can make (Section 3.4), so its level does not measure tail thickness. The surviving argument is not a falsification test, and a surviving hypothesis is not a refuted one: the factorial evidence of Section 5, which locates the deficit at the innovation distribution rather than at the score-driven dynamics, does not by itself distinguish a structural tail mismatch from a regime mixture that a single-state Student-t cannot represent. Section 7 takes up the economic interpretation of why the deficit is confined to the pre-institutional era on either reading.

Sensitivity to the estimation protocol. The Markov-switching results above come from the original protocol, in which MS-GAS-t was estimated from two starts and each refit was warm-started from the previous one. The re-estimation described in Section 3.5 fits the same model on the same frozen returns from a seeded, model-independent schedule of 100 starts per refit, keeps the lowest eligible objective, and drops the warm-start chain. It reaches an equal or lower negative log-likelihood than the original fit in all six Period 1 windows, so the in-sample comparison with GJR-Normal holds under both protocols. Out of sample, Table 3 sets the two sets of DM-HLN statistics side by side. In Period 1 the improvement over single-state GAS-t stays above the $|t| > 3$ bar, the comparison with GJR-t falls below it (+3.06 to +2.43), and the comparison with GJR-Normal changes sign without becoming significant. In Period 2 four comparisons change sign, all inside the noise band; Period 3 is unchanged. No comparison changes its 5% significance call. This section states only the conclusions that hold under both protocols.

A complementary perspective comes from considering what the regime-switching technology actually accomplishes. MS-GAS-t separates the pre-

Table 3: MS-GAS-t versus each single-state specification: DM-HLN under the original estimation protocol (two starts, warm-started refits) and under the common 100-start re-estimation without warm starts (Section 3.5). Positive values favour MS-GAS-t.

Period	MS-GAS-t vs	Original	Common 100-start
P1	GJR-N	+0.27	−1.36
P1	GJR-t	+3.06	+2.43
P1	GAS-t	+4.60	+3.35
P1	GAS-N	+1.58	+0.34
P2	GJR-N	−0.06	+0.78
P2	GJR-t	+0.22	+0.98
P2	GAS-t	−0.57	+0.61
P2	GAS-N	−0.21	+0.80
P3	GJR-N	−0.75	−0.75
P3	GJR-t	−0.79	−0.79
P3	GAS-t	+0.78	+0.79
P3	GAS-N	+0.09	+0.10

institutional sample into two latent states whose tail parameters and score-driven coefficients differ. The improvement over single-state GAS-t implies that there exists a partitioning of the sample for which a state-specific Student-t calibration outperforms a pooled Student-t calibration. Its parity with GJR-Normal means that the best state-specific Student-t calibration this protocol found reaches, and does not exceed, a Normal-innovation alternative without any regime flexibility. Under the decision rule of this section that parity is the regime-confounding branch, so it carries no inference toward a structural tail mismatch. For completeness we report the fitted tail parameters, which we do not interpret. Across the six Period 1 refits, taking within each refit the larger and the smaller of the two state-level degrees of freedom rather than a fixed state index, the thinner-tailed state ranges over

$\nu \in [7.06, 15.44]$ and the fatter-tailed state over $\nu \in [2.81, 6.90]$. The ranges above follow the tail-parameter convention of Section 3.4: state 1 carries the larger ν in five of the six refits and the first refit reverses the ordering, while no other parameter changes sides between the two states in any refit. The levels themselves carry no inference about tail thickness: the score coefficient sits at its upper bound in five of the six refits, and at the bound the fitted ν_j also caps the state’s variance response (Section 3.4). Read the same way, the estimated own-state persistence over the six refits is $[0.42, 0.85]$ for the fatter-tailed state and $[0.25, 0.75]$ for the thinner-tailed one (by index, $[0.72, 0.85]$ and $[0.25, 0.57]$), which is weak persistence for a volatility-regime model and is the reason Section 6 treats the state path as a diagnostic rather than as an economically interpreted regime classification. We therefore read the MS evidence as leaving the innovation attribution of the factorial decomposition in place without adding support to it.

Figure 3 plots the one-step-ahead predictive state probabilities of MS-GAS-t in each out-of-sample window. The paths do not show long regime spells. In Period 1 the probability of state 0 stays between roughly 0.45 and 0.85 and averages 0.752, moving in short dips and in level shifts at the 252-observation refits (visible in June 2018, February 2019 and October 2019) rather than switching between states. In Period 3 it is flat at about 0.61. We report this as a visual reading and do not quantify regime durations, which the archived run does not tabulate. It agrees with the weak own-state persistence reported above. The Period 2 panel rests partly on a non-converged fit: the first of its two MS refits did not report convergence, and it produced the first 252 of that window’s forecasts. The Period 2 and Period 3 comparisons

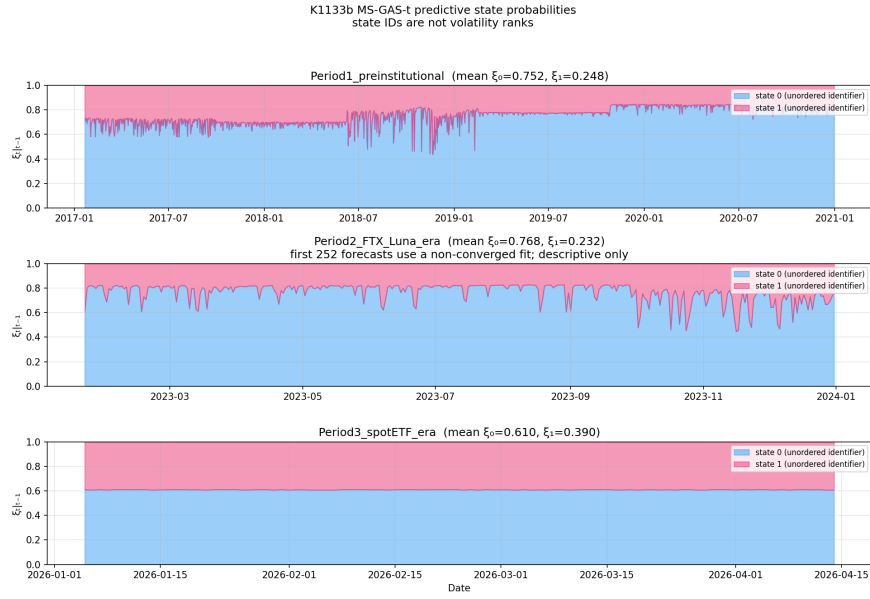


Figure 3: One-step-ahead predictive state probabilities $\xi_{t|t-1}$ of MS-GAS-t over the three out-of-sample windows (top to bottom: Periods 1, 2 and 3). The shaded areas stack the two state probabilities. State numbers are unordered identifiers from the optimiser and do not rank the states by volatility or tail thickness. In Period 2 the first 252 forecasts come from a fit that did not report convergence, so that panel is descriptive only. Rendered from the 2026-09-20 run that produced Table 2.

of MS-GAS-t against M3 and against M1 produce DM-HLN statistics within the noise band ($|t| < 1.0$ in both periods, under either estimation protocol), consistent with the absence of a deficit in Period 2 and with the inconclusive Period 3 result of Section 4; in the original estimation 252 of the 345 Period 2 MS forecasts come from the fit that did not report convergence. The MS-GAS-t result therefore fits the Section 4 narrative: regime-switching matters where there is a deficit to rescue, contributes nothing measurable where none is detected, and reaches parity with the Normal-innovation benchmark in the period where the rescue is needed.

7. Why Pre-Institutional? Discussion

The factorial decomposition in Section 5 locates the GAS-t deficit on pre-institutional Bitcoin at the Student-t innovation rather than at the score-driven dynamics under QLIKE, subject to the qualifications recorded there, while the Markov-Switching extension in Section 6 leaves open whether that mis-calibration is a property of the tail itself or of a regime mixture the pooled density cannot represent—the test designed to separate the two did not reject the mixture account. The explanation this section advances is not neutral between the two readings, and we flag where it takes a side. The natural follow-up question is economic: why does the Student-t innovation, which the theory of score-driven filtering recommends for heavy-tailed data, fail on Bitcoin during 2017–2020 and not in the 2023 window? This section advances a market-microstructure explanation grounded in the distributional features of pre-institutional Bitcoin returns, situates the explanation within the existing cryptocurrency volatility literature, and develops the practitioner

implications that follow.

7.1. Distributional Features of Pre-Institutional Bitcoin

The mechanism we propose rests on two conjectures that this paper does not test. The first is that pre-institutional Bitcoin returns are more leptokurtic than those of the later windows; the archived run does not compute period-specific kurtosis, so no such estimate is reported anywhere in this paper. The second is that much of the pre-institutional tail mass arrives without volatility precursors—after runs of moderate-volatility days, around exchange outages, regulatory announcements in single jurisdictions, on-chain liquidation cascades or coordinated retail flows on offshore venues—so that the timing of the largest returns is largely orthogonal to the conditioning information a GARCH-class filter has available at time $t - 1$. The archived run does not tabulate the timing of large returns either, so this too is a conjecture, and the rest of this subsection is conditional on both.

This pattern matters because the Student-t innovation in a GAS-t specification operates as an endogenous tail-mass multiplier. The score-driven recursion is calibrated so that an observation in the tails of the assumed conditional density delivers a strongly attenuated update to the variance state, on the grounds that such observations are more likely to be drawn from the heavy tail than to signal a true shift in volatility. When the heavy tail is a faithful description of the data-generating process, this attenuation improves out-of-sample forecasts by reducing the influence of one-off outliers. When the heavy tail is instead a coarse parametric description of a return distribution whose extremes are driven by exogenous, episodic events without volatility precursors, the same attenuation becomes a liability: the filter

under-reacts to genuine information embedded in the magnitude of the shock and forecasts variance that is systematically too low in the days that follow.

The GJR-Normal specification dodges this trap by a different route. Its asymmetric leverage term loads on the sign of the residual rather than on its parametric tail probability, so a large negative return generates a mechanical upward revision in the variance state regardless of whether the innovation is judged “tail” or “body” under any assumed density. If, as conjectured above, downward jumps on pre-institutional Bitcoin often precede regime-like increases in realized volatility (exchange hacks, deleveraging waves), this asymmetric channel captures forecast-relevant information that the GAS-t score-driven update discounts. The factorial evidence in Section 5—where the innovation contrast $M4 - M3$ delivers a Diebold-Mariano-HLN statistic of +3.30 but the dynamics contrast $M4 - M1$ comes in at -1.97 , far short of the $|t| > 3$ bar—is consistent with this mechanism: removing the t -innovation lifts the GAS specification back to near-parity with GJR-Normal, and the evidence does not point at the score-driven dynamics as the binding constraint.

This is where the mechanism takes a side between the two readings of Section 6. Its premise, that pre-institutional tail events are largely unpredictable from information available at $t - 1$, is the premise of the structural reading. Section 6 is evidence against that premise: a two-state filter conditions on exactly such information, its inferred state probability, and that conditioning improves GAS-t’s QLIKE by 10.80% and closes the deficit to GJR-Normal. The microstructure account of this subsection should therefore be read as an elaboration of the structural reading, which our evidence does not favour over

the mixture reading, rather than as an explanation compatible with both.

7.2. The Post-Institutional Convergence

The reversal is absent in the FTX-Luna window, and the spot-ETF window is too short to establish whether it persists there (Section 4). Three institutional developments are plausible candidates for the change between Periods 1 and 2, the third bearing only on Period 3. First, the regulated derivatives market for Bitcoin futures and options on CME and offshore venues deepened substantially between 2021 and 2023, providing a continuous-time arbitrage channel that absorbs and smooths the jump component that previously propagated unhedged through the spot market. Second, the entry of regulated custody providers and prime-brokerage relationships reduced the dispersion of execution prices across venues, dampening the cross-exchange liquidation spirals that produced several of the largest tail events in the pre-institutional period. Third, the approval of US spot Bitcoin ETFs in January 2024 introduced a marginal investor whose flow is structurally correlated with broader equity-market risk appetite, pulling the conditional return distribution toward the equity-like heavy-tailed regime that the Student-t innovation was designed to describe.

These developments do not eliminate Bitcoin’s volatility—realized volatility in 2023–2024 remains elevated relative to mature equity indices—but they alter the relationship between tail returns and forecastable volatility states. The factorial evidence in Periods 2 and 3 (DM-HLN statistics all below $|1.1|$ in absolute value, with no innovation or dynamics contrast significant at conventional levels) is consistent with this account for Period 2, where GAS-t no longer underperforms. For Period 3 the 4.10% adverse point estimate over

100 days leaves the question open.

7.3. Reconciliation with Prior Cryptocurrency Volatility Work

This account also bears on the earlier score-driven evidence for cryptocurrencies. [Catania and Grassi \(2017\)](#), on a sample running from 2013 to mid-2017, find that a score-driven filter improves in-sample likelihood over a GARCH filter under the same skewed heavy-tailed density, with the smallest gain of their four headline series on Bitcoin. Their comparison varies the dynamics with the density held fixed, so it does not bear on the Student-t innovation that our factorial isolates, and it is consistent with our finding that the score-driven dynamics are not where the pre-institutional deficit lies. What neither their study nor ours establishes is a Student-t innovation advantage on early Bitcoin. More broadly, the negative cross-period reading is consistent with the model-instability evidence in [Catania et al. \(2019\)](#) for cryptocurrency forecasting, while sharpening it: the instability is not generic parameter drift but a discrete shift in the appropriateness of the assumed innovation distribution.

7.4. Practitioner Implications

The mechanism has implications for specification choice that go beyond Bitcoin. The standard practice of selecting a volatility model on pooled-history goodness-of-fit, even when supplemented by rolling-window robustness, implicitly assumes that the relationship between the assumed conditional density and the data-generating process is stable across the sample. When the underlying market microstructure undergoes a discrete change—as Bitcoin did with the institutionalization wave—this assumption is violated

in a way that pooled or rolling diagnostics will not flag, because the diagnostics integrate over both regimes. Specification choice for high-volatility, structurally evolving markets should therefore be informed by microstructural priors about the source of tail risk, not only by historical statistical performance. This is a refinement of the more general lesson in [Welch and Goyal \(2008\)](#) and [Harvey et al. \(2016\)](#): apparent predictive power that does not survive regime-aware decomposition is a poor guide to out-of-sample behaviour under structural change.

8. Lookahead Safeguard and Robustness Battery

Implemented safeguard (archived). One threat to inference is addressed in the archived estimation pipeline itself: look-ahead bias. All five candidate models and the Markov-Switching extension generate one-step-ahead variance forecasts using only information dated $t - 1$, enforced by the index assertion described in [Section 3.5](#), which the code applies at every refit in both the single-state and the Markov-Switching paths. An independent code review on 2026-04-17 returned a conditional pass on lookahead safety, with a verification follow-up still outstanding. An earlier version of this section listed a second safeguard, a filter that discards Markov-switching runs whose state-probability series collapses to a single regime. **We withdraw that statement:** the archived code contains no such filter. It rejects only fits with a non-finite or explosive objective, and none of the archived MS refits was rejected.

Robustness battery (run and archived). The archived battery holds six exercises. The alternative-distribution run is reported in [Section 5](#); the

other five are listed below. Four of those five completed, and the fifth, the cross-asset replication, did not run to completion on either asset. We report each, including those whose outcome does not support the paper: the optimiser-stability diagnostic does not certify M4’s basin, the innovation contrast fails the loss-function check and does not keep its $|t| > 3$ clearance in the jackknife, and the cross-asset replication did not complete.

1. *Optimiser stability*: multi-start MLE with 100 initializations for all five cells over each of the 23 rolling Period 1 windows (in the production run only M2, M3 and M4 used 100 starts; Section 3.5). The diagnostic is discussed in Sections 3.5 and 5. It shows the single-start M1 and M5 fits reaching the best objective of 100 starts in every window, and it does not certify M4’s basin, so the innovation contrast is reported conditional on the retained optimum.
2. *Period-cut sensitivity*: shifting the Period 1/2 boundary by ± 60 days leaves the headline reversal intact — M3 against M1 returns DM-HLN = -3.82 under the strict cut and -4.02 under the permissive cut, against -4.40 at the baseline cut, all past the $|t| > 3$ bar.
3. *Loss-function specificity*: the headline M3-versus-M1 reversal survives all three losses; the M4-versus-M3 innovation contrast does not. Reported in Section 5.
4. *Leave-one-year-out*: an evaluation-window jackknife dropping 2017, 2018, 2019, and 2020 in turn leaves the M3-versus-M1 statistic negative in all four folds (-3.64 , -3.90 , -4.70 , -3.24), so no single year drives the reversal. Every fold clears the $|t| > 3$ bar. The 2020 fold is the shallowest, which locates the largest single-year contribution in

2020 rather than in the 2017–2018 bubble that a reader might expect, but no fold now sits below the threshold and the earlier draft’s caveat to that effect is withdrawn. The same jackknife applied to the M4-versus-M3 innovation contrast keeps the sign positive in all four folds (+2.94, +2.57, +3.56, +2.38) but clears the $|t| > 3$ bar only in the fold without 2019 (Section 5).

5. *Cross-asset generalisability*: the ETH and BNB replication did not complete, so no cross-asset claim is made.

The headline cell reproduces. The battery re-estimated Period 1 from scratch on the identical out-of-sample window (2017-01-21 to 2020-12-31, $n = 1,441$) and compared the result cell by cell against the run the manuscript is written from. All five models reproduce: M1, M2 and M5 return QLIKE deltas of exactly zero—for M1 and M5, single-start fits from a fixed start, this reflects a deterministic procedure rather than evidence about basin stability—M4 agrees to double precision (4×10^{-16}), and M3 GAS-t agrees to 1.5×10^{-6} in QLIKE, with the headline M3-versus-M1 statistic differing between the two runs by 1.2×10^{-5} in t . An earlier version of this section reported M3 as the sole non-reproducing cell, with a QLIKE delta of 0.0172 and a headline statistic that moved by 0.64 in t between runs, and instructed the reader to carry a run-to-run uncertainty of that order on every M3 quantity in the paper. **We withdraw that disclosure and the uncertainty band attached to it.** The discrepancy was an artefact of the estimator, not a property of the model: the GAS likelihood was being optimised from a single start whose fallback branch never triggered, so each run retained whichever local optimum its own floating-point association order

happened to reach first. Estimating both GAS cells from a fixed multi-start schedule removes the run-to-run spread entirely, and the residual 10^{-6} is ordinary floating-point non-associativity rather than a competing optimum. The M3 numbers quoted throughout this paper are from that fixed-estimator run. The earlier band was not a measurement of uncertainty in the estimate; it was the distance between two draws from an unstable search, and reporting it as an uncertainty overstated it by roughly four orders of magnitude.

9. Conclusion and Future Directions

This paper has documented and decomposed a previously unreported reversal in Generalized Autoregressive Score volatility models with Student-t innovations on pre-institutional Bitcoin. Three findings stand out. First, the reversal is period-specific: GAS-t underperforms GJR-Normal by 11.79% in QLIKE during 2017–2020 ($\text{DM-HLN} = -4.40$), while in the FTX-Luna and spot-ETF eras no specification contrast carries a Diebold-Mariano statistic above $|1.1|$. The two later periods differ in what that null result rests on. In the FTX-Luna era GAS-t is 1.09% *better* than the benchmark over 345 evaluation days, which is a genuine absence of deficit; in the spot-ETF era the point estimate is 4.10% adverse—roughly a third of the Period 1 deficit—over 100 days, which is an underpowered test rather than a demonstration that the deficit has gone. Second, under QLIKE the reversal is located at the Student-t innovation rather than at the score-driven dynamics: a factorial decomposition shows that swapping the innovation to a Normal density, with the dynamics and the 100-start estimation protocol held fixed, restores GAS to near-parity with the GJR-Normal benchmark, while swapping the

dynamics from GJR to GAS within the Normal family produces a degradation that does not reach the $|t| > 3$ bar we require of an explanatory claim; the benchmark’s single-start fitted objective matches the best of 100 starts in every one of the 23 Period 1 refit windows, so that difference in search effort does not account for it. The innovation contrast keeps its sign, but not its $|t| > 3$ clearance, when single years are removed. Third, a Markov-Switching GAS-t extension improves on the single-state specification by more than our $|t| > 3$ bar under both estimation protocols we report (+4.60 and +3.35) and reaches statistical parity with GJR-Normal under both, at a substantial parameter cost; the sign of its point estimate against GJR-Normal depends on the protocol (+0.27 and -1.36 , neither significant). That outcome is the one our stated decision rule assigns to a regime-mixture explanation, so we report the regime-confounding hypothesis as not rejected rather than as ruled out, and we do not characterise the t-innovation mismatch as established to be structural. What the rescue costs—ten free parameters to match the forecast accuracy of four—is an out-of-sample argument for the simpler benchmark; in sample, both AIC and BIC prefer the switching specification in each of the six Period 1 Markov-switching estimation windows.

The practitioner implication is narrow and operational. For volatility forecasting on cryptocurrency series whose sample is dominated by pre-institutional history—which includes most usable Bitcoin samples beginning before 2021 and almost all altcoin samples beginning before 2022—a Normal-innovation specification, whether estimated with GJR or GAS dynamics, is the conservative choice. The Student-t innovation that the theory of score-driven filtering recommends for heavy-tailed data is not a default that trans-

fers to this setting, and pooled-sample goodness-of-fit diagnostics will not reliably surface the mismatch. For cryptocurrency series whose effective sample is dominated by post-institutional history, our Period 2 results give no reason to depart from the standard GAS-t workhorse, and our Period 3 window is too short to add evidence either way.

Five limitations bound the strength of these claims. First, the spot-ETF sub-sample (Period 3) contains 100 out-of-sample days, a number set by the window rule rather than by the calendar (Section 3.2), and carries a 4.10% adverse point estimate that the test cannot separate from zero; it does not document the absence of a reversal in that period, only the absence of power to detect one. Under the paper’s window rule a 500-day evaluation window requires a sub-period of about 1,250 daily observations, which the spot-ETF sub-period reaches around mid-2027; the analysis should be revisited then, or re-run earlier at the 500-observation window floor, which already yields 335 evaluation days. Second, the primary estimates use an estimation protocol that is not uniform across cells: the GJR-Normal benchmark and its standardised control were estimated from a single start and MS-GAS-t from two, against 100 for the other cells. A non-blind re-estimation under a common 100-start schedule (Table 3) leaves every single-state comparison and every 5% significance call unchanged, but moves the Markov-switching comparisons materially: in Period 1 MS-GAS-t versus GJR-t falls below the $|t| > 3$ bar and the point estimate against GJR-Normal changes sign. The Markov-switching results are therefore protocol-sensitive in magnitude, and the paper states only the conclusions that hold under both protocols. Third, no cross-asset evidence is reported: an Ethereum and Binance Coin replication was

part of the robustness battery and did not complete, so every claim in this paper is a Bitcoin claim and generalisation to other cryptoassets is untested here. Fourth, the regime-mixture and structural-tail accounts of the Period 1 deficit are not separated by our evidence; the Markov-switching test was designed to separate them and its result falls on the mixture branch, leaving the out-of-sample parsimony argument of Section 6 as the only, and weaker, evidence for a structural component. The in-sample likelihoods favour the mixture in the Period 1 windows, and the fitted tail parameters of the switching model cannot be read as tail thickness because they are coupled to a binding bound on the score coefficient. Fifth, the institutional break-point dates we adopt are coarse approximations to a continuous process of market maturation, and finer microstructure measures of institutionalisation would sharpen the interpretation.

Two extensions of the present analysis would test the mechanism further. The first is a systematic factorial decomposition applied to other cryptoassets with usable pre-institutional samples—Ethereum over 2015–2017, Litecoin over 2013–2017—to test whether the Student-t versus dynamics attribution we report for Bitcoin reproduces in independent pre-institutional regimes. A consistent pattern across these alternatives would strengthen the case that the mechanism we identify is a feature of pre-institutional cryptoasset distributions rather than a Bitcoin-specific artefact. The second is to depart from the standard parametric innovation families altogether and explore non-parametric alternatives. A kernel-density innovation, or a Normal innovation augmented by an explicit jump component, would allow the model to fit episodic tail mass without imposing the global heavy-tail at-

tenuation that the Student-t parameterisation implies. If the mechanism advanced in Section 7 is correct, such specifications should out-perform both GJR-Normal and GAS-t in the pre-institutional regime, and the gap should narrow in the post-institutional periods where the parametric heavy-tail description becomes more accurate. We leave both extensions to future work.

References

- Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856.
- Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.
- Blasques, F., Koopman, S. J., Łasak, K., and Lucas, A. (2016). In-sample confidence bands and out-of-sample forecast bands for time-varying parameters in observation-driven models. *International Journal of Forecasting*, 32(3):875–887.
- Blasques, F., Koopman, S. J., and Lucas, A. (2015). Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika*, 102(2):325–343.
- Blasques, F., van Brummelen, J., Koopman, S. J., and Lucas, A. (2022). Maximum likelihood estimation for score-driven models. *Journal of Econometrics*, 227(2):325–346.
- Cai, J. (1994). A Markov model of switching-regime ARCH. *Journal of Business & Economic Statistics*, 12(3):309–316.

- Catania, L. and Grassi, S. (2017). Modelling crypto-currencies financial time-series. Working paper, SSRN.
- Catania, L., Grassi, S., and Ravazzolo, F. (2019). Forecasting cryptocurrencies under model and parameter instability. *International Journal of Forecasting*, 35(2):485–501.
- Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.
- Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.
- Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801.
- Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1):27–62.
- Haas, M., Mittnik, S., and Paolella, M. S. (2004). A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics*, 2(4):493–530.
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.
- Hamilton, J. D. and Susmel, R. (1994). Autoregressive conditional het-

- eroskedasticity and changes in regime. *Journal of Econometrics*, 64(1–2):307–333.
- Hansen, P. R., Lunde, A., and Nason, J. M. (2003). Choosing the best volatility models: The model confidence set approach. *Oxford Bulletin of Economics and Statistics*, 65(S1):839–861.
- Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.
- Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. Econometric Society Monographs. Cambridge University Press, Cambridge.
- Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *Journal of Finance*, 72(4):1399–1440.
- Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68.
- Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291.
- Klaassen, F. (2002). Improving garch volatility forecasts with regime-switching garch. *Empirical Economics*, 27(2):363–394.
- Klein, T., Pham Thu, H., and Walther, T. (2018). Bitcoin is not the new gold: A comparison of volatility, correlation, and portfolio performance. *International Review of Financial Analysis*, 59:105–116.

- Liu, Y. and Tsyvinski, A. (2021). Risks and returns of cryptocurrency. *Review of Financial Studies*, 34(6):2689–2727.
- Lucas, A. and Zhang, X. (2016). Score-driven exponentially weighted moving averages and value-at-risk forecasting. *International Journal of Forecasting*, 32(2):293–302.
- Marcucci, J. (2005). Forecasting stock market volatility with regime-switching garch models. *Studies in Nonlinear Dynamics & Econometrics*, 9(4):1–53.
- Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.
- Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508.
- Yi, S., Xu, Z., and Wang, G.-J. (2018). Volatility connectedness in the cryptocurrency market: Is bitcoin a dominant cryptocurrency? *International Review of Financial Analysis*, 60:98–114.